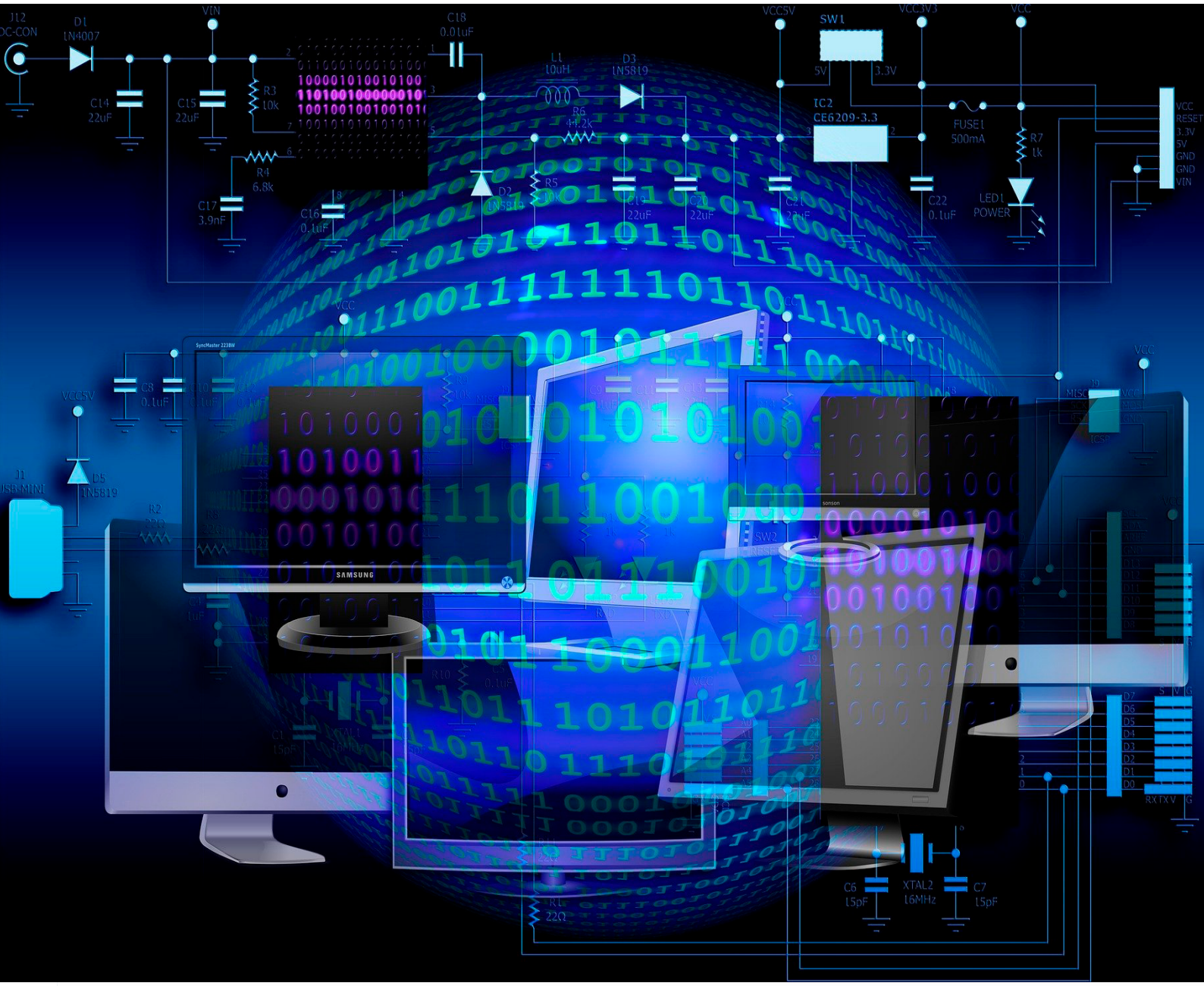
INSTITUT
POUR QUE L'AVENIR AIT BESOIN DE NOUS
SAPIENS



La responsabilité au cœur de l'IA : Du paradoxe de l'IA aux différentes stratégies pour les organisations

Etude — Mai 2021

À propos de l'auteur

Patrice Cailleba

Patrice Cailleba est professeur de management au sein de Paris School of Business et membre du PSB Research Lab. Diplômé de l'ESCP Business School, docteur en Philosophie de l'Université de la Sorbonne (Paris IV) et titulaire d'une HDR en sciences de gestion (Paris-Saclay), il est auditeur de l'IHEDN (Institut des Hautes Etudes de Défense Nationale). Ses thèmes de recherche concernent l'éthique des affaires, les comportements organisationnels et la philosophie politique.

A propos de l'Institut Sapiens

L'Institut Sapiens est un organisme à but non lucratif dont l'objectif est de peser sur le débat économique et social. Il se veut le premier représentant d'une think-tech modernisant radicalement l'approche des think tanks traditionnels. Il souhaite innover par ses méthodes, son ancrage territorial et la diversité des intervenants qu'il mobilise, afin de mieux penser les enjeux vertigineux du siècle.

Sa vocation est triple :

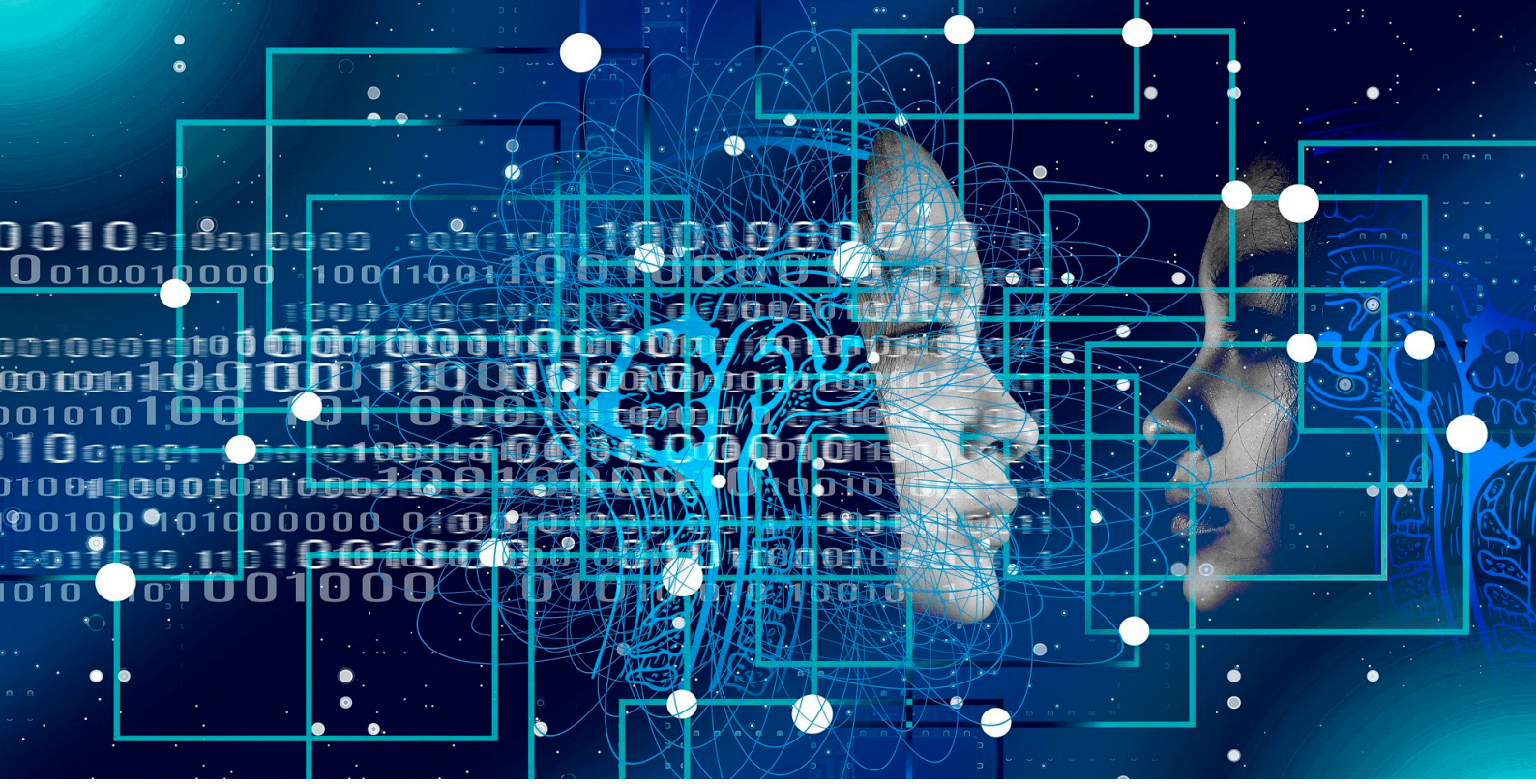
Décrypter — l'Association aide à la prise de recul face à l'actualité afin d'être capable d'en comprendre les grandes questions. L'Institut Sapiens sera un centre de réflexion de pointe sur les grands enjeux économiques contemporains.

Décloisonner et faire dialoguer — l'Association veut mettre en relation des mondes professionnels trop souvent séparés : universitaires, membres de la sphère publique, praticiens de l'entreprise ou simples citoyens, ils doivent pouvoir se rencontrer pour réfléchir et dialoguer. Afin d'être réellement représentatifs de toutes les compétences et expériences, les groupes de travail associent systématiquement des personnes d'horizons professionnels divers (de l'ouvrier au dirigeant de société cotée) et peu important leur lieu de vie (Métropole, Outre-mer).

Former — Le XXIe siècle est le siècle de l'information ; il doit devenir pour l'individu celui du savoir. Comprendre le monde implique une capacité à faire un retour sur notre histoire, à connaître le mouvement millénaire des idées, à posséder ces Humanités dont l'importance est plus grande que jamais. Parce qu'il veut faire accéder à une compréhension du monde, l'Institut Sapiens se fixe aussi pour objectif de promouvoir cette culture générale sans laquelle demain plus personne ne pourra comprendre son environnement.

Plus d'informations sur institutsapiens.fr

INSTITUT SAPIENS
POUR QUE L'AVENIR AIT BESOIN DE NOUS



Synthèse

Les travaux récents autour de l'IA commencent à peine à explorer le paradoxe qu'elle présente, mais aussi le type de responsabilité qu'elle implique. Pourtant, ils sont d'une importance cruciale dans la mesure où la stratégie des entreprises et la politique des États devront se définir à l'avenir autour de l'IA.

Il y a un caractère paradoxal, c'est-à-dire contradictoire, interdépendant et indépassable, au cœur de l'IA. Il s'agit du couple automatisation/augmentation qui la nourrit de manière dynamique et récursive. En outre, ce paradoxe ouvre sur des dialectiques subordonnées à l'idée particulière de responsabilité qu'implique l'IA.

L'IA est le produit du management dans la mesure où le management vise à rendre autonome une personne dans la réalisation de ses missions. Or, cette relative autonomie constitue bien le reproche qui lui est fait. Mais l'IA est aussi du management artificiel dans la mesure où elle accompagne et/ou remplit les fonctions managériales classiques.

Il faut donc penser l'intégration homme-machine que permet l'IA au travers de la responsabilité donnée, voire imposée, à l'individu. Ceci est d'autant plus nécessaire que cette notion paraît inextricable en raison du nombre d'acteurs impliqués et insondable en raison des capacités de l'IA à devenir autonome. Pourtant, la responsabilité est au cœur de l'IA : elle nous rend, sans le savoir, « otages d'autrui ».

A partir de la récursivité de l'IA, on peut identifier des stratégies managériales en distinguant les cercles vicieux potentiels des cercles vertueux envisageables pour les organisations qui s'y engagent.

Les cercles vicieux concernent les entreprises qui privilégient d'abord les tâches d'automatisation et de remplacement de l'activité humaine. Cherchant la rentabilité à court et moyen terme, elles surestiment les avantages financiers liés à la baisse du coût salarial, sans prendre en considération les investissements colossaux. Ces cercles vicieux concernent également les entreprises trop soucieuses de protéger des métiers qui deviennent progressivement obsolètes en les accompagnant *artificiellement*.

Les cercles vertueux passent par des travaux de prospectives par secteur ou métier qui permettront d'anticiper les besoins en formation des travailleurs et renforceront les dispositifs de sécurisation des parcours professionnels. Equilibrer l'automatisation et l'augmentation paradoxale de l'IA permettrait d'opérer une déqualification sélective de certains métiers en même temps qu'une stratégie de requalification pour d'autres en limitant – autant que faire se peut – les coûts sociaux et économiques d'une transition inévitable.

L'heure est à l'urgence tant une véritable géopolitique des IA se met déjà en place. Il faut donc mener cette bataille pour ne pas devenir une « colonie du monde numérique ».

Un investissement massif de l'État dans la recherche est nécessaire afin de permettre des « audits d'algorithme » pour comprendre ce qui se passe dans la *blackbox*. De même, la création de formations spécifiques à l'IA, la mise à jour obligatoire de certains diplômes ainsi que l'utilisation de manière prospective du Compte Personnel de Formation (CPF) constituent des facteurs clés de succès.

L'écueil du management de l'IA serait de la considérer comme une simple technique. Sous le couvert d'une « calculabilité intégrale » de son environnement, l'individu ne ferait que se déposséder de lui-même en croyant « arraisonner » le monde et en oubliant sa propre responsabilité au sein de l'IA.

Introduction

En 2018, le rapport « ***Intelligence artificielle et travail*** » de France Stratégie affirmait qu'aucun des défis que posait l'intelligence artificielle (IA) n'était nouveau. Qu'il s'agisse d'une « perte d'autonomie du salarié, soumis à un contrôle automatisé de plus en plus insidieux, avec les risques psychosociaux associés »[1] ou d'une amélioration des conditions du travail, les hypothèses optimistes (libération de certaines tâches) et pessimistes (aliénation du travail) étaient et restent toutes aussi crédibles et potentiellement concomitantes. Il y a en effet un « en même temps » de l'IA qui favorise littéralement tout et son contraire à mesure que progressent le *machine learning* (apprentissage automatique), le *deep learning* (apprentissage en profondeur), les ***réseaux neuronaux artificiels***, etc. dans un véritable élan schumpetérien de destruction créatrice.

Au-delà de la dimension technique et de ses retombées économiques et sociales, les travaux récents autour de l'IA commencent à peine à explorer le paradoxe qu'elle présente, mais aussi le type de responsabilité qu'elle implique. Pourtant, ils sont d'une importance cruciale dans la mesure où la stratégie des entreprises et la politique des États devront se définir à l'avenir autour de l'IA.

I – Le paradoxe et les dialectiques de l'IA

Comme le précise ***Marc Guyon***, « l'intelligence humaine, comme tout processus du vivant, n'est pas réductible aux capacités logico-mathématiques de l'IA » : le travail est prise de recul et mise en perspective, habileté à critiquer et capacité à créer, en bref, gestion et dépassement de la contradiction. L'intelligence est plurielle (à la fois capacité d'apprentissage, d'adaptation et de proposition de solutions dans un contexte donné) alors que l'intelligence artificielle renvoie à une raison instrumentale, à une intelligence purement cognitive, qui peut faire le lit d'une « névrose de la maîtrise ». ***Margarida Romero*** a souligné, à juste titre, que nourrir massivement de données un programme ne lui permet pas automatiquement de « porter un jugement

1 France Stratégie (2018), *Intelligence artificielle et travail*, Rapport à la ministre du Travail et au secrétaire d'État auprès du Premier ministre, chargé du numérique, S. Benhamou et L. Janin (rapporteurs), Mars.

métacognitif sur son processus et ses produits en lien à un contexte socioculturel donné ». A bien des égards et sur bien des domaines, il s'agit encore d' « intelligences très artificielles »[2].

Une fois rappelées ces limites, la proposition que fait l'intelligence artificielle à l'activité humaine est double :

- Soit elle remplace l'individu en automatisant des tâches humaines (*automation* en anglais) ;
- Soit elle l'accompagne sur ces même tâches (*augmentation* en anglais) dans la mesure où la machine prolonge et place l'individu dans une position de supervision.

Les éléments du couple augmentation/automatisation ne sont pas exclusifs l'un de l'autre. A la fois synchronique et diachronique, ce couple offre ainsi cette double possibilité d'augmentation/automatisation à un même moment, tout comme il évolue dans le temps et ce, de manière circulaire (d'abord augmentation, ensuite automatisation, puis re-augmentation, etc.). En outre, ce couple n'est pas manichéen. La complémentarité avec la machine peut être incapacitante et aliénante alors que le remplacement par la machine peut transformer positivement certains métiers et les réorienter vers des tâches de supervision et/ou créatrices de plus de valeur. Et vice-versa. Ainsi la tension au cœur de l'IA apparait-elle clairement avec force : augmentation et automatisation ne vont pas l'une sans l'autre et nourrissent une relation qu'il faut interroger et comprendre.

L'IA se caractérise par un paradoxe augmentation/automatisation qui évolue dans le temps et ce, de manière circulaire.

Dans leur article scientifique publié début 2021 sur l'IA[3], Sebastian Raisch et Sebastian Krakowski reviennent sur le saut technologique qui a marqué le passage au « second âge de la machine »[4] que nous vivons. Ils soulignent le caractère paradoxal, c'est-à-dire contradictoire, interdépendant et indépassable, au cœur de l'intelligence artificielle.

2 Dessalles, J. L. (2019), *Des intelligences très artificielles*, Odile Jacob.

3 Raisch, S. & Krakowski, S. (2021), Artificial intelligence and management: The automation–augmentation paradox. *Academy of Management Review*, 46(1), 192-210.

4 *« During this "first machine age," which started with the invention of the steam machine in the eighteenth century, mechanical machines enabled mass production by taking over manual labor tasks at scale. Today, we face an analogous inflection point of unprecedented progress in digital technology, taking us toward the "second machine age"* (Brynjolfsson & McAfee, 2014 : cf. infra). *Instead of performing mechanical work, machines now take on cognitive work, which was traditionally an exclusively human domain."* (Raisch et Krakowski, 2021 : 193).

Au-delà de son acronyme même, l'IA constitue clairement une dyade au sens d'Edgar Morin[5], à savoir une dialogique où les termes sont « complémentaires, concurrents et antagonistes ». La dialogique n'est pas une dialectique au sens hégélien car il n'y a pas de solution qui dépasse et supprime ce qui n'est pas toujours une contradiction. Il s'agit plutôt d'une *unité complexe* entre deux logiques concurrentes et antagonistes dont la clé de compréhension est à chercher, en ce qui concerne l'IA, dans une réflexion plus approfondie autour du management. Car, en effet, l'IA s'avère être à la fois :

- le produit du management dans la mesure où le management vise à rendre autonome une personne dans la réalisation de ses missions[6]. Or, cette relative et progressive autonomie constitue bien le motif du reproche fait à l'IA ;

- et du management artificiel dans la mesure où elle accompagne et/ou remplit les fonctions managériales classiques identifiées par Henri Fayol[7] (Prévoir, Organiser, Commander, Coordonner, Contrôler).

De fait, l'unité complexe de l'IA se retrouve et se prolonge dans le management qui, en miroir, présente des dialogiques qui la concernent. Ces dialogiques managériales ont trait à la dimension politique de l'activité humaine et plongent leurs racines dans la théorie des organisations et plus profondément dans la philosophie politique.

A la fin des années 1970, Peter Drucker[8] avait identifié deux dyades dans son « éthique de l'interdépendance ». Tout management étant politique (et réciproquement), il qualifia les relations essentielles qui caractérisent le management, à savoir : la relation de commandement et obéissance et la relation amicale de type « friend and friend » qui pouvait devenir une relation entre « friend » et « potential ennemies ».

Sans le savoir, Peter Drucker reprenait l'analyse de Julien Freund[9] qui, en 1965, avait déjà identifié trois dialectiques propres au phénomène politique :

- La dialectique du commandement et de l'obéissance ;

- La dialectique du privé et du public ;

- La dialectique de l'ami et de l'ennemi.

5 Morin, E. (2004), *La méthode : Éthique,* Ed. du Seuil, p.59
6 Roche, L. (2016), *La théorie du Lotissement,* Presses Universitaires de Grenoble, PUG.
7 Fayol, H. (1916), *Administration Industrielle et Générale,* Dunod.
8 Drucker, P. F. (1981), « What is 'business ethics'? », *McKinsey Quarterly*, (3), p.2-15.
9 Freund, J. (2004), *L'essence du politique,* Dalloz, postface de P.-A. Taguieff, Paris.

En ajoutant celle du public/privé, les trois dialectiques de Freund enrichissent définitivement la compréhension du paradoxe de l'IA. Ainsi peut-on comprendre que si le management est traversé par ces dialectiques, alors le paradoxe de l'IA (automatisation/augmentation) ouvre également sur ces trois dialectiques :

- Qui, entre la machine et l'individu, commande ou obéit ?
- Comment garantir l'anonymat de données qui fondent l'IA et sa réussite ? Quelles données doivent conserver un caractère privé ?
- Qui, dans l'IA, est le vrai responsable : le concepteur, l'utilisateur, le fournisseur d'IA ou la machine elle-même ?

Drucker comme Freund ne pouvaient pas anticiper l'essor de l'IA tel que nous le connaissons aujourd'hui. Pourtant, les dialectiques complémentaires qu'ils ont identifiées nourrissent le paradoxe de l'IA et demeurent essentielles pour comprendre les défis à relever en vue d'un management éthique de cette dernière.

La question de la responsabilité est au cœur de l'IA car elle subordonne les questions principales : qui commande, qui obéit ? Comment garantit-on l'anonymat des données et la protection des individus ?

Toutefois, ces dialectiques ne sont pas d'égale importance dans la mesure où l'une subordonne les deux autres. La question de la responsabilité est effectivement première tant la responsabilité de l'individu ne peut, ni ne doit s'effacer au « profit » de la machine ou au détriment des individus. Il faut donc penser l'intégration homme-machine que permet l'IA au travers de la responsabilité donnée, voire imposée, à l'individu. Ceci est d'autant plus nécessaire et pertinent que les études menées depuis vingt ans montrent que l'intégration homme-machine réduit non seulement les biais humains[10] mais aussi les biais homme-machine[11]. Ainsi l'intégration entre automatisation et augmentation permise par l'IA doit-elle absolument conserver, à défaut de renforcer, la dimension de la responsabilité humaine. Encore faut-il examiner de près ce que cette responsabilité a de particulier au sein de l'IA.

10 Larrick, R. P. (2004), "Debiasing", in D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making*: 316–338. Malden, MA: Blackwell.

11 Skitka, L. J., Mosier, K., & Burdick, M. D. (2000), "Accountability and automation bias", *International Journal of Human-Computer Studies,* 52: 701–717.

II – La responsabilité au cœur de l'IA

Dans l'introduction de son étude de 2019 intitulée « ***Responsabilité et IA*** », le Conseil de l'Europe a souligné que « le devoir de protéger les droits de l'homme appartient en premier lieu aux États. Ils doivent donc s'assurer que ceux qui œuvrent à la conception, au développement et au déploiement de ces technologies et en tirent des bénéfices répondent aussi de leurs impacts négatifs »[12]. Ainsi le Conseil de l'Europe rappelle-t-il combien la notion d'IA doit être constamment liée à celle de responsabilité. Or, les obstacles sont nombreux tant l'IA paraît rendre inextricable l'idée d'une responsabilité unique pour plusieurs raisons.

D'abord, trois types d'acteurs sont interconnectés par l'IA :

- Le concepteur/développeur de la technologie qui, seul, peut connaître le ou les algorithmes de la technologie concernée et infléchir, ce faisant, les résultats et les décisions à prendre ;
- Le déployeur/fournisseur de la technologie qui peut exploiter les données collectées à son seul avantage et, partant, manipuler les utilisateurs ;
- Et l'utilisateur final qui peut être aussi celui qui a fourni les données (dans le passé) ou qui alimente l'IA (dans le présent) en interagissant avec elle. Souvent, il oublie qu'il est autant le produit de l'IA (ses données alimentent l'IA) que sa raison d'être (le propre de l'IA est de le servir).

Chacun de ces acteurs est ici considéré en terme générique tant il recoupe une myriade d'individus qui viennent concrètement l'incarner. Le problème de cette multiplicité d'acteurs n'empêche pas de considérer les responsabilités de chacun lorsqu'ils ont contribué à un évènement indésirable : c'est l'essence même des systèmes juridiques. Néanmoins, il convient d'en discuter constamment et de clarifier à la fois les attentes de la société civile à leurs égards et les marges de manœuvres de ces mêmes acteurs.

12 Conseil de l'Europe (2019), *Management et IA*, Etude du Conseil de l'Europe DGI (2019) 05, Karen Yeung (Rapporteur), Préparée par le Comité d'Experts sur les dimensions des droits de l'homme dans le traitement des données et les différentes formes d'intelligence artificielle (MSI-AUT).

Ensuite, le fonctionnement même de l'IA, c'est-à-dire la manière dont les résultats sont obtenus, peut être opaque et impénétrable, a fortiori lorsque les technologies employées sont complexes et évolutives (voir le *machine learning* et le *deep learning* cités plus haut). Il est alors rendu difficile de rendre responsable quiconque d'un résultat lorsque nul ne peut expliquer comment ce dernier a été obtenu (biais technologique ? biais humain ? autonomie de l'IA ?) et donc remonter à la source même, à savoir dans la ***black box***. Le recours à la prise de décision algorithmique (en anglais, *algorithmic decision-making* ou ADM) dans plusieurs domaines (les réseaux et médias sociaux, l'information en général mais aussi la justice, le commerce et bientôt la santé) rend nécessaire la prise de conscience de chacun d'entre nous sur ces thématiques.

Il ne peut exister dès lors un « vide de responsabilité »[13] qui renverrait aux seuls développeurs la responsabilité des résultats de l'IA. De même, l'interaction « humain-machine » ne doit pas effacer la responsabilité essentielle des individus au risque de laisser se développer des procès où une machine, un algorithme, est déclaré, in fine, responsable et coupable. Tout comme au Moyen-Âge, on déclarait coupables des ***porcins*** lors de procès d'animaux[14].

Au sein de l'IA, la question de la responsabilité paraît inextricable tant les acteurs sont nombreux, son fonctionnement de plus en plus insondable et ses impacts vertigineux.

Enfin, la dimension spatio-temporelle de l'impact de l'IA ébranle l'idée même d'une responsabilité unique et individuelle. En effet, nulle autre technologie ne peut à ce point s'étendre aussi rapidement sur l'ensemble de la planète et modifier durablement les affaires humaines. L'interaction des systèmes algorithmiques complexifie la notion de responsabilité et rend possible deux écueils.

D'un côté, l'IA vient pulvériser cette notion de responsabilité : chaque individu ne peut, ni ne veut, assumer qu'une petite part de responsabilité en tant que partie prenante dans l'IA, à savoir comme créateur, développeur et/ou utilisateur qui rejetterait la faute sur l'autre en cas de problème. Il ne s'agit pas de dire que l'identification d'une responsabilité individuelle n'est ni

13 Conseil de l'Europe (2019), *Management et IA*, Etude du Conseil de l'Europe DGI (2019) 05, Karen Yeung (Rapporteur), Préparée par le Comité d'Experts sur les dimensions des droits de l'homme dans le traitement des données et les différentes formes d'intelligence artificielle (MSI-AUT).

14 Daboval, B. (2003), *Les Animaux dans les procès du Moyen Âge à nos jours,* Thèse pour le doctorat vétérinaire, Ecole Nationale d'Alfort, Faculté de médecine de Créteil.

souhaitable, ni atteignable. Il s'agit plutôt de considérer le caractère vertigineux d'actions individuelles combinées entre elles, dont les répercussions concernent des centaines de milliers, voire davantage, d'êtres humains.

De l'autre côté, l'inclination naturelle des sociétés humaines à chercher et à trouver des boucs-émissaires au sens de Girard[15] questionne l'arrêt possible d'une IA. Si le bouc-émissaire permet, entre autres, de se débarrasser de son sentiment de culpabilité en excluant tout prise de conscience collective, sera-t-il possible alors de modifier pratiquement certaines IA ? En même temps, se pose la question du rapport coût/bénéfice de chaque innovation dont les sociétés humaines (occidentales) ont de plus en plus de mal à accepter l'idée même de coût (voir les polémiques sur certains ***vaccins*** au nom du principe de précaution).

Pour résoudre ce problème de la responsabilité, le Conseil de l'Europe avance deux types de responsabilités pour l'IA :

- Une responsabilité rétrospective qui établit les responsabilités pour des évènements passés et ;
- Une responsabilité prospective qui définit les obligations de chaque acteur pour limiter les impacts négatifs et/ou indésirables.

Cependant, cette acception temporelle de la responsabilité échoue face au réel dans la mesure où le passé avance (aujourd'hui sera du passé demain) et l'avenir avec lui. Dès lors, rendre compte d'une responsabilité rétrospective (ou prospective) est toujours une œuvre à reconstruire : c'est d'ailleurs le propre du travail des historiens (ou prospectivistes). En outre, les récits de fiction qui aident à donner corps aux situations du quotidien illustrent à l'envi le caractère synchronique (à la fois prospectif et rétrospectif) de la responsabilité individuelle. Olivier Paquet et Anne-Caroline Paucot, dans le ***rapport Villani***, montrent brillamment comment l'IA met en jeu, de manière complexe, cette question de la responsabilité[16]. Il se dégage alors, derrière son caractère apparemment inextricable, l'idée d'une imbrication totale de la responsabilité au sein de l'IA. Qui peut alors être considéré comme responsable ?

15 Girard, R. (2014), *Le bouc émissaire*, Grasset.

16 Lire, en particulier, l'excellente fiction sur les accidents causés par les voitures autonomes dans le rapport Villani : Paucot, A.-C. (2018), « Dans l'imaginaire : questions qui tuent », *in Rapport Villani,* p. 158-161.

Seules les œuvres de fiction peuvent nous permettre de comprendre l'imbrication totale, et de plus en plus réelle, de la notion de responsabilité au sein de l'IA.

Les utilisateurs (personnes privées et/ou institutions publiques) de l'IA s'accordent sur l'idée que les personnes qui développent et déploient des systèmes d'IA doivent assumer les conséquences indésirables de leurs actes, tout autant qu'elles tirent profit de leurs services. Inversement, les concepteurs et les déployeurs de l'IA pensent de même en ce qui concerne les utilisateurs qui cèdent leurs données pour en bénéficier ultérieurement : en bref, ils sont le produit qu'ils utilisent et en sont donc directement responsables. Parce que ces considérations sont autant légitimes les unes que les autres, elles doivent être examinées ensemble. Dès lors, la notion de responsabilité au sein de l'IA doit remplir, tout à la fois, les conditions suivantes :

- la personne qui prête ses données doit confirmer qu'elle se défait temporairement de sa responsabilité qu'elle pourra reprendre le cas échéant. On rejoint ici la dialectique public/privé sur le partage des données privées ;
- l'organisation qui crée les algorithmes de l'IA doit garantir sa propre responsabilité en cas d'effet indésirable. Dans ce cas, elle s'engage à corriger le programme dans une relation dialectique de commandement/obéissance : elle décide ce qu'elle fait de l'IA (commandement) autant qu'elle obéit aux injonctions des parties prenantes lors d'effets indésirables (obéissance) ;
- l'organisation qui déploie l'IA et exploite les données qu'elle en tire doit assumer sa responsabilité dans leur bonne utilisation vis-à-vis des autres parties prenantes, mais aussi vis-à-vis de concurrents potentiels qui auront tôt fait de la délégitimer afin de la remplacer (dialectique ami/ennemi).

L'éthique de l'IA illustre la philosophie de Lévinas. Elle nous rend, sans le savoir, « l'otage d'autrui ».

On le voit, il s'agit d'une co-responsabilité de l'ensemble des acteurs, c'est-à-dire d'une responsabilité qui renvoie à une interdépendance et à une réciprocité. Cette réciprocité de la responsabilité est rendue nécessaire et elle nous ramène au philosophe Lévinas. Ce dernier écrivait que nous sommes otages d'autrui[17]. Etrangement, l'IA offre un champ d'application intéressant et quelque peu inespéré au philosophe de l'altérité. En effet, l'éthique lévinassienne fait du respect de la vulnérabilité[18] de l'individu la clé de son action. Or, une éthique de l'IA doit être résolument pensée en ces termes, c'est-à-dire entre vulnérabilité et respect, dans la réciprocité de quelqu'un qui m'est inconnu, lointain, à venir ou passé, dont j'utilise directement, mais sans le connaître concrètement, ses données :

« J'entends la responsabilité comme responsabilité pour autrui, donc comme responsabilité pour ce qui n'est pas mon fait, ou même ne me regarde pas (…). La proximité d'autrui est (…) le fait qu'autrui n'est pas simplement proche de moi dans l'espace, ou proche comme un parent, mais s'approche essentiellement de moi en tant que je me sens – en tant que je suis – responsable de lui. C'est une structure qui ne ressemble nullement à la relation intentionnelle qui nous rattache, dans la connaissance, à l'objet de quelque objet qu'il s'agisse, fût-ce un objet humain. La proximité ne revient pas à cette intentionnalité ; en particulier elle ne revient pas au fait qu'autrui me soit connu. »[19]

Une fois considérée la responsabilité partagée et réciproque qu'implique l'IA, il reste encore à travailler à l'instauration de normes et de procédures qualités qui permettent la gestion et la traçabilité des actions de chacun des acteurs impliqués. Voilà la difficulté de la tâche à venir qui fait de la réciprocité un des éléments clés de compréhension de l'éthique de l'IA. Or, comme nous l'avons vu auparavant, l'IA est à la fois le produit du management et du management artificiel. Cette relation dynamique se retrouve dans le paradoxe de l'IA qui fournit des stratégies particulières aux entreprises.

17 Levinas, E. (1961), *Totalité et infini,* La Haye : Nijhoff.
18 Antenat, N. (2003), « Respect et vulnérabilité chez Levinas », *Le Portique* [***En ligne***], n°11.
19 Levinas, E. (1982), Éthique et infini, Dialogues avec Ph. Nemo, Essais, Fayard, France Culture, p. 91-93.

III – L'IA comme levier stratégique pour les entreprises

Le paradoxe augmentation/automatisation qui caractérise l'IA est dynamique, chaque élément se répondant et se nourrissant l'un l'autre dans une boucle de rétroaction (*feedback*). Cette causalité circulaire[20] à l'œuvre s'appuie sur la rationalité limitée des acteurs (individus et, *ipso facto*, machines) qui se retrouve plus ou moins enrichie à chaque nouvelle boucle. A partir de cette récursivité, certains auteurs et chercheurs ont identifié des stratégies managériales en distinguant les cercles vicieux potentiels des cercles vertueux envisageables pour les organisations engagées dans l'IA.

Les **cercles vicieux** concernent les entreprises qui privilégient d'abord les tâches d'automatisation et de remplacement de l'activité humaine. Cherchant la rentabilité à court et moyen terme, elles surestiment les avantages financiers liés à la baisse du coût salarial, sans prendre en considération les investissements colossaux en termes d'infrastructures IT, de logistique, de relations clients, etc. mais aussi de ressources humaines (licenciements d'un côté, modalités d'accompagnement de l'autre).

Ces perspectives concernent autant les métiers faiblement qualifiés que les métiers plus qualifiés. Le résultat potentiel pour les cadres et managers qui décrochent est un déclassement ou une déqualification (« de-skilling ») brutale qui accroît le risque des inégalités dans la mesure où ces derniers se dirigent vers des métiers moins qualifiés où la concurrence fait déjà rage. A l'inverse, le résultat pour les managers et cadres qui réussissent à s'accrocher est une concurrence toujours plus dure[21].

Ces cercles vicieux concernent également les entreprises trop soucieuses de protéger des métiers qui deviennent progressivement obsolètes en les accompagnant *artificiellement* de ressources technologiques, alors que ces métiers font long feu. Ce faisant, les entreprises additionnent les coûts d'une main d'œuvre progressivement disqualifiée et démotivée avec

20 Follet, M.P. (1924), Creative Experience, NY London : Longmans, Green & Co., p.124 (document web : http://mpfollett.ning.com/mpf/follett-writings). Follet a enrichi la pensée managériale naissante en soulignant ces boucles récursives entre intention-action-résultat-nouvelle intention.

21 Voir Brynjolfsson, E. & McAfee, A. (2014), *The second machine age: Work, progress, and prosperity in a time of brilliant technologies.* WW Norton & Company.

des investissements matériels et immatériels dont le montant élevé ne masque même pas les limites de l'ambition prophylactique (adoucir l'impact de la disruption technologique). Cette ambition se révèle finalement dispendieuse financièrement (on finance et on accompagne par à-coups) et catastrophique socialement (on désespère les employés qui ne se sentent ni formés, ni accompagnés et, donc, non protégés).

Les cercles vicieux concernent les entreprises qui privilégient d'abord les tâches d'automatisation et de remplacement de l'activité humaine.

Traçant des perspectives de ruptures économiques et sociales, le *World Economic Forum* a publié en Octobre 2020, un rapport sur le ***Future of jobs*** dans lequel sont présentées, pays par pays, les anticipations de certains décideurs concernant l'impact de la pandémie sur la digitalisation de l'économie. Le rapport met en avant plusieurs aspects remarquables qui illustrent ces cercles vicieux :

- L'automatisation, en tandem avec la récession provoquée par la Covid-19, crée un scénario de "double disruption" pour les travailleurs, scénario qu'accélère la nécessaire digitalisation des activités dans nombre de domaines.
- A moyen terme, la création d'emplois devrait ralentir tandis que la destruction d'emplois devrait s'accroître.
- En l'absence d'efforts proactifs, les inégalités risquent d'être exacerbées par le double impact de la technologie et de la récession pandémique[22].
- L'apprentissage et la formation en ligne devraient augmenter, mais différemment pour les personnes qui ont un emploi et celles qui sont au chômage.
- Les possibilités de reconversion et d'amélioration des compétences des travailleurs devraient être de plus en plus limitées – au moins dans un premier temps – sur le marché du travail nouvellement restreint.

En fonction des technologies, des ressources et du marché, les métiers apparaissent, évoluent et, le cas échéant, disparaissent. Les métiers du commerce, de la logistique, de la santé ou de la banque n'ont-ils pas déjà évolué significativement plusieurs

22 Voir aussi Babeau, O. (2020), *Le nouveau désordre numérique : comment le digital fait exploser les inégalités*. Buchet/Chastel.

fois depuis le début de la révolution industrielle ? Les États qui échouent à accompagner les individus et les organisations en facilitant la création de nouveaux métiers ou de nouvelles opportunités sont fatalement voués à nourrir ces cercles vicieux ; tout comme les organisations qui oublient que ce sont elles qui doivent s'adapter à ces destructions/créations régulières.

L'existence de ces cercles vicieux ne doit pas faire oublier que les interactions humaines (organisationnelles, sociales et culturelles) comme les limites techniques rendent impossibles autant qu'illusoires l'automatisation complète de processus managériaux dans un futur proche. Surtout lorsqu'il s'agit de tâches ambiguës, complexes et d'occurrence rare.

L'IA s'attache, pour le moment, à des tâches automatisables que la gestion massive et ***stratégique*** de données rend possible. Reprenant des éléments du Conseil d'orientation de l'emploi, le ***rapport Villani*** a rappelé en 2018 les quatre critères qui aident à identifier une tâche automatisable :

- *« l'absence de flexibilité : (…) la tâche consiste à répéter continuellement une même série de gestes ou d'opérations ;*
- *l'absence de capacité d'adaptation : (...) la tâche consiste en une application stricte d'ordres, de consignes ou de modes d'emploi ;*
- *l'absence de capacité à résoudre des problèmes : lorsqu'il se produit une situation anormale, le travailleur fait appel à d'autres pour résoudre le problème ;*
- *l'absence d'interactions sociales ».*

Prenant en compte ces éléments, les entreprises qui ont compris l'urgence à considérer la tension paradoxale de l'automatisation/augmentation constitutive de l'IA peuvent enclencher des **cercles vertueux**. Le rapport ***France Stratégie*** cité en introduction esquisse les perspectives à suivre :

- *« conduire (…) des travaux de prospective (…) pour assurer un bon niveau d'information et d'anticipation des acteurs ;*
- *assurer la formation des travailleurs aux enjeux de demain (…) ;*
- *renforcer des dispositifs de sécurisation des parcours professionnels pour les quelques secteurs (…) qui seraient fortement impactés par le risque d'automatisation ;*

- *(…) ne pas sous-estimer les risques en matière de condition de travail – perte d'autonomie, intensification du travail, etc. – liés aux conditions de déploiement des outils IA. »*

L'automatisation complète de processus managériaux dans un futur proche est impossible, a fortiori *lorsqu'il s'agit de tâches ambiguës, complexes et d'occurrence rare.*

Ainsi les secteurs public, para-public mais aussi privé doivent-ils apporter un soutien plus important à la reconversion et à l'amélioration des compétences des travailleurs à risque ou déplacés en anticipant davantage ce qui est inéluctable. Les déficits de compétences ou les secteurs pénuriques en termes d'emplois restent importants tant les compétences requises pour tous les emplois vont évoluer fortement au cours des cinq prochaines années. L'impact de l'IA est déjà largement anticipé et commenté. Dans le rapport précédemment cité, on évaluait, en 2018, entre 10 et 47% les métiers qui seraient transformés en partie ou en totalité par l'IA, en particulier dans la santé, la banque de détail et les transports. En 2019, une ***étude d'IBM*** estimait que 120 millions d'emplois a *minima* devraient s'adapter, *nolens volens*, à l'IA d'ici 2022, dont quelques millions en France.

Cependant, le rapport du World Economic Forum d'Octobre 2020 sur le ***Future of jobs*** a envisagé des perspectives positives qui permettent d'illustrer ces cercles vertueux en même temps que leur caractère prometteur :

- L'avenir du travail en ligne est déjà arrivé pour une grande majorité des « cols blancs ».
- A long terme, le nombre d'emplois détruits devrait être dépassé par le nombre d' « emplois de demain » ;
- Malgré le ralentissement économique actuel, la grande majorité des employeurs reconnaissent la valeur de l'investissement dans le capital humain.
- L'IA rendra nécessaire l'investissement dans le capital humain et social en adoptant des mesures environnementales, sociales et de gouvernance (ESG) tout en les assortissant de nouvelles mesures de comptabilité du capital humain.

De leur côté, Raisch et Krakowski[23] mettent en avant les gains en matière de productivité, d'amélioration de la qualité, sans parler des innovations générées du fait de la combinaison complémentaire individu/machine. Equilibrer l'automatisation et l'augmentation paradoxale de l'IA permettrait d'opérer une déqualification sélective de certains métiers en même temps qu'une stratégie de requalification pour d'autres en limitant – autant que faire se peut – les coûts sociaux et économiques d'une transition inévitable.

Les cercles vertueux passent par des travaux de prospectives par secteur ou métier qui permettront d'anticiper les besoins en formation des travailleurs et renforceront les dispositifs de sécurisation des parcours professionnels.

La nécessité de réfléchir de manière complexe en mêlant professionnels des IT, spécialistes des comportements organisationnels, avec des collaborateurs concernés et impliqués, apparait comme la plus sûre méthode qui permettrait de prévenir les risques et d'anticiper les problèmes (humains et techniques). Dans une compétition où les premiers acteurs à être prêts seront aussi ceux qui prendront une large avance (« *the winner takes it all* »), comme c'est déjà en partie le cas (voir les GAFAM américains ou les BATX chinois[24]), le mouvement s'accélère et il est nécessaire de ne pas démarrer trop tard. Duale, l'IA présente en synthèse la possibilité de ré-humaniser certains métiers tout autant que la probabilité d'accroître certaines ***inégalités*** et discriminations[25].

23 Raisch, S. & Krakowski, S. (2021), « Artificial intelligence and management: The automation–augmentation paradox », *Academy of Management Review,* 46(1), 192-210. Leurs travaux se basent sur trois ouvrages antérieurs : Brynjolfsson, E. & McAfee, A. (2014), *The second machine age: Work, progress, and prosperity in a time of brilliant technologies.* WW Norton & Company ; Daugherty, P. R. & Wilson, H. J. (2018), *Human+ machine: Reimagining work in the age of* AI, Harvard Business Press ; Davenport, T. H. & Kirby, J. (2016), *Only humans need apply: Winners and losers in the age of smart machines,* New York, NY: Harper Business.

24 Google, Apple, Facebook, Amazon & Microsoft ; Baidu, Alibaba, Tencent & Xiami. Voir aussi : Colin, L. (2020), Google et l'Ethics As A Service (EAAS) : le grotesque et le Funeste, Institut Sapiens ; Meneceur, Y. (2020), IA et justice : le grand malentendu, Institut Sapiens.

25 Borgesius, F. (2018), Discriminations, IA et décisions algorithmiques, Rapport pour le Conseil de l'Europe.

IV – Quelques recommandations pour ne pas devenir une « colonie du monde numérique »

A l'heure de la multiplication des ***cyber-attaques***, au moment où la ***Russie*** développe un projet propre d'IA, il faut mener la bataille de l'IA pour ne pas devenir une « colonie du monde numérique » comme l'avançait Catherine Morin-Desailly dans son ***rapport au Sénat***. Car même si l'Europe et la France savent aiguiser leur pensée critique et leur créativité, cela ne peut suffire[26]. Il existe déjà une véritable ***géopolitique des éthiques de l'IA.*** Et l'Histoire a souvent montré qu'une pensée critique, même juste, pouvait être écrasée par la puissance de la technique, la force du nombre et la simplicité des arguments. Une réponse coordonnée au niveau européen, malgré quelques ***coupables*** tergiversations jusqu'à présent, doit absolument donner une signification et surtout une direction.

D'ailleurs, la ***guerre des normes*** commence à faire rage. L'introduction de cadres réglementaires, en plus des règles de gouvernance, doit permettre de formaliser la question problématique, à savoir celle de la responsabilité. Le Conseil de l'Europe[27] avertissait déjà en 2018 que « quoique bienvenus sous de nombreux aspects, les codes et normes [des acteurs des IT] en question sont généralement dépourvus de tout mécanisme de mise en œuvre et de sanction et ne peuvent donc servir de base à une véritable protection ». Les États doivent, par conséquent, veiller à la protection effective de leurs citoyens tout en favorisant les actions collectives et les voies de recours individuelles contre tout abus vis-à-vis des libertés individuelles. Le rapport du Conseil de l'Europe propose de réaliser régulièrement un « audit des algorithmes » et, partant, de financer la recherche publique pour ce faire. **La première recommandation tombe donc sous le sens : il faut un investissement massif de l'État dans le domaine de l'IA.** L'investissement dans l'éducation en général et dans l'enseignement supérieur en particulier apparaît une nécessité à l'heure où le plan de relance, français comme européen, semble sous-estimer cet aspect. Si l'injection massive d'argent frais

26 Stankovic, M. (2019), L'intelligence artificielle comme vecteur d'inégalités, partie II, Institut Sapiens.

27 Conseil de l'Europe (2019), *Management et IA*, Etude du Conseil de l'Europe DGI (2019) 05, Karen Yeung (Rapporteur), Préparée par le Comité d'Experts sur les dimensions des droits de l'homme dans le traitement des données et les différentes formes d'intelligence artificielle (MSI-AUT).

dans l'enseignement primaire et secondaire ne peut dépendre que de l'État seul, le recours à des solutions développant des partenariats publics-privés dans l'enseignement supérieur est souhaitable en même temps qu'inévitable. En mars 2018, il y a trois ans, la promesse du financement d'un plan « Intelligence Artificielle » à hauteur de 1,5Md€ était faite (***#AIFORHUMANITY***). Il sera intéressant de tirer un premier bilan à l'approche de la présidentielle début 2022.

Un investissement massif de l'État dans la recherche est nécessaire afin de permettre des « audits d'algorithme » pour comprendre ce qui se passe dans la blackbox et ne pas devenir une « colonie du monde numérique ».

Au même titre que ce que représente la sécurité sociale pour la santé des françaises et des français, doit être mise en place une sécurité sociale de la vie professionnelle. Il ne s'agit pas de garantir un emploi, et encore moins un emploi à vie, mais de garantir et d'élargir le recours à une formation tout au long de sa carrière.

Ce que les anglo-saxons qualifient de « life-long learning » n'a pas encore gagné dans la pratique l'ensemble des acteurs de la formation en France. Pourtant, l'apprentissage tout au long de la vie constitue la seule solution pour s'adapter à un environnement en constant changement. A l'image de ce qui est fait pour la validation des acquis de l'expérience (VAE[28]), **l'État doit encourager la création de cycle de formations autour de blocs de compétences en lien, entre autres, avec l'IA mais surtout avec la création/destruction de nouveaux métiers.** Cela passerait par :

- La création de titres RNCP et/ou de formations courtes sur ces domaines au sein d'organismes de formations publiques et/ou privées ;
- La mise en place de normes qualités contrôlées et validées par l'État (MESRI) sur le modèle existant (Qualiopi, CTI, CEFDG), en rendant obligatoires des analyses d'impact *de commodo et incommodo* ;
- L'obligation faite aux institutions de l'enseignement supérieur de proposer, de manière régulière, des mises à jour de leur formation professionnalisante ainsi que de leurs diplômes, comme le font déjà certaines écoles pionnières comme ***Aivancity***.

28 http://www.vae.gouv.fr/

La question de l'apprentissage de nouvelles compétences renvoie intuitivement à celle de l'emploi et des métiers. Or, ce n'est pas le métier qu'il faut conserver, mais la capacité à acquérir de nouvelles compétences. Ce n'est pas simplement ***l'employabilité*** qu'il faut développer, mais l'hybridation des savoirs et des compétences, le sens critique et l'analyse. Et ce n'est pas l'emploi qu'il faut protéger, mais la dignité de l'employé qui doit être accompagné pour se prendre, *in fine*, en main. A partir de recherches de l'ISEOR qui s'appuyait sur l'analyse de plus de 2 000 cas d'entreprises, une étude récente de l'***Institut Sapiens*** a mis en avant la création de valeur que génèrerait le développement efficace d'une formation professionnelle « utile et stratégique ». **L'utilisation de manière prospective du Compte Personnel de Formation (CPF) en constitue un des leviers principaux,** en plus du renforcement des dispositifs autour de l'apprentissage. Il s'agit ainsi d'accompagner tout un chacun à investir dans sa formation. Cela passerait, d'après nous, par :

- Un crédit d'impôt sur les prestations de formation pour les individus qui s'engagent à les suivre et à les financer. Même si à présent une minorité seulement de français s'acquitte de l'IRPP[29], les universités et les *Grandes Ecoles* trouveraient là des relais de croissance et des sources de financement.

- Le développement accru de CFA (Centre de Formation d'Apprenti) interne aux entreprises comme le permettent déjà la ***loi « Avenir Professionnel »*** de 2018 et ses ***textes d'applications***. Cela faciliterait la mise en place des formations en lien avec les besoins en compétences des entreprises et aurait l'avantage de toucher des personnes non soumises à l'IRPP.

- La relance du dialogue social autour de la formation et des métiers, au sein des entreprises, avec des interlocuteurs responsables. Dès lors, ***redonner du pouvoir aux syndicats***, comme le préconise également un avis récent du Conseil Economique Social et Environnemental sur les ***reconversions professionnelles***, tout en exigeant des contreparties (en particulier, en matière de transparence de leur financement[30]) parait une nécessité.

29 En 2017, l'IR représentait un quart des recettes fiscales de l'État français et était acquitté par 43% des contribuables (https://www.vie-publique.fr/fiches/21885-quest-ce-que-limpot-sur-le-revenu).
30 Voir le Rapport dit "Perruchot" pour l'Assemblée Nationale sur le financement des comités d'entreprise du 18 Janvier 2012, n°4186.

La création de formations spécifiques à l'IA, la mise à jour obligatoire de certains diplômes ainsi que l'utilisation de manière prospective du Compte Personnel de Formation (CPF) constituent des facteurs clés de succès.

Comme le soulignait Mary Parker Follet il y a déjà près d'un siècle, le management relève à la fois de la science et de l'art[31]. Il emprunte à la première ses méthodes scientifiques dont l'ambition d'application universelle vient se heurter à l'incontournable prise en considération d'un contexte économique, social et politique, mais aussi individuel et organisationnel. Partant, dans la pratique du management, l'art vient nécessairement suppléer la science en permettant l'adaptation et la contextualisation de principes et résultats voulus et pensés comme universels.

Car si le management est moins une science qu'un art, l'écueil du management de l'IA serait bien de la considérer comme une simple technique, au sens heideggérien du terme[32]. Sous le couvert d'une « calculabilité intégrale » qui laisse accroire à une domination à outrance de l'environnement, l'individu ne ferait que se déposséder de lui-même en croyant « arraisonner » le monde et en oubliant sa propre responsabilité au cœur de l'IA.

31 « *Business is becoming a profession and management a science and an art*. » Follet, M.P. (1987), *Freedom and coordination*, Taylor & Francis, p.41.

32 Heidegger, M. (1958), *Essais et conférences. La question de la technique*, Gallimard, Paris.