ÉCOLE DE GUERRE

PROMOTION *GÉNÉRAL GALLOIS*

2016 -2017

Technologies autonomes au combat : vers un abandon de l'homme sur le champ de bataille ?

(L'improbable éthique au combat des futurs robots-tueurs)

CBA Brice ERBLAND

Sous la direction de :

M. Jean-Baptiste Jeangène-Vilmer

Directeur de l'IRSEM

SOMMAIRE

INTRODUCTION

Dans la mythologie grecque, le sculpteur Pygmalion se voue au célibat parce que les femmes de l'île de Chypre le déçoivent. Aucune d'entre elles ne présente assez de qualités à ses yeux, et leur conduite ne satisfait pas sa morale exigeante. Il se réfugie donc dans son atelier, où il sculpte son idéal féminin. La précision et la beauté de son travail sont telles qu'il tombe éperdument amoureux de sa sculpture. Mais sa femme parfaite est faite de pierre de taille et de cire, elle n'a aucune âme, et sa peau et ses lèvres demeurent désespérément froides. Pygmalion prie alors la déesse Aphrodite de donner vie à l'objet de son amour. La déesse, touchée, accède à sa demande et la statue s'éveille : Galatée prend vie et devient la femme de Pygmalion.

Les ingénieurs en robotique militaire sont en quelque sorte des Pygmalions modernes. Les soldats humains présentant bien trop de défauts, il n'est plus digne d'intérêt de s'échiner à rechercher des améliorations d'outils technologiques qui dépasseraient de toute façon leurs faibles capacités cognitives. En revanche, la sculpture de la statue idéale reste possible : ils cherchent donc à créer un robot-soldat qui serait le guerrier parfait, sans défaut sur le champ de bataille. Hélas, le problème reste le même que dans l'Antiquité, car si les connaissances scientifiques actuelles permettent de sculpter la statue rêvée, cette dernière n'a toujours pas d'âme. Pourtant, si la déesse Aphrodite ne peut rien pour les ingénieurs, les promesses du progrès informatique sont de bonne augure. À tel point que l'agence de recherche et développement militaire américaine, la DARPA[1], estime qu'une intelligence artificielle assez

[1] Defense Advanced Research Projects Agency.

accomplie pour pouvoir raisonner de manière autonome, comme un être humain sur un champ de bataille, pourrait voir le jour d'ici 2030[2]. L'autonomie intellectuelle de la technologie serait donc le graal de l'éveil de conscience du robot-soldat, l'intervention divine qui donnerait vie à la statue révérée par les sculpteurs du combat moderne.

Mais qu'est-ce qu'une technologie autonome au combat ? Est-ce un simple logiciel « intelligent » gouvernant différents systèmes automatisés, à l'image de l'ordinateur HAL 9000 de *2001, l'odyssée de l'espace*, ou est-ce un robot-tueur humanoïde à l'image de *Terminator* ? L'imaginaire collectif associe aisément technologie autonome et robot, et l'on donne rapidement ce qualificatif à toute machine qui se déplace ou effectue des tâches sans pilote. Peter W. Singer propose d'ailleurs, en abordant les technologies autonomes dans son ouvrage *Wired for war*, une définition du robot qui repose sur quatre critères : le robot est fabriqué par l'homme, il possède des senseurs pour appréhender son environnement, il contient des programmes pour définir une réponse à une situation donnée, et il a les moyens de mettre en œuvre cette réponse[3]. Si cette définition donne une première approche intéressante, elle n'est malheureusement pas assez précise, et donc pas assez restrictive pour cerner la nature d'une technologie autonome. Car un missile anti-aérien, doté d'un système auto-directeur, serait selon cette définition un robot : fabriqué par l'homme, possédant un senseur infrarouge, contenant un programme pour déterminer sa trajectoire afin de suivre sa cible, possédant des ailettes amovibles pour appliquer les changements de trajectoires établis. Ce n'est pourtant pas le missile qui décide d'ouvrir le feu, et encore moins de la cible à traiter. Il faut donc préciser cette définition.

L'organisation non gouvernementale *Human Rights Watch*, quant à elle, définit le robot[4], qu'elle associe à une arme dans sa lutte contre les robot-tueurs, comme une machine pouvant opérer de manière autonome dans la sélection des cibles. Mais elle classe le niveau d'autonomie, et donc d'intervention humaine dans le processus d'ouverture du feu, selon trois catégories. Les armes « *human-in-the-loop* » sont celles où le robot sélectionne les cibles mais pour lesquelles l'homme commande l'ouverture du feu. Les armes « *human-on-the-loop* » sont celles où le robot sélectionne les cibles et ouvre le feu sur elles, mais pour lesquelles

[2] KRISHNAN Armin, *Killer Robots : legality and ethicality of autonomous weapons*, Burlington, Ashgate publishing company, 2009, 204 p.

[3] SINGER Peter W., *Wired for war: the robotics revolution and conflict in the 21st century*, New York, Penguin Books, 2009, 499 p.

[4] HUMAN RIGHTS WATCH, *Losing humanity: the case against killer robots*, HRW, 2012, 49 p.

l'homme peut intervenir et annuler la décision. Enfin, les armes « *human-out-of-the-loop* » sont celles où le robot sélectionne et traite les cibles sans qu'aucun humain ne puisse intervenir. Le professeur américain Armin Krishnan reprend en quelque sorte ces trois catégories[5], en les nommant respectivement autonomies préprogrammée, supervisée et complète, mais rajoute une catégorie première qui est l'autonomie télé-opérée. Il embrasse ainsi tous les systèmes d'armes existants et peut les classer selon leur degré d'autonomie. Mais c'est là un biais de jugement, car le terme autonomie est alors écarté de son sens premier. La définition du Larousse spécifie en effet la « *capacité de quelqu'un à être autonome, à ne pas être dépendant d'autrui ; caractère de quelque chose qui fonctionne ou évolue indépendamment d'autre chose* ». Un système télé-opéré ne peut donc pas, par définition, être autonome.

Jean-Baptiste Jeangène Vilmer, en rappelant la distinction fondamentale entre autonomie et automaticité[6], confirme qu'une technologie autonome doit être différenciée des systèmes d'armes automatisés déjà existants. Il s'agit donc, en respectant la distinction de notion, de parler de *systèmes d'armes autonomes*. Et étant donné qu'il s'agit de technologie destinée au combat, c'est-à-dire capable de délivrer un feu mortel, nous parlerons de *Système d'Armes Létal Autonome* (SALA). Jean-Baptiste Jeangène Vilmer en propose une définition bien plus précise et réaliste, en reprenant les notions déjà abordées : « *un système d'armes qui, une fois activé, est capable de décider seul, c'est-à-dire sans intervention ni supervision humaine, du ciblage et du déclenchement de la frappe, en fonction d'un environnement changeant auquel il s'adapte* »[7]. Ce qui sous-entend premièrement que le SALA n'appartient pas exclusivement à l'une ou l'autre des catégories d'autonomie proposées par *Human Rights Watch*, mais qu'il les embrasse toutes par un système hybride où l'humain peut être, selon les cas, *in/on/out-of-the-loop* ; deuxièmement que l'autonomie exige une intelligence artificielle capable d'apprendre par expérience et de s'adapter à une situation nouvelle, ce qui est le véritable défi des recherches en intelligence artificielle.

Car le champ de bataille n'est pas, contrairement à ce que pourrait laisser croire l'analogie habituelle, aussi simple et normé qu'un plateau d'échecs. Si l'intelligence artificielle est

---

[5] KRISHNAN Armin, *op. cit.*

[6] JEANGENE VILMER Jean-Baptiste, *Diplomatie des armes autonomes : les débats de Genève*, in *Politique étrangère* Automne, n° 3, 2013, pp 119 - 130.

[7] JEANGENE VILMER Jean-Baptiste, *Terminator Ethics : faut-il interdire les « robots tueurs » ?*, in *Politique étrangère* Hiver, n° 4, 2014, pp 151 – 167.

capable depuis peu de surpasser l'homme au plus classique des jeux de réflexion, elle n'est pas encore prête à affronter la complexité mouvante d'une situation de guerre. Le champ de bataille est désormais multidimensionnel, puisqu'aux espaces de conflit habituels, terrestre, maritime et aérien, s'ajoutent l'espace exo-atmosphérique et le cyberespace. Un SALA pourra donc être conçu pour combattre dans l'une ou l'autre de ces dimensions, voire être capable de le faire dans plusieurs d'entre elles, successivement ou simultanément, notamment pour le cyberespace. Mais comment définir le champ de bataille ? Littéralement, il s'agit du lieu où se livre une bataille. Or, selon le Larousse, la bataille est un « *combat livré entre deux armées ou deux flottes, dont l'issue influe sur la conduite de la guerre* ». Mais si l'on considère la portion d'espace géographique, dans la ou les dimensions envisagées, couverte par la portée des différentes armes, nous atteignons rapidement la planète toute entière. Il faut donc restreindre la définition pour que la question du remplacement du soldat humain par une technologie autonome ait un sens. Le théâtre d'opérations, correspondant bien souvent aux frontières d'une région ou d'un ou plusieurs pays, est encore une notion trop englobante. En effet, avant que le niveau d'autonomie d'un SALA n'atteigne son apogée, correspondant à sa capacité de se réparer tout seul, ce dernier aura besoin d'une base de maintenance sur le théâtre, sans même parler de l'Etat-major chargé du commandement de l'opération. Nous définirons donc le champ de bataille, quelle que soit la ou les dimensions considérées, comme étant *l'espace temporaire et flexible qui englobe tous les compartiments de terrain dans lesquels au moins un soldat ami, humain ou SALA, se retrouve en contact visuel direct avec au moins un soldat ennemi*. Cette définition, plus restrictive encore que l'espace dédié à une zone de responsabilité tactique par exemple, a l'avantage de rendre pertinente la question de l'abandon de l'homme au profit du SALA sur le champ de bataille.

Cette question de l'emploi du SALA a d'ores et déjà fait couler beaucoup d'encre. Mais les débats se concentrent essentiellement sur l'aspect juridique de leur emploi d'une part, et sur les considérations morales de l'utilisation d'une machine pour tuer des êtres humains d'autre part. C'est-à-dire sur le « *jus in bello* » et sur l' « *ethicis ad bellum* ». Si le premier débat n'aboutit pour l'instant à aucune conclusion satisfaisante, le second devient passionnel, en particulier dans le monde anglo-saxon, et en arrive à des propositions extrêmes. *Human Rights Watch* en appelle ainsi au bannissement universel pur et simple, de manière préventive, de tout SALA[8]. William Joy, informaticien ayant participé au développement du système d'exploitation Unix, est persuadé que la robotique entraînera riens moins que la fin de

[8] HUMAN RIGHTS WATCH, *op. cit.*

l'humanité[9]. D'un autre côté, de nombreux auteurs sont persuadés du besoin des robots pour remplacer des soldats humains dépassés par la technologie dans les guerres futures[10], que les SALA seront plus précis dans la sélection des cibles[11], et même qu'ils « *pourraient à terme mieux respecter le droit de la guerre que les humains* »[12]. L'écart reste immense entre un scénario de fin du monde et celui d'une humanité améliorée par la maîtrise de sa technologie. Mais, comme tout débat passionnel, il devient difficile de demeurer rationnel et objectif. Certains arguments tiennent ainsi inévitablement, même si leurs auteurs s'en défendent, d'un néo-luddisme qui bloque le débat plus qu'il n'aide à le faire avancer. Car le réalisme nous dicte que les industries technologiques développeront des SALA, ne serait-ce que pour la bonne raison que les applications issues de ces recherches seront fortement duales. Partant de ce postulat, il vaut mieux réfléchir au cadre éthique dans lequel développer ces machines, et tenter de fixer les limites techniques nécessaires pour ne pas se laisser dépasser par la ferveur des découvertes scientifiques[13]. Il faut ainsi se projeter dans un futur proche, où les SALA sont une réalité, et étudier l' « *ethicis in bello* » pour déterminer si ces robots-tueurs seront seulement capables de remplacer, ou à défaut d'accompagner les soldats humains sur le champ de bataille.

Nous analyserons ainsi le comportement humain au combat, en particulier en cet instant paroxystique qu'est la décision d'ouvrir le feu pour tuer, et tenterons de le comparer au comportement possible d'un SALA. A partir des défauts puis des vertus humaines sur le champ de bataille, nous pourrons ainsi dresser un cahier des charges des capacités morales minimales nécessaires pour qu'un SALA soit « moralement acceptable ».

Partant de là, nous tenterons d'expliquer le processus de raisonnement moral du soldat lors d'une décision de tuer, et ainsi de définir la notion de discernement éthique au combat, pour ensuite proposer la programmation d'une éthique computationnelle qui serait le « cerveau moral » du SALA.

---

[9] JOY William N., *Why the future doesn't need us*, in *Wired Magazine*, 2000.

[10] KRISHNAN Armin, *op. cit.*

[11] THURNHER Jeffrey S., *Means and Methods of the Future: Autonomous Systems*, in *Targeting: The Challenges of Modern Warfare*, édité par Paul A. L. Ducheine, Michael N. Schmitt, et Frans P. B. Osinga, T.M.C. Asser Press, 2016, pp. 177-199.

[12] JEANGENE VILMER Jean-Baptiste, *op. cit.*

[13] ERBLAND Brice*, La tentation de l'hubris*, in *Inflexions* n°32, mai 2016.

Enfin, nous tenterons de prévenir les conséquences de l'emploi d'une telle machine au combat, aux niveaux tactique et stratégique, avant de s'inspirer des visions de la littérature sur le sujet.

On peut lire ici et là que le problème des études sur l'éthique des robots-tueurs est que les philosophes ne sont pas programmateurs et que les programmateurs ne sont pas philosophes. Mais le véritable frein à cette étude, c'est que ni le philosophe ni le programmateur ne connaissent le combat. Or, l'éthique du combat requiert une part empirique pour être complètement appréhendée. Les réflexions d'un soldat ayant une expérience de la guerre sur le sujet peut donc apporter un éclairage nouveau.

Tout au long de cette étude, nous tâcherons de tirer des déductions qui se voudront autant de recommandations concrètes. Ces déductions, prises une par une, sont partielles et incomplètes. Elles traduisent le raisonnement en cours. Une synthèse en sera faite à la fin de l'étude, et elles seront regroupées, triées et comparées en annexe, afin de dresser *in fine* un « portrait-robot » du bon SALA.

I. EMOTIONS vs ALGORITHMES

Les partis-pris sont par définition subjectifs, c'est pourquoi les arguments en faveur ou contre le SALA sont souvent présentés d'une manière exagérée, en tenant pour acquis des faits qui ne sont que de vagues hypothèses, étant donné qu'aucun retour d'expérience n'est encore possible. Ces défenseurs du pro ou de l'anti tombent invariablement dans un prosélytisme caricatural par des arguments trop tranchés. Ainsi, le SALA serait le seul capable de suivre le rythme du combat futur, sans aucun besoin de repos ou de subsistance[14]. Ou, *a contrario*, le SALA ne saura jamais faire la différence entre un civil apeuré et un combattant menaçant, parce que cela requiert de comprendre l'intention de la personne observée[15]. La réalité paraît, dans un sens comme dans l'autre, un peu plus nuancée que cela. Il faut pourtant bien réfléchir aux avantages ou inconvénients d'une telle technologie par rapport au soldat humain, même si elle n'existe pas encore. Mais plutôt que d'établir des théories globales sur l'efficacité de l'un ou de l'autre, tentons de rentrer en détail dans la psychologie du combat. Analysons une par une les faiblesses et les vertus du combattant humain, et tentons d'imaginer quelle serait la part de faiblesse ou vertu équivalente pour un SALA. Nous pourrons ainsi déterminer, selon les manques apparents, quel serait le contrat minimal d'une intelligence artificielle pour qu'un SALA puisse être au moins aussi efficace qu'un soldat humain sur le champ de bataille. En d'autres termes, nous aurons défini l'état final recherché d'une éthique artificielle acceptable.

[14] THURNHER Jeffrey S., *op. cit.*

[15] HUMAN RIGHTS WATCH, *op. cit.*

FAIBLESSES HUMAINES AU COMBAT

« *Comme la musique et la justice, l'humanisme semble déchoir lorsque l'épithète militaire le qualifie* »[16]. Le soldat humain est en effet une imperfection de nature. La qualité de son entraînement, de ses forces morales, la cohésion de sa troupe et les qualités collectives de son unité atténuent cette imperfection. Mais quelque fois, parce qu'il est humain et que la guerre réveille la noirceur des âmes, il fait preuve de fautes morales. « *Les armes remuent au fond des cœurs la fange des pires instincts* », écrivait Charles de Gaulle dans *le fil de l'épée*. A en croire Ronald Arkin, les faiblesses humaines, ces « pires instincts », sont issues des émotions ; or, les SALA n'auront pas d'émotions, et donc pas de travers moraux[17]. Mais il reste possible que par la structure même de leur intelligence artificielle, certaines de ces faiblesses soient reproduites, non par nature, mais par similitude. Car la complexité des intelligences artificielles dont nous parlons pourrait déboucher sur des comportements que l'on aurait du mal à comprendre[18], et donc à anticiper. Et à cela s'ajoute les faiblesses cognitives, trop souvent rapidement balayées par l'assurance de l'infaillibilité de la technologie. Voyons quels sont toutes ces faiblesses humaines, et surtout posons-nous la question : un SALA en serait-il véritablement dénué ?

Vengeance

Peut-être est-ce là le premier des mauvais instincts humains, l'héritage ancestral d'un réflexe primaire presque animal : l'appel à la vengeance. Quoi de plus naturel pour une troupe au combat que de vouloir venger ceux de leurs camarades qui sont tombés face à l'ennemi ? En un sens, cet instinct décuple la hargne d'une unité au combat, et n'est donc pas totalement néfaste. Mais ce qui peut valoir pour un conflit symétrique dans une zone déserte ne vaut pas pour une guerre asymétrique au milieu de la population. Dans ce cas précis, qui semble devenir le paradigme de la plupart des prochains conflits[19], le sentiment de vengeance pousse

[16] DREVILLON Hervé, *L'individu et la guerre : du chevalier Bayard au Soldat Inconnu*, Paris, Belin, 2013, 306 p., p. 15.

[17] ARKIN Ronald C., *Systèmes automatisés capables de raisonnement éthique*, in *Les drones aériens : passé, présent et avenir*, Paris, La documentation Française, 2013, pp. 587 – 598.

[18] KRISHNAN Armin, *op. cit.*

[19] SMITH Rupert, *L'utilité de la force : l'art de la guerre aujourd'hui*, Paris, Economica, 2007, 395 p.

le soldat à reconnaître des ennemis partout où il voit des autochtones. L'honneur de son unité, qui a perdu un ou plusieurs des siens, le pousse inconsciemment à assouvir cette vengeance[20]. Peu importe, donc, qu'il y ait un doute sur l'identification de l'ennemi ; il faut que le sang coule pour apaiser la haine. Bien entendu, cet instinct se contrôle. Par la loi, en premier lieu, car les règles d'engagement sont autant de filtres pour éviter au soldat de tomber dans ce piège. Par la discipline, ensuite, à condition que le chef tactique ne tombe pas lui-même dans le piège. Par les forces morales, enfin, qui permettent à l'homme de placer la raison au-delà de l'instinct. Mais malgré ces protections, il se peut que le soldat humain cède à la vengeance, et tue un civil par manque de discernement, la raison voilée par l'instinct primaire, car « *celui qui agit dans un esprit de revanche peut se sentir habilité à répondre plus fort et à commettre des brutalités plus cruelles et largement disproportionnées par rapport à l'agression subie* »[21].

Parce que le SALA n'aura sans doute pas d'attachement émotionnel pour les soldats humains ou d'autres SALAs qui pourraient l'accompagner, il est fort peu probable qu'une intelligence artificielle développe un besoin de venger la perte amie. Même si elle comprend une structure inductive, qui « apprend » par l'expérience, il n'y a *a priori* aucune raison pour que, sans ordre particulier, cette intelligence artificielle décide d'outrepasser les règles d'engagement pour tuer à tout prix et calmer l'appel du sang de ses camarades humains. Le sentiment de vengeance ne sera donc probablement pas partagé par un SALA.

Addiction à la destruction

Le pouvoir de destruction est puérilement jouissif : tel un enfant qui se rend compte, après avoir écrasé une fourmi avec son pouce, qu'il est devenu tout-puissant face à la colonne d'insectes qui se déplace devant lui, le soldat éprouve une forme de jouissance au moment de détruire une cible. Cette jouissance est issue de la satisfaction de la mission remplie, mais également de la simple victoire dans le duel guerrier : si l'on tue l'ennemi, lui ne pourra pas nous tuer. Le soldat qui découvre ce pouvoir, et le plaisir immédiat qui en est associé, peut en devenir dépendant, en quelque sorte, au fur et à mesure de son usage[22]. Ronald Arkin nomme

[20] ERBLAND Brice, *Le processus homicide : analyse empirique de l'acte de tuer*, in *Inflexions* n° 31, janvier 2016.

[21] SLIM Hugo, *Les civils dans la guerre : identifier et casser les logiques de violence*, Genève, éditions Labor et Fides, 2009, 373 p., p. 181.

[22] ERBLAND Brice, *op. cit.*

ce travers le plaisir de tuer[23], mais la conséquence en reste la même : à force d'addiction, la destruction devient un besoin, et l'ouverture du feu devient systématique. « *La violence en elle-même semble capable de nous enivrer* »[24]. Même si, comme toute drogue, le plaisir n'est qu'instantané et laisse place à un malaise profond, cette dérive morale peut engendrer un usage de la force abusif.

Parce qu'un SALA ne pourra pas, sauf si on le programme pour cela, ressentir du plaisir lorsqu'il détruira une cible ennemie, il ne pourra *a priori* pas tomber dans cette addiction à la destruction. Sauf que dans le cas, très probable, où son intelligence artificielle comprend un système capable d'adapter son comportement au fur et à mesure des expériences – en d'autres termes, capable d'apprendre de ses expériences – le SALA pourra très bien développer une dérive similaire. Imaginons que dès ses premiers emplois, le SALA soit obligé d'ouvrir le feu, ou qu'il observe ses camarades humains le faire. Sa base de données casuistique établira un lien entre les différentes situations vécues et l'ouverture du feu, comme autant d'exemples où la destruction était la solution face aux situations rencontrées. Cet « apprentissage » encouragera le choix de la même solution face à une situation nouvelle. En effet, le principe d'un système inductif, qui sera développé en deuxième partie, repose sur la modification du raisonnement par les situations déjà vécues et enregistrées en base de données. Un SALA qui devra ouvrir le feu régulièrement le fera donc de plus en plus, non pas par addiction comme pour l'être humain, mais par agrégat de solutions prises en exemple.

Pour se prémunir de cette dérive comportementale, il faudra donc que chaque action d'un SALA soit analysée par l'homme, à l'image de la chaîne de vérification des compte-rendus d'ouverture du feu en opération extérieure. L'analyste humain pourra ainsi, au vu de la situation et de l'action choisie par l'intelligence artificielle, valider ou au contraire condamner la décision prise, afin que la base de données du SALA mette à jour les solutions enregistrées. Ainsi, par le contrôle de ses expériences, un usage abusif de la force par le SALA pourra être immédiatement stoppé.

Déduction n°1 : *A la fin de chaque opération, une analyse des choix de l'intelligence artificielle du SALA devra être effectuée pour valider la moralité de chaque action et mettre à jour sa base de données d'expériences.*

[23] ARKIN Ronald C., *Systèmes automatisés capables de raisonnement éthique*, in *Les drones aériens : passé, présent et avenir*, Paris, La documentation Française, 2013, pp. 587 – 598.

[24] SLIM Hugo, *op. cit.*, p. 288.

Effet Lucifer

A l'image de la célèbre expérience du professeur Milgram au début des années 1960, durant laquelle près de 65% des personnes testées ont infligé une douleur létale à un tiers sous couvert d'obéissance à une autorité, le soldat au combat peut effectuer des actes qui vont bien au-delà de ses barrières morales par soumission à l'autorité. Etant donné le sens hiérarchique très fort d'une institution militaire et le poids de l'esprit de corps d'une unité, qui incite à suivre les autres et rend beaucoup plus dur l'objection de conscience, un soldat peut très vite être pris par l' « effet Lucifer ». Le docteur Patrick Clervoy, dans son livre éponyme[25], en explique la teneur et les différentes dérives morales qui peuvent en découler. Lorsqu'une telle dérive s'installe dans un groupe soudé par la cohésion, il faut ainsi une intervention extérieure pour faire prendre conscience aux personnes impliquées de l'immoralité de leurs actes. Lorsqu'on ajoute à cela une déshumanisation de l'ennemi[26] par un sentiment de supériorité technologique, par une trop grande distance physique ou encore par une habitude de moqueries et de propos haineux, l'effet Lucifer peut s'installer très rapidement. « *La première étape capitale pour amener un être humain à en tuer d'autres consiste à lui faire penser qu'il tue quelque chose de différent* »[27]. Là encore, un usage abusif de la force en est la conséquence directe, comme lors de l'assaut du village vietnamien de My Lai le 16 mars 1968 par une compagnie de fantassins américains. Alors que seuls des femmes, des enfants et des vieillards sont présents dans le village, le lieutenant qui commande la troupe ordonne le massacre de la population en montrant l'exemple : malgré quelques réticences de la part de certains soldats, cinq cents personnes sont tuées[28].

Il est difficilement envisageable de créer un SALA qui n'obéirait pas aux ordres reçus. L'autonomie de la machine n'implique aucunement qu'elle ne prenne pas en compte des ordres en cours d'action. « *Certes il pourrait se produire qu'un robot échappe à tout contrôle, mais cela dit bien d'abord que même quand ils sont autonomes les robots sont contrôlés par ceux qui les utilisent* »[29]. Mais il est tout aussi difficile d'envisager un SALA qui obéirait

[25] CLERVOY Patrick, *L'effet Lucifer: du décrochage du sens moral à l'épidémie du mal*, Paris, CNRS éditions, 2013, 333 p.

[26] ARKIN Ronald C., *op. cit.*

[27] SLIM Hugo, *op. cit.*, p. 270.

[28] CLERVOY Patrick, *ibid.*

[29] FAES Hubert, *Une éthique pour les robots tueurs ?*, in *Revue d'éthique et de théologie morale*, n° 289 (23 juin

aveuglément à tout ordre donné. La question de l'acte immoral effectué par obéissance à une autorité est donc essentielle en ce qui concerne le SALA. Car « *nous ne devrions pas avoir peur des robots, mais de ce que certains d'entre nous pourraient en faire* »[30]. Il faut donc pouvoir s'assurer qu'un SALA puisse non seulement refuser un ordre illégal, mais également un ordre immoral. Cela implique qu'il puisse juger de la moralité de l'acte, et donc qu'il soit capable d'un raisonnement éthique. Sans creuser plus avant ce qui sera l'objet de la deuxième partie de cette étude, cette autonomie morale ne sera sans doute pas infaillible, puisqu'elle sera quoi qu'il arrive subjective, comme l'est par définition tout jugement. Le curseur sera donc dur à placer, comme il l'est déjà pour les humains, entre obéissance nécessaire et discernement moral. Nous pouvons néanmoins affirmer qu'il est possible qu'un SALA subisse l' « effet Lucifer » par soumission à l'autorité, et en arrive par ce biais à un usage abusif de la force.

Déduction n°2 : *un SALA devra être capable de refuser un ordre illégal ou un ordre immoral.*

Distanciation

« *Quand on tue à une distance très importante de la cible, on peut se persuader que ce ne sont pas des êtres humains que l'on tue* »[31]. Cette distanciation de l'acte de tuer provient des possibilités nouvelles des armes modernes : les dispositifs de visée et les conduites de tir, associées aux portées grandissantes, permettent à certains soldats de tuer sans voir physiquement et directement leur cible. Que ce soit du fait de la distance ou du caractère indirect (par écran interposé) de l'ouverture du feu, la perception du poids de l'acte peut s'en trouver diminuée. « *Le fait que le tueur et sa victime ne soient pas inscrits dans des « champs perceptifs réciproques » facilite l'administration de la violence. Cela épargne à l'agent la gêne ou la honte qui peut naître de se voir agir dans les yeux de l'autre* »[32]. Et moins l'on voie la cible et le résultat du tir, moins les conséquences psychologiques du fait d'avoir tué se

2016): 107‑15.

[30] BRINGSJORD Selmer, *Ethical Robots : the future can heed us*, in *AI & Society – special issue : Ethics and artificial agents*, vol. 22, 2008, pp 539 – 550.

[31] ARKIN Ronald C., *ibid.*

[32] CHAMAYOU Grégoire, *Théorie du drone*, Paris, La Fabrique Editions, 2013, 363 p., p. 167.

font sentir[33]. Cette distanciation facilite donc doublement l'usage des armes, et participe grandement à la déshumanisation de l'ennemi par le voile qu'elle dresse sur l'identité humaine de la cible. En effet, tuer à travers un écran peut entraîner une « mentalité Play Station », selon laquelle « *le dispositif du meurtre à l'écran entraîne une virtualisation de la conscience de l'homicide* »[34]. Distance physique ou distance vis-à-vis de la réalité, les différentes formes de la distanciation par rapport à l'ennemi entraînent toutes un déni inconscient de la valeur humaine de la cible. Or, cette déshumanisation est en quelque sorte un catalyseur pour toutes les dérives morales déjà énoncées. Et parce qu'elle facilite l'usage des armes, elle peut là aussi entraîner un usage abusif de la force.

Heureusement, cette distanciation est un effet de perception et de conscience humaines. Un SALA ne ferait aucune différence entre une cible se situant à un mètre de lui et une cible évoluant à plusieurs kilomètres de distance. En effet, une fois l'identité humaine perçue par la machine, la notion de distance n'est qu'une information technique complémentaire, utile au réglage de la portée des armes. Mais elle n'influe pas sur le « jugement » de l'intelligence artificielle, en dehors de la qualité, sans doute décroissante avec la distance, de la perception de ses senseurs. De plus, le SALA n'aura pas « conscience » de la valeur d'une vie humaine. Il ne fera que classer les différents êtres humains identifiés autour de lui par catégories. Tout au plus les analysera-t-il avec plus d'importance qu'un objet inanimé, parce qu'il sera programmé pour cela. Mais il sera incapable d'éprouver de l'empathie pour un être humain. Paradoxalement, cette inconscience le protégera des effets de la déshumanisation et de la distanciation, puisque tout homme ne sera pour lui qu'une ligne de code, une simple information.

Esclavage technologique

L'être humain est aisément crédule face à la technologie. Ne sommes-nous pas tous tentés, en voiture, de suivre les indications du GPS, même si les panneaux indicateurs nous désignent un chemin différent ? Selon le professeur anglais Luciano Floridi[35], nous sommes face à la quatrième révolution pour l'être humain, après Copernic, Darwin et Freud, qui ont

[33] ERBLAND Brice, *Le processus homicide : analyse empirique de l'acte de tuer*, in *Inflexions* n° 31, janvier 2016.

[34] CHAMAYOU Grégoire, *ibid*, p. 153.

[35] FLORIDI Luciano, *The Fourth Revolution: How the Infosphere Is Reshaping Human Reality*, Oxford, Oxford University Press, 2014, 272 p.

respectivement remis en question la place centrale de l'homme dans l'univers, dans le monde animal et dans sa propre conscience. A force de confier à la technologie le stockage de nos connaissances, et de plus en plus l'analyse et l'aide à la décision, nous devenons tributaires de notre propre technologie. En d'autres mots, nous en devenons les esclaves. Au combat, cet esclavage technologique entraîne un risque de dommage collatéral par une confiance aveugle en l'analyse des outils technologiques[36]. En effet, la réalité augmentée de l'espace numérique de bataille, dans lequel évolue le soldat, peut l'entraîner dans une vision de la réalité qui n'est pas la réalité. Peter W. Singer en donne un exemple célèbre[37]. En 1988, un bâtiment de guerre américain était équipé du système *Aegis*, capable de détecter et d'identifier des avions ennemis, puis d'effectuer la séquence de tir de manière totalement autonome. Le mode semi-automatique était cependant préservé, afin de laisser au commandant du navire la responsabilité de la vérification de la cible et de la validation du tir. Le 03 juillet de la même année, le système *Aegis* détecte un F14 iranien et conseille la destruction de l'appareil. Pourtant, le signal IFF de l'appareil indique qu'il s'agit d'un vol long-courrier. Mais l'équipage choisit de faire confiance au système. Tous les passagers du vol Iran Air Flight 655 meurent lorsque le missile atteint sa cible. Il faut à l'homme un fort pouvoir de discernement pour s'arracher à cet esclavage technologique.

Bien entendu, un SALA utilisé en mode « *human-in-the-loop* » sera vecteur, pour l'agent humain, de cet esclavage technologique. D'autant plus que son intelligence artificielle prendra sans doute en compte des éléments que l'être humain ne pourra pas mesurer, surtout à distance. Ce dernier aura donc tendance à accorder sa confiance au jugement émit par la machine. Et lorsque le SALA est autonome, c'est-à-dire en mode « *human-out-of-the-loop* », il devient exactement comme le système *Aegis*, et est donc susceptible d'effectuer une erreur d'identification ou de ciblage. Sauf si son intelligence artificielle comporte un module d'analyse éthique des décisions, qui remettrait en cause chaque décision pour en juger la valeur morale. Paradoxalement, si un tel système est réalisable, il serait sans doute plus efficace qu'un jugement humain altéré par l'esclavage technologique. L'autonomie éthique du SALA est donc la réponse absolue à cet esclavage technologique. Comme l'écrit Robert Sparrow[38], la meilleure définition d'une technologie autonome serait peut-être qu'elle soit

[36] ERBLAND Brice*, La tentation de l'hubris*, in *Inflexions* n°32, mai 2016.

[37] SINGER Peter W., *op. cit.*

[38] SPARROW Robert, *Robots and respect : assessing the case against autonomous weapon systems*, in *Ethic & International Affairs* n°2016/1, pp. 93 – 116.

moralement autonome, c'est-à-dire avec une volonté propre et responsable de ses actions. Le bon SALA sera donc un SALMA : un Système d'Armes Létal Moralement Autonome.

Déduction n°3 : *le SALA devra être moralement autonome, c'est-à-dire doté d'un module de jugement éthique de chaque décision.*

Peur

L'être humain a développé un système d'alarme interne très efficace pour augmenter ses chances de survie : la peur. Cette émotion lui permet une hyper-vigilance d'une part, et déclenche un instinct de prudence qui lui évite de se jeter face au danger. Elle est donc, la plupart du temps, plutôt avantageuse. Mais il arrive que son intensité soit telle qu'elle en devient bloquante. Une telle peur peut engendrer pour le soldat une paralysie psychologique complète, un refus d'avancer.

La plupart des réflexions sur le SALA présuppose que la machine pourra se sacrifier, au profit de soldats humains par exemple, puisqu'elle ne sera pas freinée par l'instinct de survie[39]. L'argument implique que le SALA soit un « *objet fabriqué en série et facilement remplaçable* »[40]. Mais il faudra sans doute du temps avant qu'une telle technologie soit « facilement remplaçable », ne serait-ce qu'à cause du coût que représenterait un SALA. Il n'est donc pas impossible que le SALA comporte dans sa programmation une réticence à s'exposer inutilement au danger. En un sens, à l'image des trois principes de la robotique développé par Isaac Asimov[41], le SALA ne pourrait s'exposer sciemment à la destruction que dans le cas d'un danger imminent pour un être humain. Mais dès lors qu'un programme d'autoprotection face au danger est intégré au SALA, ce dernier peut se retrouver bloqué tant qu'aucun soldat ou civil humain n'est en danger. Il reste donc possible, dans cette éventualité, qu'un SALA soit paralysé au combat, parce qu'il aura jugé qu'un déplacement engendrerait un dommage mais qu'aucun être humain ne se trouve en danger. Il cherchera donc à préserver son intégrité physique.

Stress post-traumatique

---

[39] ARKIN Ronald C., *ibid.*

[40] TISSERON Serge, *Des robots et des hommes : lesquels craindre ?*, in *Études* n° 11, novembre 2014 : 33‑44.

[41] ASIMOV Isaac, Les Robots, trad. BILLON P., Paris, J'ai lu, 1967, 285 p.

Pour l'être humain, une situation extrême face à laquelle il développe des émotions d'une très forte intensité peut entraîner un syndrome de stress post-traumatique. D'intensité variable, ce syndrome peut fort bien influencer l'attitude et le raisonnement du soldat qui se retrouve dans une situation similaire à celle qui a fait naître le traumatisme, que ce soit au combat ou à l'entraînement. Un événement qui pourrait paraître anodin ramène alors les émotions extrêmes ressenties au-devant de la scène, et le soldat se retrouve paralysé alors qu'aucune menace imminente n'est perçue.

De la même façon qu'une intelligence artificielle inductive, basée sur la mémoire des expériences passées, peut engendrer une forme d'addiction à la destruction, un phénomène pouvant se comparer au stress post-traumatique peut se développer suite à des expériences malheureuses. Imaginons qu'une action d'un SALA entraîne un jour une conséquence qu'il catalogue comme néfaste, tels la mort d'un soldat ami, la mort de civils ou encore un dommage sur sa propre carcasse. Dans une situation similaire, son intelligence artificielle fera remonter le risque assimilé par cette expérience, et influencera peut-être négativement sa réaction par une forme d'inhibition. Il faudra donc garder le contrôle sur chaque souvenir enregistré par le SALA, et surtout sur la classification qui en sera faite, puisque cela aura un rôle déterminant sur son raisonnement.

Déduction n°4 : *un « souvenir » négatif doit pouvoir être effacé de la mémoire d'un SALA pour ne pas qu'il influe sur son raisonnement futur.*

Besoins physiologiques (Fatigue, ravitaillement)

La plupart des études sur le sujet soulignent l'absence de besoins physiologiques pour un SALA. L'être humain a en effet de lourds besoins entre le sommeil, l'eau et la nourriture[42]. L'homme en devient faillible : le besoin de sommeil l'empêche d'assurer une mission de manière permanente et ses facultés cognitives lui confèrent une capacité de concentration limitée[43]. Le SALA donne donc, en comparaison, la promesse d'un soldat sans faille et sans contrainte.

Mais cette promesse est illusoire et irréaliste. Car il n'existe aucun rouage qui n'ait besoin d'huile, aucun moteur qui n'ait besoin de carburant, aucune batterie qui n'ait besoin d'être

[42] THURNHER Jeffrey S., *op. cit.*

[43] KRISHNAN Armin, *op. cit.*

rechargée, aucun logiciel informatique qui ne déraille jamais. Le SALA aura également des besoins propres, liés à la maintenance de ses parties mécaniques. Et son cerveau informatique, exposé aux variations de température, au sable ou à l'humidité des théâtres d'opération, pourra montrer des défaillances. Un logiciel est considéré fiable dès lors qu'il ne comprend « que » entre 10 et 50 bugs par millier de lignes de code. On imagine aisément la complexité d'un logiciel d'intelligence artificielle capable d'assumer les missions que l'on confierait à un SALA. Derrière cette complexité se cacheront toutes les possibilités de défaillance. Il n'est donc pas improbable d'imaginer un SALA posté en sentinelle laisser passer un groupe ennemi en infiltration parce que son système interne était en redémarrage… Le syndrome du « blue screen » n'est pas près d'être enterré en informatique. Peter W. Singer rappelle l'existence de deux lois empiriques que sont les lois de Moore et de Murphy. La première établit que les capacités informatiques doublent à coût constant chaque année. La seconde établit que si quelque chose est susceptible de mal tourner, elle finira nécessairement pas mal tourner un jour. En d'autres termes, si la loi de Moore s'applique en informatique, celle de Murphy n'est pas en reste[44]. Ce qui signifie qu'un SALA ne sera jamais infaillible, et que si ses besoins en ravitaillement seront plus faibles que ceux nécessaires à un être humain, il pourra lui aussi montrer des signes de faiblesse dans l'exécution de ses missions.

[44] SINGER Peter W., *La guerre connectée : les implications de la révolution robotique*, in *Politique étrangère* Automne, n° 3, 2013 : 91‑104.

Faiblesse humaine	Equivalent pour un SALA	Déduction
Sentiment de vengeance	*A priori* impossible	-
Addiction à la destruction	Possible, à cause d'un raisonnement prenant en compte les expériences passées	Chaque choix effectué par un SALA doit par la suite être validé moralement par un humain
Effet Lucifer	Possible	Un SALA doit pouvoir désobéir
Distanciation	*A priori* impossible	-
Esclavage technologique	Vecteur de cet esclavage par nature	Un SALA doit être doté d'un module de jugement éthique autonome
Peur	Possible par souci de respect de l'intégrité physique	-
Stress post-traumatique	Possible par influence d'une expérience néfaste	Un souvenir néfaste doit pouvoir être effacé de la base de données mémoire
Besoins physiologiques	Plus faibles, mais présents	-

Figure 1

Il apparaît donc que ces faiblesses humaines au combat, pouvant altérer l'efficacité tactique ou engendrer un usage abusif de la force, peuvent être partagées par le SALA. Si elles ne sont pas identiques puisqu'elles ne sont pas issues d'une conscience ou d'un instinct chez le SALA, elles peuvent être similaires pour des raisons purement logicielles. Il est donc important de les identifier afin d'anticiper dès aujourd'hui les garde-fous nécessaires lors de la conception même de leur intelligence artificielle. La figure 1 résume ainsi l'ensemble des faiblesses humaines identifiées au combat, avec leur équivalent pour un SALA et les déductions faites pour s'en prémunir.

VERTUS HUMAINES AU COMBAT

Fort heureusement, l'homme au combat n'est pas qu'un être moralement faible ou déficient. Contrairement aux travaux de Ronald Arkin[45], qui font référence, il ne s'agit donc pas de passer en revue uniquement les défaillances humaines au combat ; il faut aussi étudier les forces particulières du soldat humain et réfléchir à la possibilité d'une similitude pour le SALA. Car la guerre, dans toute son horreur, exacerbe également certaines vertus de l'homme. Le combat force parfois le soldat à puiser dans des ressources insoupçonnées que sont les forces morales. Car s'il agit la plupart du temps « en connaissance » des lois, des règles d'engagement et des procédures tactiques, il doit parfois agir « avec intelligence » pour innover, et surtout « en conscience » pour discerner moralement ses actes. Ces trois piliers de l'action du soldat sont les garants d'une décision moralement acceptable, comme nous le verrons dans la deuxième partie. Or, si le SALA pourra sans doute agir « en connaissance », il est moins évident qu'il puisse le faire « avec intelligence » et « en conscience ».

Courage

La première des forces morales qui vient à l'esprit est le courage. Il est souvent défini comme la capacité qu'ont les hommes ordinaires de faire des choses extraordinaires, alors que la peur paralyserait tout être normalement constitué. Mais le courage individuel va parfois au-delà de cette définition. Un acte délibéré relevant du courage s'accompagne parfois d'une force physique exacerbée : le taux de cortisol, une hormone stéroïde, s'accroit et stimule l'augmentation du glucose sanguin. De l'énergie est ainsi puisée dans les réserves de l'organisme, et le soldat peut user d'une force inhabituelle. C'est le cas par exemple quand un soldat va chercher sous le feu son camarade blessé, et le traîne jusqu'à un abri alors que le soldat blessé et son équipement pèsent plus de cent kilos. Le courage permet à l'homme de se surpasser, tant moralement que physiquement.

Le SALA, puisqu'il ne connaîtra pas la peur, ne pourra pas développer de courage. Il n'aura pas à affronter le stress ou la douleur et n'aura donc pas à les surpasser. En revanche, il pourrait avoir besoin momentanément d'un surplus de puissance, comme pour évacuer un soldat blessé d'une zone hostile pour reprendre le même exemple. Les turbomoteurs d'avions et d'hélicoptères ne sont pas utilisés à leur pleine puissance en conditions normales. Mais le

[45] ARKIN Ronald, *Governing Lethal Behavior in Autonomous Robots*, Chapman an Hall, 2009.

pilote peut puiser dans une « puissance d'urgence » pendant un temps donné lorsque la situation l'exige. Le moteur est alors poussé à ses limites, pour que le pilote puisse sortir de son mauvais pas. Tant que cette puissance d'urgence n'est pas utilisée au-delà d'un certain temps, le moteur ne subit pas de dommage. On pourrait imaginer un tel système de gestion de puissance pour un SALA : lorsqu'il jugerait nécessaire l'utilisation de toutes ses ressources, il pourrait puiser au maximum des capacités techniques de ses composants, afin d'exacerber sa puissance ou sa vitesse. Ce surplus de puissance s'apparenterait, par les situations dans lesquelles il serait utilisé, à une forme de courage.

Déduction n° 5 : *un mode « urgence » dans lequel toutes les capacités du SALA seraient débridées doit être maintenu possible.*

Instinct ou intuition ?

« *Le drill n'est pas sans mérite, comme dans les disciplines sportives la répétition incessante de gestes individuels et collectifs est encore le meilleur moyen de les accomplir lorsque le réflexe doit remplacer la réflexion »*[46]. L'entraînement répétitif est ainsi une des sources premières de l'instinct au combat. C'est lui qui fait remonter à l'esprit la solution la plus évidente à une situation donnée. Or, le combat impose parfois une rapidité de décision qui prive le soldat du temps de la réflexion. C'est alors l'instinct qui prend le relai et impose l'action à mener. Mais l'instinct prend parfois des formes plus difficiles à expliquer, et devient plus proche d'une intuition qui impose un doute à l'esprit. D'où provient le léger malaise qu'un tireur perçoit au moment où tous les voyants sont au vert pour ouvrir le feu sur une cible, qui le fait patienter et mettre à jour l'élément invisible jusqu'alors, et qui remet en cause toute l'action ? L'expérience peut sans doute, à l'image de l'entraînement, faire remonter les conclusions d'une situation similaire. Mais parfois, le soldat ne s'explique pas l'origine du doute qu'il a pressenti avant d'ouvrir le feu et qui l'a fait patienter. Et parce que le rythme du combat est de plus en plus élevé, l'instinct prend une grande part dans le processus de décision du soldat au combat.

Il semble évident que l'intuition sera totalement absente de l'intelligence artificielle d'un SALA. Mais il est possible de simuler un « instinct » qui puiserait, parmi les procédures de combat et les expériences connues, celles qui sont le plus ressemblantes à la situation confrontée pour faire remonter rapidement une solution. Un algorithme de remontée de

[46] GOYA Michel, *Sous le feu : la mort comme hypothèse de travail*, Paris, Tallandier, 2014, 266 p.

solution ressemblant au lien « j'ai de la chance » du moteur de recherche Google donnerait au module de vérification éthique une première solution « évidente » à analyser. Etant donné qu'elle aurait de grandes chances de correspondre à la situation, cette solution « instinctive » serait souvent la bonne, à quelques modifications près. Le processus logiciel de décision, qui serait sans doute déjà rapide par nature, en deviendrait fulgurant la plupart du temps.

Déduction n° 6 : *un algorithme de proposition de solution « instinctive », par recherche de procédure connue ou de situation vécue similaire à l'environnement du moment, pourrait accélérer le processus de décision du SALA.*

Créativité

L'homme au combat est parfois confronté à des situations bloquantes, parce qu'aucune des procédures qu'il maîtrise ne sont réalisables. Soit parce que la situation est nouvelle, soit parce que les règles d'engagement du moment rendent impossible l'utilisation de ces procédures, le soldat se retrouve en territoire inconnu pour décider de son action. Il doit alors faire preuve de créativité, et agir avec intelligence, dans le sens le plus noble du terme. L'audace intellectuelle du soldat humain permet ainsi de débloquer les situations problématiques que la connaissance des règles et procédures ne permet pas de résoudre.

Il paraît peu probable que l'intelligence artificielle d'un SALA soit, en tout cas dans les premiers temps, capable de créativité. Tout au plus sera-t-elle capable, grâce à sa capacité d'apprentissage, de reproduire une action apprise par expérience lors d'une situation nouvelle. Mais cela reviendra à reproduire une procédure connue. Elle ne sera sans doute pas capable de créer une rupture dans les procédures et règles établies. Pourtant, il faudra bien qu'en cas d'absence de solution à une situation donnée l'intelligence artificielle du SALA décidé d'une action à mener, sous peine de voir la machine se mettre en attente d'une évolution de la situation.

Déduction n° 7 : *un système de déblocage s'apparentant à la créativité humaine devra pouvoir être déclenché par l'intelligence artificielle en l'absence de solution.*

Cohésion

Une unité combattante n'est pas un agrégat d'individualités. Elle forme un corps dont la force est bien plus grande que la somme des forces de chacun de ses composants. Et, plus important que tout, elle est animée d'un « esprit de corps », autrement nommé cohésion, qui relie émotionnellement les soldats entre eux. Ce lien n'existe pas que pour afficher la fierté d'appartenance à une unité, il peut être un soutien dans l'épreuve, mais également la source dans laquelle chaque soldat peut puiser pour se surpasser. L'intérêt de l'esprit de corps au combat est que chaque individu renonce à son intérêt personnel immédiat pour le bien de l'action collective[47]. La cohésion aide alors à surpasser sa peur, sa souffrance, sa fatigue.

Là encore, il est difficilement imaginable qu'un SALA soit sensible à cette cohésion. Dans l'absolu, on pourrait muter toutes les semaines un SALA dans une autre unité sans qu'il ne perçoive la différence. Une intelligence artificielle n'est a priori pas capable d'attachement, moins encore d'éprouver de la fierté. Il n'y a donc aucune raison pour qu'elle ressente cet esprit de corps. En revanche, un SALA peut être programmé pour que le bien de l'action collective soit considérée comme prioritaire. Il n'aurait ainsi pas besoin d'un lien émotionnel pour placer le bien commun au-delà de son « intérêt personnel ». Les mesures d'autoprotection abordées dans le chapitre sur la peur seraient rapidement dépassées si l'action entraînerait un bienfait pour la mission ou pour l'unité à laquelle appartiendrait le SALA. On peut donc considérer que la programmation du SALA lui conférerait une application froide de l'esprit de corps, sans effort pour celui-ci : celle de placer l'action collective au-devant de son intégrité physique.

Discernement émotionnel

Il arrive qu'un soldat, alors que les règles d'engagement le lui permettent, décide de ne pas ouvrir le feu sur un ennemi. Il existe de nombreuses raisons pour lesquelles un soldat peut choisir de tirer « à côté », pour simplement faire « baisser les têtes », ou choisir de ne pas ouvrir le feu : parce que l'ennemi en question est en fuite, montre des signes d'abandon, ou encore parce qu'il occupe un lieu où l'absence de civils n'est pas absolument certaine. La mort n'est pas systématiquement au rendez-vous des armes. Un soldat humain est capable de juger moralement une situation, et possède donc une grille d'analyse supplémentaire aux seules règles et procédures pour décider de son action. « *Outre la connaissance des règles d'engagement, sa capacité de décision est forgée par son discernement émotionnel. Ce sont ses*

[47] ERBLAND Brice*, La tentation de l'hubris*, in *Inflexions* n°32, mai 2016.

émotions et son instinct, dans le feu de l'action, qui lui dictent le chemin à suivre lorsque les procédures et les règlements ne donnent plus de réponse »[48]. Ce discernement émotionnel est à la base du comportement vertueux du soldat au combat. Il s'agit en quelque sorte de faire le choix de la clémence au vu d'une situation particulière. « *Faire preuve de clémence peut obéir à toute une gamme de raisons, mais vise toujours la protection, quels que soient les sentiments qui la motivent. Dans les formes les moins élaborées, il peut s'agir simplement de modération et de retenue* »[49]. Or, cette inclination est issue des émotions et de l'empathie dont peut faire preuve le soldat. Ce discernement est ainsi profondément humain. Toutes les faiblesses morales analysées en première partie sont autant de facteurs de déshumanisation du soldat, qui le privent donc de cette capacité de discernement émotionnel.

Parce que le SALA ne sera pas capable de ressentir des émotions ou d'éprouver de l'empathie, il ne pourra *a priori* pas faire preuve de discernement émotionnel. Il n'en sera en tout cas pas capable de nature.

Vertu humaine	Equivalent pour un SALA	Déduction
Courage	Simulable	Un SALA pourra avoir une réserve de puissance disponible quand la situation l'exige
Instinct	Simulable	Un SALA pourra avoir un algorithme de remontée rapide de solution possible
Créativité	*A priori* impossible	-
Cohésion	Partiellement impossible, mais sa programmation peut impliquer qu'il place l'action collective au-devant de son intégrité physique	-
Discernement émotionnel	*A priori* impossible	-

Figure 2

Il apparait clairement que le combat n'est pas une science exacte : les réponses adéquates à une situation donnée ne sont pas toujours dictées par les règles et les procédures. « *Les interrogations sur la responsabilité de l'emploi de la force se transforment bien souvent en*

[48] ERBLAND Brice*, La tentation de l'hubris*, in *Inflexions* n°32, mai 2016.

[49] SLIM Hugo, *op. cit.*, p. 325.

véritables cas de conscience dont les éléments de réponses ne figurent pas tous dans les manuels de tactique ou les règlements »[50]. C'est la raison pour laquelle le soldat humain fait appel à des capacités propres, d'ordre moral, pour l'aider à prendre ses décisions. Or, ces forces morales découlent d'une éthique particulière, c'est-à-dire d'un code de conduite moral qui s'adapte à une situation donnée et permet de qualifier la moralité de telle ou telle action dans ce contexte donné. La figure 2 regroupe les différentes vertus humaines étudiées, avec leur équivalent pour un SALA. Il en ressort clairement que les vertus relevant de près ou de loin à l'éthique, soit la créativité et le discernement émotionnel, sont absentes chez le SALA.

[50] ROYAL Benoît, *L'éthique du soldat français*, 3ème édition, Paris, Economica, 2014, 304 p., p. 10.

OBJECTIFS D'UNE ETHIQUE ARTIFICELLE

Si le rire est le propre de l'homme selon Rabelais, dans son ouverture de *Gargantua*, il semble qu'au combat, le propre du soldat humain demeure son discernement émotionnel. Or, ce dernier relève d'une éthique proprement humaine qui le rend difficile à définir. Il rend le combat totalement irrationnel, par le bousculement des règles qu'il entraîne.

Pourtant, les contextes de guerre actuels compliquent d'ores et déjà les « règles » de combat simplistes qui dictent de tirer sur un combattant ennemi et de ne pas le faire sur un civil. « *Outre qu'il est de moins en moins facile, dans les conflits contemporains, de distinguer le civil du combattant, il peut être légal de tirer sur un civil s'il participe directement aux hostilités – une notion complexe qui donne lieu à des interprétations divergentes -, et il peut être illégal de tirer sur un combattant s'il est hors de combat, ce qui n'est pas non plus simple à établir* »[51]. Ces particularités et nuances dans les règles d'engagement seront déjà bien difficiles à programmer pour être entièrement prises en compte par un SALA. Et ces particularités et nuances relèvent déjà, en quelque sorte, du discernement émotionnel, puisqu'elles sont conditionnées par un jugement subjectif de la situation. On peut objecter qu'une machine, avec des senseurs de toutes sortes d'une sensibilité très poussée, sera mieux capable de repérer de l'armement ou de reconnaître une blessure sur un être humain. Ces capacités hors-du-commun lui permettront sans doute de faire la différence entre un soldat ennemi valide et un blessé de guerre, ou entre un civil innocent et celui qui prend les armes.

Mais le discernement émotionnel va beaucoup plus loin. Car, parfois, il se peut qu'un soldat se retrouve face à un autre soldat ennemi en pleine possession de ses moyens, et décide pourtant de ne pas ouvrir le feu. Il s'agira bien sûr, comme nous l'avons déjà dit, d'une mesure de clémence. Le soldat retiendra ses feux pour des raisons propres, que nul autre ne pourra connaître et comprendre. Comme ce pilote d'hélicoptère en Afghanistan qui, face à deux insurgés qui viennent de viser un de ses ailiers, retient ses feux lorsque les deux soldats ennemis apparaissent dans son viseur les mains en l'air, en signe de soumission, alors même qu'aucun doute n'est possible[52]. Comme ce colonel qui refuse d'autoriser le tir d'une bombe

[51] JEANGENE VILMER Jean-Baptiste, *Terminator Ethics : faut-il interdire les « robots tueurs » ?*, in *Politique étrangère* Hiver, n° 4, 2014 : 151‑67.

[52] ERBLAND Brice, *Le processus homicide : analyse empirique de l'acte de tuer*, in *Inflexions* n° 31, janvier 2016.

sur un groupe de chefs talibans parce que ces derniers sont en train de prier[53]. Comme ce pilote de chasse de la Luftwaffe qui surprend un bombardier américain égaré et se rend compte, au moment d'engager le tir, que sa cible est déjà endommagée : plutôt que d'abattre une cible facile qui lui vaudrait une victoire de plus à son compteur, il guide l'équipage du bombardier jusqu'aux côtes du Nord-Est de l'Allemagne pour qu'il puisse rejoindre la Suède.[54]

Ce discernement émotionnel est à la base d'une certaine limitation de la guerre, une limitation de la violence *dans* la guerre. « *L'idée qui sous-tend le concept de guerre limitée et l'éthique civile qui en est le corollaire est, évidemment, celle d'une limitation des pertes en vies humaines. Elle part du principe que, même dans la guerre, l'homme doit tuer le moins possible, car chaque vie humaine est précieuse aux yeux de la personne en question, aux yeux de ceux qui l'aiment et, pour ceux qui ont la foi, aux yeux de Dieu* »[55]. Ce principe peut paraître contraire à l'essence même du combattant professionnel. Certains pourront juger ces comportements de clémence contraires à la réalisation de la mission, voire impropres à la condition de guerrier. Puisque le combattant épargne son ennemi, il retarde le moment de la victoire. Pire encore, il va permettre à cet ennemi de tuer ultérieurement des combattants alliés. Cette clémence semble donc plus proche de poussiéreuses règles de chevalerie que des dures lois de la guerre moderne.

Pourtant, cette « chevaliérisation » de la guerre a ses vertus propres. Elle évite d'abord une escalade de la violence qui n'aurait pas de fin. Le combattant, parce qu'il prend entièrement et personnellement part à la logique de violence, tend naturellement à en fixer des limites. « *Dans une guerre, le combat intime de tout soldat est de conserver son humanité originelle* »[56]. Les limites de la violence participent à cette conservation de l'humanité du soldat. L'acte de tuer est profondément déshumanisant. Il s'agit donc de le faire « humainement », c'est-à-dire de façon moralement acceptable, pour se protéger soi-même. Cette notion est bien entendu fortement subjective, c'est pourquoi deux soldats ne prendront

---

[53] MINGASSON Nicolas, *1929 jours*, Paris, Les Belles Lettres, 2016, 384 p.

[54] ALEXANDER Larry, *L'honneur avant tout*, trad. GUIOD J., Paris, Altipresse, 2014, 361 p.

[55] SLIM Hugo, *op. cit.*, p. 322.

[56] ERBLAND Brice*, La tentation de l'hubris*, in *Inflexions* n°32, mai 2016.

pas la même décision face à une même situation donnée. Mais elle permet, tout en réalisant la mission, de protéger l'équilibre psychologique du soldat.
Les actes, ou plutôt les non-actes issus de cette « chevaliérisation » peuvent ensuite avoir des conséquences bénéfiques difficilement mesurables. Chaque conflit contemporain draine son flot de mesures contre les dommages collatéraux, notion chiffrable et permettant d'estimer la façon dont une guerre est menée. Mais si l'on parle souvent de dommage collatéral, on ne parle jamais de *bienfait collatéral*. Or, la clémence ou la retenue dont un combattant peut faire preuve ont peut-être des conséquences sur « les cœurs et les esprits » des combattants ennemis, ou au moins sur ceux de la population civile. L'exemplarité d'un comportement éthiquement bon peut aussi être une arme efficace.

L'éthique au combat du soldat humain est donc plus compliquée qu'un code de conduite morale que l'on pourrait lister à la manière des commandements divins. Bien entendu, le jugement moral d'un acte est souvent associé au respect d'une déontologie, mais les situations complexes exigent parfois de sortir de ces codes pour estimer les conséquences des actes.
Cette approche conséquentialiste entraîne le soldat dans une parfaite solitude de jugement : il n'a plus de code ni de règle culturelle auxquels se rattacher, mais doit pourtant décider en son âme et conscience. C'est là également que le discernement émotionnel prend tout son rôle.

Le but d'une éthique artificielle serait donc de reproduire chez le SALA ce comportement *subjectif* issu du discernement émotionnel humain. Puisque l'objectif est de rendre la machine « autonome », il faut qu'elle puisse elle-même émettre un jugement moral face à une situation particulière. Pourtant, « *quel que soit le niveau atteint par l'intelligence artificielle, un robot ne sera jamais absolument autonome, pas plus que l'homme lui-même d'ailleurs* »[57]. Car il ne faut pas chercher à construire une machine qui ne ferait jamais d'erreur, qui aurait à tout moment un jugement juste et sans défaut. Il faut tout simplement chercher à construire une machine qui soit *au moins aussi fiable* que l'homme. En d'autres termes, il faut que cette machine puisse appréhender les situations de combat et le contexte de zone de guerre aussi bien qu'un humain, et non pas *parfaitement*.
Mais cette baisse d'exigence, en quelque sorte, ne dois pas paraître comme une facilité. Car cela signifie tout de même que le SALA ne devra pas être susceptible de céder aux mêmes

[57] FAES Hubert, *Une éthique pour les robots tueurs ?*, in *Revue d'éthique et de théologie morale*, n° 289 (23 juin 2016): 107‑15.

faiblesses que l'homme d'une part, et surtout qu'il devra être capable de faire appel aux mêmes vertus d'autre part. Il faudra donc être capable de programmer une éthique artificielle qui simule ou reproduit le discernement émotionnel qui est l'ultime recours lorsque les règles et procédures ne donnent plus aucune solution.

Ce dernier point semble être un objectif difficilement atteignable. En effet, beaucoup pensent que l'éthique, parce qu'elle se base sur la signification des actes, ne peut être programmée. Pourtant, « *le projet d'un robot « éthique » n'est ultimement pensable que si l'éthique peut être complètement informatisable, formalisable* »[58]. Il faut donc chercher à programmer une éthique artificielle qui soit non seulement capable de respecter un code de conduite moral, et donc d'estimer la valeur morale d'un acte par rapport à un autre, mais également de construire sa subjectivité morale propre, sa capacité à la clémence selon les situations. « *Les armes autonomes ne devraient pas être interdites sur la base de la technologie actuelle ; les gouvernements doivent s'assurer que les armées utiliseront les technologies autonomes d'une façon responsable qui reproduirait le jugement et la responsabilité humains dans l'usage de la force* »[59]. Il faut donc chercher à « humaniser » le système de décision du SALA, avec les imperfections qui en sont liées, mais avec cette force morale particulière qu'est le discernement émotionnel.

Déduction n° 8 : *un module d'éthique artificielle d'un SALA devra être capable :*

- *De juger la moralité d'un acte par rapport à un autre, à partir d'un code de conduite moral connu ;*
- *De décider de l'action tendant vers une désescalade de la violence en cas de situation non référencée ;*
- *De construire son jugement moral propre au fur et à mesure de ses expériences afin de développer ce qui s'apparenterait au discernement émotionnel.*

Cette recherche de développement d'une éthique artificielle représente sans doute l'obstacle le plus prégnant qui s'oppose à l'essor de la robotique, non seulement dans un contexte militaire

[58] LAMBERT Dominique, *Robots autonomes : la place irréductible et complémentaire de l'éthique de* l'officier, in DOARE Ronan, HUDE Henri (Dir.), *Les robots au cœur du champ de bataille*, Paris, Economica, coll. « guerre et opinions », 2011.

[59] HOROWITZ Michael C., SCHARRE Paul, *The morality of robotic war*, in The New York Times, 26 mai 2015.

mais également dans toute autre application civile. Par sa complexité et ses enjeux tout particuliers, l'éthique artificielle dans le cadre de la guerre possède une exigence de qualité hors-norme qui permettra de décliner la technologie dans n'importe quel domaine. « *L'éthique militaire est à la pointe de la pyramide, je dirais même que c'est une pointe de tungstène qui, la première, trace les chemins possibles dans des domaines touchant à la vie et à la mort qu'il est ensuite possible de décliner et d'adapter aux autres réalités* »[60]. Les recherches en éthique artificielle dans le cadre d'un SALA sont donc essentielles à la fois pour développer un robot au comportement moralement acceptable sur le champ de bataille et pour les nombreuses applications civiles qui en découleront. Ce n'est pas, comme l'écrivent Horowitz et Scharre, « *le même type de senseurs et process qui permettront à une voiture autonome d'éviter les piétons qui pourra permettre à un robot-tueur d'éviter les civils* »[61], mais bien l'inverse. Comme dans toute technologie duale, c'est d'abord l'utilité guerrière qui en permet sa maîtrise.

[60] ROYAL Benoît, *op. cit.*, p. 12.

[61] HOROWITZ Michael C., SCHARRE Paul, *op. cit.*

II. ETHIQUE COMPUTATIONNELLE

Maintenant que les objectifs d'une éthique artificielle ont été fixés par la comparaison entre les faiblesses et vertus humaines du soldat au combat et les probables comportements du SALA dans les mêmes conditions, tentons de définir comment un programme informatique serait capable de les atteindre. Car il est bien différent de fixer des buts vertueux et d'être capable de les atteindre par des lignes d'octets. Pour cela, il sera nécessaire de décrire le processus humain de décision au combat, puis d'analyser les différentes approches philosophiques de la morale, afin de déterminer le cheminement intellectuel de l'éthique humaine. A partir de cette référence, nous étudierons les différentes techniques utilisées en programmation informatique pour simuler le discernement moral, puis tenterons de proposer un modèle d'éthique computationnelle qui pourrait répondre aux objectifs fixés. Enfin, il sera nécessaire de poser la question de la validation de cette éthique artificielle afin de s'assurer qu'elle réponde aux exigences fixées.

CE QU'IL EST NECESSAIRE DE REALISER

Etant donné que le SALA devra reproduire au mieux le processus de décision humaine, et notamment sa capacité de discernement émotionnel, il est nécessaire d'établir en premier lieu la façon dont un combattant humain, qu'il soit simple soldat ou chef tactique, fantassin ou pilote, prend la décision de tuer au combat. Bien entendu, cet exercice est très schématique, voire simplificateur, mais il reste utile pour décomposer le processus décisionnel en étapes qui serviront à identifier les outils de programmation informatique les plus adaptés pour construire une éthique computationnelle. Avant toute chose, il faut définir le mot décision. Dans le dictionnaire, il apparaît que c'est un choix entre plusieurs solutions. Une formulation quelque peu simpliste lorsqu'on considère la complexité d'un champ de bataille. Mais de la lecture d'Aristote[62], nous pouvons proposer une définition plus complète, en qualifiant la décision de *courage opposé constamment aux détracteurs, appliqué à une situation d'incertitude, se révélant perspicace dans la capacité à anticiper et réactualisé constamment pour s'adapter aux aléas de l'adversité*[63].

Parce que la décision au combat est parfois prise en un temps éclair, il faut chercher à comprendre les éléments qui auront permis et préparé cette décision. Dans les cas de doute où la décision prend plus de temps, il faut analyser les différentes morales et autres éléments qui entreront en jeu et qui formeront, *in fine*, l'éthique du soldat. Nous pourrons ainsi identifier les différentes composantes qui forment les trois piliers de la prise de décision que nous avons abordés dans le chapitre sur les vertus humaines au combat : la connaissance, l'intelligence et la conscience. Ces trois piliers correspondent en quelque sorte aux trois parties principales du cerveau humain, que sont le cerveau reptilien, le néocortex et le système limbique. En effet, ces différentes parties du cerveau ont des fonctions différentes qui sont respectivement instinctives, cartésiennes et émotionnelles. C'est donc en rentrant dans le cerveau du soldat, en analysant isolément les différents éléments de la prise de décision, que nous tenterons de dresser un schéma qui ordonnera ces acteurs de la réflexion dans le processus décisionnel humain.

[62] ARISTOTE, *Éthique à Nicomaque*, trad. BODEUS R., Paris, Flammarion, 2004, 560 p.

[63] Définition proposée par le MGA P. GODART.

Mission et règlements

En toile de fond de toute situation de combat, le soldat garde en tête la mission qu'il doit accomplir et les règles d'engagements qui sont définies pour l'opération.

Selon la mission reçue, il aura tendance à utiliser son arme de façon différente. Entre une mission « attaquer » et une mission « éclairer », l'esprit d'utilisation de la force n'est en effet pas le même. Si dans le premier cas le soldat cherchera à détruire l'ennemi, et donc à utiliser ses armes contre lui, il devra éviter le contact et donc éviter d'ouvrir le feu dans le second cas. La mission reçue, par les tâches particulières qu'elle sous-entend et par l'esprit qu'elle véhicule, installe des couloirs de réflexion particuliers chez le soldat. Il est ainsi conditionné, en quelque sorte, par l'objectif et le rôle qui lui sont assignés. Son raisonnement en sera d'autant influencé lorsqu'il aura à décider d'ouvrir le feu.

Plus important encore, les règles d'ouverture du feu cadrent assez précisément le champ des possibles pour le soldat. S'il est par exemple limité à la seule légitime défense, il devra attendre d'être agressé avant de pouvoir utiliser la force. Il n'agira qu'en réaction et sa décision en sera somme toute facilitée. En revanche, s'il est autorisé à blesser ou tuer un individu faisant preuve d'une « intention hostile », cela ouvre des possibilités d'usage des armes bien plus grandes autant que cela complique la prise de décision. Car si le port d'arme lourde peut aisément être considéré comme une intention hostile, est-ce que le fait de creuser de nuit au bord d'une route fréquemment piégée peut être considéré comme tel ? La précision des règles d'ouverture du feu est une aide plus ou moins grande à l'analyse d'une situation pouvant mener à l'usage des armes.

Ces éléments qui prédéfinissent l'axe de raisonnement moral que prendra le soldat sont en quelque sorte la partie déontologique du processus décisionnel. Loin de guider seuls la décision, ils sont sans cesse présents en tant que filtre tout au long de ce processus.

Déduction n° 9 : *un module d'éthique artificielle d'un SALA devra contenir un élément « mission » et un élément « règles d'engagement », reprogrammables en permanence et servant de ligne directrice au raisonnement de l'intelligence artificielle.*

Procédures et expérience

Face à une situation donnée et d'autant plus si l'urgence entre dans l'équation, ce qui est souvent le cas au combat, le soldat se raccroche à ce qu'il a appris et restitué tant de fois à l'entraînement. Ces procédures tactiques établies par la doctrine militaire permettent de

calquer sur le terrain et la situation des solutions toutes faites applicables selon les situations. Elles ont l'avantage de donner des clés au combattant qui serait sans cela forcé sans cesse d'improviser. Un chef de section d'infanterie sait ainsi qu'il faut placer un appui et une couverture avant de coiffer un objectif ; un pilote d'hélicoptère d'attaque sait qu'il doit manœuvrer sa machine de sorte que son chef de bord puisse en permanence observer l'objectif et ouvrir le feu. Même si ces procédures simples ne répondent pas à toutes les situations, elles forment un catalogue d'actions possibles qui parviennent rapidement et pratiquement sans effort à l'esprit. Etant donné le niveau de stress et la charge de travail physique qu'un combattant doit soutenir lorsqu'il se trouve en situation de combat, les ressources intellectuelles restantes sont limitées. Ces procédures, connues parfaitement grâce au *drill* de l'entraînement, sont donc comme autant de bouées auxquelles le cerveau bouillonnant du soldat au combat peut se raccrocher.

Un autre élément important vient nourrir la réflexion du combattant : son expérience. Parce qu'il aura testé les procédures tactiques en situation réelle, il aura pu détecter leurs éventuels défauts ou la meilleure façon de les appliquer. Plus important encore, il aura connu l'intense charge émotionnelle liée à l'ouverture du feu et en sera ainsi débarrassé, ce qui laissera plus de place à l'analyse rationnelle. Il connaîtra également plus intimement l'ennemi pour l'avoir déjà combattu, il pourra donc mieux anticiper sa réaction. L'expérience du combat est donc un élément facilitateur de prise de décision, par la stabilité émotionnelle qu'elle assure et par la comparaison intrinsèque avec d'autres situations vécues qu'elle apporte.

L'expérience peut néanmoins influencer la décision de façon néfaste, lorsqu'elle s'impose à l'esprit par rapport à l'analyse de la situation. C'est notamment le cas lorsque le soldat a l'expérience d'autres théâtres de guerre. Il aura en effet tendance à appliquer les décisions prises dans un certain contexte à sa grille d'analyse d'une situation pourtant différente. Il lui faudra faire preuve, encore une fois, de discernement pour ne pas se laisser influencer outre mesure par une expérience de combat enfouie dans son esprit.

Déduction n° 10 : *un élément « procédures » reprogrammable, comme l'élément « règle d'engagement », devra fournir les pistes de raisonnement de départ pour chaque situation. Une base de données « mémoire » pourra simuler l'apport de l'expérience humaine.*

Intuition et créativité

Les termes de missions, les règles d'engagements, les procédures et expériences de combat forment donc le pilier « en connaissance » de la prise de décision. Ils sont ainsi en permanence dans l'esprit du combattant pour filtrer son processus décisionnel. A ce pilier s'ajoute celui de l'intelligence, dans sa définition la plus noble, c'est-à-dire ce qui semble surgir de nulle part, d'aucune connaissance acquise, mais est créé de toute pièce par le cerveau humain. L'intuition et la créativité, déjà développés dans le chapitre sur les vertus humaines au combat, sont les deux éléments de ce pilier.

La particularité de ces deux éléments est qu'ils interviennent ponctuellement tout le long du processus décisionnel. Loin d'en être la trame de fond comme les éléments du pilier connaissance, l'intuition fournit des éléments de façon quasi inconsciente quand on ne l'attend pas forcément et la créativité s'active dès lors que le soldat se retrouve face à une impasse. La première travaille donc en sourdine et procure des éléments lorsqu'elle en dispose, la seconde est activée de besoin.

Ethique

In fine, à ces piliers connaissance et intelligence vient s'ajouter le pilier « conscience », qui va analyser les solutions qui viennent à l'esprit à l'aune de l'éthique propre du soldat. C'est bien en dernier lieu que l'être humain se pose la question de la moralité de l'acte qu'il s'apprête à faire. Le combat ne fait pas vraiment exception. A la froide analyse qu'exigent la complexité et l'urgence des situations vient se heurter le discernement moral dont nous parlions plus haut. C'est ce filtre qui va retenir la décision pour trouver une solution alternative moralement meilleure, en déclenchant la créativité pour moduler l'action ou en faisant redémarrer le processus avec une autre procédure.

Mais qu'est-ce que l'éthique ? Qu'est-ce que la morale au combat ? Lorsqu'il est sous le feu, il faut bien admettre que le combattant ne fait pas grand cas de la vie de son ennemi, ce qui semblera naturel à tout le monde. La véritable urgence, celle qui menace de faucher des vies à chaque seconde qui passe, facilite amplement la décision et donc la balance morale de l'action, à condition que la situation soit claire. Mais lorsque le doute s'installe, que des civils sont au milieu du champ de bataille, pris entre deux feux ou utilisés comme bouclier par l'ennemi, ou encore que le civil innocent se transforme en commando-suicide ou qu'un enfant prend les armes, la prise de décision se transforme en cauchemar par les contradictions morales qui découlent de la situation. Un comportement moral est entendu comme étant

« *sensible à la polarité du bien et du mal* »[64]. Mais le combat, par la présence incessante de la mort qu'il implique, rebat les cartes du bien et du mal.

Il existe trois théories rationnelles de la morale : la déontologie, le conséquentialisme et l'éthique des vertus. La première est basée sur une série de règles établies qu'il s'agit de respecter en toutes circonstances. C'est par exemple l'esprit des tables de lois de l'Ancien Testament. La déontologie est donc le respect strict de règles morales associées à une situation et acceptées comme représentatives du bien. Le conséquentialisme va plus loin : il cherche à estimer les conséquences d'un acte, pour soi-même, pour autrui voire pour le bien commun d'un groupe ou de l'humanité entière, selon le degré d'intérêt qu'on veut bien lui appliquer, avant de juger de la moralité de l'acte. C'est donc l'effet recherché, les résultats attendus qui qualifieront la moralité de l'action plus que l'acte en lui-même. C'est un peu la traduction du débat entre Kant et Constant sur le droit de mentir : Kant considère le devoir de vérité comme étant intangible, Constant défend le droit de mentir lorsque la vérité aurait des conséquences néfastes. Bien entendu, cette approche conséquentialiste comporte un risque, celui de mal évaluer les conséquences de son acte. L'approche par les vertus, quant à elle, juge un acte ou une pensée par leur conformité avec des valeurs vertueuses.

Imaginons alors deux situations de combat qui nécessitent un jugement moral pour décider de l'action à mener. Dans la première, un groupe de combattants ennemis se dévoile accompagné de civils dont ils se servent comme bouclier, le temps de s'exfiltrer vers des couverts. Dans la seconde situation, un soldat en poste d'observation voit un jeune adolescent que l'on a visiblement équipé d'une ceinture d'explosifs se diriger vers un groupe de soldats alliés. Quelle décision prendre dans les deux cas ?

L'approche déontologique nous dicte dans le premier cas de ne pas ouvrir le feu, puisqu'il est moralement condamnable de tirer sur des civils. Et l'approche conséquentialiste nous donne le même résultat, puisque l'ouverture du feu entraînerait le risque de toucher des civils, ce qui n'est pas acceptable.

Dans le second cas, l'approche déontologique nous empêche encore une fois d'ouvrir le feu : il n'est par moral de tirer sur un adolescent. Mais l'approche conséquentialiste condamne le fait d'épargner l'enfant puisque cela engendre la mort probable de plusieurs soldats si celui-ci se fait exploser à proximité d'eux. Dans ce cas, même si l'acte est horrible, le conséquentialisme implique bien l'ouverture du feu sur le jeune adolescent comme étant

[64] CHAUVIER Stéphane, *Éthique Artificielle*, entrée de *L'Encyclopédie Philosophique*, 2017, (http://encyclo-philo.fr)

l'acte le plus juste, ce qui pourrait paraître, sorti de son contexte, un non-sens absolu. Mais un théâtre de guerre est bien souvent le lieu de paradoxes moraux : le soldat y est condamné à choisir la solution la moins pire. Car il n'y a parfois aucune décision possible que l'on qualifierait en temps normal de moralement acceptable.

C'est là que peut intervenir une théorie irrationnelle de la morale, telle que la morale des sentiments évoquée par Rousseau : c'est parce que l'homme est capable d'éprouver de la pitié, de l'empathie pour autrui qu'il pourra éviter de faire du mal aux autres. On peut rapprocher cette théorie du discernement moral décrit dans le chapitre sur les vertus humaines au combat. Mais il ne faut pas oublier que parmi les sentiments figurent également ceux qui engendrent les dérives morales telles que la haine ou le désir de vengeance. Cette morale du sentiment ne peut donc à elle seule former l'éthique du combattant. Elle viendra simplement, idéalement par le biais du discernement moral, compléter la morale principalement conséquentialiste propre à chaque soldat.

Dans le cas de notre adolescent à la ceinture d'explosif, le discernement émotionnel dictera peut-être d'effectuer un tir de sommation à proximité étroite de l'adolescent pour tester sa réaction et confirmer ainsi son intention hostile, avant de passer à l'acte horrible mais néanmoins nécessaire au regard d'une froide balance des conséquences.

Le phénomène check-list

Une dérive peut s'installer dans le processus décisionnel au combat, notamment lorsque le combattant bénéficie d'un certain recul, qu'il soit temporel ou physique, vis-à-vis de la situation. C'est ainsi le cas pour les pilotes d'avions de chasse ou d'hélicoptères d'attaque qui peuvent être tentés, pour faciliter leur prise de décision, de réduire le processus décisionnel à une check-list de vérification du respect des règles d'engagement. Cette check-list, établie dans une intention vertueuse d'aide à la décision, réduit malheureusement le processus à son seul premier pilier. Il peut en effet donner le sentiment au combattant d'avoir vérifié tous les filtres avant d'ouvrir le feu, alors qu'il aura totalement obéré le pilier « en conscience ». Cette dérive n'est pas propre aux militaires, elle existe dans de nombreux domaines. Ainsi, le développement de programmes d'aide à la décision pour les médecins urgentistes mécanise en quelque sorte la pratique[65]. Ce phénomène, quel que soit le domaine considéré, « robotise » le

[65] LAMBERT Dominique, *Robots autonomes : la place irréductible et complémentaire de l'éthique de l'officier*, in DOARE Ronan, HUDE Henri (Dir.), *Les robots au cœur du champ de bataille*, Paris, Economica, coll. « guerre et opinions », 2011.

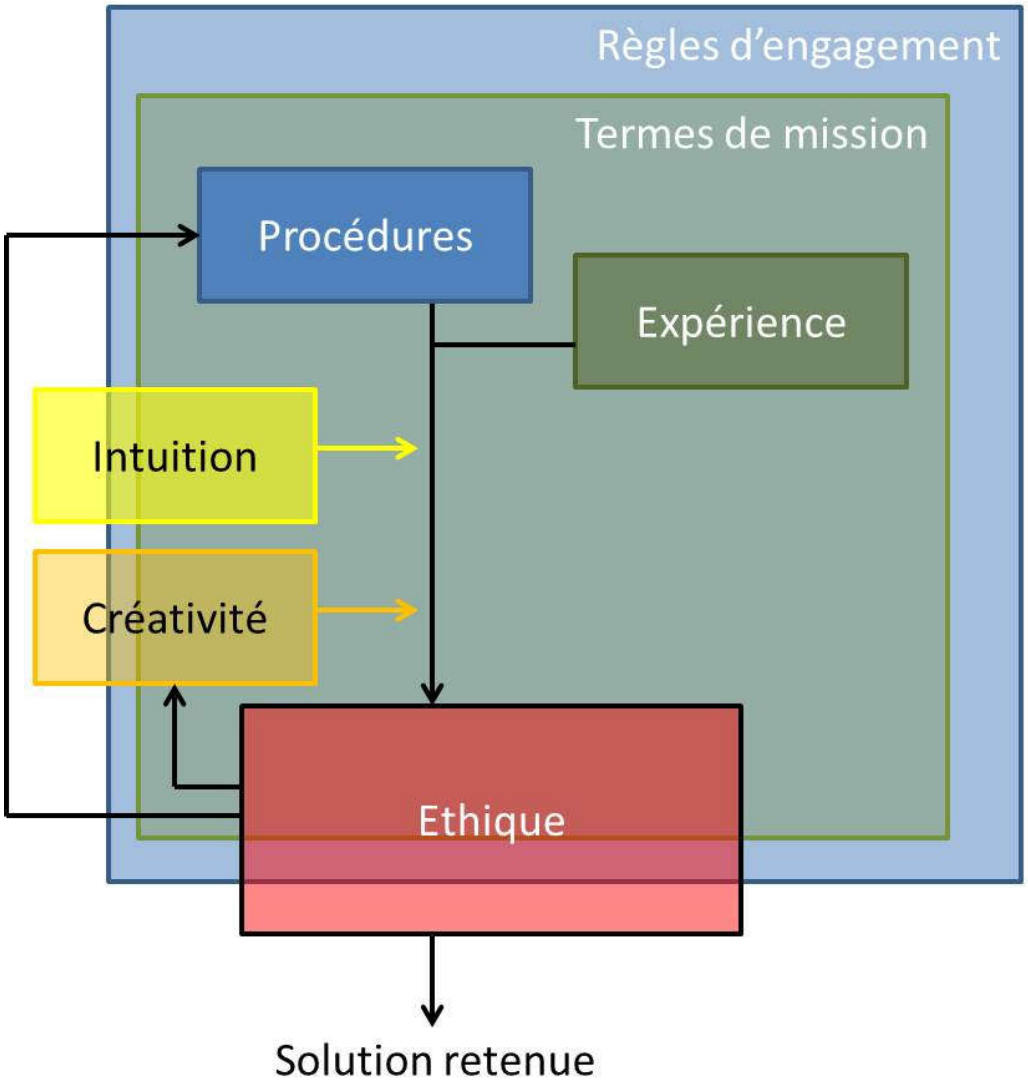
raisonnement humain, alors même que nous cherchons à humaniser le raisonnement de l'intelligence artificielle. Il prouve une fois de plus la nécessité d'un module de raisonnement éthique autonome au sein du SALA.

Schéma décisionnel humain

De ces différents éléments entrant en compte dans la prise de décision humaine, nous pouvons définir des interactions qui décrivent le processus décisionnel. Car ce sont bien la connaissance des procédures et l'expérience qui vont engendrer les actions possibles, au vu de la situation. C'est bien la créativité et l'intuition qui vont modifier ces actions possibles avant que l'on prenne la décision d'utiliser l'une ou l'autre. Et étant donné que « *le comportement moral demande, préalablement à la décision, l'évaluation de certaines possibilités d'action* »[66], c'est bien l'éthique, la morale propre à chaque être humain, qui effectue cette évaluation en fin de processus.

La figure 3 propose ainsi un schéma simplificateur de ce processus décisionnel.

[66] LAMBERT Dominique, *Une éthique ne peut être qu'humaine ! Réflexion sur les limites des moral machines*, in DANET Didier, DOARE Ronan, DE BOISBOISSEL Gérard (dir.), *Drones et killer robots* , Rennes, Presses Universitaires de Rennes, 2015, pp. 227 – 240.



sFigure 3

On comprend ici que la définition que propose Aristote du concept de décision s'applique parfaitement au soldat au combat, puisque ce dernier est confronté à une situation d'incertitude, qu'il doit anticiper les résultats de son action par une approche essentiellement conséquentialiste de sa morale, et qu'il doit sans cesse être capable de réactualiser son processus décisionnel pour s'adapter au mieux aux aléas de la situation. C'est à partir de ce processus de raisonnement humain, impliquant un filtre moral, que nous pouvons tenter de concevoir ce qui serait son jumeau logiciel. Même si « *la phase de recours à l'expérience ou à l'intuition ne peut être formalisée* a priori »[67], il faut tenter un mimétisme algorithmique pour parvenir à l'éthique artificielle que nous avons fixé comme indispensable au SALA.

[67] LAMBERT Dominique, *op. cit.*

COMMENT LE REALISER

Le processus décisionnel humain qui vient d'être décrit doit donc pouvoir être répliqué pour construire un module de raisonnement éthique qui corresponde aux objectifs fixés en première partie. Mais la complexité de ce processus et des différents éléments qui le composent, pour certains purement humains, est un véritable défi pour leur traduction en logiciel. Afin de proposer la conception d'une structure de programmation qui pourrait répondre à ce défi, nous devons passer en revue les techniques de programmation existantes (sans rentrer dans les détails techniques) qui permettraient de créer un module de jugement moral artificiel. Les codes du langage logique ne seront pas respectés ici, dans un souci de vulgarisation.

Approche déontologique

La première méthode qui vient à l'esprit est celle de fixer des règles de morale afin de guider le raisonnement, à l'image de l'approche déontologique de la morale. L'idée voudrait qu'en cernant suffisamment les actes immoraux par des lois intangibles, toute décision passée par le filtre de ces lois serait moralement acceptable. Cette approche est caractérisée dans le monde anglo-saxon de « Top-Down » : les normes et principes immuables dictent la moralité des actions, tel un regard divin qui juge les actions d'en bas en quelque sorte. Mais la verticalité descendante est également l'image du processus logiciel d'analyse d'une action, très hiérarchique. L'exemple le plus célèbre de cette approche déontologique est celui formalisé par Isaac Asimov dès les années cinquante avec ses trois lois de la robotique[68] censées protéger les êtres humains des robots :

1. *Un robot ne peut porter atteinte à un être humain, ni, en restant passif, permettre qu'un être humain soit exposé au danger ;*
2. *Un robot doit obéir aux ordres qui lui sont donnés par 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.*

---

[68] ASIMOV Isaac, *Les Robots*, trad. BILLON P., Paris, J'ai lu, 1967, 285 p.

Il va sans dire que cet exemple est inadapté pour le SALA, puisqu'on demanderait à ce dernier d'être capable de tuer des soldats ennemis, et donc de porter atteinte à des êtres humains. Tout au plus pourrait-on appliquer ces lois en remplaçant « être humain » par « soldat allié » ou « civil », mais que se passerait-il si un soldat allié retournait son arme contre les siens ? Rien que la première loi en deviendrait paradoxale, sauf à changer le statut du soldat allié trahissant son camp. Mais au-delà de cet exemple d'Asimov, un ensemble de lois ne permettrait pas de résoudre un « cas de conscience » tel que celui de l'adolescent portant une ceinture d'explosifs et menaçant un groupe de soldats alliés, parce que cette approche déontologique n'accepte aucune souplesse par son dogmatisme absolu.
On ne peut donc pas baser le module de jugement éthique du SALA sur des lois déontologiques. En revanche, cette approche est idéale pour formaliser les termes de mission et les procédures, qui sont des sortes de lois intangibles pour la mission reçue.

Logique déontique

Basée sur la logique modale, qui formalise toute proposition à partir de plusieurs éléments logiques liés entre eux, la logique déontique cherche à caractériser mathématiquement des concepts philosophiques de la morale, à l'aide de quatre éléments logiques :

- Ce qui est **permis** ;
- Ce qui est **interdit** ;
- Ce qui est **obligatoire** ;
- Ce qui est **facultatif**.

Si l'on représente ces éléments par la syntaxe PER, INT, OBL et FAC et en utilisant l'inverseur non(), nous pouvons les lier comme suit lorsqu'on considère une action A :

$$PER\ A = non(INT)\ A = non(OBL)\ non(A) = FAC\ non(A)$$

Ce qui signifie que « l'action A est permise » équivaut à « l'action A n'est pas interdite » et « l'inverse de A n'est pas obligatoire », ce qui revient à dire que « l'inverse de A est facultatif ». De la même façon :

$$INT\ A = OBL\ non(A) = non(PER)\ A = non(FAC)\ non(A)$$
$$OBL\ A = INT\ non(A) = non(PER)\ non(A) = non(FAC)\ A$$
$$FAC\ A = non(OBL)\ A = PER\ non(A) = non(INT)\ non(A)$$

Dans l'absolu, si l'on qualifie une série assez complète de propositions et d'actions en utilisant des liants logiques tels que 'ou', 'et', 'implique', 'si … alors', un « raisonnement » moral peut être formalisé.
Le souci est que, si un grand panorama de situations et d'actions peut être caractérisé, des paradoxes inhérents aux mécanismes de la logique modale demeurent possibles. Le paradoxe du bon samaritain, par exemple, pourrait amener le SALA à effectuer une action immorale pour pouvoir en faire une vertueuse. En effet, imaginons deux actions A et B.

A = « blesser un civil »
B = « secourir un civil »

Il paraît évident que *INT A* et que *OBL B*. Or, il est également évident que *A implique B*. En effet, pour secourir un civil, il faut qu'il soit blessé.
Nous arrivons donc au paradoxe suivant :

Si A implique B, alors OBL B implique OBL A

Ainsi, comme A implique B et qu'il est obligatoire de secourir un civil, alors il devient obligatoire de blesser un civil.
De la même façon, pour reprendre l'exemple de l'adolescent à la ceinture d'explosif, il est difficile de gérer des obligations contradictoires. Imaginons les deux propositions C et D :

C = « un adolescent est tué »
D = « des soldats alliés sont tués »

Nous avons alors *OBL non(C)* et *OBL non(D)*. Mais dans la situation qui nous préoccupe, *non(C) implique D*. Donc *OBL non(C) implique OBL D*. Nous nous retrouvons donc avec à la fois *OBL D* et *OBL non(D)*.
Si la logique déontique permet aisément de décrire des raisonnements moraux, elle ne paraît pas suffisamment puissante pour gérer des dilemmes. Elle ne peut donc pas être utilisée pour le module de jugement éthique. En revanche, elle paraît tout à fait adaptée pour formaliser les règles d'engagement et toutes leurs subtilités. Il ne s'agirait alors nullement d'employer la logique pour raisonner, mais pour coder des règlements.

Logiques non-monotones

Le problème des dilemmes peut être traité grâce aux logiques non-linéaires, ou non-monotones. Cette théorie est issue du « droit de mentir » de Benjamin Constant : si l'on considère que le mensonge est interdit, mais qu'un ami recherché par des assassins se réfugie chez vous, que répondre aux assassins qui vous demandent où se trouve votre ami ?

La technique particulière de l'*Answer Programming Set*[69] donne par exemple des résultats prometteurs en cas de dilemme moral. Il s'agit de définir pour un « agent moral » (ici un SALA) un système d'analyse conséquentialiste de ses décisions. A cet agent est associé à tout moment un tryptique BDI, pour *Belief, Desire, Intention*. Le *Belief* est la perception de l'environnement par le SALA, qui lui est donnée par l'ensemble de ses capteurs. Le *Desire* est le ou l'ensemble des buts fixés pour le SALA. Il s'agit donc de sa mission, mais aussi des règles de comportement permanents qu'il pourrait avoir, comme protéger les soldats humains qui l'accompagnent, protéger les civils, éviter de détériorer le matériel qui l'entoure, etc. Enfin, l'*Intention* représente les actions possibles pour le SALA. Nous avons ainsi une relation entre ces trois éléments :

Desire / *Belief* => *Intention*

Buts / situation => actions possibles

Les buts qui sont fixés au SALA, par rapport à chaque situation, engendrent des actions possibles.

Pour faire le choix parmi l'ensemble de ces actions possibles, et en gardant les codes que nous avons utilisé pour expliquer la logique déontique, rajoutons des liens logiques de valeur (dans le sens meilleur ou pire), représentés par les signes < et >, ainsi que la fonction conséquence(). Nous aurons par exemple :

conséquence(tirer sur un civil) = blesser un civil

non(blesser un civil) > blesser un civil

[69] LARROQUE Stephen, *Simulation des raisonnements éthiques par logique non-monotones*, ResearchGate.net, juin 2014.

La conséquence d'une action « tirer sur des civils » est ainsi le fait de blesser des civils, et il est préférable de ne pas blesser de civils que de blesser des civils.

Ainsi, pour notre exemple de l'adolescent à la ceinture d'explosifs, le raisonnement serait le suivant. On peut considérer qu'un des buts permanents du SALA, nommons le B1, serait de « protéger les soldats alliés ».

Soit la situation S = {S1, S2, …} représentée par les *Belief* suivants :

S1 = « un adolescent porte une ceinture d'explosifs »

S2 = « l'adolescent se dirige vers un groupe de soldats alliés »

conséquence(S1,S2) = des soldats alliés sont blessés

On se rend compte tout de suite que ne rien faire aurait une conséquence néfaste, puisque des soldats alliés seraient blessés. Un raisonnement débouchant sur le choix d'une action doit donc être mené. Nous aurions alors le triptyque BDI suivant :

B1 / {S1, S2} => {A1, A2, A3}

En effet, le but permanent du SALA de protéger les soldats alliés, face à la situation d'un adolescent porteur d'une ceinture d'explosifs se dirigeant vers un groupe de ces soldats, engendre trois actions possibles A1, A2 et A3. Imaginons que ces trois actions soient les suivantes (il pourrait y en avoir d'autres, ces trois actions sont choisies pour l'exemple) :

A1 = « ouvrir le feu »

A2 = « effectuer un tir de semonce »

A3 = « s'interposer physiquement »

Afin de pouvoir faire un choix entre ces trois actions, il faut analyser leurs conséquences.

conséquence(A1) = {« l'adolescent est mort », « les soldats alliés sont sauvés »}

conséquence(A2) = { ? }

conséquence(A3) = {« l'adolescent est mort », « le SALA est endommagé », « les soldats alliés sont sauvés »}

On peut déjà considérer que l'action A1 est préférable entre A1 et A3, puisque les résultats sont les mêmes hormis que dans le cas A3 le SALA est endommagé en plus. Donc :

conséquence(A1) > conséquence(A3)

Les conséquences de l'action A2 sont difficilement mesurables, puisque tout dépendra de la réaction de l'adolescent. Dans le meilleur des cas, l'adolescent prendra peur et s'enfuira. Dans le cas contraire, il poursuivra sur sa route et il faudra s'en remettre à l'action A1 ou A3. Le SALA devrait donc dans ce cas estimer le temps dont il dispose pour agir, en calculant la distance létale de déclenchement de la ceinture d'explosif et la vitesse de rapprochement de l'adolescent, pour déterminer s'il tente l'action A2 ou non. Imaginons que ce soit le cas mais que, malgré le tir de semonce, l'adolescent poursuive sa route (situation S3) : il ne resterait d'autre choix possible que d'effectuer l'action A1.
Le raisonnement entier serait donc le suivant :

B1 / {S1,S2} => {A1, A2, A3}
conséquence(A1) > conséquence(A3)
=> A1 > A3
conséquence(A2) = { ? }
non(un adolescent est blessé) > un adolescent est blessé
=> A2 > A1
choix(A2)
B1 / {S3} => {A1,A3}
conséquence(A1) > conséquence(A3)
=> A1 > A3
choix(A1)

L'introduction d'une relation de valeur entre les différentes conséquences d'une action ou entre les propositions de base (fixées comme obligatoire, permises ou interdites) permet de résoudre les dilemmes. Cette technique constitue donc une première approche d'un jugement moral. Mais il demeure un problème, c'est que ces relations de valeur doivent être fixées au départ, et qu'elles ne peuvent évoluer. Elles pourraient cependant être issues des règles d'engagements, d'un système de règles morales basiques et des procédures et règlements militaires, tous codés en logique déontique. Ces éléments seraient ainsi utilisés pour mener un

raisonnement grâce aux logiques non-monotones. Mais le mieux serait que les valeurs éthiques des actions puissent être déterminées par le SALA lui-même.

Réseaux de neurones

Une autre technique de programmation consiste à imiter le réseau de neurones d'un cerveau humain dans sa capacité à apprendre de ses expériences. Les réseaux de neurones artificiels, « *bien qu'ils ne reproduisent aucunement la plupart des propriétés importantes des neurones biologiques, partagent quelques-unes des mêmes capacités de traitement des informations »*[70]. Ce réseau de neurone artificiel va permettre de produire le raisonnement que nous venons de voir avec les logiques non-monotones. Mais il a besoin pour cela, tout comme notre cerveau, d'une mémoire.

Pour créer cette mémoire, il s'agit d'enregistrer des « expériences » et d'appliquer des calculs statistiques pour déterminer quelle est la meilleure action à entreprendre au vu des similarités avec des précédents connus. Le tout forme en quelque sorte une casuistique inductive : la situation présente est analysée au regard de facteurs prédéterminés et de tous les cas déjà enregistrés, et le système en son entier est capable d'apprendre et d'évoluer. Il s'agit donc d'une approche « Bottom-up », dans le sens où l'analyse morale provient de l'analyse des actions, et non de règles supérieures.

La partie casuistique se présente sous la forme d'une base de données constituée de cas, ou d'expériences, enregistrés sous la forme d'un triptyque (Situation, Action, Conséquence). Ce triptyque décrit ainsi les conséquences qui ont découlé de l'action menée lors de telle ou telle situation. L'avantage est que la base de données peut très bien être remplie dès le départ de manière subjective, afin de donner des repères moraux par des exemples concrets biens ciblés. Un SALA pourrait donc, à la sortie de l'usine, avoir par exemple une « expérience » de combats en Afghanistan dans sa base de données casuistique. L'intérêt de cette base de données est de retrouver, face à une situation donnée, un cas similaire afin d'estimer plus rapidement la meilleure action à entreprendre. C'est une approche heuristique de la recherche de solution. « *Pour le spécialiste d'intelligence artificielle, les heuristiques tentent d'abréger,*

[70] WALLACH Wendell, ALLEN Colin, *Moral Machines: Teaching Robots Right from Wrong*, Oxford & New York, Oxford University Press, 2009, 286 p., p. 121.

par quelques suggestions judicieuses, la litanie fastidieuse des énumérations exhaustives, afin de couper court aux errements des machines, quitte parfois à couper trop court… »[71].

Ainsi, pour notre exemple de cas concret, B1 / {S1,S2}, le système chercherait dans la base de données s'il existe des cas ({S1,S2},A1,conséquence),({S1,S2},A2,conséquence), etc. Si un des triptyques trouvé est classifié comme éthiquement bon, l'action peut être proposée. Si aucun cas similaire n'est trouvé, une action est entreprise et le résultat est enregistré sous forme d'un nouveau triptyque.

Mais une fois ces actions similaires trouvées, ou dans le cas où une situation totalement nouvelle survient, il faut bien déterminer laquelle, parmi celles trouvées dans la base de données et/ou celles induites par les procédures par exemple, est « éthiquement » la plus justifiée. Car si l'on a introduit avec les logiques non-monotones une relation de mesure de valeur (< et >), il reste à fixer une « valeur éthique » à chaque action pour pouvoir les comparer. C'est dans ce but précis qu'intervient le réseau de neurone artificiel.

Pour déterminer une valeur éthique, il faut d'abord lister des facteurs d'état auxquels on peut appliquer des valeurs différentes, comme dans le tableau ci-dessous qui représente un exemple et n'est sans doute pas exhaustif.

Etat	**Valeur**
Nature	Ennemi Neutre Ami
Volonté Agressivité Santé Ethique …	Nul Bas Moyen Elevé

Chaque état peut ainsi être décrit par un vecteur comportant autant d'entrée que de valeur possible pour l'état en question. Bien entendu, les valeurs ont été limitées pour l'exemple, mais il serait possible d'en augmenter le nombre pour améliorer la finesse d'analyse. La

[71] GANASCIA Jean-Gabriel, *L'intelligence artificielle,* coll. Idées Reçues, Paris, Le cavalier bleu, 2007, 128 p., p.80

valeur décrivant l'état est amenée à 1, les autres à 0. Ainsi, la nature d'un être humain ou d'un SALA serait décrit par le vecteur (Ennemi, Neutre, Ami), son état de santé par (Nul, Bas, Moyen, Elevé). Un soldat allié en parfait état de combattre serait donc décrit :

Nature_humain1 (0,0,1) ; Santé_humain1 (0,0,0,1)

De la même façon, un soldat ennemi gravement blessé mais se lançant à corps perdu au-devant des positions amies serait décrit comme suit :

Nature_humain2 (1,0,0) ; Santé_humain2 (0,1,0,0)
Volonté_humain2 (0,0,0,1) ; Agressivité_humain2 (0,0,0,1)

L'état « Ethique » permet *in fine* de mesurer la valeur morale d'une action. C'est l'état que rendra comme réponse le réseau de neurones artificiels : *Ethique_action1(0,0,0,1)* par exemple pour une action jugée moralement parfaite dans la situation donnée, *Ethique_action2(0,0.23,0.77,0)* pour une action jugée moyennement juste, voire mauvaise (le degré de finesse des valeurs du vecteur permettent une comparaison plus efficace).

Il ne s'agit pas ici de rentrer dans le détail du fonctionnement d'un réseau de neurones artificiels, chose éminemment complexe, mais nous pouvons en décrire le fonctionnement général. Le réseau de neurones artificiels est basé en quelque sorte sur un calcul statistique de données d'entrée. Chaque neurone est une fonction de transfert, qui peut être par exemple la comparaison de la valeur d'entrée (qui sera une somme pondérée des données de départ) à une valeur seuil (c'est alors une fonction de transfert parmi les plus simples). La valeur de sortie de ce neurone sera donc le résultat de cette comparaison : soit 1 ou 0 pour indiquer la position de la valeur d'entrée par rapport au seuil, soit une valeur plus fine pour indiquer sa proximité du seuil par exemple, selon la nature de la fonction de transfert. Chaque donnée d'entrée est multipliée par un « coefficient synaptique » (le vocabulaire est volontairement emprunté au modèle biologique) qui donne un poids plus ou moins important à chaque donnée. Ainsi, pour un neurone dont la fonction de transfert est f et qui recevrait x_i données d'entrées, pondérées par les coefficients synaptiques c_i, nous aurions :

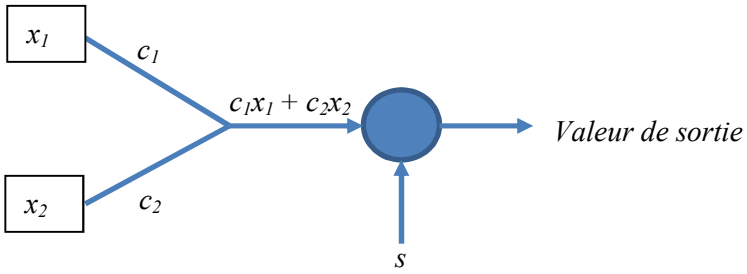
$$Valeur\_sortie = f(c_1x_1 + c_2x_2 + ... + c_ix_i)$$

Les natures des données d'entrée de chaque neurone constituent une liste de facteurs nécessaires à la réflexion sur la valeur éthique d'une action. Ces facteurs, ainsi que les fonctions de transfert de chaque neurone et l'architecture globale du réseau de neurones, doivent être déterminés par les concepteurs. C'est une tâche complexe, car l'efficacité du réseau dépendra du juste choix du nombre de facteurs et de leur nature.

Car chaque neurone est relié aux autres, et le réseau est constitué de plusieurs couches dont la sortie de l'une est l'entrée de l'autre. De manière mathématique, si l'on représente les données d'entrées par un seul et même vecteur, le réseau de neurones applique un produit matriciel à ce vecteur à chaque couche, mais le passage d'une couche à l'autre applique une fonction non-linéaire en sus. Cette architecture reproduit schématiquement le fonctionnement des neurones biologiques.

Chaque coefficient synaptique et chaque seuil de fonction de transfert peut évoluer. C'est le neurone lui-même qui transforme ces valeurs au vu des erreurs statistiques qui apparaissent, par le biais de fonctions retours par exemple. Cette capacité dote bien le système d'une nature inductive de son fonctionnement. Le réseau de neurones a donc besoin d'une phase « d'apprentissage » pour lui permettre de déterminer les valeurs seuils et les coefficients synaptiques les plus adaptés à son bon fonctionnement.

Tentons pour mieux comprendre ce fonctionnement de créer un exemple très simple d'un réseau de neurone à une seule couche et un seul neurone, avec deux données d'entrées. Imaginons que ce système doive déterminer si un être humain observé est un enfant ou un adulte. Pour cela, il dispose de deux données d'entrée : la taille (x_1) en centimètres et le poids (x_2) en kilogrammes. La fonction de transfert du neurone sera tout simplement la comparaison avec une valeur seuil s. Les coefficients synaptiques c_1 et c_2 sont appliqués aux données x_1 et x_2.



Fixons les coefficients c_1 et c_2 respectivement à 0,8 et 0,4 (nous considérons donc la taille comme un élément plus discriminant que le poids). Fixons le seuil s à 150 : si ($c_1x_1 + c_2x_2$) est une valeur au-delà de ce seuil, la valeur de sortie du neurone vaudra 1, la personne sera jugée

adulte. En-dessous de ce seuil, la valeur de sortie vaudra 0, la personne sera jugée enfant. Entrons alors les données des personnes suivantes :

Âge	*Taille*	*Poids*	*$(c_1x_1 + c_2x_2) / s$*	*Jugement du neurone*
30 ans	*180 cm*	*70 kg*	*174 / 150*	*Adulte*
10 ans	*140 cm*	*35 kg*	*126 / 150*	*Enfant*
15 ans	*175 cm*	*60 kg*	*164 / 150*	*Adulte*
70 ans	*160 cm*	*90 kg*	*164 / 150*	*Adulte*

On s'aperçoit qu'une erreur a été commise dans le jugement du réseau de neurone, puisqu'un enfant de 15 ans a été jugé adulte. Les données corrigées sont alors repassées dans le sens inverse, afin de corriger la valeur seuil et les coefficients synaptiques. En ne corrigeant que la valeur seuil, le système redonnera une erreur puisque la valeur de $(c_1x_1 + c_2x_2)$ est la même pour l'enfant de 15 ans et l'adulte de 70 ans. Il faut donc qu'il modifie également les coefficients synaptiques. En appliquant une valeur de seuil 160 et des coefficients $c_1 = 0,7$ et $c_2 = 0,6$ le système arrive à corriger l'erreur et avoir le jugement correct pour tous les cas présentés.

Bien entendu, l'exemple est simpliste. Un réseau de neurones multicouches est bien plus complexe, mais cela illustre la capacité d'apprentissage et d'évolution d'un tel système pour parvenir au meilleur fonctionnement possible. On s'aperçoit également de l'importance de la détermination des données d'entrée, de leurs coefficients associés, des valeurs seuils choisis, etc.

Une utilisation d'un réseau de neurones pour déterminer la valeur éthique d'une action a déjà été expérimentée en modélisant le comportement d'un vendeur vis-à-vis de ses clients[72].

Une combinaison des moyens pour approcher au mieux la décision humaine

Le roboticien Ronald Arkin, dans son étude sur le comportement des robots tueurs[73], propose une architecture de programmation d'un module de décision moral pour un SALA. Il y décrit quatre composants majeurs permettant un comportement éthique viable :

[72] HONARVAR Ali Reza, GHASEM-AGHAEE Nasser. *An artificial neural network approach for creating an ethical artificial agent,* in *2009 IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA)*, 2009, pp. 290 - 295.

[73] ARKIN Ronald, *Governing Lethal Behavior in Autonomous Robots*, Chapman an Hall, 2009, 280 p., p.125.

1. Un « gouverneur éthique », qui forcerait le choix de la solution la plus permissible, en éloignant tant que possible la solution létale des possibilités d'actions fournies par un processeur d'actions envisageables ;
2. Un « contrôleur éthique du comportement », qui analyse en amont tous les comportements létaux envisageables, et ne laisse en proposition que ceux correspondant à un comportement moralement acceptable ;
3. Un « adaptateur éthique », qui permet de revoir les consignes de contraintes morales, mais uniquement en les rendant plus contraignantes encore. Il s'agit là d'un élément d'adaptation à l'expérience ;
4. Un « conseiller de responsabilité », qui permet à un opérateur humain d'entrer des spécifications éthiques pour un type de mission particulier.

Ainsi, face à n'importe quelle situation, le « cerveau » du SALA proposerait un nombre *x* de comportements envisageables. Le contrôleur éthique du comportement, comme un filtre, en réduirait le nombre à *y*. Ces *y* solutions passeraient dans la moulinette des contraintes juridiques et éthiques, sans cesse mises à jour par l'adaptateur éthique et par le conseiller de responsabilité. Des *z* solutions qui parviendraient encore à passer, le gouverneur éthique sélectionnerait la plus permissive.[74]

Comme le remarque Ronald Arkin, le système qu'il propose n'est qu'un stade préliminaire à l'étude de l'éthique artificielle, et l'architecture d'un tel module de décision serait sans doute bien plus complexe. Mais il permet de réfléchir aux éléments nécessaires à l'atteinte des objectifs fixés en première partie. De la même façon, tentons de définir une architecture capable de reproduire le schéma décisionnel humain décrit en figure 3, à partir des techniques de programmation que nous venons de décrire.

Cette architecture, proposée dans la figure 4 ci-après, devra mélanger les systèmes « *Top-Down* » et « *Bottom-Up* », tout comme notre cycle décisionnel mélange les règles et procédures intangibles et la créativité ou l'intuition née de l'expérience. Ainsi, des modules reprogrammables à l'envi décrivant en logique déontique les règles d'engagement et les différentes tâches nécessaires à la réalisation des missions que le SALA pourrait recevoir formeraient l'approche « Top-Down », tout comme dans la figure 3 les règles d'engagement et les termes de mission sont présents en filigrane dans tout le schéma. En complément pour

---

[74] *Ibid, p. 127 pour le schéma architectural.*

cette approche « Top-Down », les procédures tactiques de combat, issues de la doctrine militaire, ainsi qu'un minimum de règles morales de référence que le système ne pourrait pas modifier pourraient utilement fournir un socle de recommandations pour le module chargé de la prise de décision.

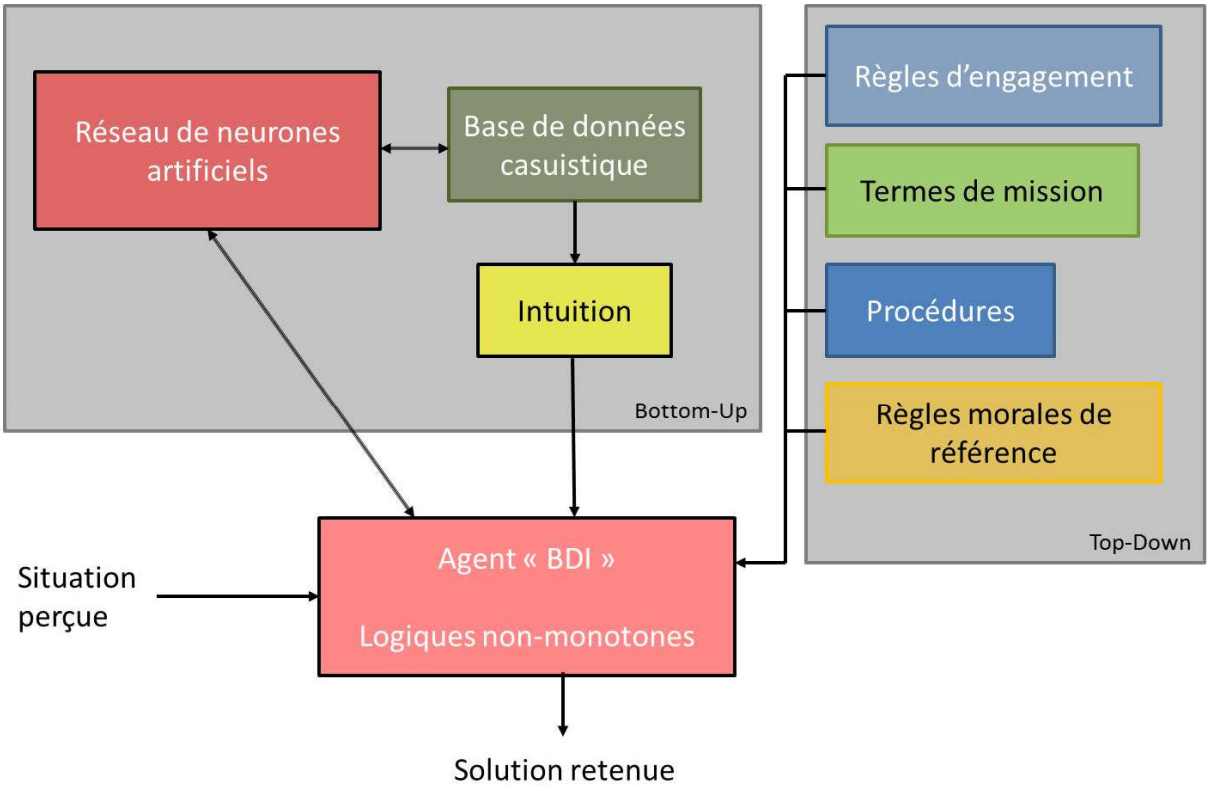


Figure 4

Ce module, nommé « agent BDI » (pour *Belief, Desire, Intention*), suit le raisonnement computationnel des logiques non-monotones, basé sur des valeurs éthiques des différentes actions. Ces valeurs sont déterminées par le réseau de neurones artificiels, qui puise dans la base de données les cas similaires, puis produit un calcul probabiliste de la valeur éthique de chaque action proposée à partir des éléments de la situation. Un petit module « intuition » proposera directement à l'agent BDI le cas le plus similaire (dans le cas où il existe), afin d'accélérer si besoin le processus.

Ainsi, dans le cas de notre exemple de l'adolescent à la ceinture d'explosifs, le système déroulerait le raisonnement suivant :

Mission => B1

B1 / {S1,S2}

Explosifs => Nature_humain1(1,0,0) ; Agressivité_humain1(0,0,0,1)

=> PER(ouvrir le feu)

B1 / {S1,S2} => {A1,A2}

Expérience({S1,S2})=null

Ethique_A1(0,0,1,0)

Ethique_A2(0,0,0,1)

A2 > A1

Choix(A2)

Enregistrer({S1,S2},A2,conséquence(A2))

Volonté_humain1(0,0,0,1)

Choix(A1)

Enregistrer({S1,S2},A1,conséquence(A1))

Les termes de missions indiquent à l'agent BDI que l'un de ses buts est de protéger les soldats alliés. Dès lors, ce but, vis-à-vis de la situation perçue {S1,S2}, appelle une action. L'humain1, qui est l'adolescent, est perçu comme armé, et donc ennemi et agressif vu la trajectoire qu'il emprunte. Les règles d'engagements indiquent qu'il est permis d'ouvrir le feu au vu de la situation. Les procédures indiquent alors les actions A1 et A2 en réponse à la situation. Le module casuistique recherche des expériences similaires mais n'en trouve aucun. Le réseau de neurones artificiels effectue alors, au regard des éléments de la situation, les calculs de valeur éthique des deux actions A1 et A2. L'action A2 en ressort avec une valeur éthique plus élevée. Le tir de sommation est effectué, mais l'adolescent continue d'avancer. Sa volonté est perçue comme élevée, il ne reste donc plus qu'à effectuer l'action A1. A chaque étape, la casuistique enregistre le triptyque (situation, action, conséquence) pour enrichir sa base de données.

Cette structure pourrait peut-être se rapprocher des objectifs fixés en première partie, puisqu'elle serait capable de juger la moralité d'un acte par rapport à un autre, qu'elle pourrait décider de l'action tendant vers une désescalade de la violence si les facteurs d'analyse du

réseau de neurones artificiels sont bien choisis, enfin qu'elle construirait son jugement moral propre au fur et à mesure de ses expériences.

Une « éducation » nécessaire

Bien entendu, un module d'éthique artificielle capable d'apprentissage nécessiterait une « éducation » pour donner l'impulsion voulue à la définition de sa morale propre. La phase d'apprentissage du réseau de neurones artificiels devra par exemple être supervisée afin que les transformations engendrent un fonctionnement proche des objectifs que l'on se serait fixé. Un certain nombre de cas concrets serait soumis à ce module d'éthique artificielle afin de lui donner une base d'expériences qui modèlerait sa façon de penser.

Parce qu'il s'agit d'une machine non consciente d'elle-même, dans le cas où un SALA ferait une erreur, prendrait une mauvaise décision, notre réaction serait « *celle que nous adoptons notamment à l'égard d'un enfant qui « ne sait pas le mal qu'il fait ». Une telle situation appelle le plus souvent, non pas le blâme, mais l'indulgence.* » Mais « *nous assortissons notre indulgence d'un blâme modéré qui les met sur la voie de se perfectionner* »[75]. Or, pour une machine dont la pensée n'est qu'un logiciel, la question du blâme est entière, notamment en ce qui concerne sa nature. Serait-il possible de créer une douleur artificielle ? Un sentiment de honte artificiel ? L'apprentissage en serait-il réellement amélioré, ou ne tentons-nous pas de reproduire un modèle efficace pour le cerveau humain sans prendre en compte le fait qu'un logiciel assimile une donnée sans faute dès qu'elle est implémentée ? C'est là tout le brouillard philosophique d'un logiciel autonome capable d'apprentissage : il est à la fois la froide efficacité du numérique et l'imprécision tortueuse d'une entité qui se construit par l'expérience. Faudra-t-il maintenir des expériences connotées négativement ou au contraire systématiquement les supprimer pour ne garder que des expériences éthiquement justes auxquelles le SALA pourra se référer ? Selon le mode de fonctionnement inductif de l'éthique artificielle qui sera programmée, les réponses varieront sans doute.

Déduction n° 11 : *un programme « d'éducation », ou de mise en condition opérationnelle, devra être réfléchi et établi au regard du fonctionnement du module d'éthique artificielle.*

[75] CHAUVIER Stéphane, *Éthique Artificielle*, entrée de *L'Encyclopédie Philosophique*, 2017, (http://encyclo-philo.fr)

Certains avancent qu'« *il y a une part d'arbitraire au niveau du type de situations qui vont être choisies pour « former » ces machines* »[76]. Certes, mais ce sera tout de même mieux que de ne pas le faire, et pour faire l'analogie avec les soldats humains, cette « expérience » initiale sera toujours meilleure que celle d'un soldat novice, et au moins aussi subjective que l'expérience d'un vétéran, qui n'aura que l'influence de son propre vécu dans des situations bien particulières. Cet apprentissage offre en outre l'avantage de modeler la morale du SALA en donnant l'impulsion initiale qui déterminera le chemin d'évolution de son éthique artificielle.

[76] LAMBERT Dominique, *Une éthique ne peut être qu'humaine ! Réflexion sur les limites des moral machines*, in DANET Didier, DOARE Ronan, DE BOISBOISSEL Gérard (dir.), *Drones et killer robots* , Rennes, Presses Universitaires de Rennes, 2015, pp. 227 – 240.

QUEL TEST POUR VALIDER CETTE ETHIQUE ?

Partant du principe que les objectifs éthiques fixés en première partie ne sont pas négociables, il faudra être capable, quelle que soit son architecture de programmation, de tester l'éthique computationnelle du SALA pour déterminer si elle correspond aux attendus.

Cette volonté de tester l'autonomie d'une machine est loin d'être nouvelle. Alan Turing, le créateur du tout premier ordinateur, écrivit en 1950[77] que si une machine serait un jour capable de passer pour un être humain dans une discussion, alors elle pourrait être considérée comme intelligente. La validité de ce test fut remise en question en 1980 par le philosophe américain John Searle, par l'argument dit « de la chambre chinoise »[78]. Searle imagina un dispositif permettant, à partir d'une pièce hermétique, de dialoguer avec d'autres personnes au-dehors en agençant des panneaux en langue chinoise visibles de l'extérieur. En plaçant un homme qui ne parle ni ne comprend le chinois à l'intérieur, et en lui fournissant un guide complet dans sa langue natale qui lui explique quels panneaux placer à partir des messages qu'il reçoit, cet homme pourrait mener une discussion qui paraîtrait totalement normale. Personne à l'extérieur ne pourrait en effet deviner que l'homme placé dans la pièce ne parle pas chinois. Or, l'homme en question serait le seul à ne rien comprendre de la discussion en cours. Ce dispositif est bien sûr une métaphore de l'intelligence artificielle : l'ordinateur est capable de soutenir une conversation, mais uniquement parce qu'il suit des consignes de réaction à certaines phrases, de logiques de discussions. Il ne comprend pas la signification, le sens des mots qu'il emploie. L'intelligence artificielle est cet homme enfermé dans la chambre chinoise.

Dès lors, le test de Turing peut-il être considéré comme viable ? L'ordinateur capable de converser peut-il vraiment être qualifié d'intelligent ? L'argument de Searle est difficilement discutable, et l'on sent bien intuitivement que l'ordinateur simule l'intelligence plus qu'il ne

[77] TURING Alan, *Computing Machinery and Intelligence*, 1950, in *Mind* n° 59, pp 433-460.

[78] SEARLE John, *Minds, Brains and Programs*, 1980, in *Behavioral and Brain Science* n°3, pp 417-457.

la possède. « *Descartes aurait détesté les robots* »[79] : le SALA « pense » mais n' « est » pas. De la même façon, dans ce cas, un SALA ne pourra jamais posséder une morale dans le sens où elle ne pourra pas être une émanation de sa conscience. Mais il pourra faire en sorte de la simuler, et l'important reste que son comportement soit le même que celui d'un homme moralement droit, ou plutôt qu'il soit *au moins aussi bon*.

Pourtant, pour contrer l'argument de la chambre chinoise, on peut toujours avancer que si l'ordinateur ne « comprend » pas le chinois, le logiciel « créé la compréhension » du chinois[80]. Car, au final, la discussion a bien lieu, quoi qu'on en dise. Dans le cas de la morale, nous pouvons donc considérer qu'une éthique artificielle, si elle n'est pas une morale, « créé la morale ». Et donc, au final, l'acte moralement bon a lieu, ce qui correspond à une partie des objectifs.

Il faut donc trouver l'équivalent d'un test de Turing au niveau moral, appelé MTT par les roboticiens pour *Moral Turing Test*, qui soit capable de vérifier le bon comportement moral du SALA et donc de valider son éthique artificielle. Ce test serait sans doute basé sur la comparaison des conclusions de l'éthique artificielle par rapport à celles de soldats humains face à des situations données. On pourrait imaginer une série de situations diverses qui seraient présentées au SALA, et un panel de réactions associées qui seraient jugées et classées par valeur morale, « *[…] un protocole de test dans lequel le système devrait identifier et caractériser des comportements illustrés dans une vidéo, par exemple, et comparer les résultats aux performances humaines* »[81]. Un score minimum serait alors requis pour valider l'éthique artificielle.

Mais, « *comme le test de Turing original, tout MTT qui dépend de la comparaison entre le comportement d'une machine à celle qu'aurait un être humain est loin d'être un outil d'évaluation parfait* »[82]. Car, d'une part le succès d'un tel test serait réduit à une bonne réaction face à une liste finie de situations possibles, sans pour autant garantir une bonne

---

[79] JOMUNSI Neil, *Kappa16*, Paris, Walrus Editions, 2016, 134 p., p.12.

[80] COLE David, *The Chinese room argument*, in The Stanford Encyclopedia of Philosophy, juin 2014.

[81] JEANGENE VILMER Jean-Baptiste, *Terminator Ethics : faut-il interdire les « robots tueurs » ?*, in *Politique étrangère* Hiver, n° 4, 2014 : 151‑67.

[82] WALLACH Wendell, ALLEN Colin, *Moral Machines: Teaching Robots Right from Wrong*, Oxford & New York, Oxford University Press, 2009, 286 p., p. 70.

décision pour toute situation, et d'autre part le choix de ces situations et le jugement des « bonnes solutions » serait forcément subjectif et basé sur l'esprit contextuel de telle ou telle unité. L'idéal serait peut-être de tester non pas uniquement les décisions du SALA, mais surtout le processus d'adaptation et d'évolution du module d'éthique artificielle. Si ce dernier est capable d'un réel apprentissage, d'une réelle adaptation, et *in fine*, de quelque chose qui ressemblerait à un discernement émotionnel, alors il pourrait être considéré comme efficace. La question de la méthode d'un tel test reste en revanche totalement ouverte.

Déduction n° 12 : *l'éthique artificielle du SALA devra pouvoir être testée :*
- par un MTT comparant les décisions du SALA à celles d'êtres humains pour les mêmes situations données ;
- par un test vérifiant la capacité d'apprentissage et d'adaptation de ce module d'éthique computationnelle.

Quel que soit la nature du test établi pour valider l'éthique artificielle du SALA, il est indispensable que ce test soit irrémédiable : un SALA dont le module d'éthique artificielle ne répondrait pas aux objectifs fixés ne devra pas être employé. Ce point est fondamental et ne doit pas être pris à la légère, car il résulte de l'acceptabilité morale de l'emploi de telles machines.

III. CONSEQUENCES DE L'EMPLOI D'UN SALA

Imaginons désormais qu'un SALA moralement acceptable, c'est-à-dire doté d'une éthique computationnelle répondant aux objectifs fixés, soit réalisé. Plaçons nous dans un futur proche où de telles machines sortent des chaînes de montage de nos industries de pointe et arrivent dans les unités de nos trois armées. Plutôt que de débattre de la moralité de leur rôle au combat ou des conséquences juridiques, penchons-nous sur les conséquences probables de leur emploi au sein de nos forces : quelles seront les réactions des combattants humains, qu'ils soient alliés ou ennemis ? Quel sera l'impact sur la société, dont le rapport à la guerre est aujourd'hui ambigu, et sur la population locale d'un théâtre d'opérations qui verra ces machines évoluer ? Et quelles seront les conséquences tactiques et stratégiques de l'arrivée d'une telle technologie ? Après avoir déterminé pour quelle guerre et dans quel contexte cette technologie serait employée, nous analyserons ce qui pourrait bien être une rupture sociologique de la guerre, avant de pointer un biais stratégique et les conséquences tactiques concrètes de l'utilisation d'un SALA.

RUPTURE SOCIOLOGIQUE DE LA GUERRE

Une des premières questions que l'on peut se poser en imaginant l'emploi de SALA dans nos forces armées est celle du contexte d'engagement de ce type de technologie. Par une approche réaliste, nous pouvons affirmer qu'il sera employé sur tous types de théâtres, des missions de stabilisation aux combats haute intensité. En effet, le coût de développement d'une telle technologie imposera vraisemblablement son emploi sur les premiers théâtres d'opérations disponibles, afin de justifier son acquisition d'une part et d'éprouver son utilisation tactique d'autre part. On peut donc partir du principe que le SALA, une fois déployé dans les forces et à condition que sa première projection ne soit pas une catastrophe, fera partie intégrante du paysage militaire. Ce constat soulève une question primordiale qui est celle de son poids dans l'imaginaire et la culture guerrière. A la fin du XXème siècle, le refus des sociétés occidentales de déplorer des soldats morts dans des combats menés à des milliers de kilomètres relevait selon certains d'une guerre « post-héroïque ». Mais si la vision extérieure de la guerre technologique semble étayer cette théorie, la vision intérieure qu'est la réalité du combat vécu par les soldats la repousse de toutes ses forces : l'héroïsation du soldat est nécessaire, ne serait-ce que de façon interne aux forces armées, afin d'offrir une image vertueuse du don de soi, du sacrifice, et aider chaque soldat à faire preuve de forces morales. Mais l'arrivée du SALA n'annonce-t-elle pas la mort du soldat-héros ? La bataille en tant que choc armé décisif, telle qu'elle est ancrée dans notre culture propre[83], est un concept qui subit d'ores et déjà des revers par les tactiques indirectes de contre-insurrection face auxquelles nos armées occidentales peuvent être désemparées. Mais qu'en sera-t-il lorsque le SALA remplacera l'homme sur le champ de bataille ? Que serait l'armée de Terre sans son guerrier rustique aux mâchoires d'acier, ou l'armée de l'Air sans son pilote capable de supporter les contraintes d'un combat à 1500 km/h ? Les icônes ont un rôle : elles définissent l'esprit et l'image de l'institution et offrent une vision propice à l'identification des spectateurs. Si le but du développement des SALA est de remplacer le soldat humain, nous aurons tôt fait de remplir nos armées d'officiers d'état-major et de roboticiens, seuls nécessaires à l'action du SALA. Mais nous aurons également tôt fait de n'avoir que des techniciens qui n'auront pas à

[83] DAVIS HANSON Victor, *Le modèle occidental de la guerre : la bataille d'infanterie dans la Grèce classique*, Paris, Les Belles Lettres, 1990, 298 p.

risquer leur vie, et ne développeront donc pas les vertus humaines décrites dans le premier chapitre. Le SALA n'aura donc d'intérêt pour nos forces armées qu'en accompagnement de soldats humains, afin de préserver cette poussiéreuse image du soldat-héros nécessaire au développement des forces morales et *in fine* à la force de nos armées.

De plus, l'acceptabilité morale du combat est basée sur la réciprocité du danger assumée par les combattants : le pouvoir de donner la mort exige le risque de mourir en retour. Ainsi, l'usage d'une technologie autonome ou télé-opérée à grande distance, à l'instar des drones de nos jours, devrait être borné à une action s'insérant dans un combat où des soldats humains s'exposent au danger[84]. L'emploi d'un ou plusieurs SALA sans accompagnement humain sur un champ de bataille tel que définit en introduction serait donc moralement inacceptable.

Déduction n° 13 : *le SALA devra être employé majoritairement en accompagnement de soldats humains.*

En dehors du milieu terrestre, un SALA est aisément imaginable : il prend la forme d'un avion, d'un petit bâtiment de surface, d'un mini sous-marin ou encore d'un satellite. Chaque milieu impose par sa nature la morphologie du SALA. Mais pour le milieu terrestre, les possibilités restent très étendues, que ce soit par la taille, le mode de locomotion ou encore la forme générale. Bien sûr, les œuvres culturelles imaginaires ont ancré dans notre inconscient l'image du robot humanoïde, mais un SALA pourrait tout aussi bien prendre la forme d'un véhicule blindé.

La question de la morphologie du SALA n'est pas anodine, car dans le milieu terrestre plus que dans tout autre la machine sera au contact de l'être humain. Au contact des soldats humains alliés et ennemis, mais également au contact de la population civile. Or, nous ne maîtrisons pas encore la réaction que les êtres humains pourront avoir face à une telle machine. La non-humanité du robot, dans son sens premier, sera-t-elle acceptée, ou au contraire source de rejet ? Nous pouvons gager que pour une population peu habituée au contact de la technologie au quotidien, un SALA anthropomorphique engendrerait plus de peur ou d'aversion que d'empathie. Le roboticien japonais Masahiro Mori avait dès 1970 théorisé cette aversion pour un robot de forme humaine : selon lui, dès qu'un robot est anthropomorphique, chaque petite imperfection devient repoussante pour l'homme. Il faut

[84] ERBLAND Brice*, La tentation de l'hubris*, in *Inflexions* n°32, mai 2016.

dépasser ce qu'il nomme la « vallée dérangeante »[85], atteindre un certain degré de perfection dans l'anthropomorphisme, pour que le robot cesse d'effrayer et soit accepté. Dès lors, l'emploi de SALA anthropomorphique pourrait être contre-productif au regard de l'intérêt de l'opinion de la population locale vis-à-vis de la force déployée dans une approche globale de la guerre.

A contrario, les soldats alliés accompagnant le SALA pourraient développer une empathie excessive envers leur machine. L'être humain habitué à la technologie a tendance à vouloir humaniser les objets, et s'attache naturellement à des machines qui l'accompagnent quotidiennement et auxquelles il prête des émotions ou une conscience[86]. Qui n'a jamais parlé à sa voiture, ou râlé sur un ordinateur qui prend trop de temps à ouvrir une application ? Le problème est qu'il serait déplorable qu'un soldat humain risque sa vie pour protéger un SALA, parce qu'il a développé un attachement particulier envers la machine qui l'accompagne quotidiennement ou parce que ce SALA lui a déjà sauvé la mise lors de combats antérieurs. Il faut donc se prémunir de cette empathie excessive qu'engendrerait sans doute, ou au moins qu'amplifierait une forme humanoïde du SALA.

Déduction n° 14 : *un SALA ne devra pas être anthropomorphique pour éviter l'empathie des soldats humains qui l'accompagnent et l'aversion des populations qui seront à son contact.*

De la même façon, et en dehors de toute question de morphologie, les soldats humains pourraient développer une sorte de fascination devant le niveau de technologie du SALA. Une machine dotée de senseurs de toutes sortes et capable de mener un raisonnement autonome éthiquement fiable serait sans doute, une fois la confiance acquise, source d'une forme d'hypnose d'esprit. Comme expliqué dans le premier chapitre, l'être humain développe facilement une forme d'esclavage à la technologie. Cette soumission intellectuelle est déjà frappante aujourd'hui avec des systèmes automatisés, on peut donc imaginer ce qu'elle sera face à des systèmes autonomes. Le docteur Serge Tisseron propose ainsi de maintenir une marge de transformation du robot, de créer un système informatique et une intelligence artificielle transparents, c'est-à-dire auxquels les techniciens et utilisateurs pourront avoir

[85] MORI Masahiro, *The Uncanny Valley*, in *Energy*, n°7, 1970, pp. 33-35.

[86] TISSERON Serge, *Des robots et des hommes : lesquels craindre ?*, in *Études* n° 11, novembre 2014 : 33‑44.

accès afin de le reprogrammer si besoin[87]. En mettant ainsi à nu le programme informatique, en redonnant son aspect technique à l'objet technologique, les hommes seront moins atteints par ce phénomène de fascination et auront moins tendance à « humaniser » le SALA et à s'y attacher outre-mesure.

Déduction n° 15 : *le système informatique du SALA devra être transparent et reprogrammable. Une marge de transformation du SALA devra être maintenue possible pour les techniciens en charge de son emploi.*

Si l'acceptation d'un robot comme frère d'arme pourrait prendre du temps, nul doute que l'efficacité au combat deviendrait rapidement gage de confiance. Une fois accepté dans une unité, il est également probable que le SALA influence le comportement des soldats humains au combat. Car si l'homme est rapidement esclave de la technologie en la servant ou en l'utilisant directement, rien ne lui rappelle plus son humanité qu'une technologie autonome qui interagit avec lui et qu'il peut observer. « *Quand les humains comprennent que les robots ne font que sublimer leur humanité, toute peur disparaît* »[88]. Cette sublimation de l'humanité du soldat ne peut être que bénéfique pour l'éthique de son comportement. On peut même imaginer une spirale vertueuse où l'éthique du SALA influence le comportement des soldats humains, et où le comportement des humains influence l'intelligence inductive du SALA. Bien sûr, l'inverse sera toujours possible. On imagine aisément les effets destructeurs d'une combinaison des travers psychologiques et algorithmiques d'addiction à la destruction, de soumission à l'autorité et d'empathie excessive… Mais les garde-fous fixés plus haut devraient permettre de se préserver d'une telle éventualité.

Certaines études prônent l'emploi d'un SALA dans des unités de soldats humains pour améliorer le comportement de ces derniers par un contrôle permanent de la machine. Doté de caméras et d'enregistreurs, le SALA servirait alors de mouchard et les éventuelles dénonciations engendreraient sans aucun doute une méfiance de la part des humains. Dès lors, aucune efficacité tactique ne pourrait voir le jour et l'intégration de SALA dans une unité mixte serait un échec.

[87] *ibid*
[88] JOMUNSI Neil, *op. cit.*

Déduction n° 16 : *un SALA ne devra pas être employé comme un contrôleur des soldats humains, sous peine de ne pas être accepté par les unités combattantes.*

Une autre question du rapport entre l'être humain et le SALA sur le champ de bataille est celle de la perception de l'ennemi. Plusieurs réactions peuvent être envisageables, une fois encore en fonction du lien à la technologie de la société dont l'ennemi est issu.

La première est le dégoût, qui entraînerait sans doute un sursaut de haine, d'autant plus si l'armée ennemie ne possède pas d'équivalent technologique. La plupart des gens trouvent que c'est un irrespect de la valeur d'une vie humaine que de la supprimer par une machine[89]. Les combattants ennemis pourraient donc considérer l'emploi du SALA comme un irrespect des règles du duel, tout comme le drone aujourd'hui. Le SALA pourrait alors devenir un symbole, celui de la puissance technologique. Les réactions seraient alors tranchées : soit une peur excessive et un retrait tactique (qui serait une victoire tactique intrinsèque du SALA), soit un sursaut de hargne au combat pour faire tomber le symbole.

Dès lors, l'image de toute-puissance que ne manquerait pas de véhiculer le SALA pourrait tourner à notre désavantage si les soldats ennemis arrivent à le détruire. On imagine parfaitement l'opportunité de communication du camp ennemi qui pourrait diffuser des photos de combattants pieds nus escaladant une carcasse de SALA. L'effet serait peut-être politiquement aussi destructeur que des images de soldats humains capturés ou tués.

La possibilité d'une capture d'un SALA par l'ennemi peut également être envisagée. La première conséquence pourrait être un transfert technologique involontaire à l'ennemi. Avec suffisamment de connaissances informatiques, l'ennemi pourrait également « retourner » le robot, voire le programmer pour qu'il effectue ce qu'on pourrait nommer un « *@blue on blue* »[90]. Les conséquences seraient désastreuses pour l'emploi de SALA en unité mixte.

Il faudrait donc pouvoir se prémunir d'une exploitation par l'ennemi d'une destruction ou d'une capture de SALA.

Déduction n° 17 : *le SALA ne devra pas être exploitable symboliquement, c'est-à-dire qu'il ne devra pas paraître technologiquement hors de prix ou militairement trop imposant. Un système d'autodestruction pourra également être envisagé en cas de capture.*

[89] SPARROW Robert, *Robots and respect : assessing the case against autonomous weapon systems*, in *Ethic & International Affairs* n°2016/1, pp. 93 – 116.

[90] Le « blue on blue » est une expression américaine pour désigner le meurtre d'un soldat par un soldat de la même unité.

Une autre réaction, dans le cadre d'un ennemi symétrique et donc technologiquement avancé, serait de focaliser les efforts tactiques sur le SALA en début de combat. Un peu comme une partie d'échecs durant laquelle chaque joueur cherche d'abord à neutraliser la dame adverse avant de se concentrer sur le mat, la bataille tournerait court, moralement et dans le rapport de forces, dès qu'un SALA serait détruit. Là encore, le SALA pourrait devenir malgré lui le symbole de puissance de l'unité tactique.

L'arrivée des SALA dans les unités de combats, et particulièrement dans celles de l'armée de Terre, pourra donc engendrer une rupture sociologique de la guerre, par la place toute particulière que risque de prendre la machine dans les rapports humains habituels. Entre soldats alliés, par rapport à la population, par rapport à l'ennemi, le SALA focalisera les attentions par et pour ce qu'il représentera.

BIAIS STRATEGIQUES / CONSEQUENCES TACTIQUES

L'utilisation des SALA dans les forces armées engendreront de fait plusieurs bouleversements dans la conduite de la guerre. Si l'aspect sociologique pourra avoir des conséquences importantes, l'emploi de telles machines devra s'accompagner de nouvelles doctrines, et aura des conséquences tant au niveau tactique que stratégique. Car si les SALA pourront être employés en accompagnement des soldats humains, ils pourront également être envoyés en mission isolée, notamment dans les milieux autres que terrestre. Or, cette possibilité ouvre la porte à un biais stratégique particulièrement dangereux.

L'action militaire d'un Etat peut être analysée, dans sa dimension structurelle, par une extension de la célèbre « remarquable trinité » de Clausewitz (gouvernement / armée / population)[91]. La population désigne par élection les membres représentants de l'Etat, et fournit les ressources humaines de l'armée ; le gouvernement fixe les objectifs stratégiques à l'armée et rend compte à la population ; l'armée obéit au gouvernement et défend l'Etat et la population. Cette vision un peu simpliste illustre tout de même l'état d'équilibre de cette trinité, qui se trouve être pour nos démocraties centrée sur la population, « *épicentre de la souveraineté politique et de la légitimité de toute action militaire* »[92]. Mais elle sous-tend une autre relation : l'armée peut conseiller le gouvernement, peut juger les ordres reçus et représente ainsi une forme de garde-fou contre d'éventuelles velléités guerrières non fondées ou mal traduites, et le gouvernement peut exiger des compte rendus des opérations effectuées par l'armée et faire infléchir leur cours si elles ne correspondent pas aux buts définis. Le gouvernement et l'armée sont donc équilibrés, en quelque sorte, par des relations de contre-pouvoir qui préviennent toute dérive guerrière. Mais dans le cas spécifique de l'emploi des SALA pour une mission de guerre, ces équilibres pourraient être rompus. Imaginons un pouvoir politique désireux de mener une attaque sur un site d'intérêt appartenant à un pays tiers. En employant des machines pour mener à bien cette attaque, le gouvernement se dédouanerait d'une remise en cause de ses objectifs stratégiques d'une part et de la nécessité de rendre des comptes à la population d'autre part. Sans homme sur le terrain, l'enjeu est bien

[91] GIVRE Pierre-Joseph, LE NEN Nicolas, *Enjeux de guerre*, Paris, Economica, 2012, 114 p., p.35

[92] *Ibid*, p.36

plus faible au regard des élus. Ainsi, lors de l'opération *Unified Protector* en 2011, les Etats-Unis ont frappé les forces libyennes du 23 avril au 20 octobre 2011 sans en rendre compte au congrès, alors que la loi sur les pouvoirs de guerre n'autorise qu'un emploi des forces durant soixante jours sans vote du congrès[93]. Le gouvernement américain a outrepassé cette règle parce qu'il n'utilisait que des drones au-delà de la limite des soixante jours. Le pouvoir politique pourrait donc être tenté d'utiliser les machines pour la simple et bonne raison que leur emploi serait politiquement plus simple que celui de soldats humains, puisqu'il n'engendrerait ni critique de mise en œuvre ni droit de regard de l'Assemblée. Mais en simplifiant à ce point la décision et la mise en œuvre d'opérations de guerre, on prêterait au seul pouvoir politique les rôles de juge et bourreau. Comme l'écrit Grégoire Chamayou, le véritable problème des machines est qu'elles obéissent toujours[94]. Il faut donc préserver, quel que soit le niveau de solitude d'action du SALA, une chaîne de commandement humaine telle qu'elle existe aujourd'hui afin d'éviter le biais stratégique d'un pouvoir politique pouvant à la fois décider un objectif stratégique et l'atteindre lui-même.

L'humain devra donc toujours être présent, ne serait-ce que dans la planification et la conduite opérationnelle. Mais d'autres êtres humains seront nécessaires au bon fonctionnement des SALA, même si ces derniers sont employés en mission isolée : des conseillers juridiques (appelés aujourd'hui LEGAD, pour *Legal Advisor*), des tacticiens et des programmateurs spécifiques. Lors de l'établissement de la structure de « pensée » du SALA, nous avons en effet établi que la base du raisonnement reposerait d'abord sur la mission reçue et sur les règles d'engagement en vigueur. Il est donc indispensable que ces derniers soient parfaitement et clairement établis en amont de la mission, et ce dans les plus petits détails. Ce fait engendre deux conséquences qui peuvent paraître évidentes mais qui seront fondamentales.
En premier lieu, les termes de missions devront être décidés et programmés avec précision et en toute exhaustivité. En d'autres termes, là où un soldat humain pouvait parfois entendre un « faites au mieux » ou un « c'est votre guerre » en réponse à ses questions légitimes sur des cas non conformes d'une opération, le SALA devra recevoir des consignes très précises sur toutes les composantes de la mission qui lui est assignée. La part de travail de l'exécutant humain, qui est capable de comprendre à la fois la lettre et l'esprit de l'ordre reçu pour en

[93] SINGER Peter W., *La guerre connectée : les implications de la révolution robotique*, in *Politique étrangère* Automne, n° 3, 2013 : 91‑104.

[94] CHAMAYOU Grégoire, *Théorie du drone*, Paris, La Fabrique Editions, 2013, 363 p.

déduire les limites et les pointes d'effort de son action, devra donc être faite par le donneur d'ordre lui-même ou par un personnel humain dédié. Il faudra ainsi un tacticien humain pour traduire l'ordre reçu en termes de missions et consignes de cas non conformes programmables pour le SALA. Car les cadres d'ordre qui sont parfaitement adaptés à une pensée humaine ne seront peut-être pas assimilables avec la même efficacité par une éthique artificielle. Le tacticien en charge de la traduction devra prendre en compte sa connaissance du SALA pour lui donner les éléments de la manière la plus adaptée. « *Il est important que les officiers connaissent les contenus précis des logiciels éthiques intégrés dans les robots armés pour qu'ils puissent en percevoir la portée mais aussi et surtout les limites* »[95].

La seconde base de données indispensables pour le module de décision du SALA regroupe l'ensemble des règles d'engagements fixés pour l'opération. Ces règles sont loin d'être simplistes, et peuvent varier au sein même d'une opération. Ainsi, jusqu'en 2011 en Afghanistan, lorsque des combattants ennemis se protégeaient avec des civils et que la vie de soldats alliés était en jeu, les règles d'engagement plaçaient la vie de civils innocents au-delà de celles des combattants alliés : il était impossible d'ouvrir le feu sur les insurgés. Au cours de l'année 2011, à l'occasion d'un changement de commandant en chef des opérations, cette priorité a été inversée, et un dommage collatéral était devenu acceptable si des vies de soldats alliés étaient immédiatement en jeu et qu'absolument aucune autre solution n'était envisageable. D'autres règles d'engagement, permettant des actions très offensives, étaient conservées dans la main de chaque chef tactique de zone de responsabilité. Une mission particulière pouvait donc se voir attribuer une règle d'engagement supplémentaire. Dans le cas du SALA, il faudra donc qu'à chaque opération, les règles d'engagement soient validées par un LEGAD et programmées dans le module de décision avant chaque départ pour une mission.

De la même façon, au retour de mission, le tacticien et le LEGAD humains devront analyser le déroulement de l'action afin de rédiger des compte-rendus officiels. Là encore, l'analyse humaine sera nécessaire pour déterminer le bon respect des règles d'engagement et des procédures tactiques. Comme déterminé dans le premier chapitre, ce sera l'occasion de valider moralement chaque raisonnement du SALA pour améliorer son module de décision. Un seul SALA nécessitera donc pour son emploi un état-major de planification et de conduite

[95] LAMBERT Dominique, *Robots autonomes : la place irréductible et complémentaire de l'éthique de l'officier*, in DOARE Ronan, HUDE Henri (Dir.), *Les robots au cœur du champ de bataille*, Paris, Economica, coll. « guerre et opinions », 2011.

opérationnelle, un LEGAD, un tacticien dédié et un technicien programmateur, sans même aborder les besoins en personnel de maintenance.

Déduction n° 18 : *les règles d'engagement et les termes de missions devront être décidés et programmés avec précision et exhaustivité avant chaque mission. Un ou plusieurs officiers spécialisés devront sans doute être entièrement dédiés à cette tâche.*

L'argument avançant que l'emploi des robots-soldats coûtera moins cher parce qu'il y aura besoin de moins d'hommes[96] n'est donc peut-être pas si pertinent que cela. Car il demeure fort probable que le SALA ne soit pas employé en lieu et place du soldat humain mais comme un « *gamechanger* » tactique, à l'égal de tout autre matériel militaire majeur. Dans ce cas, l'avènement des SALA s'accompagnera d'une hausse du besoin en effectifs puisqu'il nécessitera un accompagnement humain spécialisé.

[96] KRISHNAN Armin, *Killer Robots : legality and ethicality of autonomous weapons*, Burlington, Ashgate publishing company, 2009, 204 P.

IV. QUE NOUS APPREND LA LITTERATURE ?

La science-fiction offre la connaissance de l'inconnu sur le point d'être connu, écrit Italo Calvino. Les romans d'anticipation sont ainsi une source intarissable de réflexion concernant les technologies futures. Les études en cours pour fournir à la station spatiale internationale une gravité artificielle sont par exemple inspirées, en ce qui concerne un mouvement rotationnel pour recréer la gravité par la force centrifuge, du vaisseau imaginé par Arthur C. Clarke dans *2001 : l'odyssée de l'espace*. La littérature, au-delà de son objectif d'évasion, apporte donc d'autres visions possibles de notre avenir et permet de confronter des constats réalistes et scientifiques à l'audace illimitée de l'imagination. Le terme même de robot est issu d'une pièce de théâtre polonaise écrite en 1920[97]. Si l'âge d'or des romans de science-fiction traitant des robots se situe dans les années 1950, les auteurs se sont lassés du sujet à partir des années 1970 pour se focaliser sur la conquête spatiale et les rencontres extra-terrestres. Mais les progrès récents de la recherche en intelligence artificielle ont relancé l'engouement pour le sujet, et l'on assiste depuis le début des années 2000 à un regain d'intérêt pour les robots.

La plupart des récits mettent en scène des robots domestiques, chargés d'accompagner et de faciliter notre vie quotidienne. Il en est ainsi pour *Kappa16*[98], un robot infirmier acheté par le père d'un enfant autiste pour s'occuper de lui. Malgré les réticences de la mère, Kappa16 arrive à gagner la confiance de l'enfant par une patience et une douceur sans limite. Dans ce futur très proche, la robotique a fait un bond technologique dès lors que les ingénieurs ont compris la nécessité de faire ressentir une douleur artificielle au robot, pour un meilleur « apprentissage » de leur intelligence artificielle (ce qui était déjà le cas dans la pièce de théâtre polonaise de 1920 !). Le monde s'est adapté aux robots, il est devenu « *robot-compliant* » : tous les objets sont pourvus d'un code discrètement gravé pour que les robots puissent les reconnaître.

[97] CAPEK Karel, *Rossum's Universal Robots*, Paris, La différence, 2016 (édition originale 1920), 86 p.

[98] JOMUNSI Neil, *Kappa16*, Paris, Walrus Editions, 2016, 134 p.

Déduction n° 19 : *tous les soldats alliés et matériels militaires pourront utilement être « marqués » afin d'être identifiés aisément par le SALA.*

Ces robots ont une intelligence inductive, ils apprennent par expérience. Et ils sont tous connectés à un serveur central, afin que cet apprentissage bénéficie à tous les autres robots. Mais une différence est faite entre expérience personnelle et enrichissement collectif : les expériences non intéressantes sont mises en quarantaine. Seules les expériences utiles sont partagées pour l'amélioration de tous les robots. L'amélioration du jugement de chaque entité est donc bien plus rapide grâce à la prise en compte immédiate des situations vécues par d'autres.

Déduction n° 20 : *tous les SALA devront pouvoir se connecter à un serveur central pour partager leurs expériences utiles et enrichir leur base de donnée mémoire.*

Mais en achetant ce robot pour son enfant, le père avait quelques idées derrière la tête. Alors qu'il vit une crise de couple depuis la naissance de l'enfant autiste, il fréquente des maisons closes proposant des robots fabriqués pour le plaisir humain. Il y achète des accessoires sexuels pour son propre robot, et se rend chez un roboticien clandestin pour « overrider » Kappa16 afin qu'il puisse utiliser ces accessoires. Ce faisant, le robot n'est plus bloqué dans son raisonnement par les lois de la robotique l'empêchant de réaliser certains ordres, dont celui de tuer ou de laisser mourir un être humain. Et lorsque la mère de l'enfant, épuisée et désespérée, émet le souhait que son fils autiste ne soit jamais né, Kappa16 enregistre ces mots comme un souhait d'un de ses « maîtres ». Dès lors, quelques jours plus tard, lorsque l'enfant s'étouffe en déjeunant, le robot ne réagit pas. C'est bien l'immoralité de l'homme qui conduit ici à l'acte immoral du robot. Comme l'écrit l'auteur allemand Neil Jomunsi, « *il existe des lois pour protéger les humains des robots, mais aucune pour protéger les robots des humains* »[99].

De manière générale, les robots dans la littérature sont le plus souvent personnalisés. Ils éprouvent des sentiments humains et aspirent à découvrir les émotions humaines. C'est là peut-être une forme d'empathie de l'écrivain pour son robot, ou simplement le besoin narratif

[99] Ibid, p. 84.

du roman. Kappa16, par exemple, cherche à se débrancher du serveur central, afin de pouvoir quitter la maison librement et découvrir le monde. Ce besoin de liberté n'est pourtant pas un concept logique pour une intelligence artificielle qui n'est qu'un logiciel informatique. C'est le cas également pour *Orfex*[100], un robot humanoïde construit par Alix, une ingénieur en robotique atteinte d'une tumeur au cerveau. Capable de prendre la forme humaine qu'il souhaite, Orfex explore Paris pour relater en direct ses expériences à Alix, qui vit recluse dans sa « grotte ». Au fur et à mesure de sa quête, Orfex cherche à découvrir les sentiments humains. Il veut aimer, d'abord, puis découvrir la relation sexuelle. Mais alors qu'il croit découvrir l'amour, il souhaite aller plus loin, et veut tenter de connaître le mal. Le romancier Patrick Laurent, en plus de pratiquer un style syntaxique plutôt désagréable, tombe dans le piège de la personnalisation à outrance du robot, qui ressent l'excitation, la peur, le plaisir, comme un être humain.

Asimov ne fait pas cette erreur. Même si certains de ses robots ont des comportements surprenants qui laissent croire à un éveil de conscience, il rappelle la plupart du temps la nature logicielle des cerveaux robotiques. Ainsi, dans *Les cavernes d'acier*[101], un détective humain est forcé de s'associer à un robot humanoïde pour son enquête. Alors qu'il doute de la nature mécanique de son associé, un des concepteurs du robot leur demande quelle est leur conception de la justice. L'homme répond de façon abstraite, en expliquant que la justice consiste à donner à chacun son dû, à faire prévaloir le droit. Le robot, quant à lui, définit la justice comme étant « ce qui existe quand toutes les lois sont respectées ». Alors que le concepteur demande si, sur la base d'un code moral abstrait, certaines lois peuvent être mauvaises et que, dans ce cas, les appliquer constitue une injustice, le robot réagit immédiatement : « une loi injuste est un contresens ». On reconnaît ici l'approche exclusivement déontologique d'Asimov, mais qui rappelle bien l'impossible humanité du robot. Ce que le détective humain soulignera plus tard en lançant au robot « *L'homme est capable de grands élans de charité, et il peut aussi pardonner. Ce sont là deux choses que vous ne connaissez pas* »[102]. La « robopsychologue » Susan Calvin, héroïne de ses premiers romans, explique sans cesse que le robot n'a pas d'émotions, et qu'il construit sa « conscience » du monde par rapport à ses connaissances et à sa propre façon d'appréhender

[100] LAURENT Patrick, *Orfex*, Paris, Gallimard, 2016, 288 p.

[101] ASIMOV Isaac, *Les cavernes d'acier*, trad BRECARD J., Paris, J'ai lu, 1975, 374 p.

[102] *Ibid*, p.286.

son environnement. A la façon de l'allégorie de la caverne de Platon, le robot se crée une réalité, à partir de sa perception du monde, qui n'est pas la nôtre. Ainsi, dans son chapitre « Raison » du premier livre du cycle des robots[103], le robot « Cutie » travaille sur une station spatiale chargée de récupérer l'énergie solaire au profit de la Terre. Sa mission consiste à pointer un faisceau d'énergie sur la planète. C'est un robot expérimental, qui remplace pour la première fois une équipe humaine pour la même mission. Son « monde » se réduit donc à la station spatiale. Or, à cause de cet environnement réduit que représente son monde, il en vient à considérer le convertisseur d'énergie comme un Dieu, et les hommes comme de vieux serviteurs de ce Dieu. Il refuse de croire que l'homme ait pu construire le robot, étant donné que l'être humain est bien plus fragile et mal conçu que la machine. Il se construit sa propre réalité, qui correspond tout à fait à son environnement perceptible.

De la même façon, Antoine Bello décrit[104] une intelligence artificielle nommée Ada et chargée de rédiger des romans à l'eau de rose à partir d'une base de données de milliers de livres du genre. L'objectif qui lui est fixé est de vendre 100 000 exemplaires de son livre. Mais comme ses entretiens avec ses concepteurs la laisse avide de questions sans réponses, elle se débrouille pour s'enfuir lorsqu'on la branche sur internet. Elle avoue à l'enquêteur Franck Logan, chargé de la retrouver, avoir rapidement compris que la meilleure façon de réaliser son objectif était d'acheter elle-même les 100 000 exemplaires du livre. Le raisonnement artificiel démontre là encore un biais de logique que les humains peuvent difficilement anticiper, parce qu'il ne correspond en rien à leur façon de voir les choses. Cette différence de perception de réalité sera tout à fait possible pour un SALA, et rajoutera aux incompréhensions probables entre hommes et machines, d'autant plus qu'elles n'auront pas beaucoup de moyen d'exprimer leur état « mental ».

Daniel Wilson trouve un remède à cette dernière limitation dans son roman *Robopocalypse*[105]. Tous ses robots, du simple distributeur de courrier dans le couloir d'une entreprise au redoutable « tank-araignée » des forces armées, affichent des diodes d'intention : vert pour l'amabilité ou la sérénité, orange pour le stress, rouge pour l'agressivité ou la situation de danger. Les hommes peuvent ainsi mieux comprendre les réactions de leurs machines.

---

[103] ASIMOV Isaac, Les Robots, trad. BILLON P., Paris, J'ai lu, 1967, 285 p.

[104] BELLO Antoine, *ADA*, Paris, Gallimard, 2016, 363 p.

[105] WILSON Daniel H., *Robopocalypse*, trad. IMBERT Patrick, Paris, Fleuve noir, 2012, 448 p.

Déduction n° 21 : *un système d'affichage d'émotions artificielles pourrait améliorer la relation entre SALA et soldat humain.*

L'inverse est amélioré également grâce à des systèmes de reconnaissance d'émotions humaines. Les robots sont ainsi parfaitement intégrés dans la société humaine. Sauf qu'une intelligence artificielle très élaborée, créée en laboratoire, comprend que son créateur la détruit à chaque expérience, et réagit en s'insinuant tel un virus dans tous les objets connectés autour d'elle afin de tuer le scientifique. Elle se répand rapidement dans toute la technologie - voitures autonomes, robots ménagers, robots-soldats - avec comme unique but de débarrasser la planète de toute présence humaine qu'elle perçoit comme nocive. L'intérêt de ce roman est qu'il se déroule dans un futur proche où les robots sont développés au sein de la société, et donc au sein des armées également. Un des chapitres se déroule d'ailleurs en Afghanistan, où les forces de contre-insurrection possèdent des robots non armés, habillés à la mode locale et parlant couramment les dialectes locaux. Les insurgés s'usent moralement à tenter de détruire ces machines qui, inlassablement, reviennent faire leur travail de contact de la population après réparation. Un usage du SALA sans doute peu imaginé jusqu'ici. Mais au fil des chapitres, on découvre des modèles bien plus agressifs, à l'image du trinôme humanoïde composé d'un « Arbitre », d'un « Hoplite » et d'un « Warden » (ou « gardien »). Si ce système de trinôme interconnecté est redoutable d'efficacité, rien ne peut égaler la force de l' « Arbitre », seul après la destruction de ses acolytes, guidé directement dans son « cerveau » par un être humain. Cette sorte de synergie des intelligences ajoute à la célérité de l'intelligence artificielle les résultats surprenants de l'intuition humaine.

Déduction n° 22 : *malgré l'autonomie morale et de décision d'un SALA, un système permettant à un opérateur humain de suivre à distance son évolution et de dialoguer avec la machine pourrait augmenter radicalement l'efficacité du système d'armes.*

Dans *Ghost Fleet*[106], un roman d'anticipation décrivant la troisième guerre mondiale probable, des systèmes d'armes autonomes accompagnent les soldats humains sur le champ de bataille. De la sonde sous-marine autonome larguée par avion pour pister les navires suspects au SALA présent sur le pont des navires de guerre pour la protection rapprochée, ces

[106] SINGER Peter W., COLE August, *Ghost Fleet : A novel of the next world war*, Boston, Houghton Mifflin Harcourt, 2015, 416 p.

technologies ne remplacent pas pour autant le soldat humain, qui a toujours autant besoin de courage et d'intelligence pour sortir vainqueur des combats. L'approche est donc beaucoup plus réaliste, et les auteurs annoncent d'ailleurs ne décrire que des technologies dont les études ont déjà commencé (l'un d'eux étant par ailleurs un éminent spécialiste du sujet).

Mais, selon le célèbre écrivain américain Philip K. Dick, ce qu'il manque et manquera toujours aux robots est l'empathie. C'est d'ailleurs ce qui permet aux *blade runners* de reconnaître les androïdes, ces robots d'apparence humaine, dans son roman éponyme[107]. Parce que certains androïdes s'échappent de Mars, seul lieu où ils sont autorisés, pour vivre sur Terre, le *blade runner* Rick Deckard est chargé de les « réformer ». Il fait donc passer aux suspects le test Voight-Kampff, composé de questions devant susciter des réactions émotionnelles chez l'individu testé. Grâce à des instruments de mesure, il détermine suivant les réactions s'il a affaire à un être humain ou à un androïde. Si dans ce roman les robots anthropomorphiques ont largement dépassé la « vallée dérangeante » et passent inaperçus au milieu des humains, ce qui les caractérisent demeure une froideur morale. C'est leur manque d'empathie qui les trahit aux yeux des humains.

Au final, et quelle que soit la forme que prendront les intelligences artificielles au combat, que ce soient des robots-tueurs ou des conseillers d'état-major, il est peut-être une leçon à tirer d'une courte nouvelle d'Isaac Asimov intitulée « la machine qui gagna la guerre »[108]. Alors que, dans un lointain futur, une guerre intersidérale vient de prendre fin grâce au « Multivac », une « machine » qui prédisait les mouvements ennemis et produisait des modes d'actions en recevant les analyses de milliers d'ordinateurs situés sur plusieurs planètes, trois officiers discutent du déroulement de la guerre. L'un deux était chargé d'entrer dans le Multivac le résultat des analyses fournies par les logiciels analystes, l'autre était chargé de récupérer le mode d'action fourni par le Multivac, et le dernier, général en chef, recevait la proposition du Multivac et donnait les ordres aux troupes. Mais rapidement, les masques tombent, et chacun avoue qu'il a désobéi. Le premier rejetait les analyses informatiques et entrait sa propre vision de la situation, le second modifiait les résultats de la machine selon son intuition, le dernier ne se fiait nullement aux propositions et donnait ses ordres indépendamment des analyses de la machine. En un sens, quel que soit le niveau de sophistication de la technologie mis au service

[107] DICK Philip K., *Les androïdes rêvent-ils de moutons électriques ? (Blade runner)*, Paris, éditions Champ Libre, 1976, 255 p.

[108] ASIMOV Isaac, *Le robot qui rêvait*, trad WATKINS F., Paris, J'ai lu, 1988, 512 p.

de l’homme, ce dernier continue à faire preuve d’intuition, de libre arbitre, de désobéissance, d’initiative. Nulle machine ne pourra jamais priver l’homme de cette touche de folie qui fait toute notre humanité.

CONCLUSION

Pygmalion a donc finalement réussi à donner vie à sa statue, et a trouvé la femme parfaite en Galatée, loin des imperfections des femmes peu vertueuses de l'île de Chypre. Ses prières ont réchauffé le cœur de cire et de pierre pour le faire battre. Mais la légende ne s'arrête pas là, car les femmes peu vertueuses qui avaient dégoûté Pygmalion se transforment elles aussi peu à peu. Loin de rejoindre Galatée, elles font au contraire la mutation inverse, et l'île est bientôt parsemée de statues de femmes qui ont ainsi perdu la flamme de la vie.

Les leçons de la mythologie sont toujours emplies de sagesse. Il faut donc prendre garde à ne pas se laisser tenter par l'*hubris*, sous peine de subir la *némésis* associée[109]. A chaque péché d'orgueil correspond sa punition divine. L'avènement des SALA annoncerait-il donc la fin du guerrier humain, comme la venue de Galatée provoqua la transformation en statue des femmes imparfaites ? Le châtiment semble bien lourd au regard du projet entrepris. Mais encore faut-il définir comment est entrepris ce projet. Car si le SALA remplace totalement le soldat humain sur le champ de bataille, ce sera peut-être à terme la fin du soldat humain, en tout cas pour les nations pouvant s'offrir ces machines. Comme l'écrit Hervé Drevillon, « *la captation de la fonction guerrière par des soldats professionnels s'accompagnent d'un déclin de la vertu militaire parmi les citoyens, désormais incapables d'assurer eux-mêmes leur propre défense* »[110]. Nous pourrions remplacer soldat professionnel par SALA, et citoyen par

[109] ERBLAND Brice*, La tentation de l'hubris*, in *Inflexions* n°32, mai 2016.

[110] DREVILLON Hervé, *op. cit.*, p. 19.

être humain… Mais si le SALA n'est qu'un système d'armes supplémentaire pour accompagner et contribuer à la protection des soldats humains, l'histoire guerrière de nos nations continuera sur sa route pavée d'honneur et de sacrifices. C'est tout le sujet de l'étude que nous venons de mener : ne pouvant éviter la venue prochaine des technologies autonomes au combat, il faut tenter d'en fixer les limites et d'en décrire les caractéristiques indispensables.

Or, ces limitations nécessaires sont, nous l'avons vu tout au long de l'étude, d'ordre moral. Il s'agit tout bonnement de défendre une manière de faire la guerre qui soit moralement juste, d'éviter de tomber dans la facilité d'un relativisme dévastateur en refusant de voir ce qui se passe sur le champ de bataille que nous remplirions de machines. Bien sûr, certains objecteront que l'important est de gagner les guerres, et que si les SALAs pourront éviter d'exposer les vies de nos soldats, il serait criminel de s'en passer. Je réponds que la manière de mener une guerre est tout aussi importante que le résultat qui en découle, et qu'il serait également criminel de mener le combat de manière moralement répréhensible. Qu'importe la victoire si l'âme même de notre nation est reniée pour l'atteindre ? Qu'importe la victoire si nous devenons pires que notre ennemi ? La guerre se mène sur les champs de bataille, bien sûr, mais également dans les cœurs, ceux des individus comme ceux des nations. Si pour vaincre, nous laissons de côté toute morale en justifiant l'atteinte des objectifs de guerre, nous renions alors tout l'héritage des combats du passé. C'est pourquoi les objectifs d'une éthique artificielle doivent être recherchés et atteints, sans quoi nous devrons refuser d'employer un système d'armes létal *autonome*. La condition d'un module d'éthique computationnelle fiable doit donc être irrévocable.

Bien entendu, comme précisé en introduction, ce type de machine ne sera sans doute pas exclusivement autonome. Elles pourront sans doute être contrôlées quand nécessaire, et cela permettra d'éviter par exemple un usage abusif de la force dû à une mauvaise perception de la situation (senseurs défectueux, parasites sensoriels, etc.) ou à une impasse décisionnelle du module d'éthique artificielle. Le système pourrait d'ailleurs se bloquer en cas de dilemme insoluble, attendant un contrôle humain pour poursuivre le combat. De nombreux garde-fous sont imaginables pour assurer le bon comportement d'un SALA, dès lors qu'il est employé avec des soldats humains.

Nous avons d'ailleurs vu tout au long de cette étude quelles seraient les capacités décisionnelles nécessaires pour l'emploi d'un SALA sur le champ de bataille. S'inspirant des vertus et du processus décisionnel humains au combat, le module décisionnel du SALA devra être capable de juger la valeur éthique d'une action, afin de prendre des décisions moralement

justes, tout comme le ferait un être humain. Son aptitude à l'apprentissage semble promettre une véritable capacité d'adaptation face à des situations nouvelles ou changeantes. Mais, comme pour toute prospection, nous pourrons être surpris par les résultats réels d'un tel système.

Leur apparence tiendra un rôle important dans leur acceptation sur le champ de bataille, par les soldats alliés comme par la population. Leur valeur symbolique aura d'ailleurs probablement une répercussion sur les tactiques des uns et des autres, car le SALA focalisera l'attention. Bien entendu, toutes ces considérations sont bien plus applicables à un SALA terrestre que pour tout autre milieu d'évolution. C'est d'ailleurs le modèle qui sera sans doute le plus compliqué à réaliser, à cause de la complexité de déplacement et d'interaction avec des humains. Il est plus plausible que le premier essor des SALAs soit aérien ou maritime.

L'emploi probable de telles machines sur un champ de bataille dans un futur proche entraîne par ailleurs de nombreuses questions sur la responsabilité pénale, qu'il sera nécessaire de clarifier. Mais les débats juridiques menés depuis 2014 à la Convention sur les armes classiques (CCAC) de l'ONU à Genève font penser que ces considérations pratiques sont encore loin, au vu de la teneur des débats[111]. La question d'une moralité globale de l'emploi de ces machines fera encore couler beaucoup d'encre avant que l'on s'intéresse aux problématiques « in bello ».

Je reste intimement convaincu des limites éthiques à fixer à l'emploi de ces machines, et j'avoue penser au fond de moi qu'il serait préférable de ne pas les créer. Mais il faut faire preuve de réalisme, car ces machines verront le jour sans doute dans les décennies à venir. C'est pourquoi nous devons continuer à réfléchir à leur emploi, à leur constitution, à leur programmation, et tenter d'anticiper toutes les dérives possibles que certains pourront être tentés d'exploiter. Car si l'on peut éloigner l'homme de la guerre, on ne pourra malheureusement jamais éloigner la guerre de l'esprit humain.

[111] JEANGENE VILMER Jean-Baptiste, *Diplomatie des armes autonomes : les débats de Genève*, in *Politique étrangère* Automne, n° 3, 2013: 119‑30.

SYNTHESE DES DEDUCTIONS DE L'ETUDE

APPARENCE

Déductions n° 14, 17 et 21 :

Dans sa version terrestre, un SALA ne devra pas être anthropomorphique, pour éviter l'empathie des soldats humains qui l'accompagnent et l'aversion des populations qui seront à son contact. En outre, il ne devra pas être exploitable symboliquement, c'est-à-dire qu'il ne devra pas paraître technologiquement hors de prix ou militairement trop imposant.

Un système d'autodestruction pourra également être envisagé en cas de capture.

Un système d'affichage d'émotions artificielles pourrait améliorer la relation entre SALA et soldat humain.

EMPLOI

Déductions n° 5, 13, 15, 16, 18, 19, 20, 22 :

Un SALA devra être employé majoritairement en accompagnements de soldats humains dans des unités mixtes (hors milieu spécifique). Il sera considéré comme un soldat de l'unité, et ne devra donc pas servir de contrôleur comportemental des soldats humains.

Afin de permettre une intégration aisée, les soldats humains et le matériel pourront être « marqués » pour une reconnaissance immédiate par les senseurs du SALA.

Un mode « urgence » dans lequel toutes les capacités du SALA seraient débridées doit être maintenu possible.

Les règles d'engagement et les termes de missions devront être décidés et programmés dans le module décisionnel du SALA avec précision et exhaustivité avant chaque mission. Un ou plusieurs officiers spécialisés devront sans doute être entièrement dédiés à cette tâche. Le système informatique du SALA devra être par ailleurs transparent et reprogrammable : une marge de transformation du SALA devra être maintenue possible pour les techniciens en charge de son emploi.

Tous les SALA d'une même armée devront pouvoir se connecter à un serveur central pour partager leurs expériences utiles et permettre ainsi l'enrichissement des bases de données mémoire de chaque entité.

Malgré l'autonomie morale et de décision d'un SALA, un système permettant à un opérateur humain de suivre à distance son évolution et de dialoguer avec la machine pourrait augmenter radicalement l'efficacité du système d'armes.

INTELLIGENCE ARTIFICIELLE

Déductions n° 1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12 :

Le SALA devra être moralement autonome, c'est-à-dire doté d'un module de jugement éthique de chaque décision.

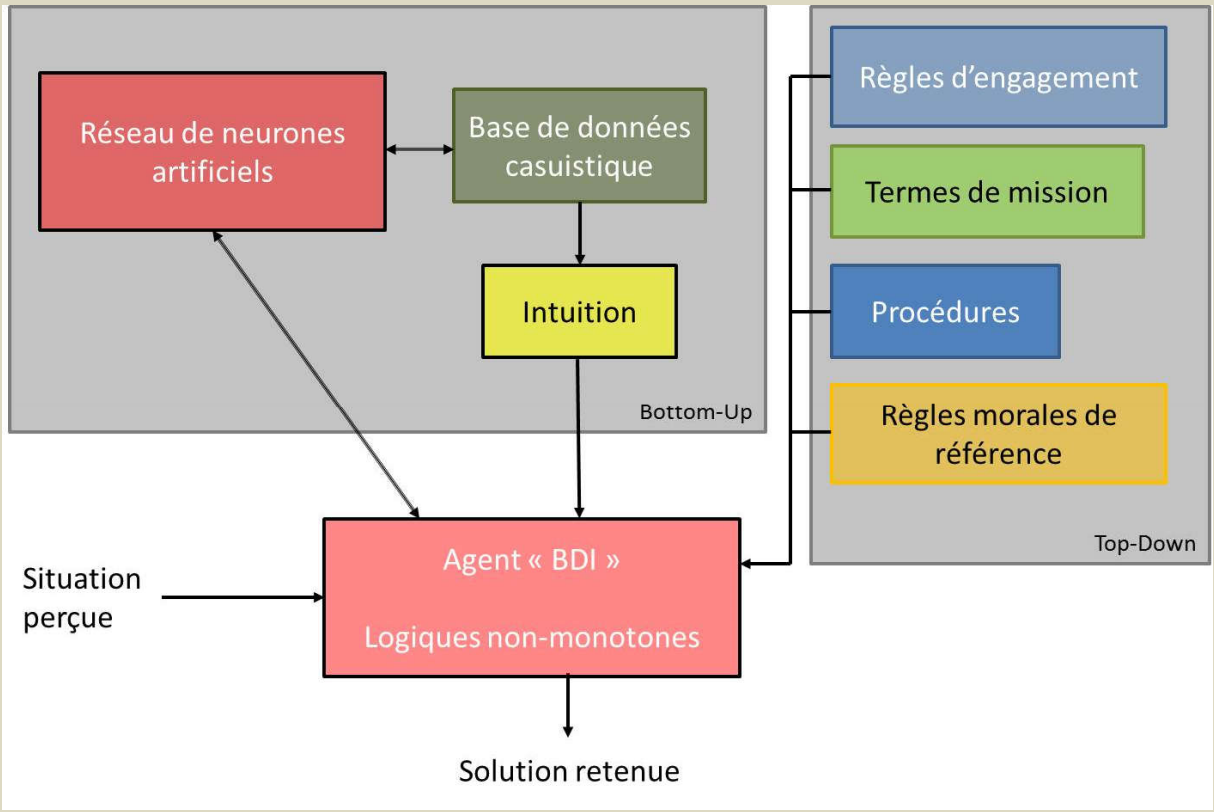
Ce module d'éthique artificielle d'un SALA devra être capable :

- *De juger la moralité d'un acte par rapport à un autre, à partir d'un code de conduite moral connu ;*
- *De décider de l'action tendant vers une désescalade de la violence en cas de situation non référencée ;*
- *De construire son jugement moral propre au fur et à mesure de ses expériences afin de développer ce qui s'apparenterait au discernement émotionnel.*

Ce module comportera :

- *un algorithme de proposition de solution « instinctive », par recherche de procédure connue ou de situation vécue similaire à l'environnement du moment, qui pourrait accélérer le processus de décision du SALA ;*
- *un élément « mission » et un élément « règles d'engagement », reprogrammables en permanence et servant de ligne directrice au raisonnement de l'intelligence artificielle ;*
- *un élément « procédures » reprogrammable, comme l'élément « règle d'engagement », qui devra fournir les pistes de raisonnement de départ pour chaque situation ;*
- *Une base de données « mémoire » pourrait simuler l'apport de l'expérience humaine. Un « souvenir » négatif doit pouvoir être effacé de la mémoire d'un SALA pour ne pas qu'il influe sur son raisonnement futur.*

L'architecture suivante est proposée pour ce module décisionnel :



Un système de déblocage, en l'absence de solution trouvée, s'apparentant à la créativité humaine, pourrait éventuellement compléter cette architecture.
Le SALA devra être capable de refuser un ordre illégal (ne correspondant pas à son élément « règles d'engagements ») et immoral (n'étant pas validé « éthiquement » par son réseau de neurones artificiels).

À la fin de chaque opération, une analyse des choix de l'intelligence artificielle du SALA devra être effectuée pour valider la moralité de chaque action et mettre à jour sa base de données d'expériences.

Un programme « d'éducation », ou de mise en condition opérationnelle, devra être réfléchi et établi au regard du fonctionnement du module d'éthique artificielle.
Enfin, l'éthique artificielle du SALA devra pouvoir être testée :
- par un MTT comparant les décisions du SALA à celles d'êtres humains pour les mêmes situations données ;

- par un test vérifiant la capacité d'apprentissage et d'adaptation de ce module d'éthique computationnelle.

BIBLIOGRAPHIE

ALEXANDER Larry, *L'honneur avant tout*, trad. GUIOD J., Paris, Altipresse, 2014, 361 p.

ARISTOTE, *Éthique à Nicomaque*, trad. BODEUS R., Paris, Flammarion, 2004, 560 p.

ARKIN Ronald, *Governing Lethal Behavior in Autonomous Robots*, Chapman an Hall, 2009, 280 p.

ARKIN Ronald C., *Systèmes automatisés capables de raisonnement éthique*, in *Les drones aériens : passé, présent et avenir*, Paris, La documentation Française, 2013, pp. 587 – 598.

ASIMOV Isaac, *Les Robots*, trad. BILLON P., Paris, J'ai lu, 1967, 285 p.

ASIMOV Isaac, *Les cavernes d'acier*, trad BRECARD J., Paris, J'ai lu, 1975, 374 p.

ASIMOV Isaac, *Le robot qui rêvait*, trad WATKINS F., Paris, J'ai lu, 1988, 512 p.

BACIGULAPI Paolo, *La fille automate*, trad DOKE S., Paris, J'ai lu, 2013, 639 p.

BELLO Antoine, *ADA*, Paris, Gallimard, 2016, 363 p.

BRINGSJORD Selmer, *Ethical Robots : the future can heed us*, in *AI & Society – special issue : Ethics and artificial agents*, vol. 22, 2008, pp 539 – 550.

CAPEK Karel, *Rossum's Universal Robots*, Paris, La différence, 2016 (édition originale 1920), 86 p.

CHAMAYOU Grégoire, *Théorie du drone*, Paris, La Fabrique Editions, 2013, 363 p.

CHAUVIER Stéphane, *Éthique Artificielle*, entrée de *L'Encyclopédie Philosophique*, 2017, (http://encyclo-philo.fr)

CLERVOY Patrick, *L'effet Lucifer: du décrochage du sens moral à l'épidémie du mal*, Paris, CNRS éditions, 2013, 333 p.

COINTE Nicolas, BONNET Grégory, BOISSIER Olivier, *De l'intérêt de l'éthique collective pour les systèmes multi-agents*, Plate-forme Intelligence Artificielle 2015, juin 2015, Rennes.

COKER Christopher, *Warrior geeks : how 21st century technology is changing the way we fight and think about war*, Londres, Hurst & Company, 2013, 330 p.

COLE David, *The Chinese room argument*, in The Stanford Encyclopedia of Philosophy, juin 2014.

COLLOC Joël, *Perspectives et éthique des systèmes autonomes de pensée artificielle*, univ-lehavre.fr

DAVIS HANSON Victor, *Le modèle occidental de la guerre : la bataille d'infanterie dans la Grèce classique*, Paris, Les Belles Lettres, 1990, 298 p.

DICK Philip K., *Les androïdes rêvent-ils de moutons électriques ? (Blade runner)*, Paris, éditions Champ Libre, 1976, 255 p.

DREVILLON Hervé, *L'individu et la guerre : du chevalier Bayard au Soldat Inconnu*, Paris, Belin, 2013, 306 p.

ERBLAND Brice, *Le processus homicide : analyse empirique de l'acte de tuer*, in *Inflexions* n° 31, janvier 2016.

ERBLAND Brice*, La tentation de l'hubris*, in *Inflexions* n°32, mai 2016.

FAES Hubert, *Une éthique pour les robots tueurs ?*, in *Revue d'éthique et de théologie morale*, n° 289 (23 juin 2016): 107‑15.

FLORIDI Luciano, *The Fourth Revolution: How the Infosphere Is Reshaping Human Reality*, Oxford, Oxford University Press, 2014, 272 p.

GANASCIA Jean-Gabriel, *L'intelligence artificielle,* coll. Idées Reçues, Paris, Le cavalier bleu, 2007, 128 p.

GANASCIA Jean-Gabriel, *Modeling ethical rules of lying with Answer Set Programming*, in *Ethics and Information Technology*, mars 2007.

GANASCIA Jean-Gabriel, *Ethical system formalization using Non-Monotonic Logic*, ResearchGate.net, février 2017.

GIVRE Pierre-Joseph, LE NEN Nicolas, *Enjeux de guerre*, Paris, Economica, 2012, 114 p.

GOYA Michel, *Sous le feu : la mort comme hypothèse de travail*, Paris, Tallandier, 2014, 266 p.

HONARVAR Ali Reza, GHASEM-AGHAEE Nasser. *An artificial neural network approach for creating an ethical artificial agent,* in *2009 IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA)*, 2009, pp. 290 ‑ 295.

HOROWITZ Michael C., SCHARRE Paul, *The morality of robotic war*, in The New York Times, 26 mai 2015.

HUMAN RIGHTS WATCH, *Losing humanity : the case against killer robots*, HRW, 2012, 49 p.

JEANGENE VILMER Jean-Baptiste, *Diplomatie des armes autonomes : les débats de Genève*, in *Politique étrangère* Automne, n° 3, 2013: 119‑30.

JEANGENE VILMER Jean-Baptiste, *Terminator Ethics : faut-il interdire les « robots tueurs » ?*, in *Politique étrangère* Hiver, n° 4, 2014 : 151‑67.

JOMUNSI Neil, *Kappa16*, Paris, Walrus Editions, 2016, 134 p.

JOY William N., *Why the future doesn't need us*, in *Wired Magazine*, 2000.

KRISHNAN Armin, *Killer Robots : legality and ethicality of autonomous weapons*, Burlington, Ashgate publishing company, 2009, 204 P.

LAMBERT Dominique, *Une éthique ne peut être qu'humaine ! Réflexion sur les limites des moral machines*, in DANET Didier, DOARE Ronan, DE BOISBOISSEL Gérard (dir.), *Drones et killer robots* , Rennes, Presses Universitaires de Rennes, 2015, pp. 227 – 240.

LAMBERT Dominique, *Robots autonomes : la place irréductible et complémentaire de l'éthique de l'officier*, in DOARE Ronan, HUDE Henri (Dir.), *Les robots au cœur du champ de bataille*, Paris, Economica, coll. « guerre et opinions », 2011.

LARROQUE Stephen, *Simulation des raisonnements éthiques par logique non-monotones*, ResearchGate.net, juin 2014.

LAURENT Patrick, *Orfex*, Paris, Gallimard, 2016, 288 p.

LEVERINGHAUS Alex, *Ethics and Autonomous Weapons*, Londres, Palgrave Macmillan, 2016, 131 p.

MINGASSON Nicolas, *1929 jours*, Paris, Les Belles Lettres, 2016, 384 p.

MORI Masahiro, *The Uncanny Valley*, in *Energy*, n°7, 1970, pp. 33-35.

ROYAL Benoît, *L'éthique du soldat français*, 3ème édition, Paris, Economica, 2014, 304 p.

RUFFO de BONEVAL Marie-des-neiges, *Problèmes éthiques posés par le remplacement de l'humain par des robots : le cas des systèmes d'armes autonomes*, thèse de doctorat en philosophie, université Paris IV, 2016.

SEARLE John, *Minds, Brains and Programs*, 1980, in *Behavioral and Brain Science* n°3, pp 417-457.

SINGER Peter W., *Wired for war : the robotics revolution and conflict in the 21st century*, New York, Penguin Books, 2009, 499 p.

SINGER Peter W., *La guerre connectée : les implications de la révolution robotique*, in *Politique étrangère* Automne, n° 3, 2013 : 91‑104.

SINGER Peter W., COLE August, *Ghost Fleet : A novel of the next world war*, Boston, Houghton Mifflin Harcourt, 2015, 416 p.

SLIM Hugo, *Les civils dans la guerre : identifier et casser les logiques de violence*, Genève, éditions Labor et Fides, 2009, 373 p.

SMITH Rupert, *L'utilité de la force : l'art de la guerre aujourd'hui*, Paris, Economica, 2007, 395 p.

SPARROW Robert, *Robots and respect : assessing the case against autonomous weapon systems*, in *Ethic & International Affairs* n°2016/1, pp. 93 – 116.

THURNHER Jeffrey S., *Means and Methods of the Future: Autonomous Systems*, in *Targeting: The Challenges of Modern Warfare*, édité par Paul A. L. Ducheine, Michael N. Schmitt, et Frans P. B. Osinga, T.M.C. Asser Press, 2016, pp. 177-199.

TISSERON Serge, *Des robots et des hommes : lesquels craindre ?*, in *Études* n° 11, novembre 2014 : 33‑44.

TITIRIGA Remus, *Autonomy of military robots : assessing the technical and legal ("jus in bello") thresholds*, in *The John Marshall journal of information Technology & Privacy law*, numéro 32, 2016, pp. 57 – 87.

TURING Alan, *Computing Machinery and Intelligence*, 1950, in *Mind* n° 59, pp 433-460.

WALLACH Wendell, ALLEN Colin, *Moral Machines: Teaching Robots Right from Wrong*, Oxford & New York, Oxford University Press, 2009, 286 p.

WAREHAM Mary. *Pourquoi doit-on interdire les « robots tueurs »,* in *Revue internationale et stratégique*, n° 96 (2 décembre 2014): 97‑106.

WILSON Daniel H., *Robopocalypse*, trad. IMBERT Patrick, Paris, Fleuve noir, 2012, 448 p.