Éthique et machines autonomes : esquisse d'un discernement

Dominique LAMBERT

Professeur, Académie royale de Belgique (classe des sciences), Université de Namur (Département de philosophie), Chercheur associé au Centre de recherche des Écoles de Saint-Cyr Coëtquidan (CREC).

Une réflexion sur le sens effectif de l'autonomie : de l'autonomie absolue à l'autonomie finalisée

De nombreuses discussions ont eu lieu ces derniers temps sur la définition de l'autonomie des machines. Ceci est dû largement au fait que le concept d'autonomie est très ambigu. Que signifie le qualificatif « autonomous » dans l'acronyme *Lethal Autonomous Weapons System* (*LAWS*). Sous ce dernier, se cachent des réalités très différentes [(1)]. Si on le considère naïvement ou étymologiquement, il caractérise des systèmes qui ont leur propre loi de fonctionnement, c'est-à-dire qui peuvent effectuer une série de tâches par elles-mêmes, sans l'être humain. Beaucoup de systèmes automatiques, préprogrammés pour réaliser des tâches définies à l'avance, répondent à cette définition. Mais la notion d'autonomie recouvre aussi le cas de systèmes qui, sans la médiation d'un sujet humain, pourraient apprendre par eux-mêmes, forger des outils de classification ou des concepts nouveaux, se reprogrammer ou prendre l'initiative d'actions. L'autonomie est ici poussée à un point tel que peuvent disparaître les liens entre les comportements de la machine et les intentions ou projets des décideurs humains. Entre les systèmes automatiques, dont les comportements sont complètement prédictibles et les machines totalement autonomes dont les actions échappent à la prédiction et à la maîtrise des humains, on peut trouver un grand nombre de systèmes intermédiaires où les machines, douées de larges degrés d'initiatives, sont néanmoins supervisées par l'humain qui, à un certain moment, peut en reprendre le contrôle.

Il faut admettre avec lucidité et réalisme qu'un système totalement autonome n'est absolument pas souhaitable. L'autonomie absolue des machines conduirait à des situations où des humains mettraient en œuvre des systèmes censés réaliser certaines tâches, mais qui pourraient, à certains moments et sans qu'ils ne l'aient ni voulu ni

(1) Pour une discussion concernant la définition de l'autonomie des robots militaires, nous renvoyons à LAMBERT Dominique, *The Humanization of Robots and the Robotization of the Human Person. Ethical Perspectives on Lethal Autonomous Weapons Systems and Augmented Soldiers (with a selection of texts from the Church's engagement on Lethal Autonomous Weapon Systems)*, Genève, The Caritas in Veritate Foundation Working Papers, 2017, 97 pages (https://fciv.org/downloads/WP9-Book.pdf).

prédit, manifester des comportements contradictoires vis-à-vis de ces tâches et des buts qu'ils leur avaient prescrits.

Remarquons que cela ne signifie nullement que l'on doive refuser tous les systèmes capables de prendre certaines initiatives (déterminer un plan de vol, effectuer certaines actions défensives, etc.). Ce qui est crucial, dans le choix de tels systèmes, c'est la **cohérence**. Ces initiatives doivent rester homogènes aux finalités globales prescrites par les autorités responsables. Pensons, par exemple, à un système permettant de reprendre les commandes et de piloter un avion dont le pilote serait privé de ses moyens. Il serait tout à fait légitime, car tout à fait cohérent vis-à-vis des finalités humaines, d'admettre des prises d'initiatives issues d'une machine autonome de pilotage. Pensons aussi à des systèmes autonomes **de défense**, placés dans des zones bien déterminées où il est clairement établi qu'il n'y a pas de non-combattants, et de nature à faire face (grâce à des « initiatives ») à des attaques massives et saturantes par des essaims de robots par exemple. Dans ces deux cas, on reste en accord avec des finalités humaines prescrites et avec les normes légales (dans la mesure où l'on respecte les règles du droit international humanitaire) et morales (dans la mesure où il s'agit d'actions qui ont comme finalité le respect de la vie et de la dignité des personnes par exemple).

Nous voyons donc que l'autonomie absolue n'est pas un but à atteindre, mais que, par ailleurs, l'autonomie peut très bien rester au service de finalités prescrites par l'humain. Parfois même, l'autonomie sera la garantie de l'exécution correcte et fiable de buts fixés par des autorités légitimes. Pensons, par exemple, à des systèmes autonomes corrigeant la sûreté de certaines machines de microchirurgie. Ici, on augmente le contrôle et la précision d'un geste crucial et vital en déléguant les initiatives à un système autonome.

La référence à l'autonomie totale des machines est un leurre sémantique. Elle ne peut être voulue et recherchée pour elle-même, ce serait de l'inconscience et de l'inconsistance. Il vaudrait mieux parler d'**autonomie finalisée**, c'est-à-dire de la caractéristique de systèmes dont les degrés de liberté sont au service, non de la machine, mais des finalités prescrites par le décideur et d'un contrôle global plus précis de l'agent humain. Si nous voulons vraiment caractériser légalement ou éthiquement les *LAWS* par exemple, il serait intéressant de dire que ces systèmes ne sont pas acceptables si leur autonomie n'est pas finalisée. On rejoint ici l'idée du contrôle humain **significatif** mais en insistant que ce contrôle doit aussi porter sur la référence concrète et le contrôle efficace des finalités (légales et morales) déterminées *a priori* par l'autorité légitime.

Une algorithmique éthique ?

Lorsque nous posons des questions éthiques à propos de systèmes possédant des capacités d'Intelligence artificielle (IA), ou à propos de robots doués d'une autonomie finalisée, il nous faut distinguer deux types de questions.

D'une part, nous devons nous poser la question de savoir si, ce que nous écrivons dans les algorithmes, reste, au niveau des contenus mais aussi de l'intention, compatible avec l'éthique que nous promouvons. Nous parlons alors d'une algorithmique

éthique. Dans ce premier cas, par exemple, le fait de ne pas tenir compte ou de minimiser, intentionnellement, dans un algorithme d'un certain type, des risques importants pour les civils, serait une faute légale et éthique. L'**écriture même d'un algorithme présente une charge éthique** : ce que l'on y écrit ou ce que l'on n'y écrit pas possède un impact moral. Un algorithme n'est jamais neutre et l'oubli de certains biais qui président à sa constitution peut être une faute éthique majeure.

Nous avons évoqué ci-dessus l'idée d'une autonomie **finalisée**. Une algorithmique **éthique** est justement celle qui commence à inscrire dans ses lignes de codes un rapport à une finalité. Certaines finalités pourraient se révéler dangereuses pour l'humain c'est évident, mais une première condition d'éthicité, une contrainte minimale de moralité, est le refus de laisser une machine en dehors des cadres d'un projet fixé par l'humain. L'algorithmique est éthique à la double condition de subordonner ses codes à un projet humain et de constituer ce dernier à la lumière de normes morales précises. Ainsi donc, opter pour une machine auto-programmable ou auto-apprenante sans supervision humaine est déjà un parti pris éthique important. C'est l'inscription *a priori* d'un retrait de l'humain, ce qui n'est pas neutre d'un point de vue éthique et juridique.

D'autre part, nous pouvons nous demander s'il serait possible de traduire des exigences éthiques dans et par un algorithme. Nous parlons alors de la question de la possibilité d'une **éthique algorithmique** [2]. Ici, le « regard » éthique ou le « contrôle » moral serait le fait non pas d'un humain mais d'un logiciel. Nous allons montrer que ceci n'est pas entièrement satisfaisant et que nous ne pouvons entièrement accepter l'introduction de « moral machines » [3].

Une éthique algorithmique ?

Il est pensable de programmer un système de telle manière qu'il contrôle la satisfaction de certaines règles. Mais la question de l'implémentation de l'éthique ne signifie pas seulement celle de règles plus ou moins complexes. Pourquoi ?

❶ Parce que l'**évaluation éthique** repose d'abord sur trois éléments majeurs : une appréciation de l'**objet** d'un acte ou d'une action, une prise en compte des éléments du contexte et enfin une appréhension de l'**intention** sous-jacente. Quels sont les éléments à prendre en compte dans un contexte complexe, inédit ? Quels sont les éléments signifiants et ceux qui ne le sont pas ? Comment évaluer une intention cachée, ambiguë ? ❷ Ensuite parce que la **décision éthique** demande que l'on reconnaisse, dans un contexte particulier, quelle règle générale va s'appliquer. Il faut effectuer une qualification des faits avant de leur appliquer la règle ou la loi. Ceci demande une interprétation assez fine.

(2) Nous nous permettons de renvoyer ici à notre article : LAMBERT Dominique, « Une éthique ne peut être qu'humaine ! Réflexion sur les limites des *moral machines* » *in* DOARÉ Ronan, DANET Didier et BOISBOISSEL (DE) Gérard (dir.), *Drones et* killer robots. *Faut-il les interdire ?* (préface de Renaud CHAMPION), Presses universitaires de Rennes, 2015, pp. 227-240.

(3) WALLACH Wendell et ALLEN Colin, *Moral Machines: Teaching Robots Right from Wrong*, Oxford University Press, 2008, 288 pages. ARKIN Ronald C., *Governing Lethal Behavior in Autonomous Robots*, CRC Press, 2009, 256 pages.

Dans ces deux cas, reconnaissance des éléments essentiels pour l'évaluation de la moralité d'un acte ❶ et qualification des faits ❷, il faut savoir « lire entre les lignes » et reconnaître ce qui n'est pas donné immédiatement dans le contenu apparent d'une situation ou dans les contours de la loi universelle. Il faut **un travail d'interprétation** qui est, pour une bonne part, intuitif et créatif, sans être arbitraire (il s'agit de faire de « l'interpolation » entre des données et non pas d'inventer sans contrôle des informations). Il n'est pas simple de programmer une machine pour faire ce genre de travail, surtout lorsqu'on est confronté à des situations inédites. On peut songer à des machines qui inventent de nouvelles règles, qui sortent des ensembles de règles connues, mais la complexité et le caractère inédit des situations imposent parfois que l'on puisse sortir (sans « règles de sortie de règles » !) de tout ensemble existant de règles « classiques », pour pouvoir sauver et faire fonctionner l'esprit général des règles et celui des lois. Dans de nombreuses situations, les machines dotées d'algorithmes implémentant des règles éthiques ou juridiques seront performantes et nécessaires [(4)], mais il existera toujours des situations dans lesquelles l'inédit forcera à inventer une solution, à créer de toutes pièces une manière de résoudre un problème éthique. Le propre de l'humain n'est certainement pas à situer dans la vitesse de raisonnement ni dans le volume d'informations à traiter, caractéristiques où les machines le dépassent allègrement. Le propre de l'humain réside dans cette capacité locale de produire un écart créateur, une rupture radicale qui, d'une certaine manière, permet de sortir d'une impasse. On oublie trop souvent, en droit ou en éthique, que la décision requiert quelque chose de l'ordre de la créativité. Celle-ci est tout autre chose qu'une production aléatoire ou débridée, mais une sorte de production d'une nouveauté qui, sans nier le cadre ancien, le déborde avec une cohérence nouvelle. Le raisonnement juridique, on le sait, est tout autre chose qu'une sorte de déduction mécanique et rigide à partir d'axiomes, comme l'ont montré les débats déjà anciens sur la possibilité d'une formalisation adéquate du raisonnement juridique par un système logique déterminé (la logique déontique par exemple) [(5)].

Le problème posé par l'idée d'une implémentation adéquate de l'éthique dans des programmes informatiques provient aussi des degrés de liberté laissés par le choix des normes, conditionné par le type de philosophie morale que l'on veut promouvoir. Or, toutes les éthiques ne sont pas « algorithmisables » de la même manière. En effet, les éthiques utilitaristes, fondées sur des principes de maximisation ou de minimisation de certaines fonctions ou grandeurs mesurables, se prêtent plus aisément que d'autres à l'écriture de programmes que celles basées sur des valeurs qualitatives (et difficilement quantifiables) par exemple. Opter pour une éthique radicalement algorithmique pourrait revenir implicitement à ne plus admettre que l'utilitarisme comme philosophie, ce qui est partial.

(4) Cf. Delmas-Marty Mireille, « La justice entre le robot et le roseau » *in* Changeux Jean-Pierre (dir.), *L'Homme artificiel*, Odile Jacob, 2007, p. 246.

(5) Perelman Chaïm et Olbrechts-Tyteca Lucie, *Traité de l'Argumentation, la nouvelle rhétorique vol. I. et vol. II.*, Puf, 1958, 350 et 384 pages. Cf. Kalinowski Georges, *La logique des normes*, Puf, 1972, 218 pages. Les objections de Perelman à la réduction du raisonnement juridique à une telle logique ont été présentées, entre autres, lors du 14e Congrès international de Philosophie à Vienne en septembre 1968 : *Akten des XIV Internationalen Kongress für Philosophie, t. II*, Vienne, Herder, 1968, p. 269-311.

On pourrait aussi « éduquer » éthiquement les machines. L'apprentissage par des machines a fait des progrès surprenant ces dernières années [6], et l'on pourrait utiliser ces techniques pour conférer aux robots des normes de comportement. Mais on sait que l'éducation dépend du milieu dans lequel elle est conférée et des superviseurs qui s'en chargent. Le choix de ces derniers, des contenus ou des temps d'apprentissage, ne peut être laissé à la machine, car elle n'a pas de véritables critères de discernement ou de moyens d'en justifier la pertinence. Ultimement, c'est une personne ou un groupe qui devra décider de la « bonne » éducation. Nous sommes de nouveau renvoyés à un choix humain crucial et à la responsabilité d'une personne ou d'une société. On pourrait aussi penser que la tendance à confier systématiquement des responsabilités morales à des machines conduira à désengager l'humain de ses responsabilités. Peut-être verra-t-on se développer des mentalités qui se soumettent de manière non-critique aux « décisions » morales des machines en perdant tout sens de leur devoir éthique propre.

La question essentielle de la responsabilité

Un robot doté d'IA pourrait être dans de nombreuses situations une aide puissante à la décision juridique ou éthique. On peut songer à des systèmes qui alertent les décideurs sur des conséquences néfastes de certaines de leurs actions en termes légaux ou de normes morales. Des algorithmes d'aide à la décision ou de contrôle éthiques sont, me semble-t-il, tout à fait importants. Mais, il ne faut jamais oublier que seul l'humain peut et doit répondre de ses actes. Dans les dilemmes classiques, les machines ne font souvent pas mieux que les humains et inversement ! Mais, la grande différence entre les premières et les seconds, c'est le fait que ces derniers prennent une responsabilité et acceptent de rendre compte de leurs actions. Ils sont, en effet, les **sujets** de leurs actions, alors que les **choses** ne peuvent jamais être dites ni sujets, ni responsables de leurs comportements. Dans les dilemmes, dans ces situations où hommes et machines sont, pour un temps qui se doit d'être très court, perdus, le sujet humain finit par trancher, en acceptant les conséquences de sa décision et en prenant sur lui tous les effets de son choix.

Il ne suffit donc pas qu'un « algorithme éthique » soit placé dans une machine pour que le problème moral de l'utilisation d'un robot autonome, armé ou non, soit réglé. Il faut encore savoir comment on peut remonter à un sujet responsable (celui qui a décidé de l'utilisation du robot autonome, celui qui a écrit l'algorithme éthique, etc.). Aujourd'hui, l'un des dangers de l'usage de la robotique et des techniques d'IA est l'usage potentiellement déviant des technologies comme écran à l'identification des responsables. La technologie et les réseaux complexes d'interactions hommes-machines peuvent donner l'impression que la responsabilité des acteurs est diluée ou même a disparu mais il n'en est rien ! Il y a toujours un ou plusieurs acteurs qui ont décidé de mettre en œuvre une technologie donnée et de l'autonomiser. Néanmoins, les

(6) LECUN Yann, « L'apprentissage profond : une révolution en intelligence artificielle », leçon inaugurale Amphithéâtre Marguerite de Navarre - Marcelin Berthelot, Collège de France, 4 février 2016 (www.college-de-france.fr/site/yann-lecun/inaugural-lecture-2016-02-04-18h00.htm).

possibilités de masquer ou de diluer les chaînes de responsabilité ont augmenté avec la sophistication technologique et la complexité des réseaux [7].

On pourrait se demander ce qui fonde notre insistance sur cette notion de responsabilité ? La réponse à cette question tient dans **un principe de cohérence humaine**. Si nous agissons en tant qu'humains, c'est pour bâtir une société, une économie, une politique, structurées par une vision déterminée de l'homme. Nos actions révèlent ce que nous sommes et ce que nous voulons. Notre humanité prend sens et se construit dans des actes que nous décidons, dont nous assumons les effets pour réaliser ce que nous croyons, individuellement ou socialement, important. Notre humanité passe et se dit par des actes dont nous répondons. C'est en les posant et en en répondant que l'humanité se construit. Parfois cette humanité se construit dans des parcours linéaires où tout semble facile à construire et à décider. D'autre fois, la grandeur de l'humain se voit dans un risque pris dans l'incertitude des situations [8], mais dans la conscience de certaines valeurs. Le droit comme l'éthique demandent un sujet responsable, car sa disparition signifierait l'évanescence de la cohérence et du contenu proprement humains de notre existence.

La fascination pour la vitesse et l'efficacité des machines autonomes dans l'exécution de certaines tâches pourrait conduire à une sorte de démission progressive de l'humain. Ces tâches se vidant alors de leur contenu réel. On le voit dans le risque des transactions financières à très haute fréquence qui conduisent à une économie vidée de sa substance [9].

Il est important de **revenir au sens** des activités. Ce qui donne sens, ce sont les finalités et les décisions humaines. L'enjeu majeur d'une réflexion sur l'utilisation des machines autonomes repose sur le choix ou non de sauvegarder le lien au sens profond des actions. Au fond, c'est le rapport ultime à des sujets responsables (capables de répondre de leurs actes et de leurs choix) qui est le chemin du sens, de la signification.

Critères pour un discernement éthique ? Le choix d'une autonomie finalisée et responsable

Esquissons, à présent, les quelques critères éthiques d'évaluation de l'usage des robots autonomes que nous avons tenté de mettre en évidence.

Il est certain, comme nous l'avons dit que l'autonomie, entendue en un sens absolu, est exclue. Nul ne voudrait d'une machine qui ne réalise pas, ou qui redéfinit, les finalités prescrites par une autorité responsable, politique ou militaire. Mais, par contre, l'autonomie n'est pas exclue radicalement, car elle peut servir les finalités

(7) Nous renvoyons ici, pour ce qui est du monde financier où des problèmes similaires se posent, à l'ouvrage de Xavier Thunis, *Responsabilité du banquier et automatisation des paiements*, Presses universitaires de Namur, 1996, 362 pages : « L'interposition d'un objet technique complexe change profondément les termes dans lesquels la responsabilité du banquier doit être posée et résolue » (p. 301).

(8) Desportes Vincent, *Décider dans l'incertitude* (2e édition), Economica, 2015, 240 pages.

(9) Cf. par exemple : Cooper Ricky, Davis Michael et Van Vliet Ben, « The Mysterious Ethics of High-Frequency Trading », *Business Ethics Quaterly*, vol. 26 n° 1, janvier 2016, p. 1-22.

humaines. Elle peut même être, dans certains cas, un gage de sécurité accrue. Ce qu'il nous faut, c'est assurer que les robots restent, dans leurs modes autonomes ou non, dans les cadres fixés par des intentions humaines précises (et bien entendu moralement et légalement défendables). Nous avons proposé de parler, dans ce cas, d'une autonomie **finalisée**. Bien entendu les finalités elles-mêmes doivent, à leur tour, faire l'objet d'un discernement. C'est ce double discernement qui doit caractériser une première évaluation de l'usage des robots autonomes.

L'analyse effective de cette autonomie **finalisée** doit être envisagée au niveau des contenus des algorithmes, mais aussi au niveau des intentions qui commandent ce qui y est écrit et ce qui n'y est pas écrit. L'approche éthique sera donc attentive également à ce qui reste oublié, à ce qui aurait pu être écrit mais qui ne l'est effectivement pas ! C'est à ces conditions que l'algorithmique sera dite « éthique ».

Nous pouvons et devons, vu la complexité des situations, penser à une éthique assistée par des machines douées éventuellement d'autonomie (c'est-à-dire de degrés de liberté importants). Cependant, une **éthique** seulement **algorithmique**, n'est pas ultimement satisfaisante, car la décision éthique n'est pas une simple application automatique de règles. Elle demande en outre une **créativité** qui permet, à certains moments cruciaux, de résoudre des problèmes en sortant (sans règles prédéterminées) des systèmes de règles pour en sauver l'esprit (et la compréhension de l'esprit des règles n'est pas facilement « algorithmisable »). L'écriture de ces algorithmes éthiques possède, quant à elle, une charge éthique. Car, décider de réduire l'éthique à des algorithmes est moralement lourd de sens, de même que le choix, souvent non explicitement justifié, du type d'éthique que l'on cherche à inscrire dans les lignes de code.

L'usage de machines pouvant opérer largement sans humains doit être enfin et surtout évalué à l'aune de **la responsabilité humaine**. La question centrale est : « Qui est responsable ? », c'est-à-dire « qui doit répondre et donner sens à l'action entreprise avec ce genre de machines ? ». Voilà une question centrale pour le discernement. Là où disparaissent et se brouillent, sous des écrans technologiques, les chaînes de responsabilité, les risques sont grands de voir se commettre des exactions. Un des critères de discernement pourrait donc être celui de la recherche d'un usage **finalisé** et **responsable** de l'autonomie. Mais au fond, ces deux qualificatifs « finalisé » et « responsable » traduisent la même exigence de maintien du lien de l'action avec la volonté humaine.

On ne pourrait alors, suivant ce critère, mettre en action un système technologique doué d'autonomie que dans la mesure où il resterait cohérent avec les finalités humaines. Une condition nécessaire de cette **cohérence anthropologique** serait assurée chaque fois que l'on serait en mesure d'identifier clairement la personne susceptible de répondre (responsabilité vient du verbe latin *respondere* !) de l'utilisation de la machine (en explicitant ses intentions) et d'en assumer les conséquences. Le défi éthique se fait ici technologique et organisationnel : comme maintenir, dans une procédure concrète de déploiement de machines douées d'autonomie, un rapport clair à une autorité responsable et à une exigence de conformité à des finalités fondées juridiquement et éthiquement. ◆