L'IA hier et demain

Audition de Jean-Gabriel Ganascia

- 18 janvier 2021 -

Jean-Gabriel Ganascia est professeur d'Informatique à Sorbonne Université, spécialiste d'intelligence artificielle et de sciences cognitives. Il a créé et dirigé en 1993 le Programme de recherches coordonnées « Sciences Cognitives » pour le compte du ministère de la recherche, puis le Groupement d'intérêt scientifique « Sciences de la cognition » (Pour le ministère de la recherche, le CNRS, le CEA, l'INRIA, et l'INRETS) (de 1995-2000). Depuis 2016, il est président du Comité d'éthique du CNRS, mais également membre depuis 2012 de la CERNA (Commission de réflexion sur l'éthique de la recherche en sciences et technologies du numérique d'Allistene). Il est aussi membre du tout jeune Comité pilote de l'éthique du numérique, créé fin 2019 par le Comité consultatif national d'éthique.

Il a notamment publié les ouvrages suivants :

Ganascia, J. G. (2022). *Servitudes virtuelles*. Éditions du Seuil.

Ganascia, J. G. (2017). *Le Mythe de la Singularité. Faut-il craindre l'intelligence artificielle ?* Le Seuil.

Ganascia, J. G. (2017). *Intelligence artificielle : vers une domination programmée ?* Le Cavalier Bleu Editions.

Nevejans, N., Hauser, J., Ganascia, J. G., & Édition, L. E. H. (2017). *Traité de Droit et d'éthique de la robotique civile*. LEH édition.

Braly, J. P., & Ganascia, J. G. (2017). *Le temps des robots est-il venu ? Découvrez comment ils transforment déjà notre quotidien.* Editions Quae.

Ganascia, J. G. (2006). *Les sciences cognitives.* Le pommier.

Ganascia, J. G. (1993). *L'intelligence artificielle.* Flammarion (Collection Dominos).

L'audition a été menée par Mehdi Khamassi :

Mehdi Khamassi est directeur de recherche en sciences cognitives au CNRS, rattaché à l'Institut des systèmes intelligents et de robotique de Sorbonne Université, Paris. Il est également co-directeur des études pour le master de sciences cognitives de l'École normale supérieure / EHESS / Université de Paris. Après un diplôme d'ingénieur à l'ENSIIE (anciennement sous la tutelle du CNAM) à Evry, il a effectué un DEA de sciences cognitives (Cogmaster) puis une thèse entre l'Université Pierre et Marie Curie et le Collège de France dirigée par Agnès Guillot et Sidney I. Wiener sur l'apprentissage en situation de navigation chez les animaux et les robots.

I. Historique de l'IA

Mehdi Khamassi

Bonjour Jean-Gabriel Ganascia. Tu es professeur d'informatique à Sorbonne Université, spécialiste d'intelligence artificielle et de sciences cognitives. Tu as dirigé différents programmes de sciences cognitives en France. De façon assez importante pour les questions que nous nous posons dans TESaCo, tu es très impliqué sur beaucoup de questions éthiques depuis quelques années. Tu es notamment, depuis 2016, Président du Comité d'éthique du CNRS, et d'autres comités que je ne vais pas tous citer. Mais je suis également très intéressé par le fait que tu es aussi membre du tout jeune Comité pilote du numérique qui a été créé fin 2019 par le Comité consultatif national d'éthique. Ce n'est un secret pour personne, les questions d'éthique et de sciences en société t'importent beaucoup, et c'est à ce titre qu'il est important pour nous d'avoir ton regard et ta vision des choses, et qu'on en discute dans le cadre des technologies émergentes et en particulier de l'intelligence artificielle.

Historique et évolution de l'intelligence artificielle

Mehdi Khamassi

Je voulais te demander d'abord de resituer les choses sur les débuts et l'évolution de l'intelligence artificielle, tels que tu les vois. Je voudrais notamment partir d'un passage d'un de tes ouvrages récents, *Le mythe de la singularité*, où tu rappelles qu'en 1955 la naissance de l'utilisation du terme "Intelligence artificielle" par John McCarthy, Marvin Minsky, Nathaniel Rochester et Claude Shannon, vient d'un projet pour un workshop à Dartmouth College qui a lieu l'année suivante, dont l'idée est de décomposer chaque aspect de l'apprentissage ou de tout autre caractéristique de l'intelligence en petits modules qu'on pourrait décrire si précisément qu'une machine pourrait être programmée pour les simuler. Pour moi, ce qui est frappant là-dedans c'est : d'une part le fait que ce n'est pas forcément le but de faire une intelligence artificielle générale, globale, il s'agit plutôt de s'intéresser vraiment à des fonctions cognitives distinctes ; d'autre part, je trouve que l'objectif de décrire et de simuler pour essayer de mieux comprendre des fonctions cognitives est très proche d'objectifs qu'on peut voir en psychologie cognitive ou en neurosciences.

Comment vois-tu cette évolution, c'est-à-dire cette éventuelle séparation qui a eu lieu, avec aujourd'hui l'apprentissage machine qui cherche rarement la bio-inspiration, mais qui est plutôt intéressé par l'efficacité statistique pour extraire de la connaissance de masse de données ? Comment vois-tu éventuellement à nouveau des échanges aujourd'hui ?

Jean-Gabriel Ganascia

D'abord bonjour et merci de m'avoir invité. Effectivement, quand on regarde le début de l'intelligence artificielle et le projet qui a été soumis par John McCarthy, Marvin Minsky, Nathaniel Rochester et Claude Shannon à la Rockefeller Foundation afin d'obtenir un financement pour une école d'été, l'ambition est très proche de celle des sciences cognitives. Il n'y a pas de différences

fondamentales. Cela puise aussi son inspiration dans les travaux qui avaient été conduits quelques années auparavant dans le cadre de la cybernétique, même si les orientations ne sont pas tout à fait les mêmes. Nous avons dans le projet, c'est dit explicitement, l'objectif de fonder une science nouvelle qui permettrait de mieux comprendre l'intelligence en procédant à des simulations avec ces outils assez étonnants pour l'époque qu'étaient les ordinateurs. Il faut bien comprendre que Marvin Minsky et surtout John McCarthy étaient des mathématiciens. Ils avaient découvert l'informatique au cours de leurs études et ils pensaient qu'elle ouvrait des pistes nouvelles. Il en allait de même pour tous les autres participants de l'école d'été de Dartmouth College, qu'il s'agisse d'Herbert Simon, qui deviendra prix Nobel d'économie, d'Allen Newell, de Ray Solomonoff, de John Holland, d'Oliver Selfridge ou a fortiori de Nathaniel Rochester, directeur scientifique d'IBM et de Claude Shannon l'homme de la théorie de l'information. Ils étaient fascinés par les ordinateurs, qui ouvraient un champ de potentialités extrêmement grand.

En même temps, quand on lit bien les textes, en tout cas ceux qui apparaîtront après, on voit que très tôt les précurseurs de l'intelligence artificielle ont eu le sentiment qu'il y avait deux perspectives qui n'étaient pas nécessairement antagonistes, mais complémentaires. L'une était de mieux comprendre la nature de l'intelligence, en procédant à des simulations de différentes fonctions cognitives, puis à confrontations empiriques avec les facultés cognitives humaines, de façon à valider ces modèles informatiques. Il faut bien comprendre que le concept d'intelligence est un très vieux concept en philosophie, et qu'il a été utilisé il y a longtemps, par exemple, par Giambattista Vico au début du XVIIIe siècle, avant d'être quasiment abandonné par les philosophes, qui ne l'emploient presque plus. Désormais, ils parlent plutôt d'esprit et de théorie de l'esprit que d'intelligence. Ceux qui, parmi eux, mentionnèrent encore le terme « intelligence », comme Hyppolite Taine, auteur d'un traité paru en 1870 et intitulé « De l'intelligence », le firent parce qu'ils étudiaient les problématiques philosophiques, en particulier les théories de la connaissance, de façon empirique, avec des méthodes expérimentales issues des sciences de la nature, en particulier de la physique, et donc de façon très interdisciplinaire. Il s'ensuit qu'à partir du XIXe siècle, le concept d'intelligence s'employa surtout en psychologie, entendu comme la discipline scientifique qui aborde des problématiques philosophiques avec les méthodes des sciences de la nature. C'est là une première perspective sur l'intelligence artificielle. Mais, il en existe une autre qui apparaît très tôt et qui se révèle être pratique. Dès ses premiers articles sur l'intelligence artificielle, Herbert Simon explique qu'outre la perspective scientifique susmentionnée, l'intelligence artificielle contribue à la résolution de problèmes, ce qui serait susceptible d'avoir de nombreuses applications pratiques fort utiles dans bien des secteurs. Avec le temps, les aspects plus fondamentaux, scientifiques, ont été repris entre autres par les sciences cognitives, qui viennent, il est important de le rappeler, de l'intelligence artificielle. Plus exactement, les sciences cognitives trouvent leur origine avec la cybernétique, avec l'idée qu'existe un parallèle entre des systèmes de traitement de l'information et la cognition. Mais le terme « sciences cognitives » a été introduit par l'intelligence artificielle sémantique, au milieu des années 1970, et a ensuite été repris et généralisé dans le milieu des années quatre-vingts pour englober des études sur le cerveau. Au départ, il y avait assez peu de neurosciences dans les tout premiers travaux de sciences cognitives ; celles-ci se centraient sur la linguistique, la psychologie, l'intelligence artificielle et l'informatique. Les choses sont donc un peu surprenantes parce qu'avec le temps, cette perspective scientifique semble être un peu sortie du champ de l'intelligence artificielle, surtout en France, mais peut-être un peu moins aux États-Unis. Sans doute car aux États-Unis, les liens entre l'intelligence artificielle et les sciences cognitives ont plus subsisté ; pour des raisons pratiques aussi, du fait d'un engouement pour l'intelligence artificielle, et de l'existence de sources de financements considérables en l'intelligence artificielle. Mais ce n'était pas le cas au tout début des sciences cognitives. Il est toujours très

intéressant d'avoir une vue généalogique des disciplines car on comprend mieux leurs racines et, par conséquent, leur assise épistémologique.

Place centrale de l'apprentissage de l'IA

Mehdi Khamassi

Oui, tout à fait. Comme tu l'as mentionné, très vite une des perspectives a été de s'intéresser à la résolution de problèmes – d'ailleurs considérée comme jusqu'alors réservé à l'homme, – et aux questions de raisonnement. Mais dès le départ, on a l'impression qu'il y avait une place centrale pour le rôle de l'apprentissage, au point même qu'aujourd'hui on a l'impression que dans la tête de nombreuses personnes, intelligence artificielle et apprentissage machine se confondent un peu. Qu'en penses-tu ? Au fond, c'est peut-être important et peut-être partages-tu cette idée selon laquelle, de toute façon on ne pourra atteindre d'intelligence artificielle que lorsque des systèmes seront capables d'apprendre par eux-mêmes. Mais est-ce que cela va plus loin ? Ce rôle d'apprentissage est-il à ce point central pour toi ?

Jean-Gabriel Ganascia

Ça a été central depuis le début de l'intelligence artificielle. Même avant. Les travaux de Turing dans son célèbre article « Computing Machinery and Intelligence », dans la revue *Mind*, évoquent déjà l'apprentissage car il pressent qu'il est essentiel, si on veut fabriquer des machines qui semblent penser, qu'elles se fondent sur une immense quantité de connaissances. Quand il conçoit son fameux "Test de Turing", qui est plus une expérience de pensée qu'un projet de réalisation matérielle, il n'avait pas l'idée qu'on irait, un jour, jusqu'à fabriquer une machine qui le mettrait en œuvre. Plus exactement, ce n'était pas son objectif premier dans l'article. Il dit que pour arriver à conduire un dialogue avec un homme et/ou une femme, et prêter au change, c'est-à-dire faire illusion, il faut qu'une machine dispose d'un nombre considérable de connaissances. À cette fin, on peut soit transférer les données directement dans la machine, ce qui est extrêmement long et fastidieux, soit essayer de faire en sorte que la machine les acquiert d'elle-même ; c'est ce qui fait l'objet de l'apprentissage. Il donne d'ailleurs plusieurs métaphores de ce que pourrait être l'apprentissage machine. Il imagine que ça pourrait ressembler à la façon dont un individu apprend ; on dresse là une analogie entre la machine et de psychisme humain ; il imagine aussi que cela se présente comme avec un cerveau, ou comme des populations dont le patrimoine génétique évolue, ou encore, en usant d'un parallèle avec la culture collective, comme les idées qui naissent dans l'intérieur des sociétés. Tout cela a été pensé avant que le terme « intelligence artificielle » n'ait été inventé. Dans le texte introductif de l'école de Dartmouth College, il est dit, comme tu l'as rappelé tout à l'heure, que l'intelligence artificielle aspire à modéliser toutes les fonctions cognitives, qu'elles relèvent de l'apprentissage, voire même des apprentissages, tant l'apprentissage est multiple, ou de toutes les autres caractéristiques de l'intelligence, raisonnement, mémoire, décision, réflexion, etc. Ici, l'apprentissage est même mis en premier, avant les autres aspects de l'intelligence. Les organisateurs de l'école d'été de Dartmouth College savaient qu'il y avait un soucis. En effet, Marvin Minsky était lui-même cybernéticien. Il avait travaillé, au cours de sa thèse sur des réseaux de neurones formels et avait essayé de faire de l'apprentissage machine, mais en vain. Il s'était heurté à de très grandes difficultés qu'il n'était pas parvenu à surmonter. Il savait donc à quel point c'était problématique. Quand il a écrit avec Seymour

Papert en 1969 l'ouvrage intitulé Perceptrons, pour faire une critique de l'algorithme d'apprentissage du perceptron de Rosenblatt, il avait lui-même une très bonne connaissance de l'apprentissage dans les réseaux de neurones formels.

Indépendamment de l'importance des mécanismes d'apprentissage, il y a l'idée assez naturelle que, si on veut être capable de faire des choses qui soient vraiment nouvelles, il ne faut pas simplement déployer un enchaînement mécanique de raisons, mais construire automatiquement de nouvelles connaissances. Les premiers travaux d'Arthur Samuel sur l'apprentissage par renforcement datent d'ailleurs de 1959 : il était arrivé, avec les techniques qu'il avait conçues, à ce que le programme de jeux de dames qu'il avait réalisé s'améliore automatiquement, jusqu'à battre l'un des meilleurs joueurs aux États-Unis. Donc dès le départ l'apprentissage accompagne l'intelligence artificielle. Toutefois, même si l'apprentissage y prend une part importance, ne recouvre pas l'intégralité du champ de l'intelligence artificielle. Ainsi, dès 1956, Herbert Simon simule sur des ordinateurs la résolution de problèmes en transférant, manuellement, des connaissances dans les machines.

Aujourd'hui ce qui est vraiment particulier, c'est qu'on a tendance à réduire l'intégralité de l'intelligence artificielle à l'apprentissage, et surtout à une seule forme d'apprentissage, l'apprentissage supervisé, et parmi les techniques d'apprentissage supervisé, à une seule technique, l'apprentissage profond, alors qu'au début il était question de tous les apprentissages possibles, c'est-à-dire de toutes les capacités qu'aurait une machine de construire des connaissances en s'inspirant de différents modèles. On observe donc une réduction de l'empan de l'intelligence artificielle aujourd'hui et, en quelque sorte, à une régression. D'autant plus, et bizarrement, que les concepts présents aujourd'hui apparaissent relativement anciens, même s'ils ont subi des évolutions, même s'il y a un grand nombre d'astuces et de tours de mains déployés pour mettre en œuvre ces technologies, enfin, même si les progrès des ordinateurs font qu'on est capable de mettre en œuvre des choses qui auraient été impossibles auparavant.

En résumé, on observe toujours une prééminence de l'apprentissage aujourd'hui en intelligence artificielle. Moi-même j'ai commencé à travailler sur des systèmes à base de connaissances. C'était les premiers systèmes experts, et assez vite, dès que j'ai eu ma thèse de Docteur ingénieur, j'ai poursuivi sur l'apprentissage car j'ai eu l'intuition que c'était central. Il s'agissait d'apprentissage symbolique, donc pas tout à fait d'apprentissage avec les réseaux de neurones formels, mais tout de même je m'y intéressais, je regardais ce qui se faisait. À l'époque on m'a appris que les algorithmes employés pour l'apprentissage neuronal étaient instables. Il a fallu attendre 1986, avec les algorithmes de rétro-propagation de gradient qui généralisaient les algorithmes d'apprentissage du perceptron, pour avoir des résultats tangibles. Mais ceux-ci demeuraient encore très ténus.

Jusqu'où va l'opposition entre les approches symboliques et l'apprentissage statistique ?

Mehdi Khamassi

Alors tu as anticipé ma question suivante. Comme tu es spécialiste de l'apprentissage machine, il me semble qu'au début de tes travaux, comme tu l'as dit, tu étais plutôt sur des approches symboliques. Ce qui me frappe c'est qu'il semble y avoir très vite une opposition entre les apprentissages

dans les réseaux de neurones distribués, avec quand même des tentatives de formalisation de choses qui viennent des statistiques, mais comme s'il y avait une opposition avec l'apprentissage symbolique. Aujourd'hui, on a presque l'impression avec le deep learning ou autre qu'on est purement dans l'efficacité de l'apprentissage distribué sans chercher à faire un retour vers la construction de symboles qui vont permettre le raisonnement. Je me demande donc dans quelle mesure une hybridation, où à un moment donné de faire communiquer ces différents niveaux, est nécessaire pour une cognition satisfaisante.

Jean-Gabriel Ganascia

C'est toujours apparu nécessaire. Il y a une expression courante, reprise parfois par des philosophes peu soucieux de philologie et d'histoire des sciences, pour décrire les approches symboliques de l'intelligence artificielle, GOFAI "Good Old Fashioned Artificial Intelligence". Cette expression laisse entendre qu'au départ l'intelligence artificielle était symbolique et puis qu'ensuite elle est devenue numérique. C'est une contre-vérité. Il suffit de lire les papiers des pionniers, par exemple le projet d'école d'été de Dartmouth College tel qu'il a été soumis par John MacCarthy et Marvin Minsky à la Rockefeller Foundation : la référence aux réseaux de neurones y est explicite. Il y a donc eu, dès le départ de l'intelligence artificielle, une composante numérique importante. De plus, à l'époque, les chercheurs étaient d'abord des numériciens, avec une formation d'ingénieurs spécialisés dans le traitement du signal, etc. Moi-même, quand j'ai commencé, je connaissais très bien le traitement du signal. J'ai découvert la logique après, en faisant de l'intelligence artificielle. Il y eu, avec le temps, un abandon des approches numériques en l'intelligence artificielle, mais très vite on s'est rendu compte qu'il fallait coupler les deux.

Rappelons d'abord que, dans son livre publié en 1969 et intitulé *Perceptrons*, Marvin Minsky explique que les techniques d'apprentissage sur des perceptrons multicouches se heurtent à des obstacles quasiment insurmontables. Soit on a affaire à un perceptron à deux couches, auquel cas l'apprentissage se limite à des classes de fonctions très pauvres, soit on a affaire un perceptron à trois couches et plus, et la complexité algorithmique des procédures d'apprentissage devient rédhibitoire. On sait depuis le tout premier article princeps de Walter Pitts sur les réseaux de neurones, qu'à partir de trois couches, un réseau de neurones est universel, il implémente n'importe quelle fonction logique. Au départ, cette propriété se limitait à des fonctions booléennes, c'est-à-dire, à des fonctions logiques qui donnent des valeurs 1 pour vrai ou 0 pour faux ; plus tard, on a étendu ce résultat d'universalité aux fonctions continues. Or, l'algorithme d'apprentissage du perceptron de Frank Rosenblatt était limité à des perceptrons à deux couches ; son extension à des perceptrons à plus de deux couches paraissait à Marvin Minsky très difficile, voire impossible. Dans le même temps, Marvin Minsky expliquait que, pour résoudre ce problème, il faudrait ajouter des traits sémantiques dans la représentation. L'idée de couplage était donc déjà présente.

Je me souviens moi-même que, quand j'ai travaillé avec Yves Kodratoff sur l'apprentissage symbolique dès le début des années quatre-vingts, nous avons eu très vite des échanges avec des spécialistes de l'analyse de données en particulier avec des chercheurs qui venaient de l'école du mathématicien Jean-Paul Benzécri, fondateur de méthodes d'analyse de données comme l'analyse factorielle par correspondance. Moi-même, je connaissais bien Edwin Diday qui a été dans mon jury de thèse d'État. Il y avait toute une école qui travaillait sur l'hybridation des méthodes et on avait organisé des journées "symbolique / numérique". Encore une fois, l'idée du couplage des deux est présente très tôt. Dès 1983, nous avons organisé des journées communes avec Françoise Fogelman où l'on

comparait les différentes techniques d'apprentissage, dont les techniques d'apprentissage neuronal. Moi-même, plus tard, j'ai eu un projet dans les années 1990 sur l'extraction de règles à partir de réseaux de neurones formels avec une étudiante, Irina Tchoumatchenko et un collègue, spécialiste des réseaux de neurones, Maurice Milgram. Notre approche consistait à « durcir » progressivement les poids synaptiques des réseaux de neurones artificiels au cours de l'apprentissage pour être en mesure d'interpréter les résultats à l'issue de l'apprentissage. Il existait d'autres travaux semblables aux États-Unis, par exemple ceux de Jude Shavlik. Dès le début, cette jonction du symbolique et du numérique est apparue comme quelque chose d'extrêmement important. Aujourd'hui, c'est encore plus important d'une certaine façon, parce qu'on a beaucoup développé les méthodes d'apprentissage numérique avec les méthodes des réseaux de neurones formels, en oubliant d'ailleurs un peu toutes les autres méthodes qui avaient été développées avec les SVM, les machines à noyau, toute les théories formelles de l'apprentissage etc. Aujourd'hui, on dispose d'algorithmes d'apprentissage supervisé remarquablement efficaces. Mais, les classifieurs résultant de cet apprentissage s'imposent comme des espèces d'oracles aveugles, c'est-à-dire comme des boîtes noires dont on ne sait pas expliquer pourquoi elles produisent telle ou telle sortie. Dans certaines situations, cela ne pose pas trop de problème. S'il s'agit, par exemple, d'examiner des grains de beauté et de faire un diagnostic médical pour savoir s'ils sont potentiellement malins ou bénins, cela ne pose pas trop de problème parce que de toute façon ça va aider le médecin à faire un premier diagnostic qu'il confirmera ensuite avec d'autres moyens. Pour d'autres applications c'est plus problématique parce qu'on a absolument besoin de justifier la réponse. Il ne faut pas se contenter de dire « oui » ou « non ». On doit fournir ce qu'on appelle des explications en indiquant, sur le cas particulier en cours d'études, les éléments d'informations qui ont conduit à telle ou telle conclusion. Si par exemple, on a un système informatique qui aide la banque à décider d'un prêt bancaire et que celui-ci dit à un client "Non, on vous refuse votre prêt. » il faut être capable de lui en indiquer les raisons, par exemple parce qu'il est trop âgé, qu'il ne gagne pas assez, qu'il dépense trop etc. Si le système répond juste "Non, parce que votre score est à 0,75 et qu'il faut 0,80 pour vous donner le prêt." ce n'est pas du tout acceptable. Bien sûr, ces questions ne sont pas neuves… Mais, cela ne veut pas dire que nous sommes capables de les résoudre.

Symbolisme, explications et argumentation

Mehdi Khamassi

C'est très intéressant que tu parles de la notion d'explication car au fond quand on pense à nous humains, on est, la plupart du temps ou du moins de temps en temps, capable de dire "J'ai abouti à telle solution parce que j'ai raisonné comme ceci, comme cela". On est capable d'ailleurs d'échanger et de partager pour ensuite corriger le raisonnement de l'autre et d'avancer dans ce sens. Pour l'instant, il me semble que les machines développées et les programmes ne sont pas encore capables de donner ces explications elles-mêmes. Pour pouvoir donner des explications, il faut peut-être quelque part extraire, simplifier des choses sur le processus et le rendre un peu symbolique, il me semble.

Jean-Gabriel Ganascia

C'est exactement ce qu'on essaye de faire aujourd'hui, car on se rend compte qu'autrement on sera conduit à une impasse. Il y a aura peut-être certaines applications en reconnaissance d'image qui

fonctionneront, mais tout un tas d'autres applications qui sont extrêmement bénéfiques en pratique, ne seront pas acceptées socialement. Donc cela est tout à fait essentiel. Concernant le deuxième point important que tu évoques, il y a non seulement les explications mais aussi l'argumentation. Il faut être capable, lorsqu'une machine nous donne un certain nombre de réponses, de mettre en cause certains éléments, de dire "voilà tel argument est compréhensible, mais il semble qu'il y ait un contre-argument à cet argument". Des travaux commencent à se déployer aujourd'hui pour mettre en évidence ces phénomènes d'argumentation collective. Ça peut être une argumentation entre des hommes ou entre des hommes et des machines. Et, bien sûr, lorsque cette argumentation se déploie entre des hommes et des machines, les arguments de la machine peuvent être construits à partir des résultats de l'apprentissage. Cela ouvre bien sûr beaucoup de possibilités. Ce qui est important c'est que l'ordinateur ne deviennent pas un oracle dont on se sente forcé d'accepter les sentences, sans rien dire. C'est la raison pour laquelle il faut instaurer un dialogue entre des hommes et des machines d'un côté, et entre des hommes par l'intermédiaire des machines.

II. La notion d'IA forte a-t-elle un sens ?

Mehdi Khamassi

Alors cela me donne envie qu'on discute un petit peu de la notion d'IA forte. Comme tu l'as souligné plusieurs fois dans tes ouvrages et notamment dans Le mythe de la singularité, une partie de ceux qui annoncent l'avènement de l'IA forte, enrobent ça de plein de choses, comme l'idée d'une singularité dans l'évolution des sociétés humaines. Il y a beaucoup d'arguments farfelus, alarmistes mais on va y revenir. Mais au fond, à vouloir dire [comme tu le fais], et c'est important, "Attention il faut décomposer ces raisonnements-là et les ramener à des choses réalistes basées sur un raisonnement scientifique", parfois tu dis à du "catastrophisme éclairé" comme le propose Jean-Pierre Dupuy, il arrive qu'on se demande si tu es sceptique sur l'avènement d'une IA forte qui dépasserait l'homme ou alors si tu considères que c'est éloigné dans le temps.

Jean-Gabriel Ganascia

Moi je trouve que le concept même d'IA forte est très problématique en lui-même parce qu'il repose sur un article de foi, tandis que l'intelligence artificielle depuis qu'elle se développe, est d'abord une discipline empirique, de modélisation et de confrontation [des performances des machines et des hommes, ou des machines entre elles]. On est capable de mieux en mieux comprendre certains aspects de l'intelligence, certaines fonctions cognitives en les simulant et en mesurant l'écart entre la simulation et la réalité. Ça peut être la cognition humaine ou la cognition animale. C'est quelque chose que je trouve extraordinaire. C'est d'ailleurs pour cela que j'ai commencé à faire de l'IA, parce que je me demandais ce qu'était l'intelligence. Quand j'étais enfant, on faisait passer [à tous les écoliers] des tests de QI qui demeuraient très mystérieux pour moi. J'ai essayé de comprendre quelle en était la signification et les limites. C'est quelque chose de passionnant. […]

L'IA forte c'est autre chose. C'est l'idée qu'à un moment donné, la machine va s'emballer et prendre le pouvoir sur nous. [Cela m'apparaît invraisemblable. C'est un article de foi, pas une hypothèse scientifique.] Paradoxalement, la notion d'IA forte a été introduite par John Searle pour criti-

quer un certain nombre de philosophes, pas du tout pour…

Mehdi Khamassi

.. pour annoncer l'avènement.

Jean-Gabriel Ganascia

Oui c'est ça. En réalité, il a essayé de radicaliser la position de ses collègues [, les philosophes cognitivistes, avec qui il était en désaccord]. Il l'a fait au milieu des années 1980-85, en introduction à son expérience de pensée sur la chambre chinoise. Là, il explique qu'il n'a rien contre l'intelligence artificielle, que ça le fascine, qu'il est certain que cette discipline va faire des progrès considérables et que ce n'est pas la puissance empirique des machines qu'il met en cause. Il va même jusqu'à affirmer qu'on sera peut-être capable, un jour, de fabriquer une machine qui simulera complètement le vivant. Là n'est pas le problème. Ce qui lui paraissait extrêmement douteux, c'est l'idée selon laquelle quelques manipulations de symboles suffisent à reproduire des phénomènes sémantiques. Selon lui, il faudrait, si on voulait simuler le vivant, qu'il y ait un degré de finesse qui soit analogue à celui de la chimie. La grossièreté des moyens mis en œuvre par l'intelligence artificielle permettra, certainement, de rendre des services pratiques, mais on n'arrivera jamais à accéder aux phénomènes producteurs du sens, à la sémantique. Pour caractériser cela, il l'appelle « IA forte » une IA qui accèderait au sens et « IA faible » une IA qui se contente de procéder à des manipulations de symboles et de faire illusion. Curieusement, la notion d'« IA forte » a été reprise quelques années plus tard par des spécialistes d'intelligence artificielle, en particulier par un roboticien qui s'appelle Hans Moravec, et par d'autres qui ont tous affirmé qu'ils étaient en train de réaliser l'IA forte. En fait, ce qu'ils voulaient dire à propos des méthodes symboliques que critiquait John Searle, c'est que l'intelligence artificielle d'aujourd'hui — nous étions au milieu des années quatre-vingts — sera capable d'aller au-delà puisqu'elle ne se limite pas aux méthodes symboliques. C'est pour ça qu'on a inventé l'idée de GOFAI. Le problème c'est que cette critique faite au milieu des années 1980 pour critiquer l'intelligence artificielle du début des années 1980 et de la fin des années 1970 n'a plus de pertinence aujourd'hui, 35 ans plus tard, car cette nouvelle approche de l'intelligence artificielle, qu'on a appelé « la nouvelle intelligence artificielle », est presque aussi vieille que l'autre. En plus de ça, elle passe sous silence le fait qu'il y avait déjà du numérique auparavant, au tout début de l'IA et, surtout avant, avec la cybernétique. Ce qui est très curieux donc, c'est que ce sont des arguments contextuels qui auraient dû disparaître avec le temps. À l'époque, pour donner plus de poids à ses arguments, Hans Moravec prophétise qu'on va arriver à l'IA forte, celle dont John Searle disait qu'elle n'était pas accessible à partir des méthodes symboliques classiques. Et il explique que l'approche sub-symbolique qu'il déploie y arrivera [, puisqu'elle s'affranchira des limitations de l'approche symbolique]. C'était quand même d'une très grande naïveté parce que, bien sûr, ça n'atteignait pas le « niveau de finesse » de la chimie qui seul aurait permis, selon Searle, d'accéder aux phénomènes sémantiques. Et, bien évidemment, ce n'est toujours pas le cas aujourd'hui, et l'on en est toujours extrêmement éloigné. Il ne s'agit pas ici de critiquer des méthodes qui se révèlent très puissantes, mais simplement de dénoncer des prétentions excessives. Dit autrement, il y a une démarche scientifique rationnelle, reposant sur des preuves empiriques et/ou mathématiques — c'est la science telle qu'elle existe — et puis il y a des « croyants » qui affirment avec une forme de dévotion que l'intelligence artificielle adviendra bientôt. Et, ici, on sort du champ de la rationalité ; cela relève de la foi, puisqu'on nous explique que ça va arriver, parce que ça doit arriver. [C'est de cette idée dont il faut se départir.]

Une dernière chose. Que les machines soient plus performantes que les hommes dans l'exécution de certaines tâches, c'est évidemment vrai depuis longtemps, en particulier depuis le début des ordinateurs. Ceux-ci font bien mieux les multiplications que nous ; et, il en allait ainsi avant même les premiers ordinateurs, avec les machines à calculer qu'Alan Turing a fabriqué pendant la seconde guerre mondiale. Donc, que les machines nous dépassent, on le sait depuis très longtemps. La question c'est de savoir si elles vont — et c'est cela qu'on craint avec l'IA forte — prendre le pouvoir pour nous asservir, au sens où on n'aura plus de capacités à nous autodéterminer. Je pense qu'il n'y a aucun argument sérieux derrière cette idée.

Quand les succès de l'IA marquent son éloignement des modèles de la cognition humaine, est-ce un échec ?

Mehdi Khamassi

Justement sur cette construction de symboles et notamment pour raisonner ou pour ensuite faire du langage, même s'il me semble que les méthodes ont beaucoup évolué, avec des méthodes numériques et des réseaux profonds ou autres, ce qui est fascinant c'est qu'encore aujourd'hui il y a toujours une négligence de certains processus comme l'importance du corps dans la construction d'un symbole, qui me semble pourtant quelque chose de central en sciences cognitives. Certes on a des algorithmes de plus en plus efficaces, par exemple, pour faire de la traduction automatique ou pour générer du texte. Ces textes peuvent paraître surprenants car on aurait presque l'impression qu'ils sont générés par un humain étant donné qu'ils respectent des régularités statistiques observées dans des grandes quantités de données. Mais au fond, ça ne veut pas dire qu'il y a compréhension des termes manipulés. On a l'intuition qu'éventuellement des travaux qui pourraient être faits – sur du très long terme – prenant en compte les informations sensori-motrices, le corps, et donc peut-être plus du côté de la programmation de l'intelligence artificielle sur des robots, pourraient être plus prometteurs. Ne serait-ce que, dans mon intérêt, pour comprendre comment chez l'être humain l'émergence de symboles est ancrée dans le sensori-moteur. Peut-être que de ce point de vue là, on pourrait arriver à des choses intéressantes, à une prochaine étape.

Jean-Gabriel Ganascia

Oui tout à fait. Je crois utile de se pencher de nouveau sur les premiers travaux de Turing. Avant d'écrire son article sur l'intelligence computationnelle dans la revue Mind, il avait commencé par d'autres articles, qui étaient des essais préliminaires sur cette question. Dès le départ, il était bien conscient de la complexité de la notion d'intelligence ; il ne souhaitait pas la réduire ; mais, dans un premier temps, il essaya d'éliminer un certain nombre de dimensions, en particulier tout ce qui était ancré dans le monde, parce qu'il jugeait que c'était trop difficile à simuler. Ce n'est pas parce qu'il pensait que ça ne relevait pas de l'intelligence. Plus tard, beaucoup de gens ont critiqué son approche en disant "C'est trop abstrait". Pourtant, il n'a jamais restreint l'intelligence à des abstractions. Il a simplement préféré, dans un premier temps, s'occuper de choses simples, pour obtenir des premiers résultats. Cela le conduisit à isoler un certain nombre de tâches intellectuelles qui ne requéraient pas de perception et de présence au monde, et qui peuvent donc se faire de façon totalement abstraite. Il en voit plusieurs : il pense aux jeux, notamment au jeu d'échecs, ensuite à la cryptographie et à la

traduction automatique.

Mehdi Khamassi

Ah. Alors là je serais moins d'accord.

Jean-Gabriel Ganascia

Peu importe ce qu'on en pense aujourd'hui, mais c'est révélateur du point de vue de Turing. Ensuite, après avoir envisagé la traduction automatique, il songe au dialogue. Cela va le conduire à la deuxième version du test dit de Turing, qui passe par une simulation de dialogue alors que la première était un jeu d'échecs. Et, comme il se rend compte que, pour conduire un dialogue, il faut disposer d'une immense quantité de connaissances sur le monde, il réfléchit à leur acquisition, ce qui l'amène à réfléchir aux différentes techniques d'apprentissage automatique que devrait concevoir. Pour la traduction, il la voit comme purement statistique. Rappelons que l'histoire de la traduction automatique commence sur ces bases-là en 1952, avant l'intelligence artificielle. Les premiers projets de machines à traduire datent même de 1949 avec Warren Weaver, juste après les théories de l'information de Claude Shannon. Les premières machines à traduire se basent sur des gros dictionnaires. En 1964, le rapport ALPAC [Automatic Language Processing Advisory Committee ; rédigé par un comité de sept scientifiques menés par John R. Pierce met en cause ces efforts de traduction automatique. Des linguistes affirment que c'est vain ; qu'on ne parviendra jamais à traduire de cette manière, sans compréhension, sans connaissances linguistiques, qu'elles soient syntaxiques, sémantiques ou pragmatiques. Dans les quarante ans qui suivirent, on n'a pas arrêté de dire qu'il fallait se référer à différents niveaux linguistiques (lexical, syntaxique, sémantique, pragmatique, etc.) et introduire des connaissances spécialisées relatives à chacun pour faire de la traduction. Je me souviens des cours de Jacques Pitrat sur le langage naturel où il expliquait qu'il ne fallait surtout pas utiliser les anciennes méthodes des années 1950. Et en 2005, tout change. Apparaît alors une nouvelle hallucinante : Google affirme — même si d'autres personnes avaient commencé à le suggérer — qu'il suffit de faire des statistiques. Que les connaissances linguistiques sont inutiles. Et, les résultats le prouvent ! Les performances des systèmes de traduction statistique sont excellentes, bien meilleures que celles des systèmes fondés sur l'utilisation de connaissances linguistiques. Pour le futur on ne sait pas dire. Il y a certainement des subtilités dont on ne parviendra pas à rendre compte si on n'approche pas les aspects syntaxiques et sémantiques, etc. Pour l'instant, ces outils de traduction automatique marchent bien. Mais on ne sait pas pourquoi ça marche, ça reste mystérieux. Il faudrait vraisemblablement faire appel à d'autres domaines du savoir pour les améliorer. Ce qui est difficile avec l'intelligence artificielle, c'est qu'il y a des choses dont on se dit qu'elles devraient se faire comme ça, car elles sont faites comme ça, et pourtant on obtient de meilleurs résultats avec des méthodes qui sont totalement différentes. D'un côté, on peut dire c'est la puissance de l'intelligence artificielle, et d'un autre côté, c'est son échec.

Je vais essayer de m'expliquer sur cet échec en prenant pour illustration justement, et sans jeu de mots, le jeu d'échecs. On a commencé à vouloir faire des machines qui jouent aux échecs très tôt. Par exemple, Torres y Quevedo [réalisa dans les années 1920 un automate qui jouait face à un adversaire humain des finales Roi et Tour contre Roi seul]. Claude Shannon et Alan Turing ont écrit des programmes informatiques du jeu d'échecs, parce que les échecs paraissaient emblématiques de l'intelligence. Le film de Bergman, Le 7ème sceau, met en scène le diable, intelligence abstraite [et désincarnée], qui joue aux échecs. Cela paraissait crucial. On se disputait sur les performances des

machines. Certains prétendaient qu'un jour une machine battrait un homme, un grand maître, tous les hommes… D'autres se refusaient à l'admettre. En 1968, David Levy, grand maître international, paria qu'aucune machine ne le battrait d'ici dix ans. En 1978, il gagna son pari, mais comprit que des machines l'emporteraient bientôt. Ce fût fait en 1989 avec la machine Deep Thought qui le battit. Ensuite, je passe les détails, mais en 1996 une nouvelle machine, Deep Blue, l'emporta sur le meilleur joueur au monde, Garry Kasparov. [Que ce soit Deep Thought ou Deep Blue,] la machine joue avec ce qu'on appelle la « force brute », à savoir à l'exploration d'un grand nombre de possibilités, du moins grosso modo, car il y a tout de même, dans ces programmes des connaissances sur les ouvertures de parties, des catalogues de fins de parties, des stratégies variables selon la phase de jeu, etc. Mais c'est la puissance de calcul qui l'emporte, ce qui veut dire que la machine ne joue pas comme les hommes, même s'il ne faut pas caricaturer, car ces programmes emploient tout de même quelques connaissances humaines, comme nous venons de le voir. Mais, le succès de Deep Blue contre Kasparov a donné un coup d'arrêt aux recherches conduites auparavant par les chercheurs qui essayaient de modéliser la façon dont les joueurs humains jouaient. En particulier, Herbert Simon et Fernand Gobet avaient fait des études formidables où ils avaient utilisé les travaux d'un psychologue, Adriaan de Groot, lui-même joueur d'échec, et qui avait interviewé beaucoup de grands maîtres aux échecs afin de comprendre comment ils se représentaient le jeu, comment ils mémorisaient la partie en cours, comment ils décidaient de jouer, quelles étaient leurs stratégies, etc. Or, tous ces travaux n'intéressaient plus personne, car ce qui les justifiait c'était qu'une machine l'emporte sur le champion du monde au jeu d'échecs. Donc d'un côté, c'est une victoire de l'intelligence artificielle capable de l'emporter sur Kasparov, le meilleur joueur d'échecs au monde, et en même temps c'est un échec, parce que ce n'est pas la modélisation des facultés cognitives humaines qui a permis de l'emporter. On retrouve ça avec le jeu de go et un peu partout. C'est en effet un paradoxe de l'intelligence artificielle. Lorsqu'elle arrive à réaliser des tâches de façon extrêmement efficace avec des méthodes qui ne sont pas les mêmes que celles qu'emploient les hommes, elles ont d'un côté des résultats qui sont pragmatiquement époustouflants, mais d'un autre côté comme ce sont des méthodes différentes des humains, cela joue en sa défaveur. En tout cas pour la traduction automatique c'est ça. Donc, il est vrai que cela permet de faire des progrès si on est capable de poursuivre dans la modélisation. Mais bien sûr, si on ne fait plus de modélisation comme cela se produit aujourd'hui en traduction automatique, d'une certaine façon cela a un côté stérilisant pour les sciences du langage, en tout cas pour l'étude de la traduction, puisqu'on ne va plus s'intéresser à des subtilités liées à la façon dont les hommes traduisent.

Mehdi Khamassi

Et dont la compréhension pourrait peut-être aider à plus long terme l'intelligence artificielle elle-même, qui sera à ce moment-là sur des subtilités.

Jean-Gabriel Ganascia

Tout à fait.

III. Quelle forme d'autonomie peuvent atteindre les systèmes artificiels ?

Mehdi Khamassi

Dans cette veine, on pourrait se poser la question : est-ce qu'une machine, un système artificiel, peut être autonome ? Je trouve que c'est une question importante. Tu le soulignes, notamment dans Le mythe de la singularité, qu'il faut faire une distinction entre d'une part une autonomie technique - par exemple apprendre à une machine de guerre à décider si elle tire ou non - ce qui est juste la base de l'autonomie. D'autre part, l'autonomie philosophique qui est beaucoup plus large et qui consiste à se définir ses propres buts, donc un autre cadre. J'ai l'impression que, même s'il y a peu d'équipes qui sont en train de le faire, qu'il n'y a rien qui me semble empêcher des machines d'être programmées pour avoir un niveau d'autonomie intermédiaire, avoir des fonctions de récompenses, avoir des objectifs de maximiser une quantité d'énergie, d'obtenir de l'information pour construire un modèle ou une représentation de l'espace. En même temps, à partir de ces fonctions de très haut niveau qui pourraient, elles, être fixes, on pourrait leur permettre de définir elles-mêmes des sous-buts et de définir ce qu'elles doivent faire à un moment donné comme mission par rapport à une autre. Cela serait, il me semble, un autre degré d'autonomie. Déjà, es-tu d'accord avec ça ?

Jean-Gabriel Ganascia

Qu'on puisse avoir des machines qui se comportent de façon automatique, ce qu'on appelle "autonome" — par exemple, un véhicule autonome qui soit extrêmement élaboré —, et qui, vues de l'extérieur, nous semble présenter une certaine autonomie ça ne pose pas de problème. Mais que des machines soient autonomes, au sens moral, c'est une autre affaire. Pour bien comprendre, prenons la notion de conscience. Que voudrait dire "être conscient" pour une machine ? Pour qu'on puisse y répondre, il faudrait préciser ce que l'on entend par conscience. Or, il y a plusieurs choses derrière ce terme. Dans les Systèmes Intentionnels (1978), Daniel Dennett, s'intéresse à ce qui se produit de l'extérieur, lorsqu'on attribue des intentions à la machine parce qu'au vu de son comportement elle fait illusion ; tout se passe là comme s'il y avait une entité consciente, un agent rationnel. Ce dernier est un investissement qu'on fait ; on prête à la machine des intentions et des connaissances. Ce prêt est rentable, si l'attribution de buts et de connaissances à cet agent permet d'anticiper le comportement ultérieur de la machine. Cette stratégie est très générale ; elle vaut aussi pour un animal ou n'importe quoi d'autres. Lorsqu'on projette des émotions sur un animal, c'est juste une hypothèse que nous faisons ; nous ne savons rien de ce qu'il éprouve vraiment. Mais ça nous permet d'anticiper assez bien les dispositions dans lesquelles il se trouve et, par-là, son comportement ultérieur. On peut faire la même chose avec les machines. Il n'y a aucun doute sur la capacité qu'on aura, dans le futur, à fabriquer des machines sur lesquelles on pourra faire des projections de ce type-là. Il y a une deuxième acception de la notion de conscience relative à la capacité des machines à s'auto-observer. Quand Stanislas Dehaene nous dit qu'il est capable d'approcher la conscience, c'est de cette dimension dont il parle, à savoir de la faculté de s'observer soi-même. Dans le futur, on sera certainement capable de fabriquer des machines qui le font. D'ailleurs, parmi les techniques d'apprentissage, il y a eu, dans les années quatre-vingts, des chercheurs qui abordaient ces problématiques. C'est ce qu'on appelait l'apprentissage par explication. La machine examinait ses comportements passés et, à partir de ses observations, elle essayait de faire en sorte de ne pas reproduire les mêmes errements, et de

devenir plus efficace. Elle regardait son parcours et se disait “j’ai été dans telle direction mais ça ne m’a pas servi à grand-chose ; la prochaine fois, j’irai plus droit au but avec les questions analogues”. Nous avons là une modélisation de la réflexion. Rien ne s’oppose donc à ce qu’une machine puisse être dotée de cette conscience réflexive. Ça lui donnerait un degré d’autonomie, en tout cas d’automaticité, encore plus fort. Ensuite, bien sûr, demeure la question de l’autonomie : est-ce possible ? C’est une question déjà mystérieuse pour nous. Que veut dire pour nous être autonome ? Sommes-nous vraiment autonomes ? Ça ramène à des questions éthiques anciennes. Dans les comités d’éthique de l’intelligence artificielle, on affirme, en s’inspirant des grands principes de bioéthique, qu’il faut préserver à tout prix l’autonomie de la personne humaine. Or, l’autonomie n’est jamais donnée ; elle se conquiert, par la volonté ; et, il en va de même de la liberté. D’ailleurs, toutes les éthiques ne reposent pas sur l’autonomie de la volonté. Ce sont surtout les éthiques des Lumières, au XVIIIème siècle, et à partir de penseurs comme Rousseau, Kant, etc. qui fondent la loi morale sur l’autonomie de la volonté, c’est-à-dire sur la capacité à se donner, soi-même, ses propres règles, sans se soumettre aux traditions ou aux préceptes de la religion. Un sujet de raison s’impose de lui-même un certain nombre de maximes auxquelles il soumet son comportement. Bien sûr, je crois nécessaire de le répéter, cette autonomie n’est pas donnée. Kant explique très bien que, par exemple si on a trop faim, on ne peut pas obéir à ses propres règles. À un moment, le corps se rappelle à nous et nous empêche d’être des sujets autonomes. L’autonomie relève d’un idéal ; un être de raison devrait se comporter de façon autonome ; mais, nous ne sommes pas que des êtres de raison ; nous sommes aussi des êtres de chair et de sang, nous avons faim, nous éprouvons des désirs, etc. Dire qu’une machine peut être autonome en ce sens-là, cela fait peur, car cela suppose qu’elle a des intentions propres qui nous échappent. Lorsque des gens nous disent que tout d’un coup la machine va prendre le pouvoir, c’est l’idée qu’elle va avoir son propre objectif et qu’elle va nous l’imposer. On n’a aucun élément qui nous permette de suggérer ce genre de choses. En revanche, qu’une machine tout d’un coup prenne des décisions qui vont à l’opposé de ce que l’on souhaite et qu’on ne sache pas l’arrêter, ça c’est pas du tout impossible. Ça se voit tous les jours.

Mehdi Khamassi

C’est déjà arrivé oui.

Jean-Gabriel Ganascia

Même si elles font exactement ce qu’on leur a dit de faire, ça ne correspond pas tout à fait à ce qu’on veut qu’elles fassent au moment où elles le font.

Mehdi Khamassi

Je suis totalement d’accord. Au fond, sans forcément impliquer que des machines atteignent l’autonomie idéale, - vu qu’on n’est pas sûr de cette autonomie idéale chez l’humain, - il n’y a pas de raison, qu’on n’arrive pas à programmer des machines avec des modèles qui soient assez proches de ce qu’on comprend de l’autonomie décisionnelle telle qu’on l’observe chez l’humain.

Jean-Gabriel Ganascia

Tout à fait de ce point de vue-là. C’est juste que la notion d’autonomie morale a quelque chose

de très particulier. Il y a, dans la tradition philosophique contemporaine, des personnalités comme Emmanuel Levinas qui s'opposent à Kant sur ce point, parce que justement ce n'est pas le sujet autonome lui-même qui est moral, c'est au contraire dans le rapport à l'autre, l'ouverture au regard de l'autre, au visage de l'autre, - aujourd'hui avec les masques, cette question est d'une douloureuse actualité... je l'ai relu récemment Levinas pour ça. Il parle de "l'épiphanie de l'autre". Effectivement, c'est dans cette altérité radicale de l'autre que tout d'un coup il y a une naissance de la conscience éthique, qui est très différente de l'idée kantienne. Il ne faut pas mélanger les choses et je crois qu'il faut être assez prudent. L'étymologie d'autonomie, auto-nomos, renvoie à nomos, la loi ; cela signifie ce qui se donne sa propre loi. Or, les machines n'ont pas d'intentions propres, même si elles peuvent se comporter de l'extérieur comme si, en effet, elles avaient des intentions. Ça c'est important parce que du point de vue philosophique, il y a des différences. Il y a toujours eu des incompréhensions. Quand Daniel Dennett a travaillé sur les systèmes intentionnels, il a bien souligné que ce n'était pas une question métaphysique. Pour lui, c'est toujours dans une stratégie de compréhension des machines, qu'on leur attribue des intentions et des connaissances. C'est de ça dont il est question.

Implications sociétales des discours annonçant la singularité

Mehdi Khamassi

J'aimerais qu'on aboutisse maintenant à la question sociétale de l'implication de ces discours qui vont annoncer notamment la singularité. Comme tu le soulignes, certaines personnes qui l'annoncent se basent non pas sur un vrai raisonnement scientifique, éclairé, mais plus sur des choses de l'ordre du mythe, des peurs et des émotions. En même temps, une partie de ces gens, qui peuvent parfois avoir des liens forts avec des géants du web, sont finalement ceux aussi qui peut-être bénéficieraient du fait de l'avènement de ça et de ces technologies. Est-ce que tu penses qu'il s'agit en partie d'une opération marketing, qui a moins comme visée de nous éclairer dans le débat, de nous aider à mieux comprendre, mieux réagir et nous adapter dans le futur à la présence d'intelligences artificielles, que d'en faire un certain beurre ? Peut-être qu'il ne faut pas schématiser comme ça. Qu'est-ce que tu en penses ?

Jean-Gabriel Ganascia

C'est parce que j'ai entendu tout un tas de déclarations de personnalités reconnus, d'anciens ingénieurs comme Ray Kurzweil, de scientifiques éminents comme Stephen Hawking ou Frank Wilczek, prix Nobel de physique, d'hommes d'affaires célèbres comme Elon Musk, même de chercheurs très renommés en intelligence artificielle comme Stuart Russell qui tous alertaient le monde entier en expliquant qu'il y avait quelque chose d'inquiétant dans le développement des machines, que j'ai essayé de lire en détail leurs arguments. Au début, je trouvais cela ennuyeux. La lecture de leurs écrits m'a confirmé dans mon intuition première : scientifiquement il n'y a rien là de tangible. L'histoire de la loi de Moore ça n'est quand même pas très sérieux. Ça ne veut pas dire que la loi de Moore ne va pas se poursuivre, mais on n'en sait rien. Et quand bien même elle continuerait, ça ne signifie pas que les machines deviendront conscientes et qu'elles nous deviendront hostiles. De même la "superintelligence" de Nick Bostrom n'est pas définie ; elle est affirmée comme devant advenir, puis

différents scénarios sont dressés. J'ai essayé de lire son bouquin ; c'est d'une très grande opacité. Il faudrait revenir à la question initiale : qu'est-ce que l'intelligence ? C'est ce qui m'a motivé au départ pour faire de l'intelligence artificielle. Ça reste un mystère. On ne se sait pas ce que c'est. On l'approche en la réduisant à un ensemble de fonctions cognitives. Il y a peut-être quelque chose d'autre qui échappe à la modélisation, par exemple la création de concepts, mais peu importe, là n'est pas la question. En tout cas, on ne peut pas réduire l'intelligence à une vitesse de calcul d'un processeur. C'est d'un ordre différent. J'ai regardé en détail les articles, les déclarations et les ouvrages des tenants de la Singularité technologique ou de la super-intelligence et j'ai été très surpris parce qu'il n'y avait rien d'autre que des vues très naïves sur l'intelligence et de la science-fiction assez datée. Ce qui est curieux c'est que les géants du web, qui utilisent beaucoup l'intelligence artificielle, nous expliquent en même temps qu'elle fait peur. On se trouve dans une situation très paradoxale. C'est comme si un marchand de tabac nous expliquait que la cigarette est dangereuse. La cigarette est dangereuse, je n'en doute pas, mais que le marchand de tabac essaie de m'en convaincre, sans que je ne lui aie rien demandé, cela laisse suspecter quelque chose de louche. Les GAFA nous expliquent que l'intelligence artificielle est dangereuse et, pourtant, ils la développent et s'en font les champions. Quel danger annoncent-t-il ? Ils nous racontent une fable incroyable qui se retourne toujours : d'un mal, dont ils prétendent nous avertir, ils tirent un bien. Quand, par exemple, Ray Kurzweil nous dit qu'il faut faire attention parce que la singularité est proche, en même temps il nous dit qu'on va être sauvés par la singularité puisqu'on va pouvoir se survivre et devenir immortel en téléchargeant notre conscience sur les machines. Promesse d'apocalypse ! On sort du champ de la rationalité. De même, quand Elon Musk affirme que l'intelligence artificielle est dangereuse, et qu'il prétend nous sauver en nous aidant à rivaliser avec l'intelligence artificielle grâce aux puces qu'il mettra dans nos têtes pour augmenter nos facultés cognitives. Je ne suis pas un spécialiste de neurosciences mais concernant la capacité qu'on aurait à augmenter nos capacités mnésiques en ajoutant des dispositifs électroniques de stockage, c'est outrancièrement réducteur ! On sait tous que la mémoire, ce n'est pas seulement le stockage, c'est quelque chose de beaucoup plus passionnant que ça.

Mehdi Khamassi

Oui, le tri par exemple.

Jean-Gabriel Ganascia

Tout ce qui est de l'ordre du codage, de ce qui se passe pendant le sommeil, le rêve, la consolidation, ce qui est lié à l'accès, à la réminiscence, tout est évacué ! Cette vision est tellement simpliste qu'elle en est risible. Pourquoi ai-je écrit ce livre sur le mythe de la singularité ? Parce que je me suis rendu compte qu'il y avait beaucoup de discours là-dessus et beaucoup d'arguments qui ne tenaient pas vraiment ; qu'on nageait en pleine confusion. J'ai alors réalisé que c'était à moi, en tant que scientifique et spécialiste d'intelligence artificielle, de relire ces choses et d'expliquer, dans les promesses, celles qui sont tenables et celles qui ne le sont pas. Ça conduit alors à expliquer au grand public ce qu'est la loi de Moore, l'exponentiel et pour quelles raisons les arguments qui se fondent uniquement là-dessus n'ont rien de scientifique. On ne peut pas anticiper ce qui va se passer à partir de la loi de Moore, car ce n'est qu'un résumé d'observations de l'évolution technologique à un moment donné de l'état de la technologie, ce qui fait que, lorsque la technologie évolue, on n'a plus de certitude sur sa viabilité. Mais après, une fois qu'on a montré l'inanité de ces affirmations, ce qui est intéressant, c'est de se demander pourquoi des personnalités reconnues, alors qu'elles n'ont pas l'air d'être idiotes, allèguent des théories aussi farfelues, au risque de se décrédibiliser ?

Mon hypothèse est que plutôt que de nous éclairer, ils essayent de jeter un écran de fumée sur ce qui est en train de se produire. Pourquoi ? Parce que derrière ils ont une stratégie qui, au-delà d'être économique, est politique. Ces grands acteurs de l'internet veulent essayer de prendre le pouvoir au-delà des États, d'assumer à leur place un certain nombre d'attributs de la souveraineté. Ils nous racontent donc une fable qui a pour visée de nous endormir. Mais en réalité, derrière, ils veulent aller très loin. Bien sûr, si on a peur de la singularité, on ne va pas voir ce qui se produit vraiment. Ce qui se produit vraiment c'est non seulement une puissance économique considérable mais aussi des attitudes qui défient en permanence les États. On l'a vu, par exemple, avec Apple. Au moment du premier confinement contre la CoViD, l'accord entre Apple et Google, le protocole qu'ils ont essayé d'imposer aux États européens, est tout à fait révélateur de ce type d'attitudes. Et dernièrement, Twitter a décidé de supprimer le compte de Donald Trump. Je n'ai aucune sympathie avec Donald Trump, mais il s'agit là d'une prise de pouvoir vraiment considérable : s'opposer au président de la république de son propre pays ! Cela pose tout un tas de questions. À partir du moment où les outils de communication sont essentiels pour les citoyens, puisque ça fait partie de leur vie civique, on ne peut pas leur supprimer l'accès à ces outils de façon autoritaire, sans procès, sans instruction du dossier, simplement parce que le propriétaire du réseau social l'a décidé. Cela aurait dû passer par la loi. Les réseaux auraient dû appliquer le droit. De même, en France la question se trouve posée avec la loi Avia. Sur le principe, le gouvernement avait raison, il faut imposer des limites à ce qu'on est peut dire sur les réseaux sociaux : la diffamation, l'incitation à la haine et à la violence sont inadmissibles ; mais il y a des lois qui établissent, avec précision, ces limites ; il suffit de les appliquer. Le problème de la loi Avia, et c'est pour cette raison que le Conseil Constitutionnel ne l'a pas acceptée, c'est qu'elle proposait de confier aux réseaux sociaux la responsabilité de décider de ce qui est acceptable ou non. Cela leur donne à mon sens un pouvoir de censure vraiment considérable. J'ai eu le sentiment que derrière les affabulations des GAFAM au sujet de la Singularité, il y avait des stratégies politiques passant par l'obscurcissement, le brouillage de ce qui en train de se tramer en secret. Je dois dire que j'ai été très étonné de voir à quel point ces fantasmagories avaient du poids.

Difficulté à publier en anglais une critique de la singularité

Jean-Gabriel Ganascia

Pour donner un exemple, un peu avant que j'écrive *Le mythe de la Singularité*, un de mes anciens élèves qui avait un poste important chez Axa m'avait posé des questions pour savoir comment la Singularité changerait le futur et quels étaient les risques. Comment est-ce possible qu'une société aussi importante se soit posée des questions là-dessus ? Comment se fait-il qu'ils aient été impressionnés ? À cela il est important d'ajouter que la puissance de ces grands acteurs de l'internet fait que même les philosophes du domaine étaient très réticents à critiquer la Singularité technologique. Ainsi, au plan personnel, je m'étais intéressé, depuis quelques années, aux questions philosophiques liées à l'informatique et à l'intelligence artificielle, en particulier à la modélisation des raisonnements éthiques avec des outils d'intelligence artificielle, autrement dit à ce que l'on appelle l'éthique computationnelle, et je continue de travailler là-dessus. C'est à ce moment que, constatant qu'il y avait trop de discours absurdes sur la Singularité, j'ai ressenti le besoin d'écrire sur le sujet. J'ai donc écrit un article pour une conférence où il y avait trois pelés dans la salle. Après moi, dans la même session, il y avait quelqu'un qui disait des âneries sur l'intelligence artificielle. C'était assez pitoyable. Puis, j'ai écrit un article pour le publier dans l'édition spéciale de la revue qui faisait suite à la conférence.

Il y a eu deux relecteurs. L'un, très positif, m'a demandé de changer quelques détails, comme il est d'usage. L'autre m'a répondu qu'il ne pouvait pas être d'accord avec moi sur certaines critiques que j'adressais à Singularité et il m'a demandé de citer 4 fois un article avec le même premier auteur. J'ai alors supposé qu'il s'agissait du premier auteur de l'article. J'ai regardé sur le web ; il se trouve que je le connaissais et qu'il était régulièrement invité à faire des exposés dans les universités de la Singularité ! On a donc donné un article critique sur la Singularité à relire à un tenant de la Singularité. Puis, la personne chargée d'éditer le numéro spécial m'a répondu que, comme nous étions le relecteur et moi, sur deux positions opposées, et que mon article était bien évalué, qu'elle allait lui demander de rédiger une réponse. Cela me réjouissait. Mais, quelques mois plus tard, alors que j'avais oublié l'article entre temps, je reçois une revue d'un troisième relecteur, vraiment furieux, qui proposait de vider l'article de son contenu. J'étais un peu étonné. Je connaissais bien le rédacteur en chef de la revue. Je lui ai demandé si je pouvais réécrire mon article. Il m'a découragé en m'expliquant qu'il y avait très peu de chances que l'article soit finalement accepté et publié dans cette revue, quoi qu'il advienne. Puis je l'ai revu après et il m'a avoué qu'il était difficile d'écrire quelque chose contre la singularité car les enjeux étaient trop importants. J'ai donc compris qu'il y avait une difficulté à parler de ces choses-là en anglais. J'ai donc décidé d'écrire un livre en français. Aujourd'hui, le bouquin est traduit en japonais, en coréen, en portugais, en arabe mais pas en anglais. Fort de cette expérience, je crois pouvoir affirmer qu'il y a sur ces questions-là une ligne extrêmement forte dans les pays anglo-saxons.

Ces théories de la singularité viennent de la science-fiction

Jean-Gabriel Ganascia

La dernière chose étonnante est que ces théories de la singularité viennent de la science-fiction. Si on regarde Elon Musk, ses grands projets industriels en viennent. Prenons par exemple SpaceX : le projet vient d'histoires un peu ridicule de petits hommes verts rencontrés lors d'expéditions sur Mars. Aujourd'hui, il est impensable d'envoyer un homme sur Mars. Un robot ou une fusée, oui. Mais ce serait d'une extrême cruauté que de laisser quelqu'un trois ans dans une fusée dans des conditions difficiles. De même, Neuralink est emprunté à la science-fiction, de l'écrivain Iain Banks par exemple, etc. On a donc l'impression que les grands entrepreneurs de l'internet empruntent à la science-fiction un certain nombre de thématiques en expliquant qu'ils vont les réaliser. Ce faisant, on inverse la fonction de la science-fiction. Au début, songeons à Jules Verne, elle prenait acte des réalisations des scientifiques et en nourrissait un imaginaire littéraire, tandis qu'aujourd'hui, on part de l'imaginaire de la science-fiction et on l'introduit dans les projets technologiques des entreprises privées. Bizarrement, tant la Commission européenne que les organismes de financement de la recherche aux États-Unis, souhaitent qu'on décrive de la science-fiction dans les projets de recherche ; ce qu'on appelle un "breakthrough", une promesse de percée technologique, condition nécessaire pour qu'un projet soit financé, tient justement à un conte de science-fiction auquel on doit se référer. Je suis sûr que toi aussi quand tu as écrit tes papiers ou tes demandes de financements, on t'a demandé de mettre en avant ces "technological breakthrough".

Mehdi Khamassi

Oui, souvent.

Jean-Gabriel Ganascia

En réalité ce que demandent les bureaucrates européens c'est qu'on promette de réaliser des rêves de science-fiction, de la science-fiction qu'ils connaissent, c'est-à-dire de science-fiction un peu éculée, d'histoire de robots soldats, de voyages sur Mars, de mise en réseau des cerveaux... On se retrouve alors, en tant que scientifiques, dans une situation déconcertante où, face aux besoins de financements de nos équipes de recherche, nous nous prêtons à ce jeu. C'est le cas pour nous en tant que chercheurs ; ça l'est aussi pour les industriels. C'est grâce à ça qu'ils décrochent des contrats et qu'ils séduisent les investisseurs. Si maintenant on examine ce que fait Elon Musk, qui est un entrepreneur extrêmement avisé — aujourd'hui certains disent même que c'est l'homme le plus riche du monde —, on constate que ses entreprises sont comme les fusées, à deux étages. Par exemple, avec Neuralink il y a un premier étage qui porte sur les implants neuronaux, dont on sait qu'ils peuvent avoir des applications thérapeutiques extrêmement utiles, et un deuxième étage complètement aberrant qui porte sur cette augmentation de la mémoire et de l'intelligence en connectant les cerveaux au réseau. Régulièrement, il envoie des communiqués pour annoncer que son projet progresse et qu'il obtient des résultats extraordinaires. Mais si on regarde de près, on constate qu'il reproduit des expériences relativement classiques, comme des interfaces ordinateur-cerveau avec des porcs ou des singes, et éventuellement des implants thérapeutiques, encore qu'il n'a pas fait beaucoup parler de ses réalisations sur le sujet. Là-dessus, on sait qu'on peut progresser. En revanche, sur le deuxième étage, on ne progresse pas. Et, rien de ce qui est présenté aujourd'hui n'aborde ce point. Le deuxième étage est conçu non pour être réalisé, mais pour faire rêver et obtenir des financements qui lui permettent de développer le premier. Même chose pour SpaceX, il nous dit qu'on va pouvoir aller sur Mars, mais il n'est pas du tout en train d'aller sur Mars. Il est en train de faire des navettes spatiales ou des lanceurs satellites qui sont en train de damer le pion à Arianespace. Alors d'un point de vue industriel c'est extrêmement utile mais justement il faut bien voir que ça a un rôle qui est de brouiller les pistes et de stimuler l'intérêt des investisseurs.

IV. Sommes-nous dans un processus d'intelligence collective sur les questions d'éthique et d'enjeux sociétaux liés à l'IA ?

Mehdi Khamassi

Au fond, comme tu l'écris, cela produit une sorte de captation de l'attention vers des scénarios catastrophistes qui ne sont pas vraiment basés sur un raisonnement éclairé et qui quelque part, j'ai envie de dire, nous empêchent de bien penser ce que l'intelligence artificielle va pouvoir apporter à notre société. Et, comme tu le soulignes, l'intelligence artificielle peut nous apporter beaucoup de choses aussi. Par exemple, elle peut nous aider à faire sens des nombreuses données qu'on a, pour mieux comprendre le monde, prendre de meilleures décisions, ce qui serait un objectif humaniste justement. Et puis on pourrait aussi réfléchir aux risques qui nous semblent probables et à ce qu'on doit faire. Je voulais du coup te demander ton point de vue, avec ton expérience dans des comités d'éthique et

tes interventions, penses-tu que la façon dont les États, l'Europe, mais aussi les sociétés savantes, essayent de gérer cette poussée de l'intelligence artificielle, relève quelque part d'un processus d'intelligence collective ? Y a-t-il des choses à améliorer ? Est-ce que les comités d'éthique suffisent ? Qu'est-ce qu'il manque dans tout ça ?

Jean-Gabriel Ganascia

Il y a plusieurs choses. Il y a d'abord le besoin d'une stratégie. Les États essaient d'avoir une stratégie parce qu'ils sentent que c'est déterminant pour leur futur. Je ne suis pas sûr que l'Europe ou que la France y parvienne, parce que souvent l'intelligence artificielle s'est mise en place à partir de petites structures et puis ensuite avec la capacité de lancer de très grandes compagnies. Or, en Europe on a du mal, d'autant plus que comme le monde européen est ouvert sur l'extérieur, dès que de petites sociétés commencent à prendre de l'importance, elles peuvent être rachetées par ces grosses sociétés. On comprend bien sûr que les ingénieurs qui lancent des jeunes pousses gagnent de l'argent au moment où celles-ci sont rachetées. C'est d'ailleurs leur seule chance de gagner un peu d'argent. Donc ils ne vont pas s'y opposer, c'est difficile de le leur demander. En plus, l'Europe n'encourage pas vraiment la création de très grands groupes, et même elle les décourage de multiples façons. Ça c'est un problème de stratégie industrielle qui est énorme. D'autant plus que les financements européens sont donnés avec un mécanisme d'évaluation de projets ouvert à tous les partenaires extra-européens. Les grands groupes internationaux ont accès à toutes les informations européennes puisque c'est totalement ouvert. Ils ont un pouvoir de lobbying extrêmement fort. Si on regarde, que ce soit à la Commission européenne ou au Parlement européen, ils sont en permanence présents dans les réunions. On est dans une situation qui empêche toute stratégie industrielle autonome.

Quand on lit les travaux des comités d'éthique, on l'a vu par exemple avec le livre blanc sur l'intelligence artificielle qui a été fait au plan européen, ce qui est très étonnant c'est qu'on nous explique que l'intelligence artificielle est essentielle, que l'Europe s'engage à faire un effort considérable et à mettre de l'argent pour devenir un champion du monde dans ce secteur. Or, l'enveloppe du budget promis est 10 fois moins importante que celle que mettent les États-Unis, et 50 fois moins que la Chine. Donc ce n'est pas sérieux. Surtout, on ne nous explique pas, ce qui est extraordinaire, tout ce que tu évoquais tout à l'heure sur les perspectives qu'on peut ouvrir avec l'intelligence artificielle, tout ce qui pourrait motiver les européens, ce qui fait qu'il y a un futur. Et c'est particulièrement vrai en France où on est très pessimistes par rapport à ça. En revanche, on nous parle de tous les dangers, de toutes les limites qu'il faut imposer, de toutes les restrictions nécessaires, etc. Bref, on montre tous les aspects négatifs et toutes les précautions qu'on prend, mais jamais des aspects positifs…

Quant aux comités d'éthique que l'on met en place, ce sont bien plutôt des comités de régulation. En tout état de cause, ce ne sont pas vraiment des comités de réflexion sur l'éthique ; l'enjeu n'est pas d'ordre philosophique ; il est d'abord et essentiellement juridique. On a fait ça avec le RGPD, qui a ses qualités et ses défauts. Quand on le regarde en détail, il s'avère très lourd, très coûteux et très ambiguë. D'ailleurs, les différents États européens ne l'interprètent pas tous de la même façon. Et puis sur d'autres choses aussi, on n'engage pas les réflexions qui devraient avoir lieu. Il faudrait former l'ensemble des citoyens et des ingénieurs à une réflexion ouverte à chaque moment, parce qu'on ne peut pas anticiper ce qui se passe avec des principes abstraits et contradictoires. Le problème des technologies de l'information c'est que les modes d'appropriation sont toujours différents de ceux qu'on imagine. Qui aurait anticipé qu'avec des réseaux de copains, que sont les réseaux sociaux, on arrive à transformer la politique dans le monde et à créer à la fois des empires financiers immenses mais surtout des transformations complètes de la politique ? Donc ça, ce sont des choses qu'on a dé-

couvert petit à petit. On ne peut pas exiger dès le départ de savoir ce qui va être bénéfique ou néfaste. Quand on regarde les résultats des rapports qui ont été rédigés par les différents comités d'éthique, ils partent de grands principes de bioéthique qui sont tous un peu discutables et qui, de toute les façons, ne s'appliquent pas vraiment au numérique. En effet, à mon sens les questions d'éthique du numérique ne s'imposent pas dans les mêmes termes que les questions de bioéthique, parce qu'elles ne portent pas sur un individu à un moment donné. Ainsi, en bioéthique, on le sait depuis Hippocrate, le médecin ne doit pas nuire, il doit procurer des remèdes qui soignent et bannir le poison. Peut-on transposer ça et prétendre qu'on ne peut autoriser que des développements technologiques bénéfiques et non maléfiques ? Cela n'a pas de sens ! Était-il possible, en 2007, à la naissance de Facebook, de savoir si les réseaux sociaux seraient bons et pas mauvais ? Je crois vraiment qu'il faut qu'il y ait une réflexion plus en profondeur, une discussion vraiment philosophique, et pas seulement des comités à vocation règlementaire qui rédigent des textes et des recommandations insipides, plus nuisible qu'autre chose.

On l'a vu d'ailleurs dans les régulations du Parlement européen, par exemple, le fait qu'on a eu cette résolution sur la personnalité juridique des robots. Si on réfléchit un peu, ça n'a pas grand sens parce que cela veut dire qu'on va avoir un fond d'assurance pour indemniser les victimes. On a dit que c'était une fiction juridique analogue à la fiction d'une personnalité morale des sociétés. Mais le parallèle ne tient pas : une société possède naturellement de l'argent, c'est sa fonction, alors que, dans le cas des robots, il n'y en a pas. D'autre part, s'il y a un accident mortel, ça relève du pénal, et ça échappera à ce type de législation. Ça veut dire que la cible est extrêmement étroite. En plus de ça, s'il y a un accident qui a des répercussions financières, le fait d'indemniser la victime sans faire d'enquête, en attribuant la responsabilité au robot et enjoignant à l'assurance d'effectuer le remboursement, est très dommageable. Cela permet de faire l'économie de l'enquête alors qu'il conviendrait de pousser les investigations jusqu'au bout pour déterminer les responsabilités exactes et pour améliorer la technologie. Donc, à mon sens, il ne faut surtout pas accepter une personnalité juridique des robots. Je suis donc très gêné face à cette résolution qui m'apparaît éthiquement condamnable.

De même sur les systèmes d'armes autonomes, il y a eu beaucoup de bêtises qui ont été dites, car on suppose, — c'est ce qu'on nous explique dans les projets de moratoires, — qu'ils constituent une révolution dans l'art de la guerre qui serait analogue celle qu'ont introduite la poudre à canon et la bombe atomique. Or ça n'a rien à voir parce que ce n'est pas d'une puissance de feux dont il est question avec les armes autonomes. De plus, il n'y a pas révolution. Ce sont des armes automatiques — c'est pour ça que la différence entre automatisme et autonomie est si importante, — c'est donc une continuation de ce qui s'est toujours fait. Entre une mine et une arme autonome il n'y a pas de différence fondamentale, si ce n'est que la mine normalement est immobile (encore qu'il puisse y avoir des mines mobiles, par exemple en mer). Sous la pression d'un certain nombre de groupes d'influence, cela a conduit le Parlement européen, en février 2019, à voter une résolution qui demande à la Commission européenne de ne pas financer les industriels qui fabriqueraient des systèmes d'armes avec de l'intelligence artificielle. Or, aujourd'hui, tout système d'arme moderne contient de l'intelligence artificielle. Cela revient donc à demander à la commission européenne de renoncer à financer une industrie militaire européenne. En conséquence, cela revient à interdire à l'Europe d'avoir une défense souveraine. Quand on analyse la généalogie de ce type de résolution, on voit qu'elle vient de groupes de pression anglo-saxons, à la fois américains mais surtout anglais. Il faut savoir que c'était à l'époque où les anglais faisaient encore partie de l'Europe mais allaient quitter l'Europe. Or les anglais ont une industrie de l'armement. Donc ça les arrangeait vraiment que l'Europe ne finance pas d'autres industries de l'armement. Ça veut dire que les législations qui sont en cours d'adoption, ou disons qui ont fait l'objet de résolutions au niveau du Parlement européen, se révèlent délétères.

On invoque souvent l'éthique, mais, je ne suis pas sûr que ce soient des questions d'éthique ou

des considérations éthiques qui motivent ces décisions. L'éthique c'est d'abord — et surtout sur des domaines nouveaux comme ceux-là, — une réflexion sur ce qui est en train d'advenir, une réflexion vive et ouverte à toutes les nouveautés. Et pas simplement des corpus de règles contraignantes qui lient les mains.

Stratégie européenne sur l'IA ?

Mehdi Khamassi

Alors il y a plusieurs éléments : il y a les questions d'éthique et puis cette réflexion en profondeur vraiment à un niveau philosophique, et puis tu parles, et c'est important, de stratégie au niveau européen. Dans ce sens-là, est-ce qu'une intelligence artificielle européenne est quelque chose qui a du sens et un avenir selon toi ?

Jean-Gabriel Ganascia

Ça devrait avoir un sens ! En tout cas, je crois qu'on a besoin, dans le domaine du numérique, de travailler à une échelle plus grande qui est celle des nations telles qu'elles existent aujourd'hui, en particulier des États-nations européens. Ce qui fait la puissance des États-Unis c'est qu'ils ont un énorme marché intérieur, avec en plus tout le marché européen. Il me semble que si on veut être capable de mettre en place des très grandes sociétés, il faudrait qu'on puisse avoir un marché de taille équivalente, parce que les données sont proportionnelles à la taille et que la puissance d'un réseau de communication est une fonction quadratique du nombre de nœuds de ce réseau. Donc, plus il y a d'échanges, plus il y a de données et plus on peut développer les techniques d'intelligence artificielle. Je crois donc qu'il est absolument essentiel d'avoir des stratégies européennes. Peut-être que le problème jusqu'à présent, est qu'il n'y a pas de conscience suffisante de cette nécessité en Europe. Sans doute aussi parce que l'Europe est en situation de faiblesse ; elle se prend à son propre piège : elle veut être universelle, ouverte au monde entier et donc elle accueille tous les acteurs d'internet qui font leur propre lobbying et qui sabordent un certain nombre d'initiatives strictement européennes. Qui plus est, ils émargent souvent, à travers leurs différentes filiales, à des programmes européens dont ils pompent les financements. Ils recrutent aussi les meilleurs étudiants dont ils aspirent le savoir-faire. Donc, on se retrouve dans une situation où l'Europe est ponctionnée par différents acteurs. Je pense qu'un jour il y a aura une conscience de ces enjeux-là et à ce moment les choses se transformeront. Mais aujourd'hui on se retrouve dans une situation très difficile de ce point de vue-là.

Questions éthiques pour le chercheur en IA/robotique qui signe des pétitions sur ces sujets

Mehdi Khamassi

Du coup une question que peut se poser le chercheur dans ses actes individuels du quotidien, dans son éthique, dans ses tentatives de contribuer à des choses et parfois juste à une petite échelle : parfois on est exposé à des pétitions qu'on est invité à signer. Notamment, ces dernières années, il y en a qui

ont été initiées par des instituts comme, par exemple, le Futur of Life Institute, dont tu rappelles dans tes ouvrages qu'ils sont financés en lien avec ces géants du web, des choses liées à la singularité. Alors au départ, il y avait des premières pétitions et des concepts un peu fumeux effectivement, de la singularité ou autre, sur lesquels je restais sceptique. Et puis, je me rappelle il y a quelques années, peut-être 2017-2018, il y a avait cette pétition qui disait qu'il fallait faire quelque chose par rapport aux armes autonomes. Je me suis retrouvé face à cette pétition à me dire : dois-je me méfier des personnes qui appellent à signer ? En même temps, le contenu me paraissait important pour qu'on prenne conscience de ça et pour que le débat soit ouvert. Et, par exemple, cette pétition-là je l'ai signée, comme un certain nombre de collègues. Mais avec tout ce que tu racontes, il y avait peut-être des intentions derrières. Que doit-on faire ?

Jean-Gabriel Ganascia

Alors c'est très difficile. Justement sur les armes autonomes, au moment où je l'ai vue, j'ai quand même pris la peine de regarder en détail. Je me suis rendu compte qu'une analyse vraiment approfondie des termes montrait que c'était absurde. Souvent on nous explique que c'est une troisième révolution dans l'art de la guerre. On nous donne des exemples pathétiques avec ces quadricoptères dévastateurs qui attaqueraient des étudiants sans défense ; il y a eu des films à gros budget qui ont été faits pour nous marquer. Quand on les regarde, au premier abord on est effrayé, on a tendance à signer. Mais si l'on fait une analyse plus approfondie, si l'on discute avec des spécialistes de ce genre de choses, si l'on en parle avec des militaires, on se rend compte des affabulations. Il y a des enjeux très particuliers et il faut être très vigilant. C'est la difficulté parce qu'effectivement au départ, dans nos milieux et nos générations, on n'est en général pas très favorables aux industries de l'armement et on n'aime pas tellement la guerre. Donc effectivement lutter contre ça peut sembler relever d'un mouvement spontané de protection contre les méfaits de la guerre. Je conçois qu'on soit contre les armes et contre la guerre, mais lorsqu'on cible les armes autonomes, la portée est toute autre. Quand on lit le texte lui-même, il n'apparaît pas du tout hostile à la guerre. Il affirme même le contraire en expliquant que c'est inévitable car les États sont en train de les développer. En parallèle, il affirme que ce sont des armes seront peu chères, ce qui est contradictoire. Il fait mine d'ignorer que, contrairement à ce qui se produisait dans les années 1950, ce ne sont plus les militaires qui financent les industries de l'armement — c'est beaucoup plus compliqué que ça, — c'est l'industrie grand public et en particulier sur ces dispositifs, c'est plutôt le jeu vidéo qui va financer des recherches qui ensuite pourront être exploitées par les militaires. On est donc dans un cas de figure totalement différent de celui qui existait auparavant. C'est l'absence d'analyse de ce type de choses qui me semble très problématique. Je crois qu'il faut être très vigilant. J'ai écrit, il y a quelques années, un article sur le sujet. J'ai examiné les enjeux européens autour de ces questions. J'ai constaté qu'il y a des mouvements très forts pour imposer un moratoire au développement de ces armes autonomes. Et, derrière, il y a des enjeux d'ordre politique et économiques qui demanderaient à être élucidés avec une grande perspicacité.

Questions éthiques liées à un certain nombre d'applications de l'IA

Mehdi Khamassi

J'ai envie de te poser une dernière question qui suit cette direction. Au fond l'objectif d'une in-

telligence artificielle éthique bienfaisante, qu'on pourrait développer au niveau européen et pousser dans ce sens, semble, en même temps tel que tu le décris, en tension avec les réalités géostratégiques, peut-être avec les ambitions du gouvernement américain, des grandes entreprises comme les géants du web et avec les ambitions de la Chine. Comment faut-il selon toi aborder le problème ?

Jean-Gabriel Ganascia

Il faudrait d'abord une éducation de l'ensemble de la population. C'est la très forte demande d'un côté et le refus d'un certain nombre d'applications jugées néfastes de l'autre, qui aideront. Mais, pour cela, il faut expliquer les enjeux. Prenons l'exemple de la reconnaissance faciale et la reconnaissance de postures. Certains veulent l'interdire. Moi je pense que ce serait dommage parce que cela peut avoir des applications extrêmement utiles. Il existe aujourd'hui un patrimoine cinématographique énorme. La reconnaissance faciale peut aider à l'indexer automatiquement. On ne va pas, sous prétexte qu'on peut l'utiliser de façon malfaisante, se priver de ce genre de techniques qui sont essentielles pour mettre en valeur notre patrimoine et notre culture. En même temps, il faut refuser des applications comme celles qu'on développe en Chine, avec un marquage des individus à la culotte. Je crois que ça ne pourra se faire que s'il y a une société qui est aux aguets et qui s'inquiète à juste titre des applications des technologies. Notre rôle de scientifiques est de bien mettre l'accent sur ce qui est vraiment problématique et sur ce qui l'est moins. Il faut être très attentif parce qu'il y a des mécompréhensions. Au moment de la covid avec les systèmes de traçage il y a eu des inquiétudes qui lorsqu'on les analyse dans le détail n'étaient absolument pas justifiées, parce qu'il y avait une très grande attention, notamment entre autre de la part de chercheurs de l'INRIA, à la protection des données personnelles. Beaucoup craignaient qu'avec les applications de traçage, toute les données de santé individuelles deviennent accessibles à l'État, sans songer qu'elles le sont déjà ! Si on décide de ne jamais mettre les données individuelles de santé à disposition de l'État, il faudrait exiger le démantèlement de la sécurité sociale. On doit être raisonnable ! Il faut essayer d'expliquer les différents risques et les mesures de prévention, puis de les mettre en rapport les unes avec les autres, en expliquant la nature des compromis que l'on choisit, collectivement, de faire. Je pense que c'est ça qui est très difficile dans le monde actuel. Je crois que ce qu'il faut c'est qu'il y ait d'abord une conscience des dangers et des enjeux, et en même temps un projet commun, qui sera aussi une motivation, qui ne soit pas simplement un projet frileux, sinon on n'y arrivera jamais, et que l'ensemble de la population européenne et en particulier française s'intéresse à ces choses-là. Or, malheureusement quand on interview des gens dans la rue, on voit qu'il y a surtout des craintes. Les comités d'éthique que l'on a réunis font surtout écho aux craintes. Ils sont là pour dire "On va y aller mais à reculons". Alors qu'il faut y aller de façon enthousiaste et, en même temps, être capable de mettre en place des garde-fous contre des utilisations qui seraient extrêmement dommageables. J'ai mentionné tout à l'heure la notion de "crédit social" telle qu'elle existe en Chine. Je trouve cela terrifiant. Mais il y a plein d'autres applications qui pourraient être problématiques pour notre liberté si l'on n'y prend pas garde. Pour le résumer en un mot, ce qui m'embête, c'est que ces comités d'éthique qui se réunissent depuis 4-5 ans et qui dictent un tas de chartes, de principes et de recommandations, ne se sont jamais opposés au projet d'Elon Musk dans sa société Neuralink, et à sa volonté de connecter tous les cerveaux au réseau internet. D'un point de vue éthique, il n'y a rien de pire : il va nous mettre ce qu'il veut dans la tête. Si cela devait vraiment advenir, ça serait la pire dictature qu'on puisse imaginer. Ce qui m'étonne est que ces comités d'éthique ne s'y sont jamais opposés, alors qu'on aurait dû interdire à Elon Musk de proposer de tels projets. De même, Mark Zuckerberg dit qu'il veut développer des interfaces cerveau-ordinateur. Les interfaces cerveau-ordinateur peuvent jouer un rôle majeur dans

beaucoup de domaines, comme la rééducation après un AVC, etc. Mais Zuckerberg souhaite faire ça dans le but de communiquer entre réseaux sociaux et individus, pour percevoir leurs désirs. C'est ici aussi la pire des choses au point de vue éthique. La morale c'est d'abord d'instituer une distance avec son désir, qu'on est capable de réfréner. Là encore je n'ai pas vu d'opposition par rapport à son projet. C'est cela qu'il faut réussir à dénoncer, et que l'ensemble de la population dise "voilà ce qu'on veut faire voilà ce qu'on ne veut pas faire".

Mehdi Khamassi

Un grand merci, car tous ces éléments sont vraiment éclairants, et vont alimenter nos réflexions dans TESaCo. Et au plaisir de rediscuter de tout ça avec toi.

Jean-Gabriel Ganascia

Au plaisir. Sur les questions de technologies, enfin de neurotechnologies cognitives, je voulais qu'on fasse un avis avec le comité d'éthique de l'Inserm. Je ne sais pas si on en aura le temps. Mais ce sont vraiment des questions qu'il faudrait se poser. Il y a beaucoup de jeunes gens hyper sympas qui montent des petites boîtes développant des technologies neurocognitives ; ils n'ont pas d'intentions machiavéliques. Bien au contraire, ils souhaitent améliorer le bien-être. Mais il faut se demander quels effets induits cela peut avoir, par exemple avec la neurostimulation cognitive. Je pense que dans chaque cas de figure, il faut regarder avec attention ce qui est proposé et être capable d'alerter si nécessaire, tout en faisant en sorte de rassurer lorsque c'est possible. J'ai lu par exemple la loi sur la sécurité globale. J'ai ensuite été interviewé, mais je n'osais pas trop m'exprimer, car c'était une émission à la radio, au moment de la covid, à l'aveugle et je ne voyais ni l'interviewer ni les autres interviewés. De plus, le producteur diffusait des informations très hostiles à la loi. Beaucoup ont cristallisé leurs critiques sur l'article 24, dont ils craignaient qu'il interdise la photographie de policiers au cours de manifestations. Il y a peut-être eu plusieurs rédactions de cet article, mais dans celle que j'ai eu, il n'y avait vraiment pas de quoi s'offusquer. Plus précisément, l'article 24 comprend deux sous-articles. Le premier stipule qu'on n'a pas le droit de publier sur les réseaux sociaux des photos qui permettent de reconnaître des agents des forces de l'ordre. Beaucoup ont interprété ça comme une interdiction de prendre en photo les forces de l'ordre. Or, ce n'est pas le cas. En effet, le deuxième sous-article explique qu'il est possible, si ça se justifie, d'envoyer à la justice des photos qui permettent de reconnaître des policiers qui se seraient mal conduits. En cela, la loi vise à interdire l'appel public à la haine sur les réseaux sociaux, comme cela s'est produit avec Samuel Paty. Cela apparaît justifié. Il faut l'expliquer. Il y a une mauvaise compréhension. Pourtant, il suffit de lire la loi pour s'en convaincre. Je ne suis pas certain que la première rédaction ait été aussi limpide que celle que j'ai lue, mais celle-là était vraiment très claire. C'est notre rôle je crois d'essayer d'expliquer et de démêler le vrai du faux. Malheureusement, c'est difficile parce que nous sommes dans un monde de technologies de l'information où toutes sortes d'informations circulent et où nos facultés d'attention sont saturées. On n'a pas les capacités cognitives de tout lire. C'est un problème de communication. Dans la communication, il y a l'émetteur, le récepteur, le canal et le code. Et, trop souvent, face à la profusion, le récepteur ne prend, dans le flot immense d'information qu'il reçoit, que celles qui l'intéressent.

Mehdi Khamassi

Et c'est un exercice très délicat pour le chercheur parce que des fois on peut avoir l'air de dire

“attention il y a cet aspect positif, ou il y cela qu’il faut nuancer” mais on peut, si on n’a pas bien expliqué certaines choses, être vite mis dans le même sac “Alors lui, il justifie le système”. Cela donne l’effet inverse et il n’y a pas plus beaucoup d’écoute.

Jean-Gabriel Ganascia

Sur la reconnaissance faciale, j’ai pris ça comme exemple car c’est vrai que c’est extrêmement intéressant de faire l’indexation des vidéos. On ne va pas s’en priver. Il y a plein d’autres applications d’authentification. L’authentification on peut dire que c’est un peu sécuritaire. En revanche, là, l’indexation de données multimédia n’apparaît pas vraiment sécuritaire ; c’est d’un autre ordre. Pour des raisons culturelles, on doit l’encourager. Notre rôle, en tant que chercheurs, est d’expliquer et d’essayer de faire passer ces idées dans la population en éduquant.

Mehdi Khamassi

C’est une belle conclusion.