Vol. 30, nº 3

Le nouvel âge de l'intelligence artificielle : une synthèse des enjeux éthiques

Jocelyn Maclure* et Marie-Noëlle Saint-Pierre**

---

* Professeur de philosophie à l'Université Laval et président de la Commission de l'éthique en science et en technologie.

** M.A., conseillère en éthique, Commission de l'éthique en science et en technologie.

[Note : cet article a été soumis à une évaluation à double anonymat.]

RÉSUMÉ

Rejetant la théorie catastrophiste du robot dominant l'humain, les auteurs discutent de risques potentiels du développement de l'intelligence artificielle et de certains enjeux éthiques qu'ils soulèvent ; de la transparence des algorithmes à la protection de la vie privée, en passant par les impacts de l'automatisation par l'IA sur le monde du travail.

MOTS-CLÉS

discrimination ; données personnelles ; droit du travail ; droits de la personne (vie privée) ; éthique ; intelligence artificielle

INTRODUCTION

Le terme « Intelligence artificielle (IA) » aurait été formulé pour la première fois en 1955 par l'informaticien américain John McCarthy. Pour lui, on peut attribuer une forme d'IA à une machine si cette dernière se comporte d'une façon que l'on qualifierait d'intelligente si les comportements étaient attribuables à un être humain. Dans les dernières décennies, c'est principalement dans le domaine des jeux de stratégie que le concept frappe l'imaginaire du grand public. En 1997, le superordinateur DeepBlue (IBM) met échec et mat le champion d'échecs russe Garry Kasparov. L'ordinateur Watson a également ses cinq minutes de gloire en 2011 lorsqu'il gagne au jeu télévisé *Jeopardy*. Enfin, six ans plus tard, ce sera au tour du logiciel Alphago (Google) de devenir célèbre en battant M. Ke Jie, le champion de go ; un jeu naguère jugé trop intuitif pour être conquis par un ordinateur. Mais l'IA ne se contente pas de jouer à des jeux. Ces programmes informatiques s'immiscent maintenant dans les sphères principales de la vie humaine.

Dans notre vie privée tout d'abord. L'IA se déploie par le biais de plusieurs fonctionnalités sur nos téléphones dits « intelligents », comme la reconnaissance faciale ou la commande vocale. Elle analyse de façon continue nos habitudes de navigation et nos comportements sur les réseaux sociaux afin de nous proposer des contenus ciblés. Elle permet aussi l'organisation du transport en commun d'une ville ou de s'orienter à travers ses rues. L'IA s'invite aussi sur la route, où sont élaborés des véhicules « autonomes » qui promettent une réduction significative du nombre d'accidents. Outre de telles commodités, les IA peuvent effectuer des tâches complexes aux retombées majeures. C'est le cas dans le domaine de la santé, où l'IA permettra vraisemblablement des diagnostics plus précoces et précis en médecine, comme le repérage de tumeurs cancéreuses. En droit, le recours à l'IA est utile pour la recherche juridique, en particulier pour repérer des informations dans des documents volumineux. Plus largement, elle permet de trier de l'information à grande échelle et de la rendre à la fois accessible et utilisable. C'est d'ailleurs ce qui est illustré par les

prouesses liées à l'assistance virtuelle, aux programmes de traduction ou aux outils de planification pour entreprises.

L'IA refaçonne aussi la vie sociale et politique. Le Web est un lieu de communication et d'information, de mobilisation et d'action politique, de socialisation et de séduction, de surveillance et de voyeurisme, de commerce et d'affaires, etc. Il devient de plus en plus nécessaire d'être branché pour participer à la vie sociale et au débat public. Les journaux publient la majorité de leurs contenus sur le Web et c'est sur des interfaces de partage en ligne qu'ont lieu nombre de débats publics. Ainsi, pour y participer, il faut d'emblée accepter les conditions d'utilisation qui impliquent, de manière générale, la collecte de nos données qui iront par la suite nourrir les données massives, carburant essentiel à de nombreux algorithmes d'IA.

Notre vie professionnelle n'est pas épargnée. Un grand nombre de chercheurs, d'investisseurs, d'entrepreneurs et de décideurs publics acceptent en effet l'hypothèse selon laquelle l'IA, les données massives et l'Internet des objets seront à l'origine d'une Quatrième Révolution industrielle[1]. Le terme désigne la transformation profonde du marché de l'emploi et de la productivité engendrée par la nouvelle phase d'automatisation permise par les progrès en robotique, l'utilisation d'algorithmes d'apprentissage automatique et l'analyse des données numériques en grande quantité dans une large gamme de milieux de travail.

Aucune technoscience ne suscite actuellement autant de craintes et d'espoirs que les machines dotées d'IA. C'est à l'image de ce contraste que se forgent en grande partie les perceptions à l'égard de l'IA : soit l'IA génère un enthousiasme jusqu'à preuve du contraire excessif, soit un catastrophisme. Il s'agit là d'un schème de compréhension à partir duquel il est difficile de cerner et de définir les véritables enjeux éthiques de l'IA. Bien que les réserves exprimées

1. Après l'invention de la machine à vapeur à la fin du 18e siècle, l'électrification et l'avènement des chaînes de montage au début du 20e siècle, et enfin, l'électronique et la robotique au tournant des années 1970, l'automatisation intelligente et l'arrivée du numérique viendra, selon les promoteurs de ce point de vue, remanier les façons de faire, les systèmes, les processus de travail, les modes de gestion et le marché de l'emploi. C'est cette transformation où le numérique s'immiscera de plus en plus dans la réalité matérielle qui caractérise principalement cette période. Le terme de Quatrième Révolution industrielle a été donné en 2016 par Klaus Schwab, président et fondateur du World Economic Forum pour illustrer la force des changements qu'auront les technologies émergentes « sur la transformation de l'économie, des sociétés et de nous-mêmes, en tant qu'êtres humains ». Klaus SCHWAB, « Klaus Schwab : comment façonner la quatrième révolution industrielle », *La Tribune*, janvier 2018, en ligne : <https://www.latribune.fr/opinions/tribunes/klaus-schwab-comment-faconner-la-quatrieme-revolution-industrielle-764814.html>.

à son sujet soient de plus en plus fondées sur l'état actuel de son développement, les craintes les plus vives tombent fréquemment du côté de la spéculation proche de la science-fiction. Les thèmes de la « singularité », de la « superintelligence », du « risque existentiel », du « téléchargement de l'esprit », de la création d'un « cerveau complet artificiel », de la « fin du travail » ou de l'« obsolescence de l'être humain » sont fréquemment abordés par les médias. Certaines des meilleures productions télévisuelles et cinématographiques actuelles explorent la relation humain-IA (*Westworld, Black Mirror*, *Her*, *Ex Machina*, etc.) en donnant vie à des IA fortes et générales égales ou supérieures aux êtres humains. Or, une réflexion collective sur l'IA qui repose trop lourdement sur ces scénarios futuristes s'interdit de voir les risques et les bénéfices des applications actuelles de l'IA.

Après avoir brièvement présenté les causes des progrès récents en IA, nous discuterons, dans cet article synthèse, de la possibilité de la création d'une « superintelligence » artificielle et du « risque existentiel » qu'elle pourrait poser. Nous soutiendrons que les raisons qui pourraient nous motiver à prendre cette menace au sérieux sont largement insuffisantes et que la priorité doit être accordée aux enjeux engendrés par l'état actuel de la science et par son évolution prévisible. Nous passerons ensuite en revue un certain nombre des risques éthiques associés aux progrès de l'IA, en particulier eu égard aux principes de responsabilité, d'égalité et de non-discrimination, d'autonomie et de protection de la vie privée et de justice distributive[2].

1. LA RENAISSANCE DE L'IA

Les progrès en IA comptent au moins trois conditions conjointement nécessaires : (1) un changement de paradigme technoscientifique, où l'« apprentissage machine » ou l'« apprentissage automatique » (*machine learning*) est devenue l'approche dominante ; (2) l'augmentation continue de la puissance de calcul des ordinateurs en vertu de la Loi de Moore[3] ; et (3) la disponibilité de données numériques

2. Cette synthèse ne prétend pas à l'exhaustivité. Des questions générales comme les conséquences anthropologiques de la coévolution humain-machine et des questions spécifiques comme l'utilisation de systèmes d'IA par les forces armées sont, par exemple, laissées de côté.
3. En 1965, l'ingénieur Gordon E. Moore énonce ce qui deviendra la « Loi de Moore » à savoir que le nombre de transistors par circuit de même taille doublerait tous les ans, à coûts constants, augmentant la puissance des ordinateurs de manière exponentielle. En clair, la miniaturisation permettrait d'augmenter le nombre de transistors sur un microprocesseur, augmentant ainsi la puissance des appareils. Il ajusta en 1975 sa loi, portant à 18 mois le rythme de doublement. Bien que plusieurs soutiennent que le rythme de l'augmentation ne pourra être maintenu, la loi n'aurait pas encore été démentie.

en grande quantité (données massives). L'approche ayant permis les avancées récentes relève du projet visant à permettre aux ordinateurs d'apprendre de façon automatique et s'appuie sur un modèle qui consiste à voir les algorithmes d'IA comme des « réseaux de neurones artificiels » s'inspirant vaguement du fonctionnement neuronal des cerveaux animaux[4]. Cette approche, d'abord élaborée dans les années 1940 et 1950, exigeait à la fois des ordinateurs plus puissants et l'accès à de larges jeux de données d'entraînement pour déployer tout son potentiel. Contrairement à l'approche classique, dite « symbolique », en IA, qui misait sur l'application, par un programme informatique, de commandes et de règles logiques à des situations particulières, l'apprentissage automatique cherche à permettre à l'algorithme d'apprendre par lui-même à partir des exemples qui lui sont présentés. « Machine learning », écrivent Pedegrosa *et al.*, « is about learning some properties of a data set and applying them to new data »[5].

Un réseau de neurones artificiels consacré à la reconnaissance visuelle peut ainsi parvenir à reconnaître correctement, de façon inductive et sans connaissances préalables, un enfant sur une photo après un entraînement en vertu duquel des photos identifiées comme « contenant un enfant » ou « ne contenant pas d'enfant » lui ont été présentées. C'est ce qui est appelé l'« apprentissage supervisé ». L'apprentissage « non supervisé » est aussi possible : le réseau de neurones artificiels analyse un très grand nombre de photos non identifiées, dont plusieurs contenant un enfant[6]. Étant conçus pour déceler des régularités ou des modèles (*patterns*) dans les jeux de données, les algorithmes d'apprentissage automatique arrivent à repérer les corrélations qui sont associées, selon des probabilités élevées, à la présence d'un enfant sur une photo. Au final, ces algorithmes arrivent à identifier correctement les enfants sur des images sans avoir la moindre compréhension de ce qu'est un enfant. L'apprentissage automatique des machines s'appuie sur une approche entièrement corrélationiste et probabiliste. Les nouveaux systèmes d'IA permettent, bien davantage que les méthodes d'analyse statistique conventionnelles, d'établir des corrélations dans de grands jeux de données et d'émettre des jugements probabilistes. Si cette approche a permis des progrès majeurs, le type même de rationalité qu'elle

4. Pour une introduction à ce modèle, voir Peter NORVIG et Stuart RUSSELL, « Artificial Intelligence: A Modern Approach », *Pearson Education*, 2009 ; Yann LECUN, Yoshua BENGIO et Geoffrey HINTON, « Deep Learning », *Nature*, nº 521, mai 2015, p. 436-444.
5. Cité dans Meredith BROUSSARD, « Artificial Unintelligence: How Computers Misunderstand the World », *The MIT Press*, 2018, p. 92.
6. Jerry KAPLAN, « Artificial Intelligence: What Everyone Needs to Know », Oxford University Press, 2016, p. 30.

déploie fait dire à certains, peut-être nostalgiques du GOFAI (*Good Old-Fashioned AI*), que l'apprentissage automatique ne relève tout simplement pas de l'*intelligence*[7]. En outre, les réseaux de neurones artificiels sont pour l'instant incapables de reconnaître une relation de causalité entre deux variables[8] et exigent un entraînement considérable pour s'acquitter de tâches cognitives qu'un jeune enfant peut accomplir aisément comme reconnaître un chat sur une photo ou saisir les implications de la loi de la gravité sur le plan du comportement des objets inanimés.

2. L'IA « FORTE » ET LA MENACE EXISTENTIELLE : UN POINT DE VUE DÉFLATIONNISTE

Pour un petit nombre de personnes influentes, le plus important risque éthique inhérent au développement de l'IA est la possibilité de la conception d'IA fortes égalant et même dépassant l'intelligence humaine[9]. C'est sans doute au philosophe Nick Bostrom que l'on doit l'explication la plus riche et détaillée de ce qu'il appelle le « risque existentiel » posé par l'IA[10, 11]. Actualisant le point de vue du statisticien britannique Irving J. Good sur la probabilité d'une « explosion » sur

7. Le chercheur en IA Hector Lévesque s'emploie ainsi, par différents exemples, à démontrer que la machine montre sa « stupidité » dès lors qu'elle doit faire preuve de sens commun pour répondre à des questions simples, mais inédites ou surprenantes telles qu'« un crocodile peut-il courir un steeple-chase ? ». La machine propose des équipes de sport ayant pour nom Crocodile ou Gators, mais n'arrive pas à comprendre le sens de la question et à y répondre correctement. La machine n'apprend que des vérités statistiques sur la base des données analysées. Il s'avère ainsi difficile d'anticiper sa réponse lorsqu'elle sera confrontée à une situation inattendue. Or, le raisonnement éthique consiste souvent en l'application de concepts abstraits comme des valeurs à des situations particulières, souvent inédites. Hector J. LÉVESQUE, « Common Sense, the Turing Test, and the Quest for Real AI », *The MIT Press*, 2018 ; Hector J. LÉVESQUE, « On our best behaviour », *University of Toronto*, 2013, en ligne : <http://www.cs.toronto.edu/~hector/Papers/ijcai-13-paper.pdf>.
8. Judea PEARL et Dana MACKENZIE, « The Book of Why: The New Science of Cause and Effect », *Penguin Books*, 2018 ; Judea PEARL, « Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution », *University of California Technical Report*, juillet 2018, en ligne : <http://ftp.cs.ucla.edu/pub/stat_ser/r475.pdf>.
9. Pour John Searle, une IA « forte » est dotée de l'équivalent d'un esprit humain conscient. Une IA « faible » simule la cognition humaine. John SEARLE, « Minds, Brains and Programs », *The Behavioral and Brain Sciences*, vol. 3, Cambridge University Press, 1980.
10. Nick BOSTROM, « Superintelligence: Paths, Dangers, Strategies », Oxford University Press, 2014, p. 140.
11. Max TEGMARK, « Life 3.0: Being Human in the Age of Artificial Intelligence », Knopf, 2017, 280 p.

le plan du développement de l'IA[12], Bostrom croit qu'il faut prendre au sérieux la possibilité de l'émergence de « superintelligences artificielles », largement plus puissantes que l'intelligence humaine, et que le contrôle de ces systèmes d'IA en vienne à échapper aux êtres humains[13]. Selon Bostrom, une IA, bien que générale et forte, pourrait poursuivre de façon trop mécanique et inflexible les buts qui lui auront d'abord été donnés par les programmeurs, si bien qu'il pourrait y avoir un conflit entre les buts de l'IA et les intérêts des êtres humains[14]. Un autre danger fréquemment évoqué est celui de systèmes d'IA fortes capables de programmer de façon autonome des algorithmes d'IA toujours plus performants, créant ainsi une distance supplémentaire entre l'IA et l'être humain. L'atout dans le jeu des êtres humains est, selon Bostrom, que « nous jouons en premier ». Selon lui, des recherches sont dès maintenant nécessaires pour que

12. « Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an « intelligence explosion, » and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control. » (Irving John GOOD, « Speculations Concerning the First Ultra-intelligent Machine », *Advances in Computers*, vol. 6, Academic Press, 1965) L'ingénieur chez Google Ray Kurzweil croit pour sa part qu'une IA forte et générale sera créée dans un horizon d'une douzaine d'années, mais que cela sera bénéfique pour l'humanité. La fusion entre l'humain et la machine – la « singularité » – inaugurera une nouvelle époque dans l'histoire de l'être humain. Voir Ray KURZWEIL, « The Singularity Is Near: When Humans Transcend Biology », Penguin Books, 2006, 651 p. Kurzweil et Bostrom sont tous les deux associés au mouvement transhumaniste. Pour une critique, voir Jean-Gabriel GANASCIA, « Le mythe de la singularité », Éditions du Seuil, 2017, 144 p.
13. Nick BOSTROM, *supra*, note 10, chapitres 6 et 9.
14. Bostrom insiste sur la nécessité d'inculquer une éthique à l'IA parce que, selon lui, aucun ordre ou objectif donné à un système d'IA n'est complètement inoffensif. Pour illustrer son propos, il décrit une expérience de pensée où l'humain donne à un système d'IA l'objectif de maximiser le nombre de trombones (*paperclips*) dans le monde. Ainsi orientée, la machine considérerait que pour atteindre son objectif, elle doit se défaire de tout ce qui pourrait l'arrêter. Sa première action serait donc de détruire tous les humains, s'assurant ainsi qu'ils ne l'empêchent pas d'atteindre son but en essayant de l'arrêter. En plus, les humains, composés d'atomes, pourraient même servir de matière à la fabrication de trombones. Leur anéantissement serait donc « logique » pour atteindre l'objectif de la machine. En réponse à la situation illustrée par cette expérience de pensée de Bostrom, Stuart Russell avance le problème de l'alignement des valeurs (*value alignment problem*). Celui-ci pose que pour éviter le genre de dérive décrite par Bostrom, on doit s'assurer que des valeurs concordantes aux valeurs des humains sont intégrées dans les machines d'IA dès leur développement (p. ex. : tu ne tueras point) pour que leurs objectifs et comportements respectent ces valeurs de base. Une lettre ouverte, portée par le Future of Life Institute et signée par plus d'une centaine de chercheurs en IA, explique et appuie cette proposition. Voir en ligne : <https://futureoflife.org/ai-open-letter/>.

des règles – s'inspirant peut-être des lois d'Asimov[15] – soient opérationnalisées et que des mesures de sécurité soient mises en place. Cherchant visiblement l'effet dramatique, Bostrom ajoute toutefois que notre désavantage est que « nous n'aurons peut-être pas la chance de jouer un deuxième tour ». C'est sur la base de raisonnements de ce genre que l'astrophysicien Stephen Hawking a déclaré :

> [j]e pense que le développement d'une intelligence artificielle complète pourrait mettre fin à l'humanité. Une fois que les humains auraient développé l'IA, celle-ci décollerait seule, et se redéfinirait de plus en plus vite. Les humains, limités par une lente évolution biologique, ne pourraient pas rivaliser et seraient dépassés.[16]

Des entrepreneurs et investisseurs comme Elon Musk et Bill Gates, le physicien Max Tegmark et l'intellectuel public Sam Harris partagent aussi le point de vue de Hawking.

Bien que cela exigerait une démonstration plus étayée que celle que nous pouvons proposer ici, cette perspective inflationniste nous semble injustifiée. Disons simplement que le fait qu'un scénario soit concevable ne rend pas son occurrence probable et vraisemblable. Comme nous l'avons vu plus haut, l'apprentissage automatique et les réseaux de neurones artificiels permettent aux machines de repérer des corrélations dans de vastes jeux de données et de généraliser, mais il est difficile de voir comment cette approche peut permettre la création d'agents parfaitement autonomes, soucieux de leur auto-préservation et capables d'actions mettant en péril la civilisation humaine et ses institutions. Non seulement rien ne permet de penser que quelque chose comme une conscience pourrait émerger en l'absence de tout substrat biologique, mais rien ne laisse croire qu'il soit possible de dissocier l'intelligence multidimensionnelle et générale comme celle démontrée par les êtres humains du corps dans sa totalité, des émotions, des désirs, des relations sociales et de l'inscription au sein de ce que les phénoménologues appellent un

---

15. En 1942, Isaac Asimov énonçait dans sa nouvelle « Runaround », les trois lois de la robotique :
    1. un robot ne peut porter atteinte à un être humain ni, en restant passif, permettre qu'un être humain soit exposé au danger ;
    2. un robot doit obéir aux ordres qui lui sont donnés par un être humain, sauf si de tels ordres entrent en conflit avec la première loi ;
    3. un robot doit protéger son existence tant que cette protection n'entre pas en conflit avec la première ou la deuxième loi.
16. Rory CELLAN-JONES, « Stephen Hawking warns artificial intelligence could end mankind », *BBC News: Technology*, 2 décembre 2014, en ligne : <https://www.bbc.com/news/technology-30290540>.

« monde vécu ». Les inflationnistes doivent nous donner des raisons de prendre ce scénario au sérieux, ce qu'ils n'ont pas réussi à faire jusqu'ici. C'est pour des raisons de ce genre que le point de vue que nous adoptons dans ce texte est déflationniste.

La perspective que les machines dotées d'IA dominent le monde et anéantissent l'humanité nous apparaît donc trop improbable pour être centrale dans la réflexion éthique et juridique. Il n'en demeure pas moins que l'IA, de par son mode de conception et de fonctionnement, entraîne des risques. Ces risques soulèvent des enjeux éthiques parce qu'ils mettent en péril des valeurs se trouvant au fondement de l'État de droit démocratique. C'est donc sur ces risques que nous souhaitons porter une attention dans le reste du texte.

Avant de ce faire, notons toutefois une difficulté d'ordre général pour les chercheurs et ingénieurs qui œuvrent à rendre les systèmes d'IA « éthiques ». Plusieurs affirment qu'il faut apprendre aux machines à devenir éthiques en leur inculquant les valeurs que nous croyons qu'elles doivent respecter. En plus de nous rapprocher de l'approche symbolique critiquée par les défenseurs de l'apprentissage automatique, ce projet d'éducation morale des machines semble destiné à se buter à ce que l'on pourrait appeler le « problème aristotélicien/wittgensteinien ». Les valeurs sont des raisons d'agir, parfois abstraites et générales, devant guider l'action ou la décision. Or, l'un des enseignements les plus durables de la théorie aristotélicienne du raisonnement pratique, et en particulier du raisonnement éthique, concerne le rapport entre les normes abstraites et les situations concrètes et singulières. Les principes généraux sont nécessaires, mais il est parfois difficile d'identifier la meilleure façon de les appliquer dans des situations particulières inédites. Pour Aristote, si le législateur doit élaborer de bonnes lois d'application générale, le décideur doit faire preuve d'une forme de sagesse pratique (*phronesis*) permettant de les appliquer d'une façon telle que leur esprit sera respecté, et parfois permettre les exceptions. Si on peut programmer un algorithme pour qu'il respecte des règles comme immobiliser un véhicule à un panneau d'arrêt, comment lui inculquer un sens commun et une sagesse pratique lui permettant de faire les bonnes inférences dans des situations spécifiques ? Que faire si le sens d'une indication routière est vague ou ambiguë[17], ou si deux indications rou-

17. « A rule stands there like a sign-post. Does the sign-post leave no doubt open about the way I have to go? Does it show which direction I am to take when I have passed it; whether along the road or the footpath or cross-country? But where is it said which way I am to follow it; whether in the direction of its finger or (e.g.) in the opposite one? And if there were, not a single sign-post, but a chain of adjacent ones or of chalk marks on the ground; is there only one way of interpreting them?

tières semblent à première vue se contredire ? Comme Wittgenstein l'a avancé, l'activité qui consiste à interpréter des règles et des concepts n'est pas toujours elle-même régulée par des règles[18] ; il faut parfois décider en fonction de notre meilleur jugement, aidé de notre « sens commun ». Le problème est encore plus grand lorsque les règles sont des raisons d'agir abstraites comme le sont les valeurs.

3. À QUI LA FAUTE ? LA RESPONSABILITÉ DES ALGORITHMES

Les machines dotées d'IA sont entraînées pour prendre des décisions, faire des choix, agir. Par exemple, lorsqu'on entraîne une machine à distinguer les essences de bois par la reconnaissance d'image, elle sera en mesure de reconnaître et ainsi trier les billots qui lui seront présentés selon leur essence. C'est ce même type de mécanisme de reconnaissance d'image qui sert à trier à haute vitesse et avec beaucoup de précision des images de strates de cerveau et d'y repérer, le cas échéant, des tumeurs cancéreuses. La machine agit avec une certaine autonomie, en extrapolant à partir de recoupements d'information. Dans la majorité des cas, pour l'instant, il est envisagé d'utiliser des systèmes d'IA qui agiront en complémentarité avec l'humain, c'est-à-dire en faisant des recommandations qui devraient nécessairement être validées par des humains. Il est toutefois possible que plus l'utilisation deviendra régulière et les résultats probants, plus on aura tendance à se fier aux décisions que ces machines prendront. De plus, dans d'autres cas, par exemple celui des véhicules autonomes, il est envisagé de conférer une autonomie décisionnelle complète aux machines.

Cela soulève ainsi la question de l'imputabilité pour les décisions prises par des agents artificiels[19]. En effet, qui sera responsable des mauvaises décisions prises par des systèmes d'IA et des inévitables défaillances technologiques[20] ? Par exemple, qui devrait porter

---

So I can say, the sign-post does after all leave no room for doubt. Or rather: it sometimes leaves room for doubt and sometimes not. And now this is no longer a philosophical proposition, but an empirical one. » (Ludwig WITTGENSTEIN, « Philosophical Investigations », Wiley-Blackwell, 4e ed., 2009, par. 85).

18. *Ibid.*, par. 84.
19. Virginia DIGNUM, « Responsible Artificial Intelligence: Designing AI for Human Values », *ITU Journal: ICT Discoveries*, Special Issue no 1, septembre 2017, p. 4-5, en ligne : <https://www.itu.int/en/journal/001/Pages/01.aspx>.
20. Les systèmes d'IA présentent également un risque qui pourrait provoquer des défaillances, soit celui d'être vulnérable aux cyberattaques, malgré toutes les précautions prises par les manufacturiers. Cette vulnérabilité aux piratages des systèmes et donc aux cyberattaques pose de graves problèmes de sécurité, de protection de la vie privée et des renseignements personnels, en plus de soulever

la responsabilité d'un accident imputable à un véhicule autonome ? L'équipe d'ingénieurs derrière la création des algorithmes décisionnels ? Ceux en charge de l'entraînement de la machine et de la préparation des données ? Ceux qui ont créé les senseurs qui devraient permettre au véhicule de percevoir correctement son environnement et de se diriger ? Le constructeur automobile qui a mis en marché le véhicule ? Le propriétaire, qui doit assumer une partie des risques inhérents à l'utilisation de son véhicule ? Le gouvernement qui a mis en œuvre un cadre réglementaire possiblement lacunaire ? Nous nous trouvons ici devant une série de questions qui exigent potentiellement l'élaboration d'une nouvelle conception de la responsabilité morale et juridique, une responsabilité adaptée aux machines autonomes et apprenantes, ainsi qu'au cadre réglementaire en vigueur[21,22].

4. LORSQUE LA MACHINE EST BIAISÉE. JUSTICE, ÉQUITÉ ET DISCRIMINATION

L'IA « apprend » à l'aide des données qu'elle doit traiter. Les données qui nourrissent l'algorithme influencent les résultats et décisions de l'IA, créant parfois des biais potentiellement discriminatoires à l'encontre des membres de certains groupes. Les biais peuvent être introduits par deux sources : par la programmation ou par les données utilisées pour l'apprentissage.

Les données sont triées par un algorithme programmé pour une certaine fonction. L'équipe d'ingénieurs entraîne l'algorithme pour accomplir certaines tâches, le menant à des réponses que les ingénieurs considèrent adéquates en vue du travail que l'algorithme doit accomplir (sa fonction d'utilité) et, souvent de façon inconsciente, de leurs croyances et valeurs. Les algorithmes peuvent ainsi assimiler ou reproduire les valeurs des équipes de programmeurs[23].

---

des questions sur la responsabilité. Il ne semble pas possible de blinder un système électronique contre toutes les attaques et les enjeux peuvent être dramatiques si ces attaques visent des infrastructures névralgiques qui miseront sur l'IA, comme le réseau électrique des villes ou les hôpitaux.

21. Luciano FLORIDI, « Faultless responsibility: on the nature and allocation of moral responsibility for distributed moral actions », *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 28 décembre 2016, p. 374.
22. Louis CHARTRAND, « Agencéité et responsabilité des agents artificiels », Éthique Publique, vol. 19, nº 2, 2017, en ligne : <https://journals.openedition.org/ethiquepublique/3068>.
23. Brent Daniel MITTELSTADT, Patrick ALLO, Mariarosaria TADDEO, Sandra WACHTER et Luciano FLORIDI, « The ethics of algorithms: Mapping the debate », *Big Data & Society*, juillet-décembre 2016, p. 7, en ligne : <http://journals.sagepub.com/doi/abs/10.1177/2053951716679679>.

Il y a biais de programmation lorsque ces valeurs reproduites par l'algorithme incluent des croyances produisant de l'exclusion, de la stigmatisation ou de la discrimination[24]. Dans certains cas, les biais de programmation sont introduits volontairement. C'était le cas notamment de l'algorithme de livraison d'Amazon, qui n'offrait pas le service 24h dans les quartiers à prédominance noire, tels que le Bronx ou Roxbury[25]. Les raisons derrière cette exclusion d'un territoire de livraison peuvent être diverses, mais le résultat n'en est pas moins discriminatoire, soulevant une question d'équité. À la suite de la publication de l'article de Bloomberg, le maire de la ville de Boston a exigé d'Amazon que le service 24h soit disponible sur l'entièreté du territoire, y inclus le quartier noir de Roxbury.

Dans d'autres cas, les biais sont des résultats fortuits, introduits involontairement. C'est le cas par exemple lorsque le traducteur en ligne de Google ajoutait en anglais un pronom masculin (« he ») au mot médecin et un pronom féminin (« she ») à infirmière, alors que la langue d'origine n'avait pas de pronoms genrés[26]. Cette introduction d'un biais sexiste, bien qu'involontaire, reproduit un stéréotype qui renvoie à une hiérarchie entre les genres, renforçant l'inégalité entre les hommes et les femmes où les femmes ont une valeur moindre – ici un métier moins bien rémunéré que les hommes. Une piste de solution évoquée pour éviter les biais de programmation involontaires est de diversifier les équipes de conception des algorithmes afin d'y inclure des personnes de différents groupes et de multiplier les points de vue. Des formes d'audit sont présentement élaborées afin de détecter les biais algorithmiques volontaires et involontaires[27].

Aussi, bien que l'algorithme soit capital dans le traitement des données, ces dernières sont fondamentales à son entraînement et fonctionnement. En fait, un algorithme d'apprentissage automatique moins performant, mais lié à une base de données massives importante sera souvent plus précis et fiable que le meilleur algorithme avec une base de données restreinte. Ce sont les données, la

---

24. Alex CAMPOLO, Madelyn SANFILIPPO *et al.*, « AI Now 2017 Report », p. 16, en ligne : <https://ainowinstitute.org/AI_Now_2017_Report.pdf>.
25. David INGOLD et Spencer SOPER, « Amazon Doesn't Consider the Race of Its Customers. Should It? », *Bloomberg*, avril 2016, en ligne : <https://www.bloomberg.com/graphics/2016-amazon-same-day/>.
26. Mathew HUTSON, « Even artificial intelligence can acquire biases against race and gender », *Science*, avril 2017, en ligne : <http://www.sciencemag.org/news/2017/04/even-artificial-intelligence-can-acquire-biases-against-race-and-gender>.
27. Jessi HEMPEL, « Want to prove your business is fair? Audit your algorithm », *Wired*, 9 mai 2018, en ligne : <https://www.wired.com/story/want-to-prove-your-business-is-fair-audit-your-algorithm/>.

variété et la qualité de celles-ci qui rendent l'algorithme capable d'un meilleur discernement. Des données peu nombreuses ou relatant des pratiques discriminatoires peuvent reproduire des biais ou en créer, par exemple, en faisant des corrélations entre des éléments qui ne devraient pas être liés[28]. On a notamment repéré des biais discriminatoires dans un algorithme créé pour évaluer les potentialités de récidives de criminels et leur admissibilité à une libération conditionnelle. L'algorithme employé pour traiter ces demandes évaluait systématiquement les demandes provenant d'Afro-Américains comme présentant des risques de récidive supérieurs à la réalité, alors qu'il minimisait – à tort – les risques de récidive des Caucasiens, notamment parce que les données ayant servi à l'apprentissage de l'algorithme faisaient ressortir un plus haut taux de criminalité dans les communautés noires[29]. Similairement, un algorithme utilisé pour trier des CV dans une firme de recrutement sélectionnait des hommes blancs dans une proportion beaucoup plus grande que des femmes ou des hommes de couleur pour pourvoir aux postes clés. Ici encore, les données fournies à l'algorithme étaient basées sur l'historique de recrutement de grandes organisations où, traditionnellement, des hommes blancs occupaient les postes de pouvoir. Les biais algorithmiques peuvent ainsi perpétuer des injustices déjà existantes, menant à des impacts négatifs sur le plan du respect des principes d'égalité ou d'équité entre les personnes.

Enfin, la complexité de l'enjeu est encore plus grande lorsque les données qui sont fournies à la machine proviennent des interactions avec les utilisateurs et permettent à l'algorithme de continuer à apprendre en cours d'utilisation réelle. C'est le cas par exemple de Netflix, qui adapte ses propositions de contenu en fonction des choix faits par l'utilisateur. Plus celui-ci utilise la plateforme, plus les propositions s'affinent pour correspondre aux contenus déjà visionnés. Cette manière de faire soulève des problèmes auxquels nous reviendrons plus loin, mais notons seulement ici que, dans ce cas, la plateforme reproduit en quelque sorte les préférences de l'utilisateur, personnalisant son expérience. Certaines issues sont toutefois moins heureuses. Tay, le chatbot de Microsoft, n'aura été sur Twitter que 16 heures avant d'être désactivé. Prévue pour « devenir plus savante au fil des interactions », elle a plutôt subi des attaques concertées par des internautes qui l'ont, par leurs interventions, rendue raciste,

28. Brent Daniel MITTELSTADT, Patrick ALLO, Mariarosaria TADDEO, Sandra WACHTER et Luciano FLORIDI, *supra*, note 23, p. 5-8.
29. Andréa FRADIN, « États-Unis : un algorithme qui prédit les récidives lèse les Noirs », *Le Nouvel Observateur*, mai 2016, en ligne : <https://www.nouvelobs.com/rue89/rue89-etats-unis/20160524.RUE2964/etats-unis-un-algorithme-qui-predit-les-recidives-lese-les-noirs.html>.

antisémite et misogyne[30]. Mais comment prévenir la manipulation des IA lorsqu'elles sont placées en contexte réel ? Comment s'assurer que les algorithmes ne subissent pas une influence indue qui viendrait modifier la trajectoire décisionnelle souhaitée, introduisant des biais nocifs ? Ce sont notamment des enjeux de confiance et de fiabilité dans les décisions prises par les machines qui sont ici mises en cause. Peut-être serait-il souhaitable que les systèmes qui utilisent l'IA pour prendre des décisions fonctionnent à l'intérieur de balises où des mécanismes de surveillance sont prévus avant l'exécution de ces décisions.

5. LE PROBLÈME DE LA BOÎTE NOIRE : L'INTERPRÉTABILITÉ DES DÉCISIONS DES SYSTÈMES D'IA

Rendre les processus décisionnels des algorithmes d'apprentissage automatique plus transparents, compréhensibles et corrigibles est l'un des grands défis des chercheurs de l'IA[31]. Les algorithmes issus du paradigme de réseaux de neurones artificiels et l'apprentissage automatique, et en particulier de l'apprentissage profond, sont par définition opaques. Il est généralement impossible, même pour l'ingénieur qui a programmé la machine, d'expliquer le chemin pris par la machine pour arriver à une conclusion particulière. Le mathématicien et député Cédric Villani, qui s'est vu confier la mission d'éclairer le gouvernement français sur l'IA, le confirme dans son rapport « Donner un sens à l'IA : Pour une stratégie nationale et européenne » :

> Une grande partie des considérations éthiques soulevées par l'IA tiennent à l'opacité de ces technologies. […] il est souvent très difficile d'expliquer leurs décisions de manière intelligible par le commun des mortels. C'est le fameux problème de la boîte noire : des systèmes algorithmiques dont il est possible d'observer les données d'entrée (input), les données de sortie (output) mais dont on comprend mal le fonctionnement interne.[32]

30. Elle HUNT, « Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter », *The Guardian*, mars 2016, en ligne : <https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter>.
31. Virginia DIGNUM, *supra*, note 19, p. 6.
32. Cédric VILLANI, Marc SCHOENAUER, Yann BONNET, Charly BERTHET, Anne-Charlotte CORNUT, François LEVIN et Bertrand RONDEPIERRE, « Donner un sens à l'intelligence artificielle : pour une stratégie nationale et européenne », France, mission confiée par le premier ministre Édouard Philippe, p. 140-141, en ligne : <https://www.aiforhumanity.fr/pdfs/9782111457089_Rapport_Villani_accessible.pdf>.

Une certaine opacité est parfois souhaitée par l'industrie pour préserver le secret commercial, l'exclusivité d'un contenu. En effet, lorsqu'une entreprise développe un système d'algorithme pour lequel elle espère tirer des profits, il va de soi qu'elle cherche à en préserver la propriété intellectuelle[33]. Reste que lorsque la machine est destinée à prendre des décisions qui ont un impact direct et déterminant sur la vie des gens (accorder ou non un prêt, sélectionner des candidats pour une entrevue d'embauche, permettre ou non une libération conditionnelle, poser un diagnostic, etc.), il est essentiel de pouvoir faire confiance aux décisions de la machine qui traitera les données. Villani encore :

> […] l'explicabilité des systèmes à base d'apprentissage constitue donc un véritable défi scientifique qui met en tension notre besoin d'explication et notre souci d'efficacité.[34]

Mais expliquer, diagnostiquer et corriger ce qui pose problème dans l'algorithme est souvent impossible : la structure sous-jacente qui conduit la machine à un résultat est complexe, en particulier lorsqu'il s'agit de réseaux de neurones artificiels « profonds », c'est-à-dire comptant plusieurs couches reliées de « nœuds » qui traitent des informations différentes et qui les agrègent ensuite. Contrairement aux systèmes d'IA raisonnant de façon logico-déductive, il n'est pas possible d'extraire un arbre de décision clair et logiquement cohérent. Plusieurs s'entendent toutefois sur la nécessité d'exiger une explicabilité ou une interprétabilité des algorithmes, qui est fondamentale à la confiance des utilisateurs[35,36]. Des solutions pour assurer la traçabilité des calculs effectués cherchant à rendre la « boîte noire » plus transparente sont avancées. Certains proposent que, dès la conception de l'algorithme, celui-ci soit en mesure d'expliquer son processus décisionnel en même temps qu'il émet sa réponse[37,38]. Pour d'autres, il faut envisager que la machine puisse être soumise à une

33. *Ibid.*, p. 143.
34. *Ibid.*, p. 141.
35. « For artificial intelligence to thrive, it must explain itself », *The Economist*, février 2018, en ligne : <https://www.economist.com/science-and-technology/2018/02/15/for-artificial-intelligence-to-thrive-it-must-explain-itself>.
36. Nicolas BOUSQUET, Gill MORISSE et Jean-Mathieu SCHERTZER, « Un aspect fondamental du rapport Villani : l'explicabilité de l'IA », *Quantmetry*, avril 2018, en ligne : <https://www.quantmetry.com/single-post/2018/04/26/aspect-fondamental-du-rapport-Villani-explicabilite-intelligence-artificielle>.
37. David GUNNING, « Explainable Artificial Intelligence (XAI) », *DARPA*, en ligne : <https://www.darpa.mil/program/explainable-artificial-intelligence>.
38. Ilknur Kaynar KABUL, « Interpretability is crucial for trusting AI and machine learning », SAS, décembre 2017, en ligne : <https://blogs.sas.com/content/subconsciousmusings/2017/12/18/interpretability-crucial-trusting-ai-machine-learning/>.

forme d'audit certifié en cas de litige sur la décision[39,40], soit une forme « d'interrogatoire » où la machine serait soumise à diverses variables en intrants, pour vérifier la validité des extrants. Cependant, des chercheurs affirment qu'il est aussi difficile et coûteux de tenter de comprendre et d'expliquer le fonctionnement de ce qui est fait présentement en IA que de travailler à faire de nouvelles découvertes. Les ressources en temps et en argent pour créer et déployer ces systèmes d'audit seraient considérables et moins profitables que la conception d'un nouvel algorithme d'IA.

6. À QUOI CONSENTONS-NOUS EXACTEMENT ? AUTONOMIE ET PROTECTION DE LA VIE PRIVÉE

Comme nous l'avons vu, les progrès en IA reposent fortement sur l'accès aux données numériques, incluant des données personnelles. Or, l'hébergement, le partage et l'analyse des données personnelles posent des risques sur le plan de la protection de la vie privée des producteurs de données. Il est déjà possible en recoupant les données partagées, même anonymisées, et l'information disponible publiquement au sujet d'une personne, de connaître plusieurs éléments de sa vie privée (état de santé, allégeance politique, habitudes d'achat, salaire, âge, lieux fréquentés, adresse, propriétés immobilières, etc.). Comme l'a montré l'affaire impliquant la firme Cambridge Analytica, nous savons qu'en combinant ces divers éléments, des algorithmes peuvent, à travers les réseaux sociaux ou les abonnements aux médias, mieux cibler le public susceptible d'être réceptif aux contenus promus (nouvelles, concours, publicités, etc.)[41]. Cette manière de cibler l'information formatée et transmise spécifiquement à une personne en fonction des intérêts qu'elle a déjà manifestés pourrait, nous l'avons dit précédemment, être perçue comme une forme de biais souhaité par l'utilisateur pour correspondre à ses goûts. Cela a toutefois également pour conséquence que l'utilisateur ne reçoit plus que ce qui confirme ses convictions, créant ainsi une sorte de « bulle informationnelle »[42]. N'étant alors plus confrontée à des idées différentes

39. Alice PAVALOIU et Utku KOSE, « Ethical Artificial Intelligence – An Open Question », *Journal of Multidisciplinary Developments*, vol. 2, nº 2, 2017, p. 20, en ligne : <https://arxiv.org/pdf/1706.03021>.
40. Cédric VILLANI, *supra*, note 32, p. 143.
41. Eitan D. HERSH, « Hacking the electorate: How campaigns perceive voters », Cambridge University Press, 2015.
42. Une « bulle informationnelle », « bulle d'information numérique » ou « bulle de filtre » est la situation dans laquelle se trouve une personne dont l'environnement numérique ne lui présente que certaines informations filtrées, généralement sur la base de ses préférences inférées à partir de son comportement en ligne et de ses interactions avec les autres utilisateurs. Ce phénomène renforce les convic-

ou divergentes qui pourraient faire évoluer sa pensée, cette façon de faire vient alors renforcer sa position au point d'en faire disparaître les nuances. La vision globale d'une situation est alors balisée de sorte que la personne, évoluant dans un éventail de choix restreints, peut plus difficilement faire preuve d'ouverture d'esprit et de pensée critique. On peut ici penser aux opinions politiques confortées par les articles de journaux reçus partageant les mêmes orientations, aux publicités ciblées en fonction de nos habitudes d'achat ou de la période de paye (on fera la promotion d'items plus chers le jour de la paye, par exemple), ou encore à Netflix et Amazon qui proposent des séries, films, documentaires et des livres en fonction des préférences déjà exprimées, rendant ainsi la découverte de contenus variés plus difficile.

Ainsi, l'utilisation qui est faite de l'information personnelle enregistrée par toutes les applications et outils Web a des impacts sur l'autonomie de l'humain, sur ses choix disponibles et conséquemment, sur ses décisions. Pour pallier ces utilisations « abusives », il est demandé aux utilisateurs de protéger leurs données, de lire les directives de consentement, de s'assurer d'être vigilants dans leur utilisation des outils et applications Web. La responsabilité de protéger ses données est attribuée à l'utilisateur, qui n'a généralement pas la détermination et les connaissances requises pour comprendre les formulaires de consentement et acquérir les compétences technologiques pour le faire[43]. Le pouvoir de l'utilisateur est mince en comparaison de celui de l'industrie qui collecte et monnaye les données. Cette asymétrie des pouvoirs et des connaissances rend quasi impossible la protection des données, et incite plusieurs analystes à prôner une réglementation juridique plus serrée de l'utilisation des données personnelles.

7. UNE RECONFIGURATION MAJEURE DU MARCHÉ DU TRAVAIL ?

À toutes ces préoccupations liées au fonctionnement des systèmes d'IA fondés sur l'apprentissage automatique s'ajoute une grande incertitude quant à l'impact qu'aura le déploiement de l'IA sur le marché du travail. Ce déploiement devrait certes augmenter la productivité, mais pourrait aussi aggraver des injustices sociales existantes. Les experts ne s'entendent pas sur la portée et le degré de

tions et les biais des utilisateurs en ne les exposant pas à ce qui diffère de leurs orientations idéologiques.

43. Alex CAMPOLO, Madelyn SANFILIPPO *et al.*, *supra*, note 24, p. 30.

disruption qu'entraînera l'arrivée des machines pour accomplir des tâches de plus en plus complexes. Peut-on espérer une augmentation de la productivité à un moindre coût, un remplacement d'emplois aliénants et répétitifs par des emplois plus stimulants pour l'humain, tout cela pour une augmentation substantielle de la qualité de vie des travailleurs ? Ou doit-on plutôt craindre la disparition rapide d'emplois, c'est-à-dire un « chômage technologique » qui entraînerait une augmentation du nombre de personnes sans emploi, une précarité croissante des travailleurs, un déséquilibre entre les entreprises qui pourront s'automatiser rapidement et celles qui n'auront pas accès à la technologie, et une répartition de la richesse encore plus contrastée que celle qui prévaut actuellement ? Cela demeure difficile à prévoir. Les analystes élaborent des scénarios allant d'une transformation autorégulatrice à un changement radical s'apparentant à une Quatrième Révolution industrielle.

8. TRANSFORMATION AUTORÉGULATRICE

Le premier scénario ne prévoit pas de transformation brusque du monde du travail par l'IA, avançant plutôt l'idée que le marché de l'emploi et l'économie devraient se réguler d'eux-mêmes. L'IA ferait certes disparaître des emplois, mais en créerait d'autres, plus stimulants et payants. La majorité des emplois survivrait, seules les tâches répétitives de certains emplois seraient exécutées par des machines, laissant à l'humain les tâches faisant appel à la créativité, au sens commun et à la sensibilité. La transformation vers une plus grande place des machines se ferait à un rythme régulier et prévisible dans certains secteurs précis, permettant de voir venir la vague et de mettre en place des structures d'accompagnement pour soutenir les organisations et les travailleurs[44]. Enfin, certaines tâches exigeant de l'empathie, de la créativité ou une forme de pensée critique, des capacités que l'on considère propres aux humains, pourraient ne pas pouvoir être accomplies par les machines[45]. Les métiers requérant ces attributs ne seraient donc pas menacés par les transformations liées à l'automatisation.

---

44. Matthias OSCHINSKI et Rosalie WYONCH, « Le choc du futur ? Les répercussions de l'automatisation sur le marché du travail au Canada », *Institut C.D. Howe*, mars 2017, p. 14, en ligne : <https://www.cdhowe.org/public-policy-research/le-choc-du-futur-les-r %C3 %A9percussions-de-lautomatisation-sur-le-march %C3 %A9-du-travail-au-Canada>.
45. PEW RESEARCH CENTER, « AI, Robotics, and the Future of Jobs », Août 2014, p. 7, en ligne : <http://www.pewinternet.org/2014/08/06/key-insights-expert-views-on-artificial-intelligence-robotics-and-the-future-of-jobs/>.

9. VERS UNE QUATRIÈME RÉVOLUTION INDUSTRIELLE ?

Le scénario plus pessimiste pose plutôt que les avancées en IA entraîneront une Quatrième Révolution industrielle qui aurait des effets dévastateurs sur l'emploi. Pour Klaus Schwab, fondateur et président exécutif du Forum économique mondial de Genève, les bouleversements anticipés seront nettement plus dommageables que les révolutions industrielles précédentes en raison de la rapidité avec laquelle les changements s'opéreront, du type d'emplois menacés par l'automatisation intelligente et aussi parce que les systèmes d'IA viendront modifier la structure même de l'emploi[46].

Contrairement aux précédentes révolutions industrielles qui visaient principalement l'automatisation d'emplois du secteur tertiaire, cette Quatrième Révolution industrielle vise des secteurs d'emplois variés ; de l'analyste au médecin, du traducteur au journaliste. Il est possible que certains secteurs d'emploi soient plus sévèrement touchés par la transition et la rapidité de celle-ci, permettant difficilement une reconversion de tous les employés et accentuant les inégalités sociales. Le filet social disponible pour les personnes (et leurs familles) œuvrant dans les secteurs les plus sensibles ne sera peut-être plus suffisant ou adéquat pour prévoir et mitiger les effets néfastes des pertes d'emplois massives dans une région ou un secteur d'emploi précis.

10. UN ENTRE-DEUX

La réalité se situera peut-être entre les deux. Plusieurs experts considèrent en effet qu'on ne devrait pas tant craindre la disparition d'emplois autant que la modification des tâches par l'intégration d'IA ou l'automatisation. Selon le McKinsey Global Institute[47], en utilisant la technologie disponible en 2016, il est possible d'automatiser 45 % des tâches faites par les humains, et 60 % des emplois pourraient voir plus de 30 % de leurs tâches faites par des machines. Dans la même veine, l'OCDE estime que 9 % des emplois présentent un risque élevé

46. Klaus SCHWAB, « The Fourth Industrial Revolution: what it means, how to respond », *World Economic Forum*, janvier 2016, en ligne : <https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond>.

47. « Jobs lost, jobs gained: workforce transitions in a time of automation », McKinsey Global Institute, décembre 2017, en ligne : <https://www.mckinsey.com/~/media/McKinsey/Global%20Themes/Future%20of%20Organizations/What%20the%20future%20of%20work%20will%20mean%20for%20jobs%20skills%20and%20wages/MGI-Jobs-Lost-Jobs-Gained-Report-December-6-2017.ashx>.

d'automatisation[48]. Bien que moins perturbatrice que la prévision précédente, celle-ci soulève tout de même certains risques. En effet, dans la majorité des secteurs, les projections montrent que l'humain devra apprendre à concilier son travail à celui des machines qui auront pris le relais pour une partie plus ou moins grande de ses tâches. Le travail de l'humain deviendrait alors complémentaire à celui des machines qui serait, lui aussi, complémentaire à celui de l'humain. Ainsi, les définitions de tâches étant modifiées, les aptitudes requises avant et après l'intervention de l'IA ne seront pas les mêmes. Si l'on peut penser que le travail de la machine pourra dans plusieurs cas permettre à l'employé de se consacrer à des tâches plus gratifiantes, on ne peut exclure qu'il ne soit plus en mesure de maîtriser ou d'être performant dans son propre travail[49], voire ne pas aimer les tâches qui lui incomberont désormais. Devoir « partager » son expertise avec une machine, ne plus être complètement adéquat ou stimulé par le travail pour lequel il a été formé, devoir faire des mises à niveau régulières pour se réapproprier son métier sont quelques-uns des éléments qui peuvent créer de l'incertitude chez le travailleur.

CONCLUSION

De nombreux regroupements de praticiens, chercheurs et industriels de l'IA font paraître des rapports insistant sur la nécessité d'une vigilance éthique accrue dans le développement de l'IA. C'était le cas notamment pour l'Institute of Electrical and Electronics Engineers (IEEE), qui publiait en 2016 un rapport[50] offrant des recommandations pour « privilégier le bien-être de l'humain dans le développement et l'utilisation de l'IA et des machines ». Un autre regroupement encadré par le Future of Life Institute publiait au début de 2017 les 23 principes d'Asilomar[51] ayant pour but de servir

48. C'est-à-dire que 70 % des tâches composant ces métiers peuvent être faites par des machines. En deçà de 70 % de tâches automatisables, on estime que l'emploi serait conservé et plus ou moins redéfini. « Automatisation et travail indépendant dans une économie numérique », *OCDE*, mai 2016, en ligne : <https://www.oecd.org/fr/els/emp/Automatisation%20et%20travail%20ind%C3%A9pendant%20dans%20une%20%C3%A9conomie%20num%C3%A9rique.pdf>.
49. « Skill shift, automation and the future of the workforce », McKinsey Global Institute, *McKinsey & Company*, mai 2018, en ligne : <https://www.mckinsey.com/~/media/McKinsey/Featured%20Insights/Future%20of%20Organizations/Skill%20shift%20Automation%20and%20the%20future%20of%20the%20workforce/MGI-Skill-Shift-Automation-and-future-of-the-workforce-May-2018.ashx>.
50. « Ethically Aligned Design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems », IEEE: Creative Commons, décembre 2016, en ligne : <http://standards.ieee.org/develop/indconn/ec/ead_v1.pdf>.
51. « Asilomar AI Principles », *Future of life Institute*, 2017, en ligne : <https://futureoflife.org/ai-principles/?cn-reloaded=1>.

de base pour l'élaboration d'un guide de référence pour le développement éthique de l'IA. Plus près de nous, en octobre 2017 se tenait à Montréal le Forum sur le développement socialement responsable de l'IA. De celui-ci est sortie la Déclaration pour l'IA responsable, laquelle est discutée, augmentée et bonifiée, dans un processus itératif de coconstruction dont un premier bilan des discussions est paru en juin 2018[52]. Aussi, une initiative conjointe des Fonds de recherche du Québec (FRQ) et du ministère de l'Économie, de la Science et de l'Innovation (MESI) est en développement. Elle vise à mettre sur pied un Observatoire international portant spécifiquement sur les impacts sociétaux de l'IA et du numérique[53]. Enfin, des partenariats bilatéraux sur l'IA ont été conclus entre la France et le Canada ainsi qu'entre le Québec et la France. En effet, par une lettre d'intention commune sur l'IA, la France et le Canada s'engagent à créer un Groupe international d'Étude sur l'IA qui aurait pour objectif de fournir une expertise internationalement reconnue du sujet[54]. Dans une Déclaration d'intention conjointe, le Québec et la France s'engagent, quant à eux, à renforcer les partenariats et ententes rendus possibles par l'Accord économique et commercial global entre l'Union européenne et le Canada. Le Québec et la France ont également convenu de se mobiliser de concert sur une approche visant à « évaluer de façon méthodique et sans parti pris les informations d'ordre éthique, scientifique, technique et socio-économique essentielles au développement harmonieux de la technologie » lors de la prochaine conférence Neural Information Processing Systems (NIPS) qui aura lieu à Montréal en décembre 2018[55].

Ces diverses initiatives semblent indiquer une véritable volonté des parties prenantes d'encadrer adéquatement le développement de l'IA. Pour l'instant, toutefois, cette réflexion multiforme n'a pas

52. « Bilan des délibérations citoyennes de la déclaration de Montréal IA responsable », Déclaration de Montréal, IA responsable, juin 2018, en ligne : <https://docs.wixstatic.com/ugd/ebc3a3_b3b089b2f8174625a12e60051560f7c3.pdf>.
53. « Appel à propositions : création d'un Observatoire international sur les impacts sociétaux de l'intelligence artificielle et du numérique », Scientifique en chef, 4 mai 2018, en ligne : <http://www.scientifique-en-chef.gouv.qc.ca/nouvelles/appel-a-propositions-creation-dun-observatoire-international-impacts-societaux-de-lintelligence-artificielle-numerique/>.
54. « Le Canada et la France signent une lettre d'intention commune sur l'intelligence artificielle », Gouvernement Français, Services culturels français, 28 juin 2018, en ligne : <https://francecanadaculture.org/fr/le-canada-et-la-france-signent-une-lettre-dintention-commune-sur-lintelligence-artificielle/>.
55. « Déclaration d'intention conjointe entre le premier ministre du Québec et le président de la République française », Gouvernement du Québec, Site du premier ministre du Québec, 7 juin 2018, en ligne : <https://www.premier-ministre.gouv.qc.ca/actualites/communiques/details.asp?idCommunique=3416>.

engendré une révision des normes étatiques s'appliquant à l'industrie. L'essentiel reste à faire sur le plan de l'élaboration de propositions rigoureuses et spécifiques concernant l'encadrement éthique des systèmes d'IA.