La révolution de l'Intelligence artificielle (IA) en autonomie

Jean-Gabriel Ganascia

Professeur à Sorbonne Université, membre de l'Institut universitaire de France et directeur de l'Équipe Acasa (Agents cognitifs et apprentissage symbolique automatique) du LIP6 (Laboratoire informatique de Paris 6).

Agents autonomes

L'adjectif autonome est de plus en plus souvent employé dans le secteur de l'ingénierie pour désigner des machines dont le comportement se révèle à la fois imprédictible et efficace. Ainsi, parle-t-on aujourd'hui de voitures, de robots, d'armes ou même d'agents autonomes. Ceux-ci défraient régulièrement la chronique par leur nouveauté et surtout par l'inquiétude qu'ils suscitent : on craint qu'ils s'émancipent de la tutelle des hommes qui les ont créés et se retournent contre eux. À titre d'illustration, on lit dans les journaux que les voitures autonomes seront vraisemblablement programmées pour sacrifier la vie de leurs passagers afin de sauver celles des passants, plus nombreux, qui traversent malencontreusement au feu vert [(1)]. Dans un ordre d'idées analogue, mais plus inquiétant encore, des lettres ouvertes [(2)] signées par des dizaines de milliers de personnes demandent un moratoire sur les armes autonomes considérées comme la troisième grande révolution dans l'art de la guerre après la poudre à canon et la bombe atomique ; faisant suite à cette crainte, le Parlement européen a voté, le 12 septembre 2018 [(3)], une résolution demandant aux gouvernements européens de prendre une position commune pour réguler, voire interdire, les armes autonomes.

De plus, on établit généralement un lien entre, d'un côté, les progrès de l'Intelligence artificielle (IA) et, surtout, de l'apprentissage machine et, d'un autre, le déploiement de plus en plus massif d'objets qualifiés par leur comportement d'autonomes.

Cela laisse entendre qu'adviendrait une nouvelle génération de dispositifs autonomes conçus grâce à l'IA et que ceux-ci prendraient une part déterminante dans

(1) Rozières Grégory, « Les voitures autonomes devraient sacrifier leur conducteur pour sauver des passants, mais… », *Huffington Post*, 24 juin 2016 (www.huffingtonpost.fr/).

(2) Cf. AI & Robotics Researchers, « Autonomous Weapons: an Open Letter », 28 juin 2015 (https://futureoflife.org/open-letter-autonomous-weapons/) ou Guerrier Philippe, « Stop les robots tueurs : ces experts français de l'IA qui s'engagent avec Elon Musk », *IT Espresso*, 22 août 2017 (www.itespresso.fr/).

(3) Parlement européen, *Résolution du 12 septembre 2018 sur les systèmes d'armes autonomes (2018/2752(RSP))* (www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P8-TA-2018-0341+0+DOC+XML+V0//FR).

beaucoup d'activités humaines, ce qui s'apparenterait alors à ce que l'on a coutume de qualifier de révolution technologique, à savoir de changement radical dans la conception des machines. Nous allons voir que s'il y a peut-être « révolution en autonomie », et si cette révolution tient bien à l'intelligence artificielle, il ne s'agit pas pour autant d'une révolution technologique.

Autonomie et automatisme

Pour bien comprendre ce qu'il en est, il convient d'abord de revenir à l'étymologie : autonomie vient du grec *autos*, soi-même, et *nomos*, loi, règle. Est autonome ce qui se donne ses propres lois. Cela s'applique d'abord à une Cité ou à un État qui se dote de sa propre législation, avant de concerner, à partir des philosophes des Lumières du XVIII^{e} siècle, des sujets qui s'affranchissent de la religion et de la tradition pour décider eux-mêmes des maximes auxquelles ils s'engagent à soumettre leur comportement. Cela s'oppose à l'hétéronomie, à savoir à la soumission à des lois imposées de l'extérieur, par exemple, pour un État, à se laisser imposer des lois par un autre État, ou, pour un sujet, à subordonner son comportement à des contraintes physiologiques, par exemple à la faim, à la soif ou à la souffrance, ou encore à la colère et à la perte de contrôle de soi-même. Selon cette acception, l'autonomie de la volonté, à savoir la capacité d'un individu à décider de lui-même des lois qu'il se donne, correspond à un idéal jamais pleinement atteint.

Au sens strict, si l'on appliquait cette définition à un dispositif matériel, cela signifierait qu'il déciderait de lui-même, sans le secours de personne et en fonction de sa volonté propre, de ce qu'il ferait. Ainsi, une voiture autonome, en ce sens premier, ne vous conduirait pas où vous le souhaitez, mais là où elle déciderait. De même, toujours en ce sens premier, une arme autonome choisirait d'elle-même sa cible selon ses propres critères qu'elle ne soumettrait à personne. Bien évidemment, de tels « machins » seraient inopérants, puisqu'ils seraient imprévisibles et que, de ce fait, on ne pourrait pas les soumettre à nos propres objectifs.

Aujourd'hui, lorsque l'on parle d'autonomie pour une voiture ou un robot, on signifie généralement autre chose ; il s'agit d'un abus de langage pour désigner un automatisme, à savoir, au sens étymologique, quelque chose qui se meut de soi-même, sans le secours d'un agent extérieur. Pour un homme, un comportement automatique est un enchaînement d'actions qui se fait par-devers lui, sans qu'intervienne de décision consciente ; pour une machine, cela signifie qu'il y a un enchaînement de causalités matérielles dans laquelle n'intervient aucune présence extérieure, en particulier aucun être humain. En ce sens, une voiture autonome détermine, à partir de l'objectif qu'on lui a fixé – comme aller à la piscine – et des informations qu'elle a glanées, la séquence des actions qui vous conduiront à bon port, alors qu'une voiture autonome au sens premier risquerait de remettre en cause l'objectif, se refusant à aller à la piscine, prétextant par exemple que le parking n'y est pas commode ou qu'elle a mieux à faire… Rapportée à un système d'armes, au sens second, une arme autonome exécuterait d'elle-même les objectifs qui lui auraient été donnés, par exemple atteindre tout ce qui présente une signature radar caractéristique, alors qu'au sens premier, elle n'obéirait pas

aux ordres et choisirait seule ses cibles. De ce fait, et contrairement à ce qu'affirment les rédacteurs des lettres ouvertes susmentionnées, aucun militaire un tant soit peu responsable ne voudrait utiliser d'armes aussi imprévisibles.

Intelligence artificielle et agent autonomes

Rappel sur l'histoire et l'empan de l'intelligence artificielle

L'intelligence artificielle est une discipline scientifique qui a vu officiellement le jour en 1956, au *Dartmouth College* de Hanover, dans l'État du New Hampshire (États-Unis), lors d'une école d'été organisée par quatre chercheurs américains : John McCarthy (1927-2011), Marvin Minsky (1927-2016), Nathaniel Rochester (1919-2001) et Claude Shannon (1916-2001). Cette discipline vise à décomposer l'intelligence en fonctions élémentaires, puis à construire des machines pour les simuler, une à une. Ainsi, si l'on décompose les capacités cognitives humaines en cinq grandes catégories :

1) perception des signaux envoyés par nos organes des sens, c'est-à-dire construction d'une représentation qui agrège les informations qu'ils nous transmettent ;
2) mémoire, à savoir représentation – au sens étymologique de re-présentation, c'est-à-dire de restitution de la présence d'une chose en son absence –, et exploitation de ces informations avec l'apprentissage ;
3) pensée, c'est-à-dire calcul sur les représentations ;
4) fonction communicative, à savoir capacité à échanger entre machines ainsi qu'entre hommes et machines ;
5) enfin, fonctions exécutives pour qu'une machine prenne des décisions et les mette en œuvre.

Durant les soixante dernières années, de nombreuses fonctions cognitives ont été simulées, que l'on pense à la perception pour les images ou les paroles, à la modélisation de mémoires sémantiques, à l'apprentissage supervisé qui exploite les immenses quantités de données stockées, au traitement et à la compréhension du langage naturel ou encore à la prise de décision. Ces simulations aident à mieux comprendre l'intelligence, qu'elle soit humaine ou animale. Elles permettent aussi de réaliser des automatismes, par exemple des voitures dites autonomes ou des robots qui prennent, d'ores et déjà, une place si importante dans l'industrie actuelle.

Notions d'agent autonome

Il est d'usage de caractériser les automatismes programmés à l'aide de techniques d'IA comme étant des « agents autonomes » en cela que ce sont des entités agissantes, à savoir des *agents*, et qu'ils sont *autonomes* au sens second que nous avons évoqué précédemment, car ils sont mus par un enchaînement de causalités qui va de la prise d'information à l'action, sans qu'intervienne aucune présence humaine.

Plus précisément, un agent au sens technique est une notion très générique renvoyant à toute chose qui dispose de capteurs, de capacités d'actions, de procédures

de décision et éventuellement de buts ou, à défaut, d'une fonction de récompense. Dans le cas d'une voiture autonome, les capteurs peuvent être des caméras et un *GPS* ; les actions : des coups d'accélérateur ou de frein, des changements de vitesse et des petits mouvements du volant à gauche ou à droite ; les buts : des lieux à atteindre ; et les procédures de décision : des algorithmes qui déterminent les actions à effectuer pour rapprocher la voiture de sa destination en tenant compte de la situation de la voiture telle que les capteurs permettent de l'apprécier. L'IA aide à la fois à interpréter les signaux fournis par les capteurs pour identifier la route, les trottoirs, les panneaux de circulation, les piétons, les autres voitures, etc. et à décider des actions à accomplir dans chaque situation.

Si l'agent était autonome au sens premier, et pas seulement au sens second, il ne se contenterait pas de choisir une action pour parvenir à réaliser les buts qu'on lui a fixés ou pour optimiser la somme des récompenses qu'il espère obtenir, mais il déciderait de lui-même des objectifs qu'il se fixe. Aujourd'hui, les agents autonomes qu'ils soient programmés avec des techniques d'intelligence artificielle ou non, ne choisissent pas d'eux-mêmes leurs cibles ; sans doute agissent-ils d'eux-mêmes et choisissent-ils leurs actions mais ils le font au regard des buts qu'on leur a fixés, pour les résoudre. Si l'on se réfère à l'étymologie, on devrait donc parler d'automates, autrement dit d'entités qui font effort d'elles-mêmes, et non d'agents autonomes.

À ce premier paradoxe qui fait qu'un agent autonome n'est pas proprement autonome, on doit en ajouter un second : un agent autonome n'est pas un agent au sens philosophique. En effet, eu égard à la théorie philosophique de l'action, un agent est ce qui est à l'origine de l'action. Cela s'oppose à une chose qui, soumise à des forces matérielles, serait mue par elles. Or, les agents autonomes que l'on fabrique ne font qu'exécuter les instructions qu'on a écrites pour eux ; ils se réduisent à des séquences de causalités matérielles parfaitement identifiées ; il n'est donc pas juste de dire qu'ils initient des actions d'eux-mêmes.

Bien étrange notion que celle d'agent autonome qui n'est ni un agent au sens philosophique, ni vraiment autonome, et qui, de ce fait, et à première vue, s'apparente quelque peu au couteau de Lichtenberg à savoir, selon l'aphorisme du philosophe, écrivain et physicien allemand du XVIII^{e} siècle Georg Christoph Lichtenberg, à « un couteau sans lame auquel ne manque que le manche ». Pour autant, cette notion n'est pas vide ; loin de là, elle recouvre des réalités technologiques actuelles et tangibles. On ne saurait dénier leur existence qui est patente. Quant à leur dénomination, elle est attestée, et l'on ne saurait pas plus la mettre en cause. Il importe toutefois de bien comprendre la signification exacte des termes employés pour éviter tout malentendu. La notion d'agent recouvre en intelligence artificielle une réalité qui a été rappelée plus haut ; et l'autonomie correspond ici à la seconde définition que nous avons donnée, à savoir à une séquence de causalités qui ne fait pas intervenir d'agent humain.

Malentendus sur les « robots tueurs »

Les *Systèmes d'armes létaux autonomes*, en abrégé les Sala, appelés plus communément les « robots tueurs » correspondent à des agents dont l'autonomie devrait être

entendue non au sens second, comme dans une voiture autonome, mais au sens premier d'entité déterminant par elle-même ses objectifs. En effet, selon la définition que l'on en donne couramment, les Sala [4] sélectionneraient d'eux-mêmes leur cible et engageraient le feu sans intervention humaine. Toute l'ambiguïté de cette définition repose sur le verbe « sélectionner ».

Soit cette sélection résulte d'un processus de classification automatique à partir d'un objectif fixé à l'avance, par exemple atteindre un char ou un homme, ou encore un visage dont la signature radar, visuelle ou infrarouge, aurait été caractérisée par apprentissage machine supervisé. Dans cette conception, l'arme identifierait une cible dans un flux d'information, puis engagerait bien le tir sans intervention humaine mais cela n'aurait rien de nouveau. En effet, d'ores et déjà, une mine sélectionne sa cible et engage le feu sans intervention humaine. De plus, il n'y a généralement pas là, avec les mines, d'intelligence artificielle. De même, à la frontière de la Corée du Nord et de la Corée du Sud, on dit qu'il existe des armes qui sonnent l'alerte, lancent des sommations puis engagent le tir dès qu'elles détectent des mouvements. Il n'y a pas là non plus d'IA à proprement parler, mais simplement des capteurs et des automatismes.

Cependant, si l'on en croit les lettres ouvertes qui ont été signées en juillet 2015, puis en juillet 2017 lors de deux conférences internationales d'intelligence artificielle, les très récents progrès de l'IA conduiraient, d'ici peu et inéluctablement, à la réalisation de nouveaux systèmes d'armes autonomes constituant, dans les affaires militaires, une révolution analogue à celle qu'a provoquée la poudre à canon ou à celle qui s'est produite plus tard avec la bombe atomique. Cette révolution dans l'ordre de l'autonomie tiendrait à ce que les Sala ne se contenteraient pas de catégoriser leur cible, à partir de critères prédéfinis, car si tel devait être le cas, il n'y aurait rien là de radicalement neuf, mais qu'ils la détermineraient d'eux-mêmes. Il y aurait donc un basculement qui ferait passer d'une autonomie au sens second, c'est-à-dire d'une séquence de causalités où n'intervient aucun humain, à une autonomie au sens premier, en l'occurrence au choix, par la machine, de sa cible, et non à la simple catégorisation des flux d'informations entrant sur un critère prédéfini.

Nous avons donc deux conceptions de l'autonomie en matière de systèmes d'armes. Selon la première, nous nous inscririons dans la continuité des automatismes existants et il n'y aurait alors pas de rupture. Selon la seconde, nous parviendrions effectivement à des systèmes d'armes totalement différents de ceux qui existaient auparavant. Cependant, l'état de développement des technologies en IA n'autorise pas à comprendre comment de tels systèmes d'armes seraient construits et cela ne permet pas non plus d'appréhender la nature de ces systèmes. Bref, la crainte de voir une nouvelle génération de Sala révolutionner la guerre paraît bien peu fondée.

Au reste, il se peut que des systèmes d'armes fassent appel à de l'intelligence artificielle pour aider l'opérateur à déterminer sa cible ou pour donner l'alerte, auquel

(4) Pour une analyse des appels aux moratoires sur les armes autonomes, consulter GANASCIA Jean-Gabriel, TESSIER Catherine et POWERS Thomas M., « On the Autonomy and Threat of "Killer Robots" », *APA Newsletter on Philosophy and Computers*, vol. 17, n° 2, été 2018, p. 3-9 (https://c.ymcdn.com/).

cas ils ne seraient pas automatiques et encore moins autonomes, puisqu'ils se mettraient au service d'un opérateur ; ce ne serait donc pas des Sala. De même, il se pourrait que des systèmes d'armes fassent appel à de l'intelligence artificielle pour brouiller les communications, auquel cas, ce serait bien des systèmes d'armes utilisant de l'IA, mais ils ne seraient pas à proprement parler létaux. Ainsi, y a-t-il dans l'esprit du grand public un lien entre IA et Sala qui est doublement discutable, d'une part parce que la notion de Sala ne recourt pas nécessairement à l'intelligence artificielle et d'autre part, parce que l'IA ne conduit pas inéluctablement à la réalisation de Sala et *a fortiori* d'une nouvelle génération de Sala qui révolutionnerait l'art de la guerre.

Nous devrions en conclure que de révolution de l'IA en autonomie, il n'y a point. Pourtant, à bien y regarder, les choses ne sont pas si simples.

Automatismes au service de l'autonomie

L'autonomie, au sens philosophique, recouvre, comme nous l'avons déjà dit, l'idée de liberté telle que l'ont définie des philosophes des Lumières comme Jean-Jacques Rousseau ou Emmanuel Kant, c'est-à-dire la capacité à obéir aux maximes que l'on s'est données. Autrement dit, est libre non celui qui décide à tout moment de faire ce qui lui plaît, car celui-ci serait esclave de son désir instantané, mais celui qui a adopté sciemment des règles et qui s'y soumet.

Or, nous savons tous que cette autonomie du sujet est un idéal impossible à atteindre. En effet, il est bien des situations où les conditions extérieures troublent notre esprit, obscurcissant notre perception de la réalité ou nos facultés de raisonnement. C'est tout particulièrement le cas dans des situations extrêmes comme l'on en rencontre souvent dans le domaine militaire. Par exemple lorsque, dans un avion de chasse, le pilote est soumis à des accélérations qui lui font perdre conscience ou lorsqu'une surcharge cognitive prive l'opérateur de discernement et de clairvoyance. Dans ces éventualités, il nous arrive, sans que nous le souhaitions, d'agir en contradiction avec nos engagements personnels.

Afin de lutter contre les conséquences néfastes de pertes de contrôle du pilote d'un processus, il peut être utile de recourir à des automatismes qui conservent la maîtrise de la situation tant que l'opérateur se trouve en situation délicate. Ainsi, pour reprendre l'exemple d'une cabine de pilotage d'avion, un agent autonome de pilotage automatique continue de maintenir le cap et l'altitude le temps que le pilote reprenne conscience. Dans de telles éventualités, l'agent autonome aide l'opérateur à être plus fidèle à sa volonté, autrement dit, si l'on se reporte à ce qui vient d'être dit, à être plus autonome. Ainsi, l'agent autonome ne serait pas autonome en lui-même, ou du moins, il ne le serait qu'au sens second que nous avons donné, mais il aiderait surtout l'opérateur à être plus autonome au sens premier.

⁂

En conclusion, la vraie révolution en autonomie ne consisterait pas à fabriquer des dispositifs matériels qui se substitueraient à nous, du fait de leur supposée autonomie de comportement, mais à concevoir des automatismes qui prendraient le relais, en cas de défaillance humaine causée soit par une situation physique éprouvante, soit par une saturation des facultés d'attention, nous aidant ainsi à demeurer fidèles à nos engagements en dépit de nos absences passagères, et donc à rester plus autonomes. ◆