Jean-Gabriel Ganascia

PEUT-ON CONTENIR L'INTELLIGENCE ARTIFICIELLE ?

En même temps qu'elle fascine, l'intelligence artificielle fait peur. On le constate tous les jours, non seulement dans les enquêtes d'opinion, mais aussi chez des autorités incontestées du monde contemporain qui sonnent le tocsin et parfois battent déjà en retraite, en prophétisant la fin de l'humanité. Cela arrive aux États-Unis, avec des hommes d'affaires comme Bill Gates ou Elon Musk, ou encore des scientifiques comme Frank Wilczek, prix Nobel de physique, au Royaume-Uni, où le regretté Stephen Hawking s'inquiétait des avancées de l'intelligence artificielle, qui constituait selon lui un risque existentiel pour l'humanité, ou en France, avec des philosophes comme Jean-Michel Besnier et des personnalités comme Laurent Alexandre qui font tous d'inquiétantes annonces sur l'évolution du monde. Comment se prémunir contre les dangers que l'intelligence artificielle nous fait encourir ? Telle est la question que nous abordons ici. Mais avant, il convient de préciser la nature de ces dangers et la possibilité même de les combattre, lorsque ceux-ci relèvent plus de craintes illusoires que de réelles menaces.

Les trois dimensions de l'IA

Rappelons tout d'abord que le terme « intelligence artificielle » a été inventé en 1955 par un jeune mathématicien de 28 ans, John McCarthy, pour désigner une discipline scientifique neuve destinée à mieux comprendre l'intelligence en la décomposant en facultés cognitives si élémentaires que l'on devrait être en mesure de fabriquer des machines pour les simuler. De la sorte, le raisonnement, la perception, le calcul, la mémoire, voire la découverte scientifique ou la créativité artistique, devraient être décrits avec une précision telle que des ingénieurs seraient

en mesure de les reproduire à l'aide d'un ordinateur. Soixante-quatre ans plus tard, beaucoup de progrès ont été réalisés dans cette perspective. L'intelligence artificielle et les sciences cognitives ont permis de mieux comprendre notre intelligence en en simulant de nombreuses facettes. Sans doute conserve-t-elle encore bien des mystères. Ainsi ne connaît-on pas encore la fonction du rêve dans la consolidation de nos souvenirs. De même, le rire demeure incompréhensible. Quant à l'organisation des informations dans notre cerveau, en dépit des progrès accumulés grâce à l'imagerie fonctionnelle cérébrale, elle demeure encore bien obscure. Mais, indéniablement, même si elles sont loin d'épuiser la cognition humaine, ces modélisations ont déjà fait faire de grands progrès à notre compréhension.

En outre, le coût des ordinateurs ayant diminué considérablement, et leur puissance s'étant accrue dans les mêmes proportions, ceux-ci se sont disséminés partout, en particulier dans tous les objets quotidiens : voitures, bicyclettes, téléphones, fours, etc. Il s'ensuit qu'ils prennent une part de plus en plus grande dans toutes les activités humaines. De plus, grâce aux progrès de l'intelligence artificielle, la simulation de différentes fonctions cognitives – par exemple, la reconnaissance des visages ou les empreintes digitales, la reconnaissance de la parole, la traduction automatique, etc. – permet d'automatiser des tâches qui, jusque-là, auraient requis une présence humaine. On conçoit que cela puisse avoir un effet conséquent sur le travail dans nos sociétés et que cela fasse craindre le chômage de la même façon, voire peut-être plus, que les révolutions technologiques d'antan. À cela on peut ajouter que les échanges interhumains se passant en grande partie par le truchement de flux d'information, l'intelligence artificielle aide non seulement à les réguler mais aussi à les anticiper, de façon à les dominer.

Enfin, les performances des machines laissent songeur : une machine a vaincu, à plusieurs reprises, le champion du monde en titre au jeu d'échecs et même, plus récemment, l'un des meilleurs joueurs au monde au jeu de go ; d'autres démontrent ou aident à démontrer des théorèmes mathématiques ; on construit automatiquement des connaissances à partir de masses immenses de données (*big data*). Grâce à cela, des automates reconnaissent la parole articulée et « comprennent » des textes écrits en langage naturel, c'est-à-dire les traduisent dans des formalismes logiques ; des voitures se conduisent seules ; des robots font la guerre à la place des hommes ; certains scientifiques cherchent même à vaincre la mort en déterminant les mécanismes du vieillissement… Non seulement la plupart des dimensions de l'intelligence – sauf peut-être l'humour – font l'objet

d'analyses et de reconstructions rationnelles avec des ordinateurs, mais de plus les machines outrepassent nos facultés cognitives dans la plupart des domaines, ce qui fait craindre à certains un risque pour le futur de l'humanité, risque qui fait écho à des mythes très anciens, comme celui du Golem ou du robot.

En résumé, le terme d'intelligence artificielle renvoie à trois choses : une discipline scientifique qui vise à mieux comprendre l'intelligence en en reproduisant certaines fonctionnalités sur des ordinateurs ; les dispositifs matériels qui simulent certaines fonctions cognitives et qui sont amenés à prendre une place de plus en plus conséquente dans la vie quotidienne des hommes de notre temps ; enfin, l'inquiétude suscitée par les progrès fulgurants de ces techniques qui bien souvent dépassent l'entendement humain et qui pourraient, si cela se poursuivait, conduire, à terme, à une domination de l'humanité par les machines.

Ces trois dimensions associées à l'IA suscitant à la fois espoirs et inquiétudes, il est légitime de se demander si l'on peut contenir les progrès de façon à maîtriser les évolutions à venir et à éviter le pire. Cependant, on conçoit que, ces trois dimensions étant hétérogènes, elles conduisent à poser les questions dans des termes différents et donc à être envisagées successivement.

Se prémunir contre une « IA forte »

Comment et pourquoi vouloir limiter les progrès d'une discipline scientifique ? L'expérience passée montre qu'outre son caractère réactionnaire et rétrograde une telle attitude serait illusoire, car jamais un décret n'a stoppé l'avancement des connaissances. De plus, qu'est-ce qui justifierait un tel dessein ? Nous avons vu que l'intelligence artificielle, en tant que discipline scientifique, vise à mieux nous connaître, ce qui n'a rien de répréhensible ni de condamnable, ni *a fortiori* de dangereux, bien au contraire. Comme nous le verrons plus loin, certains s'inquiètent, à juste titre, d'applications délétères qui découlent de tels progrès. Pour autant, cela ne justifie pas un moratoire sur les recherches fondamentales, à moins que l'on craigne la mutation de cette discipline scientifique de nature pragmatique et expérimentale en une science au caractère maléfique et irrépressible, autrement dit le passage de ce que certains appellent une intelligence artificielle faible (*weak artificial intelligence*) ou étroite (*narrow artificial intelligence*), qui se contente de simuler des facultés cognitives spécifiques comme la reconnaissance de la parole, la compréhension du langage naturel ou la conduite automobile, à une

intelligence artificielle dite générale (*artificial general intelligence*) ou forte (*strong artificial intelligence*), qui reproduirait un esprit, voire une conscience, sur une machine. Or, aujourd'hui, il est courant d'affirmer que ce dernier type d'IA adviendra bientôt et qu'il aura des répercussions majeures, tout à la fois positives et négatives, sur le devenir de l'espèce humaine.

Pour être en mesure d'apprécier l'éventuel bien-fondé des perspectives inquiétantes qui se profilent avec l'IA forte et qui légitimeraient un moratoire sur l'IA en général, il convient d'abord de faire l'archéologie de la distinction entre IA forte et IA au sens classique, qualifiée aussi d'IA faible.

L'origine de cette différenciation remonte au début des années 1980, avec les travaux d'un philosophe américain, John Searle, qui mettaient

en cause les théories des philosophes dits cognitivistes selon lesquelles le fonctionnement de l'esprit serait en tout point analogue à celui d'un ordinateur et pourrait être reproduit intégralement à l'aide de techniques d'intelligence artificielle. Comme il professait une grande admiration pour les réalisations pratiques de l'IA et qu'il souhaitait pourtant en montrer les limites intrinsèques, John Searle distingua deux formes d'IA, celle des ingénieurs susceptible de reproduire un grand nombre de fonctions cognitives avec une manipulation symbolique d'information, et celle des philosophes qui prétendaient restituer avec les techniques d'IA l'esprit dans son intégralité, et en particulier la conscience.

Il appela la première l'« intelligence artificielle faible » (IA faible) et lui concédait tout : selon lui, elle parviendrait à des réalisations techniques inouïes. En revanche, la seconde, celle des philosophes cognitivistes, il la qualifia d'« intelligence artificielle forte » (IA forte) tout en la discréditant, car selon lui elle serait incapable d'atteindre ses objectifs. Pour le montrer, il fit appel à l'expérience de pensée dite de la chambre chinoise[1].

Dans les années qui suivirent son introduction par John Searle, la notion d'IA forte eut tant de succès qu'on lui assimila souvent l'IA dans son ensemble. Cela se produisit tout particulièrement chez des philosophes

1. Un Américain ne parlant que l'anglais est enfermé dans une chambre où il doit, s'il veut manger, choisir, en fonction des caractères qu'on lui présente et selon des règles précises inscrites dans un catalogue, des carreaux de céramique dans un panier et les exhiber à une lucarne. Observant de l'extérieur les réponses à leurs questions exprimées en chinois, les passants peuvent éventuellement penser, si les règles de manipulation des carreaux sont bien conçues, que l'Américain « comprend » le chinois, car il répond avec pertinence. Pourtant, d'après Searle, il n'en ira pas ainsi. Et, toujours selon lui, cette prison chinoise est une bonne métaphore de ce qu'est l'IA : elle fait illusion, mais n'accède jamais au sens…

peu soucieux de philologie qui inventèrent la notion de « bonne vieille IA » (GOFAI, pour *good old-fashioned artificial intelligence*) afin de caractériser ce qu'ils déclarèrent avoir été l'ambition démiurgique de l'IA des origines. Ils voyaient alors le texte de John Searle comme une critique des ambitions excessives de l'IA elle-même et des méthodes qu'elle était supposée employer.

De façon assez paradoxale, des scientifiques, comme le roboticien Hans Moravec, leur emboîtèrent le pas vers la fin des années 1980 en reprenant à leur compte le concept d'IA forte pour affirmer que les méthodes de l'intelligence artificielle un peu renouvelées – ce qu'ils appelèrent la « nouvelle IA » en référence à la « nouvelle cuisine » – conduiraient à la construction de machines totalement intelligentes faisant écho aux machines ultra-intelligentes de la science-fiction.

Quelques années plus tard, au tout début du XXIe siècle, se fit jour un autre courant qualifié d'intelligence artificielle générale (AGI, pour *artificial general intelligence*), qu'il ne faut surtout pas confondre avec l'IA des origines, née cinquante ans auparavant. Ses promoteurs, parmi lesquels on peut citer entre autres Ben Goertzel, Marcus Hutter ou Jürgen Schmidhuber, désirent refonder l'IA sur des bases mathématiques solides, équivalentes en certitude à celles sur lesquelles s'appuie la physique.

Sans nous étendre sur les fondements et les justifications génériques de l'IA générale, indiquons que tandis que l'intelligence artificielle forte trouve son origine dans les réflexions de philosophes, l'intelligence artificielle générale vient de travaux de physiciens théoriciens reconvertis. Même si elle reprend à son compte le projet de l'IA forte, elle s'appuie sur des théories mathématiques plutôt obscures et, parfois, sur des réalisations informatiques, alors qu'initialement l'intelligence artificielle forte reposait uniquement sur une justification d'ordre discursif.

Ajoutons que, quoique, de par leur dénomination, l'IA dite forte ou générale d'un côté et l'IA au sens premier – que l'on a rebaptisée IA faible – de l'autre apparaissent sœurs, tant la finalité que les méthodes de l'une et de l'autre diffèrent radicalement : là où nous avions une discipline scientifique fondée sur des simulations informatiques et sur leur validation expérimentale, nous trouvons des approches philosophiques ou mathématiques fondées uniquement sur des argumentations théoriques, sans vraie contrepartie empirique ; de plus, on insistait pour l'une sur la décomposition de l'intelligence en fonctions élémentaires reproductibles sur des ordinateurs, alors qu'on appuie pour l'autre sur la recomposition totale d'un esprit et d'une conscience à partir de fonctions cognitives élémentaires.

Reconsidérons maintenant, à la lumière de ce qui vient d'être dit, les pronostics sur la réalisation d'une IA forte qui viendrait supplanter l'IA faible : là où l'on était capable de mesurer, pas à pas, les progrès d'une discipline scientifique, en confrontant des simulations avec des observations empiriques, l'IA forte et/ou générale s'imposent, l'une et l'autre, comme des professions de foi auxquelles on adhère plus par conviction que par raison… En conséquence, le danger ou l'espoir que représente l'IA forte est plus imaginaire que réel. Il n'apparaît donc pas plus nécessaire de la « contenir » que l'IA faible.

Fin de l'humanité

Le 1er mai 2014, une tribune alarmiste parue dans le journal *The Independent* et signée par quatre éminents scientifiques, Stephen Hawking, astrophysicien fameux, Stuart Russell, professeur à l'université de Californie à Berkeley et auteur d'un manuel sur l'intelligence artificielle qui fait autorité, Max Tegmark et Frank Wilczek, tous deux physiciens et professeurs au Massachusetts Institute of Technology, nous alertait des dangers que l'IA fait courir à l'humanité. Selon ces quatre personnalités, nous atteindrions très bientôt un point de non-retour au-delà duquel nous irons inéluctablement à notre perte sans jamais pouvoir revenir en arrière. Aujourd'hui, il serait encore temps ; demain, plus rien ne sera possible !

Cet appel à la vigilance fut suivi de beaucoup d'autres lancés par les mêmes, par exemple par Stephen Hawking à la BBC, par Stuart Russel, qui a parrainé en 2015 deux lettres ouvertes publiées sur le site internet du Future of Life Institute, l'une sur les dangers de l'IA, l'autre sur les méfaits potentiels des armes autonomes, ou par d'autres, qu'il s'agisse de philosophes comme Nick Bostrom ou d'hommes d'affaires très en vue comme Elon Musk et Bill Gates. Partout dans le monde, ces personnalités font des émules. En France, c'est le cas d'un médecin qui est en même temps un homme d'affaires très médiatique, Laurent Alexandre, et d'un philosophe qui s'est spécialisé dans le transhumanisme, Jean-Michel Besnier.

Ces déclarations publiques annoncent toutes un événement majeur et inquiétant consécutif à l'utilisation massive des technologies de l'information. Elles pointent les conséquences dramatiques de cet événement pour l'humanité, son inéluctabilité et son imminence. Ces trois points méritent d'être examinés l'un après l'autre.

Commençons par les supposées conséquences néfastes pour l'humanité dans son ensemble. Aux dires de Stephen Hawking, le déploiement des

technologies d'IA sur des ordinateurs hyperpuissants constituerait « notre plus grande menace existentielle », car les humains ne pourront plus rivaliser avec des machines devenues plus intelligentes qu'eux. Cela sous-entend que ces machines, du fait de leur ultra-intelligence, entreraient en rébellion contre nous, prendraient le pouvoir et nous réduiraient en esclavage, ce qui signifie que les machines que nous fabriquerons auront des désirs, des aspirations, des besoins distincts des nôtres et de ceux que nous leur avons insufflés, autrement dit qu'elles se constitueront en sujets autonomes agissant pour eux-mêmes et donc doués d'une conscience et d'une volonté propres. Or, pour l'instant, les scientifiques ne savent pas comment procéder pour concevoir de telles machines, et l'on est loin de comprendre les mécanismes à l'origine de la conscience et de la volonté. D'après leurs auteurs, ces prédictions reposent sur le degré de complexité des machines, calculé en nombre de composants, qui avec le temps deviendrait équivalent, puis supérieur à celui du cerveau humain. Cependant, même si l'accroissement des capacités de calcul des machines a permis, ces dernières années, à l'IA de réaliser des prouesses sur des tâches spécifiques, comme les jeux (jeux d'échec, jeu de go ou poker), la reconnaissance faciale, la reconnaissance de la parole ou la conception de voitures autonomes, la complexité, la capacité de stockage d'information et la rapidité de calcul ne produisent pas à elles seules de l'intelligence, tant s'en faut ! De nombreuses facultés cognitives – par exemple, le rire et le rêve – demeurent encore très difficiles, voire impossible à simuler sur des ordinateurs, quand bien même il n'existe pas d'argument tangible permettant d'affirmer qu'elles sont à jamais hors de portée de l'intelligence artificielle.

Ajoutons que ces annonces supposent aussi que les machines formeront une coalition hostile aux hommes, ce qui, là encore, ne repose sur aucun fondement et, en conséquence, paraît difficile à admettre et, plus encore, à prouver…

Le deuxième point porte sur l'inéluctabilité de cet événement catastrophique, ce qui suppose, implicitement, que la technologie se déploie de façon autonome, indépendamment de nous. Cette inéluctabilité se conjugue avec l'imminence, à savoir avec le troisième point, qui découleraient l'un et l'autre, selon certains auteurs, d'un calcul mathématique issu de l'extrapolation de la loi de Moore. Rappelons que cette loi, émise en 1964 par Gordon Moore, le fondateur de la société Intel, est une loi d'observation qui constate que, depuis 1959, les performances des processeurs doublent tous les deux ans. Si l'on prolongeait indéfiniment cette loi, nous obtiendrions à terme des ordinateurs infiniment puissants, ce

qui paraît tout à la fois contraire à l'intuition et contraire aux anticipations scientifiques des physiciens. C'est pourtant sur l'extrapolation de cette loi d'observation que certains ingénieurs se fondent pour affirmer que les ordinateurs nous dépasseront bientôt. C'est ce qui justifie l'affirmation selon laquelle cette catastrophe que l'on appelle la « Singularité technologique », avec un « S » majuscule pour signifier son unicité et son exceptionnalité, serait à la fois inéluctable et imminente.

En dépit de l'autorité des scientifiques qui, comme Stephen Hawking, se prononcent là, et de la célébrité de personnalités qui, comme Elon Musk ou Bill Gates, leur emboîtent le pas, on doit considérer avec circonspection ces prédictions en se demandant ce qui les justifie sur le plan scientifique. Au sein du grand public, beaucoup de personnes impressionnées par la renommée de ces grands personnages pensent qu'ils disposent d'informations confidentielles motivant leurs inquiétudes. Or, à supposer que les informations dont ils ont connaissance soient secrètes, au point qu'ils préféreraient les tenir scellées quand bien même la survie de l'humanité serait en jeu, cela voudrait dire qu'ils en seraient d'une façon ou d'une autre les bénéficiaires et que, dans cette éventualité, leur attitude devrait être suspectée puisqu'ils nous dissimuleraient des informations essentielles pour notre devenir collectif. En revanche, si ce n'était pas le cas et s'il n'y avait rien là de caché, ils devraient être en mesure d'expliquer clairement ce qui justifie leurs craintes, ce qu'ils ne font pas de façon convaincante dans leurs déclarations, qui demeurent elliptiques. Bref, nous ne devons pas renoncer à notre sagacité critique et nous laisser abuser par des déclarations « apocalyptiques », au sens étymologique, c'est-à-dire qui prétendent révéler une catastrophe imminente, qu'aucun argument rationnel ne justifie. Nous devons, moins encore, tenter de contenir l'intelligence artificielle, car les risques d'emballement, de dépassement et d'assujettissement de l'homme par les machines relèvent plus de la fable que d'arguments scientifiques fondés avec rigueur.

Éthique des applications de l'IA

Qu'on la considère comme une discipline scientifique ou que l'on s'inquiète des prouesses qu'elle réalise et de la transformation du statut de l'humanité qui en résulterait, l'IA ne saurait être « contenue », au sens où toute interdiction apparaîtrait chimérique, soit qu'elle soit vaine, si elle faisait l'objet d'une interdiction de procéder à des recherches, soit qu'elle n'ait aucun objet, dans l'éventualité où il s'agirait de lutter contre

une pseudo-IA forte et de s'opposer à l'autodéploiement de la technologie ou à une prétendue Singularité au-delà de laquelle la civilisation disparaîtrait. Notons le caractère particulièrement absurde de cette dernière éventualité, puisqu'elle repose en grande partie sur le postulat d'un déterminisme technologique qui, s'il était avéré, rendrait vaine toute intervention humaine.

Il n'en demeure pas moins que l'IA prend une part de plus en plus grande dans la vie de tous les jours et qu'elle transforme la société. Cela correspond au second volet que nous avions évoqué dans l'introduction. Dans ce contexte, il apparaît essentiel de contenir non l'IA en tant que telle, mais le déploiement des applications de l'IA de façon à les mettre au service d'un projet social et à éviter l'asservissement des femmes et des hommes aux impératifs des technologies ou, pire, aux désirs des maîtres de ces technologies, en l'occurrence des géants du Web. Cela signifie qu'il faut s'assurer de la compatibilité de l'usage du numérique et de l'IA avec les règles de vie que nous choisissons, autrement dit avec notre morale collective. La difficulté tient à ce qu'avec le numérique tout ce qui fait la trame du tissu social, à savoir ce à partir de quoi s'établissent les relations entre les hommes, évolue grandement. Prenons quelques-uns des éléments à partir desquels se bâtissent les liens interpersonnels. L'amitié, notion ancienne s'il en est, est réécrite avec les réseaux sociaux qui proposent une forme d'amitié nouvelle en ligne. Non seulement celle-ci diffère quelque peu de l'amitié traditionnelle, mais de plus cette amitié sur les réseaux a un effet en retour sur l'amitié au sens traditionnel. De même, la confiance se transforme : elle ne disparaît pas dans le monde numérique, mais elle s'établit différemment. Ce n'est plus la présence d'un témoin humain qui la fonde, ce n'est plus l'écrit, c'est la machine ou, plus exactement, la *blockchain*[2], c'est-à-dire une procédure que seuls des réseaux de machines peuvent mettre en œuvre. La réputation elle-même ne se construit plus désormais sur le seul qu'en-dira-t-on, mais aussi sur un score numérique qui agrège un grand nombre d'informations relatives à des infractions mineures. En Chine, celles-ci font référence à des manquements qui nous paraîtraient véniels, comme traverser au feu vert pour un piéton ; chez nous, cela peut résulter, pour une banque, du calcul du nombre de jours de découverts ou d'autres

2. On appelle *blockchain* (littéralement « chaîne de blocs ») une technique cryptographique établissant la confiance collective sans recourir à un garant, par exemple à un État. Illustration la plus populaire, le Bitcoin, monnaie fondée sur la *blockchain*, n'a pas besoin d'un hôtel des monnaies ou d'une banque centrale ou encore d'un prince pour garantir sa valeur.

informations personnelles comme notre salaire, notre âge, etc. Enfin, la communauté, au sens contemporain, n'est plus cette communauté de destin que des personnes dissemblables (jeunes et vieux, pauvres et riches, malades et bien-portants, etc.) étaient condamnées par leur proximité de vie à partager, et qui de ce fait les amenaient à construire des réseaux de solidarité. Désormais, les communautés regroupent des personnes du monde entier qui se retrouvent sur la Toile parce qu'elles ont une passion et un intérêt communs.

Les relations entre les hommes s'étant totalement modifiées, on ne peut se reposer ni sur les traditions ni sur les habitudes communes pour fonder une morale. Les religions ne sont pas non plus d'un grand secours, car cette morale doit valoir, si ce n'est à l'échelle planétaire, tout du moins dans des pays où, du fait à la fois de la décléricalisation et des migrations, des personnes d'obédiences très diverses se croisent. Il devient alors essentiel d'amorcer une réflexion éthique qui aide à fonder des règles de vie commune dans la société numérique et, par là, une morale. Des groupes de réflexion ont été mis en place par des États, des entités supra-étatiques comme la Commission européenne, des universités, par exemple l'université de Montréal, des organisations internationales comme l'Unesco, des sociétés savantes comme l'Institute of Electrical and Electronics Engineers, ou encore des compagnies privées. Ils ont rédigé des rapports et fait des recommandations fondées sur des principes louables, comme la défense de la dignité humaine, de l'équité ou de l'autonomie de la personne.

Sans entrer dans le détail de ces travaux qui préconisent un certain nombre de prescriptions tout à fait légitimes, quoique souvent contradictoires et difficiles à mettre en œuvre, il convient ici de nous demander quels en sont les fondements nécessaires.

En premier lieu, rappelons que, pour qu'une éthique soit possible, autrement dit pour qu'il y ait une réflexion sur l'agir individuel, il faut que les hommes à qui l'on s'adresse soient en mesure de décider par eux-mêmes et rationnellement. Il faut aussi qu'ils puissent agir librement. Ce sont les conditions implicites de toute éthique. Or les dispositifs d'intelligence artificielle qui apprennent sur de nombreux exemples peuvent être bien plus efficients que nous-mêmes. Il se peut, dès lors, que la pression sociale exige des hommes qu'ils se conforment aux prescriptions de ces dispositifs et renoncent de ce fait à leur liberté de jugement et, par là, à prendre leurs responsabilités. Imaginons, par exemple, qu'un outil entraîné par apprentissage automatique sur de très grandes masses d'exemples fasse statistiquement des diagnostics médicaux considérés meilleurs que ceux

des médecins usuels. Les assurances sociales risquent alors de menacer de sanctions des médecins qui proposeraient un diagnostic différent de celui de la machine. Il en résulterait une absence de liberté qui ferait que, pour se couvrir vis-à-vis des assurances, ces derniers ne prendraient plus en considération les personnes malades qui se présentent à eux, mais uniquement des indicateurs statistiques fournis par la machine. C'est là un risque qu'il faut éviter à tout prix. Pour cela, il importe de considérer les logiciels construits avec des techniques d'intelligence artificielle comme des partenaires capables d'éclairer les décisions, et non comme des automates décidant d'eux-mêmes. Cela signifie que l'action doit toujours relever de la capacité d'initiative de femmes et d'hommes sans que ceux-ci puissent éluder leurs responsabilités en arguant des réponses de machines. Or cela n'est possible que si ces dernières assortissent leurs conclusions de justifications claires. De ce point de vue, contenir l'IA exige que les résultats soient argumentés en termes compréhensibles, afin d'être discutés et non imposés.

Bref, contenir l'IA ne passe ni par un moratoire sur les recherches, ni par des déclarations prophétiques sur d'illusoires prises de pouvoir des machines, mais par une réflexion éthique sur la pertinence des applications de l'IA, des conclusions auxquelles elles parviennent et de l'utilisation qui en est faite.

RÉSUMÉ

Le terme « intelligence artificielle » renvoie au moins à trois choses bien distinctes : une discipline scientifique, les applications de cette discipline, et les craintes de voir une machine prendre son essor de façon indépendante de nous. Nous montrons ici que seules les applications nécessitent d'être contenues et que cela passe par une réflexion et une éthique permanentes.