



Activités

17-1 | 2020

IA, robotique, automatisation : quelles évolutions pour l'activité humaine ?

Évolutions de l'Intelligence Artificielle : quels enjeux pour l'activité humaine et la relation Humain-Machine au travail ?

*Evolutions of Artificial Intelligence: What issues for human activities and
Human-Machine relationships at work ?*

Moustafa Zouinar



Édition électronique

URL : <http://journals.openedition.org/activites/4941>

DOI : 10.4000/activites.4941

ISSN : 1765-2723

Éditeur

ARPACT - Association Recherches et Pratiques sur les ACTIVités

Référence électronique

Moustafa Zouinar, « Évolutions de l'Intelligence Artificielle : quels enjeux pour l'activité humaine et la relation Humain-Machine au travail ? », *Activités* [En ligne], 17-1 | 2020, mis en ligne le 15 avril 2020, consulté le 20 avril 2020. URL : <http://journals.openedition.org/activites/4941> ; DOI : <https://doi.org/10.4000/activites.4941>

Ce document a été généré automatiquement le 20 avril 2020.



Activités est mis à disposition selon les termes de la licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Pas de Modification 4.0 International.

Évolutions de l'Intelligence Artificielle : quels enjeux pour l'activité humaine et la relation Humain-Machine au travail ?

*Evolutions of Artificial Intelligence: What issues for human activities and
Human-Machine relationships at work ?*

Moustafa Zouinar

NOTE DE L'ÉDITEUR

Article soumis le 8 janvier 2019, accepté le 20 janvier 2020

Introduction

- ¹ Depuis quelques années, l'Intelligence Artificielle (IA) fait l'objet d'une médiatisation et d'une attention sans précédent¹ et suscite beaucoup de promesses, mais aussi des craintes dont certaines reposent sur des visions très spéculatives ou très lointaines des capacités des machines. Ce fort regain d'intérêt pour l'IA est notamment lié à d'importantes avancées technologiques qui ont permis d'accroître de façon considérable les performances des ordinateurs dans de nombreux domaines comme la reconnaissance automatique de la parole ou la vision par ordinateur. Ces avancées ont ouvert de vastes perspectives d'introduction de l'IA sous différentes formes (applications, robots, chatbots, etc.) dans les situations de travail. Un point particulièrement notable est que de plus en plus de secteurs sont concernés (industrie, santé, agriculture, finance, banque, assurance, transport, etc.). L'IA est ainsi sur le point

de prendre une place de plus en plus importante dans les organisations et les systèmes de production. Les champs d'application de l'IA dans ces secteurs ne cessent de se multiplier (automatisation de tâches, relation client, logistique, analyse prédictive, diagnostic, analyse de grandes bases de données, etc.). Dans ce cadre, l'IA est le plus souvent vue comme un ensemble de technologies pouvant produire de nombreux bénéfices, notamment en termes de performance (optimisation de processus internes, rapidité d'exécution de tâches, accroissement de la productivité, etc.) et parfois en termes de facilitation du travail voire de réduction de la pénibilité en permettant l'automatisation des tâches fastidieuses ou répétitives. Certains discours avancent même l'idée qu'en transférant ces tâches à des systèmes d'IA, il sera possible d'orienter le travail des salariés qui seront « libérés » de celles-ci vers des activités à plus forte « valeur ajoutée ». Ces perspectives d'application ont fait l'objet de multiples rapports et d'ouvrages sur les incidences (potentiellement) positives et négatives de l'IA sur le travail et l'emploi.

- 2 L'objectif de cet article est d'examiner ces perspectives et les réflexions qu'elles suscitent d'un point de vue ergonomique, en particulier sous l'angle de l'activité. Plus précisément, dans la mesure où l'IA ne constitue pas une nouveauté technologique, il s'agit ici de traiter les questions suivantes : quels sont les nouveaux enjeux et les nouvelles questions soulevés par ces évolutions de l'IA dans le domaine du travail ? Quels sont les thèmes émergents ? Quelles sont ou vont être les conséquences de ces évolutions sur les activités humaines dans les situations de travail ?
- 3 D'un point de vue méthodologique, cet article s'appuie sur la littérature actuelle (scientifique et « grise² ») qui traite de l'IA et du travail, à la lumière des travaux et des connaissances accumulées depuis plus de 40 ans en ergonomie et dans des disciplines ou courants connexes (par exemple, facteurs humains, sociologie, anthropologie) sur les incidences de l'automatisation et l'introduction de systèmes « intelligents » sur l'activité humaine au travail. Autrement dit, il s'agit ici de proposer un exercice de réflexion et de prise de recul par rapport aux développements actuels de l'IA et à la manière dont ils sont appréhendés pour penser les transformations du travail.
- 4 Dans la suite de cet article, une première partie présente une brève histoire de l'IA, en mettant l'accent sur les étapes majeures de son développement ainsi que sur ses évolutions récentes. Cette section nous permettra de préciser ce que nous entendons par IA. Nous verrons que le regain d'intérêt actuel pour celle-ci est en particulier lié au développement de l'apprentissage automatique ou machine (*machine learning*) qui est devenu le paradigme dominant dans ce champ. Les deux parties qui suivent sont consacrées aux questions, enjeux soulevés par les évolutions récentes de l'IA en rapport avec le travail. Une première série de questions qui sera traitée dans la deuxième partie concerne d'une part les conséquences de l'automatisation du travail par l'IA, et, d'autre part, la division du travail et la relation entre Humains et IA. Nous montrerons qu'une grande partie de ces questions et les cadres conceptuels qui sont mobilisés pour penser cette relation ne sont pas nouveaux. Ceci nous permettra également de re-questionner à nouveaux frais ces approches. Nous aborderons également un point qui nous paraît relativement nouveau et qui concerne l'interaction au travail avec des machines « intelligentes » qui présentent des caractéristiques anthropomorphes. La troisième partie est consacrée à « l'explicabilité » ou « interprétabilité » des systèmes d'IA, qui renvoie de façon générale à la compréhension ou intelligibilité de leur fonctionnement et de leurs « décisions » ou actions³. Nous verrons que cette question, qui est en

particulier soulevée par la complexité des techniques d'apprentissage machine, mais qui n'est pas non plus nouvelle, peut-être importante à prendre en compte dans le cadre d'activités de travail qui impliquent des interactions avec des systèmes qui fonctionnent selon ces techniques⁴.

1. Une brève histoire de l'IA : genèse et évolutions récentes

- 5 L'IA a été définie de multiples manières et reste de ce fait un domaine dont il n'est pas facile de circonscrire précisément les contours. La pertinence du terme d'IA fait même débat et il est régulièrement remis en question⁵. Une définition assez large est celle proposée par Shapiro (1992) qui considère l'IA comme le domaine de la science et de l'ingénierie qui traite de la compréhension, à l'aide de l'ordinateur, du comportement intelligent et de la création de systèmes artificiels qui reproduisent ce comportement. D'autres définitions mettent plutôt l'accent sur l'aspect scientifique de l'IA, par exemple, Levesque (2013) définit l'IA comme le domaine qui étudie le comportement intelligent en termes computationnels. Selon Nilsson (2005), pour de nombreux chercheurs du domaine, l'objectif scientifique à long terme de l'IA en tant que discipline est la « mécanisation » de l'intelligence humaine (*mechanization of "human-level" intelligence*). Il convient de noter que si l'humain a constitué la référence principale pour le développement de machines intelligentes, les chercheurs du domaine se sont également inspirés des comportements d'autres organismes vivants comme les insectes sociaux. Dans la suite de cet article, l'expression « IA » se référera aux machines, algorithmes, ou programmes qui s'inspirent ou tentent de reproduire des facultés humaines comme la compréhension du langage naturel, la reconnaissance d'objets visuels ou le raisonnement dans ses différentes formes. Mais, pour lever toute ambiguïté, notons dès à présent que, sous réserve que l'on dispose d'une bonne compréhension de ce qu'est l'intelligence humaine, la reproduction « informatique » de cette intelligence reste un horizon lointain. Certains spécialistes reconnus du domaine considèrent même que l'IA n'est pas si « intelligente » qu'on le dit (Lecun, 2019 ; Julia, 2019). Selon Julia (2019), ce n'est pas l'intelligence qui caractérise les systèmes d'IA d'aujourd'hui, mais leur capacité de reconnaissance grâce à l'apprentissage machine sur lequel nous reviendrons⁶.

2. Évolutions de l'IA : le retour en grâce des réseaux de neurones artificiels

- 6 L'acte de naissance officielle de l'IA est généralement situé dans les années 50, lors de la conférence Macy qui avait regroupé différents chercheurs (informaticiens, mathématiciens, psychologues) qui s'étaient donnés comme objectif d'explorer l'hypothèse selon laquelle il est possible de simuler de manière informatique l'apprentissage et l'intelligence humaine :

« (...) tous les aspects de l'apprentissage ou toute autre caractéristique de l'intelligence peuvent en principe être décrits avec une telle précision qu'une machine peut être construite pour les simuler⁷. » McCarthy, Minsky, Rochester, & Shannon (1955)

- 7 Cette hypothèse a été appliquée à différents domaines comme la résolution de problèmes, la compréhension du langage, la démonstration de théorèmes mathématiques, ou encore la perception (par exemple, la reconnaissance de formes visuelles). C'est dans ce cadre que le projet de simuler l'intelligence humaine a occupé une place centrale dans les recherches en IA.
- 8 À partir des années 50, l'IA s'est structurée autour de deux courants principaux, l'approche symbolique et l'approche connexionniste. La première, qui a été dominante jusqu'aux années 80, consiste à envisager l'intelligence en termes de manipulation de symboles sur la base de règles formelles ; c'est l'hypothèse des « systèmes symboliques physiques » formalisée par Newell et Simon (1976). La seconde approche, qui remonte aux années 40 avec les travaux pionniers de McCulloch et Pitts (1943), consiste à s'inspirer de la structure et du fonctionnement des réseaux de neurones biologiques. Cette approche repose ainsi sur l'idée que la perception ou le comportement intelligent résultent de l'interaction entre différentes unités computationnelles (neurones formels) interconnectées et régies par des règles d'apprentissage. L'un des exemples emblématiques de réalisation de cette approche est le perceptron, un algorithme d'apprentissage supervisé (cette notion est expliquée plus bas). Le connexionnisme a fait l'objet d'importantes critiques qui ont freiné son développement jusqu'aux années 80, notamment suite à la publication de l'ouvrage de Minsky et Papert (1969) où les auteurs montrent les limites calculatoires des réseaux de neurones artificiels développés dans les années 60, en particulier le *perceptron*⁸. Enfin, signalons qu'un autre courant, l'Intelligence Artificielle Distribuée (IAD), s'est développé à partir des années 70. L'IAD développe et étudie des systèmes multi-agents, c'est-à-dire des systèmes composés d'agents computationnels intelligents, actifs et autonomes qui interagissent entre et/ou avec leur environnement (Ferber, 1995 ; Weiss, 1999). S'inspirant notamment des comportements des insectes sociaux, une des particularités de ce courant est de simuler des interactions sociales (coopération, compétition, etc.) qui sont supposées mieux adaptées pour résoudre certains types de problèmes de manière plus efficace et plus rapide. Dans ce cadre, des comportements émergents d'apparence complexe sont simulés à partir de règles « simples ».
- 9 Les travaux réalisés en IA, symbolique notamment, ont progressivement abouti à des applications pratiques pour les situations de travail, en particulier sous la forme de systèmes dits « experts⁹ ». Un système expert est supposé reproduire l'expertise cognitive d'un humain, en particulier ses raisonnements et ses connaissances, dans un domaine particulier. L'expertise est modélisée comme un ensemble de règles de type « si-alors ». Ces systèmes ont été développés dans de nombreux secteurs (médecine, industrie, etc.) avec différentes finalités (diagnostic, résolution de problèmes, prise de décision, etc.) Le premier système développé est DENDRAL, qui a été conçu dans les années 60 pour aider les chimistes à identifier la structure de molécules organiques. Les concepteurs du système ont mis en œuvre ce qu'ils ont appelé une « ingénierie de la connaissance » (*Knowledge engineering*) qui consiste à « extraire » les connaissances des experts (en l'occurrence des chimistes) pour fabriquer les règles. DENDRAL a été ensuite utilisé comme base pour la conception d'un autre système, MYCIN, qui visait à assister les médecins dans le diagnostic médical. À partir de ces travaux pionniers, de nombreux systèmes experts ont été développés dans d'autres secteurs, par exemple, dans l'industrie métallurgique pour la conduite de Hauts Fourneaux, dans les transports pour le diagnostic de pannes ou encore dans la finance et la banque. Les

usages de l'IA dans les entreprises ou de façon plus générale dans les environnements de travail existent donc depuis longtemps. Notons dès à présent que de nombreux travaux d'orientation ergonomique ont porté sur ces systèmes, ou de façon plus générale, se sont intéressés à l'IA, dans un cadre de conception d'outils d'aides à l'activité ou d'utilisation de celle-ci comme outil de simulation (par exemple : Amalberti, & Deblon, 1992 ; Benckroun, Pavard, & Salembier, 1994 ; Cacciabue, Decortis, Drozdowicz, Masson, & Nordvik, 1992 ; Dugdale, & Pavard, 2000 ; Haradji, Guibourdenche, Reynaud, Poizat, Sabouret, Sempé *et al.*, 2018 ; Salembier, & Zouinar, 2004 ; Woods, Roth, & Bennett, 1987 ; Zouinar, 2000).

- 10 Bien que les recherches n'aient pas cessé depuis la conférence de Macy, il est coutume de dire que l'IA a traversé plusieurs « hivers » (entre les années 70 et les années 90) au sens où elle a rencontré d'importants obstacles et des échecs (par exemple, dans les domaines de la traduction automatique et la compréhension de la parole) qui ont conduit à une baisse importante des financements et de l'intérêt des acteurs institutionnels pour ce domaine. Ces échecs ont notamment soulevé de nombreux débats et critiques sur les présupposés théoriques sur lesquels s'appuyaient l'IA, notamment l'idée cognitive que d'une part, toute forme d'intelligence (raisonnement, compréhension du langage, résolution de problèmes, etc.) consiste en une manipulation de symboles suivant des règles formelles, et d'autre part, que l'action se détermine par des « représentations mentales » ayant le statut de « plans » (voir par exemple les critiques de Searle, 1980 ; Dreyfus, 1972 ; et Suchman, 1987, et plus récemment, Collins, 2018). Ainsi, les critiques ont notamment mis en évidence l'absence de prise en compte du rôle majeur du contexte (matériel, social, culturel) et du corps dans la cognition et l'action humaines¹⁰. Ces critiques ont donné lieu au développement de nouveaux courants anti-cognitivistes avec, par exemple, les travaux du roboticien Rodney Brooks sur une intelligence artificielle « sans représentation », c'est-à-dire qui fonctionne de façon décentralisée et où l'intelligence émerge continuellement de l'interaction du système avec le monde sur la base de comportements « simples » (Brooks, 1991). Un autre courant qui s'est développé à partir de ces critiques et qui rejoint celui développé par Brooks est « l'IA incarnée » (*Embodied AI*) qui s'appuie sur l'idée de cognition incarnée, idée qui, rappelons-le, rejette le dualisme cognitiviste consistant à concevoir la cognition indépendamment du corps et de l'environnement. En IA, cette idée consiste par exemple à développer des robots dont les actions sont contrôlées par des boucles sensori-motrices ou à concevoir des morphologies spécifiques qui rendent possibles des déplacements comme la marche.
- 11 Depuis les années 2010, l'IA suscite une forte attention avec les performances obtenues grâce à l'apprentissage machine et, plus particulièrement, l'apprentissage profond (*deep learning*) qui repose sur des réseaux de neurones artificiels (approche connexionniste). De façon générale, l'apprentissage machine, parfois appelée apprentissage statistique, consiste à apprendre à une machine à réaliser des tâches (par exemple : reconnaissance d'objets dans des images) en l'entraînant sur des données du domaine concerné (par exemple, des images qui comportent des voitures). Dans ce domaine, on distingue habituellement les algorithmes des modèles d'apprentissage. Un modèle résulte de l'apprentissage qui se construit à partir de l'application des algorithmes aux données utilisées dans la phase d'apprentissage. Ce modèle peut-être vu comme un arrangement et une mise en relation mathématique complexe de paramètres ou coefficients. Autrement dit, un modèle d'apprentissage, c'est « l'algorithme + les données d'apprentissage ». Il existe plusieurs types d'algorithmes

d'apprentissage : les « classifieurs » qui reçoivent une donnée en entrée (par exemple une image) et produisent une « sortie » (par exemple une catégorie comme « oiseau » qui est reconnue dans l'image) ; et les algorithmes qui élaborent à partir de données d'apprentissage un modèle qui est utilisé par le classifieur pour produire ses réponses. L'apprentissage peut être de type supervisé (à partir de données fournies et étiquetées par un humain), non supervisé (l'algorithme doit découvrir par lui-même la structure plus ou moins cachée des données), ou par renforcement (l'algorithme apprend un comportement à partir d'observations et ses actions sur l'environnement produisent des valeurs de retour (*feedback*) qui le guident). L'apprentissage profond, dont on parle beaucoup aujourd'hui, est une technique qui se distingue par l'utilisation de plusieurs couches de réseaux de neurones artificiels qui extraient et traitent de manière successive des informations spécifiques d'une entrée (par exemple, une image). Cette technique n'est pas nouvelle puisque son développement est amorcé à la fin des années 80 avec notamment les travaux de Lecun sur les réseaux *convolutifs*¹¹, mais a été délaissé entre le milieu des années 90 et 2012 du fait de la difficulté à la faire fonctionner avec les ordinateurs de cette époque, car ils n'étaient pas assez puissants. L'année 2012 est importante, car c'est le moment où ces réseaux de neurones ont permis d'atteindre des performances inédites (dans la reconnaissance vocale et la vision par ordinateur), notamment grâce à l'augmentation importante des capacités de calcul des machines informatiques et la disponibilité de volumes de plus en plus grands de données (*Big Data*). Cette technique, qui prend une place prépondérante en IA, est de plus en plus utilisée dans la conception d'applications ou de produits¹² (par exemple, les assistants vocaux, les véhicules autonomes, les robots dits « sociaux¹³ », ou les systèmes d'analyse d'images comme dans l'imagerie médicale). Enfin, avec les performances atteintes par les machines apprenantes actuelles ressurgit le débat concernant la possible émergence d'une IA générale (Goertzel, 2014) voire même « consciente », autrement dit qui serait la même que celle de l'humain (*Human-level Artificial Intelligence*) ou la dépasserait¹⁴ (« *Superintelligence* », Boström, 2014). C'est ce qui est appelé l'IA « forte » par opposition à l'IA « faible » qui, elle, renvoie aux systèmes actuels qui sont spécialisés dans des domaines de tâches particuliers. Pour certains auteurs comme Lecun (*Les Échos*, 2019), une telle intelligence générale ne peut être atteinte qu'avec l'apprentissage non supervisé, ce qui signifie des systèmes capables d'apprendre de façon autonome, en interagissant avec le monde. Il considère qu'il faudrait doter les machines de « sens commun » pour elles soient véritablement intelligentes, un objectif qui a déjà fait l'objet de tentatives dans le cadre de l'approche symbolique, en particulier celle de Lenat, Guha, Pittman, Pratt et Shepherd (1990) et qui consiste à spécifier ce sens sous forme de connaissances et de règles. À la différence de cette tentative qui a fait l'objet de nombreuses critiques, notamment l'impossibilité de modéliser le sens commun de manière symbolique, l'apprentissage autonome est supposé remédier à cette limite. Tenant compte de l'émergence de l'IA forte comme un scénario probable, Russell (2019) suggère de repenser l'IA d'une nouvelle façon. Selon cet auteur, jusqu'à présent le développement de ce champ s'est principalement appuyé sur la reproduction de l'intelligence humaine considérée comme la capacité à réaliser des buts. Cette approche devient problématique dans le scénario d'une IA forte. Pour Russell, dans ce cas, une machine au moins aussi « intelligente », voire plus « intelligente » que l'humain pourrait en principe chercher par tous les moyens possibles de réaliser un but qui lui a été fixé, sans tenir compte de l'ensemble de ses (l'humain) préférences relativement à la manière de l'atteindre. Une telle machine

pourrait par exemple chercher à réaliser le but y compris en commettant des actions qui mettent en danger la vie d'autres humains. Or, en tant qu'humain, lorsqu'on assigne (sans mauvaises intentions) un objectif à quelqu'un, cela sous-tend qu'on ne lui demande pas de le réaliser à n'importe quel prix (par exemple, nuire à une personne ou mettre sa vie en danger). C'est ce que Russell appelle la « structure de préférences » qui est donc communiquée de façon implicite. Pour faire face à ce problème du contrôle de la réalisation du but, Russell considère qu'il est nécessaire de développer des machines « compatibles avec l'humain » (*Human Compatible*), c'est-à-dire qui apprennent nos structures de préférence et nos buts à partir de l'observation de nos actions, grâce notamment à l'apprentissage par renforcement inverse (apprentissage des buts qu'un humain cherche à réaliser à partir de l'observation de ses comportements). Les actions des machines seront ainsi mieux alignées avec les objectifs et les préférences des humains. De façon plus générale, Russell plaide pour une IA qui reste sous le contrôle de l'humain.

- 12 Concernant le débat sur l'IA faible vs forte, deux points liés intéressants sont à noter. Un premier est que cette distinction entre IA faible et IA forte a été utilisée pour la première fois, mais d'une manière un peu différente par Searle (1980). Pour cet auteur, selon les tenants de l'IA faible, l'intérêt de prendre l'ordinateur comme modèle dans l'étude de la cognition (par exemple, pour faire des simulations ou modélisations) est que cela permet de développer un outil « cognitif » puissant ; en revanche, selon la position forte, il n'y pas de différence de nature entre l'humain et un ordinateur au plan du fonctionnement cognitif¹⁵. Autrement dit, il serait possible de développer des machines qui ont strictement le même niveau d'intelligence que l'humain. Le deuxième point est la condition proposée par Nilsson (2005) pour déterminer si l'IA forte est atteinte. Pour cet auteur, cette IA adviendra lorsque les machines pourront réaliser la plupart des activités de travail ordinaires accomplies par les humains, autrement dit lorsque ces activités seront totalement automatisées. C'est ce qu'il appelle le « test de l'emploi » (*employment test*), qu'il considère moins problématique que le test de Turing qui avait été également proposé comme un test de « mesure » de l'intelligence des machines :

« Pour réussir le test de l'emploi, les programmes d'IA doivent être en mesure d'effectuer les tâches habituellement exécutées par des humains. Les progrès vers une IA équivalente à l'intelligence humaine [IA forte] pourraient alors être mesurés en fonction de la fraction de ces tâches qui peut être effectuée de façon acceptable par des machines¹⁶. » (p. 68)

- 13 On voit ainsi que le travail est au cœur de l'IA, pour au moins certains auteurs. L'automatisation du travail occupe une place centrale dans les débats sur l'IA, comme nous allons le voir.

3. L'automatisation par l'IA : une menace pour le travail humain ?

- 14 L'automatisation informatique des tâches professionnelles constitue l'une des conséquences de l'IA les plus débattues, notamment parce qu'elle touche de plus en plus de secteurs (secteur des services, industrie, agriculture, éducation, finance, médecine, etc.). Cette évolution est vue comme une troisième vague d'automatisation qui se caractérise principalement par l'extension de plus en plus grande de celle-ci aux

aspects cognitifs des activités humaines comme le raisonnement ou la prise de décision, les deux premières vagues étant respectivement l'automatisation des tâches manuelles « sales » / « dangereuses », puis celle des tâches répétitives, monotones (Davenport, & Kirby, 2016). L'IA est également considérée comme participant d'une nouvelle ère industrielle, « l'industrie 4.0 » (Wisskirchen, Thibault Biacabe, Bormann, Muntz, Niehauss, Soler *et al.*, 2017). Si l'automatisation de tâches par des systèmes d'IA n'est pas nouvelle puisqu'il s'agit d'une tendance qui remonte aux systèmes experts (cf. plus haut), il devient aujourd'hui possible de l'étendre à des domaines de tâches de plus en plus larges grâce à l'apprentissage machine, en particulier l'apprentissage profond. Cette extension de l'automatisation par l'IA est aussi rendue possible par la généralisation dans les situations de travail de l'usage d'outils informatiques qui permet aujourd'hui de disposer des données nécessaires à l'apprentissage¹⁷ (Crowston, & Bolici, 2019).

- 15 Les débats actuels se cristallisent ainsi autour de la question du remplacement généralisé des humains par des systèmes d'IA. Cette question sur les effets de l'automatisation, qui s'est posée à de multiples reprises depuis la révolution industrielle, a généré de nombreuses discussions et ouvrages plus ou moins alarmistes ; certains prévoient en effet une disparition massive des emplois (Brynjolfsson, & McAfee, 2012 ; Ford, 2016). Le détonateur de ces débats a notamment été l'étude réalisée par deux chercheurs, Frey et Osborne (2013), qui ont analysé les risques d'automatisation d'activités de travail (jobs), en s'adressant à des experts en apprentissage automatique. De leur étude, il ressort qu'en probabilité, 47 % des emplois existants aux États-Unis sont menacés d'automatisation totale. Les métiers les plus à risque se situent dans les domaines du transport, de la logistique, du support administratif ou encore des services. Selon cette étude, même les métiers « très qualifiés » seront également touchés (par exemple, médecin, comptable ou juriste). En outre, ces auteurs considèrent que les tâches non-routinières (« manuelles » ou « cognitives ») pourraient également être automatisées grâce à l'IA. C'est notamment le cas du dépannage de machines, la maintenance, la surveillance/détection des anomalies ou la prise de décision. Mais cette étude a été très critiquée, notamment au plan de la maille d'analyse des métiers utilisée par Frey et Osborne. Ces derniers se sont en effet concentrés sur le métier pris de façon globale ; or, dans de nombreux cas, tous les aspects des activités relatives à un métier ne sont pas nécessairement automatisables (Arntz, Gregory, & Zierahn, 2016). Négliger ce point conduit à une surestimation du nombre de métiers réellement automatisables dans leur totalité (Arntz *et.*, 2016). Partant d'une approche centrée sur les tâches qui composent un métier, Arntz, Gregory et Zierahn ont ainsi trouvé qu'en moyenne 9 % seulement des emplois dans plusieurs pays de l'OCDE étaient entièrement automatisables. Les emplois qui seront cependant les plus massivement touchés sont ceux qui sont les « moins qualifiés ». D'autres études ont porté sur les incidences de cette automatisation en termes de temps de travail. Ainsi, selon une étude récente du forum économique mondial¹⁸, la part globale du « travail » en nombre d'heures réalisé par les machines va progressivement augmenter pour atteindre 52 % (48 % seront réalisés par l'humain) alors qu'aujourd'hui la répartition est de 71 % pour l'humain et de 29 % pour les machines. Ceci pose la question importante de savoir à quoi peut être réaffecté le temps de travail libéré, comment repenser le contenu des activités. Mais, toutes ces études restent des projections hypothétiques. Ce qui semble faire consensus, c'est que l'introduction de l'IA dans les situations de travail se traduira par des reconfigurations

plus ou moins profondes des activités humaines. Les points importants sont de comprendre quelles formes prendront ces configurations, comment les anticiper et les orienter de façon pertinente. En outre, l'un des enseignements que l'on peut tirer des recherches passées sur l'automatisation est que l'humain reste indispensable dans les environnements de travail, en particulier pour gérer des événements imprévus ou lorsque les systèmes dysfonctionnent. Bainbridge (1983) faisait ainsi remarquer que plus un système de contrôle automatisé est avancé (techniquement) plus la contribution de l'humain peut-être cruciale. Un autre enseignement est que l'automatisation implique surtout une transformation plus ou moins profonde des activités de travail et l'émergence de nouveaux rôles (Dekker, & Woods, 2002). Et cette conséquence est souvent peu prise en compte dans la mesure où l'automatisation est souvent pensée comme une simple « substitution ». C'est le « mythe de la substitution » (Carr, 2017 ; Dekker, & Woods, 2002). Sur cette base bien établie, il semble donc que la nouvelle (et peut-être future ?) génération de systèmes d'IA impliquera surtout des formes d'automatisation partielle (au moins à court moyen terme) et par conséquent une mutation des activités humaines au travail. Le défi qui se pose est comment penser la reconfiguration des activités humaines et quelle place doit occuper l'IA vis-à-vis de l'humain et réciproquement. Des analyses intéressantes de cette reconfiguration se développent. Par exemple, Seidel, Berente, Lindberg, Nickerson et Lyytinen (2019) montrent comment l'usage d'outils d'IA qui génèrent des solutions de conception transforme de manière importante l'activité de design et requiert des compétences nouvelles. Ainsi, avec ces outils, l'activité du designer n'est plus une activité de conception au sens « classique » (fabriquer, modifier et ainsi de suite), mais est de plus en plus structurée par les pratiques suivantes : définir les paramètres/contraintes que le système doit suivre, évaluer et tester les solutions générées par ce dernier, ajuster les paramètres ou les algorithmes en fonction de l'évaluation. L'usage de ces outils implique ainsi une bonne compréhension de leur fonctionnement notamment pour comprendre et le cas échéant modifier les solutions qu'ils proposent.

- 16 Parallèlement à ces débats portant sur les conséquences de l'IA sur l'emploi, une autre série de questions soulevées qui s'inscrivent dans une logique de complémentarité plutôt que de substitution concerne la manière dont il faudrait envisager la division du travail et la relation entre humains et IA : quel rôle doit tenir un système d'IA par rapport à l'humain et réciproquement ? De quelle nature doit-être leur relation ? Ce sont ces questions que nous allons maintenant examiner, car elle concerne directement la transformation des activités humaines au travail puisqu'elles impliquent de travailler avec des IA.

4. Travailler avec des systèmes d'IA : quelles divisions du travail et relations entre humains et IA ?

- 17 Cette partie examine les modèles de complémentarité Humains-IA qui sont mobilisés aujourd'hui pour définir le rôle de la nouvelle génération de systèmes d'IA vis-à-vis de l'humain (et réciproquement) dans le cadre d'activités de travail. Nous allons voir que ces modèles s'inscrivent dans la continuité d'approches de la relation humain-machine développées depuis longtemps. Nous verrons ensuite que le travail avec des systèmes d'IA anthropomorphes, c'est-à-dire dotés de caractéristiques « physiques » humaines, implique de nouvelles formes de relation aux machines.

4.1. Les modèles de complémentarité Humains-IA

- 18 Pour de nombreux auteurs, les nouvelles perspectives d'automatisation ouvertes par l'IA nécessitent de repenser la division du travail entre humains et machines, car les progrès réalisés dans ce domaine font que les systèmes d'IA deviennent plus performants que les humains dans de nombreux cas comme le diagnostic ou l'analyse prédictive. Par exemple, on nous annonce régulièrement que des systèmes d'IA obtiennent de meilleures performances que les médecins dans la détection ou l'estimation de probabilités d'apparition de maladies¹⁹. Partant de ce constat de supériorité grandissante de systèmes d'IA, McAfee et Brynjolfsson (2017) soutiennent qu'il est nécessaire de repenser le « partenariat » (*partnership*) Humains-Machines. Dans cette perspective, ils considèrent qu'il faut s'orienter vers une complémentarité entre le « jugement humain » et les algorithmes en inversant ce partenariat. Plutôt que de considérer les machines comme fournisseuses de données qui vont être utilisées par l'humain pour prendre une décision ou porter un jugement sur une situation, ils proposent en effet que le jugement de l'humain, ses « intuitions » servent de données pour les algorithmes. Deuxièmement, ils considèrent qu'il faut déléguer aux machines toutes les tâches (par exemple, de traitement de l'information ou de prise de décision) susceptibles d'être automatisées, quitte à mettre l'humain « hors de la boucle » (de la prise de décision). Ils envisagent de laisser la possibilité à l'humain de contourner ou passer outre les décisions du système lorsque cela est utile (par exemple, dans le cas de situations inhabituelles ou de conditions nouvelles non anticipées ou non gérables par la machine). Pour justifier cette philosophie de répartition des rôles entre l'humain et la machine qu'ils envisagent donc comme un « partenariat », outre l'amélioration des performances des machines informatiques, les auteurs s'appuient également sur les limites de la cognition humaine, en partant de la distinction entre deux systèmes de pensée (1 et 2) qui a été développée par Kahneman (2012). Le système 1 renvoie à la pensée rapide, intuitive, où les émotions jouent un rôle important ; le système 2 correspond à la pensée basée sur le raisonnement logique ; cette pensée est contrôlée, réfléchie, et se caractérise par sa lenteur. Selon McAfee et Brynjolfsson (2017), étant donné que le système 1 est entaché de nombreux biais, défauts qui sont source d'erreurs et « d'imperfections » cognitives, il est préférable de favoriser le système 2, mais dans sa version numérique, c'est-à-dire développer des systèmes qui fonctionnent de façon logique, sans ces biais²⁰, et ce, dans tous les domaines où l'IA s'avère plus performante que l'humain. Ainsi, dans cette approche, les systèmes d'IA apparaissent en réalité comme des « prothèses » qui visent à pallier les déficiences cognitives humaines²¹. Or, comme cela a été montré par des études passées sur les systèmes experts, cette approche prothétique est problématique lorsqu'elle met l'humain dans une posture passive (Roth *et al.* 1987). Par exemple, elle peut se traduire par une perte d'expertise ou de capacité à développer une compréhension de la situation en situation d'activité. Et c'est précisément contre cette approche que des perspectives qui donnent un rôle plus actif à l'opérateur ont été développées, par exemple en considérant les machines intelligentes comme des « outils cognitifs » ou « instruments²² » (Roth *et al.*, 1987). Qui plus est, McAfee et Brynjolfsson mettent également en avant les gains en performance (productivité, efficacité, etc.) comme deuxième critère principal qui justifie le transfert des tâches à des machines lorsque cela est possible. Ainsi, de façon générale, bien que les auteurs envisagent la relation Humains-IA comme un

« partenariat », leur approche semble plutôt privilégier la performance et suivre une logique de compensation des défaillances cognitives de l'humain.

- 19 L'idée consistant à distribuer des tâches entre Humain et machine selon leurs forces et leurs faiblesses n'est pas nouvelle ; elle fait écho à une approche ancienne de l'automatisation qui remonte au travail réalisé par Fitts et ses collègues dans les années 50 (Fitts, 1951). Cette approche repose sur le principe suivant : si une machine surpasse l'humain dans une « fonction » donnée, la « fonction » doit être automatisée ; si cela n'est pas possible, l'automatisation n'a pas de sens (de Winter, & Dodou, 2014). Ancrée dans le paradigme de l'Homme comme système de traitement de l'information, cette approche propose une liste qui compare les capacités perceptives et cognitives de l'humain par rapport à celles des machines²³. Par exemple, elle indique que l'humain est capable de détecter de très petites variations de stimulations visuelles ou acoustiques, d'improviser ou d'utiliser des procédures flexibles, ou encore de raisonner de manière déductive ; la machine est capable de répondre très rapidement à des signaux, de réaliser des tâches répétitives, ou encore de raisonner de façon déductive et traiter des opérations complexes (faire plusieurs choses à la fois). Cette approche a fait l'objet de nombreuses critiques (Dekker, & Woods, 2002 ; de Winter, & Dodou, 2014). Par exemple, son caractère statique ; le fait qu'elle suppose que les capacités de l'humain sont fixes et qu'elle repose sur une théorie de l'activité humaine qui a été largement remise en question (celle qui considère l'Homme comme système de traitement de l'information) ; le principe même d'une comparaison « mécaniste » est remise en question, c'est-à-dire comparer deux types d'entités qui sont ontologiquement différentes suivant un langage purement « fonctionnaliste » et, enfin, son caractère trop général, c'est-à-dire le fait qu'il ne permet pas de déterminer les tâches qui doivent être automatisés²⁴. Enfin, un autre point problématique dans l'approche proposée par McAfee et Brynjolfsson (2017) concerne le rôle de l'humain dans la prise de décision. Comme nous l'avons vu, selon eux, il n'est pas nécessaire et il est même préférable de mettre ce dernier « hors de la boucle de décision » lorsque le système est (supposé être) plus performant. Or, les travaux réalisés en ergonomie/facteurs humains ont depuis longtemps mis en évidence les conséquences délétères d'un tel principe sur l'activité (par exemple, la réduction de la compréhension de la situation ou la difficulté de reprise en main du système en cas d'imprévus).
- 20 Une autre approche de la complémentarité consiste à penser la relation entre Humain et IA en termes d'augmentation ou de collaboration²⁵. Dans cette perspective, l'IA est vue comme une aide à l'activité et non pas comme une technologie substitutive. Norman (2017) considère ainsi qu'à mesure que les technologies d'IA se développent il faut penser cette relation en termes de « travail d'équipe » (*teamwork*). Selon cet auteur, lorsqu'il s'agit de concevoir un système intelligent, il faut les considérer comme des collaborateurs (*collaborators*), et non pas des substituts à l'humain. L'un des exemples de « travail d'équipe » qu'il donne est celui d'un système qui aide les designers à explorer des solutions de conception. Exploitant les techniques d'IA les plus récentes, ce système génère des solutions sur la base de contraintes/paramètres qui sont fixés par l'utilisateur. Pour Norman, cette association IA-designer est de type « travail d'équipe » puisque chacun réalise une partie du travail de conception. Mais on peut se demander si ce n'est pas juste un exemple d'assistance plutôt qu'une véritable collaboration. Nous reviendrons sur ce point. Dans la même veine, Davenport et Kirby (2016) considèrent que la stratégie qu'il faut adopter vis-à-vis de ce « nouveau » contexte technologique

n'est pas l'automatisation dans sa version substitutive, mais « (...) de considérer les machines intelligentes comme des partenaires, des collaborateurs dans le travail du savoir » (p. 53). Partant de cette idée, ils distinguent quatre types de positionnement de l'humain par rapport aux machines intelligentes. Ces types reposent notamment sur une distinction entre les capacités des humains et celles des machines (on retrouve l'approche de Fitts). Par exemple, dans le positionnement « parallèle », l'humain garde des activités qui nécessitent des « aptitudes » (par exemple, la créativité) qui ne sont pas (encore parfaitement) réalisables par des machines ; dans le positionnement « inclusif », l'humain s'appuie sur les ordinateurs pour optimiser ces décisions (augmentation) et est capable de les superviser. Daugherty et Wilson (2018) adoptent une approche augmentative similaire, mais incluent également dans leur modèle ce que l'humain peut faire pour « augmenter » l'IA. Autrement dit, la relation Humain-IA est pensée de manière réciproque. Ces auteurs partent de l'hypothèse que les professions qui requièrent du leadership, de la créativité ou une capacité à évaluer les situations de façon globale ou holistique continueront d'être réalisés par des humains. En revanche, ils considèrent que les métiers qui nécessitent l'utilisation de modèles prédictifs, des actions routinières ou des adaptations au contexte qui peuvent être spécifiées par des règles ou des modèles, seront de plus en plus réalisés par des machines. Entre ces deux extrêmes, il existe cependant plusieurs formes de complémentarité qui peuvent être envisagées et qu'ils appellent *missing middle*. Cette partie est divisée en deux. Une première se réfère aux formes où l'humain « augmente » la machine en l'alimentant en données, en définissant la bonne stratégie d'apprentissage qui lui permettra d'atteindre le niveau de performance recherché, en lui faisant des retours sur ces performances, ou en expliquant ses actions (ceci ne va pas de soi comme nous le verrons dans la troisième partie). Cette idée d'augmentation de la machine par l'Homme, qui est relativement nouvelle et qui est de plus en plus envisagée (voir également Dellermann, Calma, Lipusch, Weber, Weigel, & Ebel, 2019) laisse cependant une question ouverte, à savoir si elles s'ajouteront aux activités habituellement réalisées par les opérateurs, ce qui soulèverait la question de l'augmentation de la charge de travail de l'humain. La seconde partie du modèle regroupe les formes dans lesquelles l'humain est « augmenté » par la machine. Trois types d'augmentation sont envisagés : l'amplification cognitive (par exemple, la possibilité de traiter de vaste corpus de données rapidement), l'extension des capacités d'interaction (par exemple, avec les machines via la parole et le geste), et l'incorporation physique (les auteurs donnent comme exemple les exosquelettes). Un exemple donné par les auteurs pour illustrer « l'amplification » est le système Illumeo développé par Philips. Ce système vise à aider les radiologues dans leur travail d'une part en enregistrant et en reproduisant des protocoles d'affichage personnalisés (adaptation aux pratiques de l'utilisateur), et d'autre part, en leur permettant de consulter de manière contextualisée des informations importantes sur le patient (antécédents et diagnostics du patient, résultats de laboratoire, comptes rendus d'imagerie antérieurs, etc.).

- 21 Cette perspective complémentaire est même érigée en objectif politique stratégique du développement de l'IA. Par exemple, dans l'un des rapports publiés par le *National Science and Technology Council* des États-Unis (*Strategy 2: Develop Effective Methods for Human-AI Collaboration*²⁶), trois formes de « collaboration » entre l'IA et l'humain :

- l'IA prend en charge des tâches périphériques qui soutiennent l'humain dans la prise de décision, par exemple, des tâches d'analyse prédictive ou de récupération d'informations ;

- l'IA intervient lorsque l'humain se retrouve dans une situation de charge de travail élevée, en réalisant des tâches complexes de supervision/contrôle, de prise de décision ou de diagnostic lorsque l'humain a besoin d'assistance ;
 - l'IA réalise des tâches à la place de l'humain lorsque ses capacités sont limitées ou insuffisantes pour les accomplir, par exemple quand il faut agir vite, traiter un grand volume de données, réaliser des opérations mathématiques complexes ou intervenir dans des environnements hostiles.
- 22 Dans ce modèle, on retrouve là encore une division du travail qui rappelle l'approche de Fitts dans la mesure où elle se fonde sur une comparaison entre les (in)capacités de l'humain avec celles des machines « intelligentes ». Les auteurs du rapport ajoutent que certains principes bien connus de la « conception centrée humain » doivent être mobilisés pour permettre à l'humain d'avoir une bonne compréhension des capacités du système, de ce qu'il peut faire et ne pas faire. Les principes retenus auxquels on ne peut que souscrire sont de :
- concevoir des interfaces utilisateurs « intuitives », qui facilitent l'interaction avec l'IA ;
 - garder l'humain informé des actions et des états du système ;
 - former régulièrement l'utilisateur pour lui permettre d'acquérir les compétences/ connaissances nécessaires à l'interaction avec le système, notamment celles concernant son fonctionnement (logiques et algorithmes utilisés) et ses défaillances possibles ;
 - flexibiliser l'automatisation, c'est-à-dire : 1/ le cas échéant, donner à ceux qui le souhaitent la possibilité de choisir d'utiliser le système ou non ; 2/ concevoir des systèmes capables de s'adapter de façon dynamique à l'activité humaine, par exemple intervenir lorsque l'opérateur se retrouve dans une situation de charge de travail ou de fatigue excessive.
- 23 De façon générale, les approches complémentaires constituent des directions qui font globalement sens d'un point de vue ergonomique dans la mesure où elles apparaissent comme étant par certains côtés centrées sur l'Humain, autrement dit « anthropocentrées » (Rabardel, 1995). L'IA est pensée comme devant être au service de l'opérateur (sauf dans le cas de McAfee et Brynjolfsson dont la posture est ambiguë) qui garde un rôle actif vis-à-vis du système (par exemple, en termes de contrôle, de prise de décision, de choix des actions, etc.). Mais ces approches ne sont pas nouvelles. Elles sont apparues dans les années 60 dans un contexte où le développement de l'informatique et l'intelligence artificielle a conduit des chercheurs à tenter de définir le rôle des ordinateurs dans les activités humaines dans un cadre d'assistante et non de substitution. L'idée d'augmentation a été ainsi développée dès les années 60 par Engelbart (1962). Elle consiste à mobiliser les ordinateurs comme une ressource pour augmenter les capacités intellectuelles humaines. Plus précisément, il s'agit d'accroître les capacités de compréhension et de résolution de problèmes de l'humain à l'aide de ces machines²⁷. Les travaux de Licklider dans les années 60 sur la « symbiose Humain-ordinateur » (*Man-Computer Symbiosis*) ont également été à l'origine de ces approches. Pour cet auteur, la symbiose correspond à un « couplage », une « coopération » entre humains et machines électroniques dans la prise de décision et la gestion de situations complexes²⁸. Dans ce cadre, il définit une répartition du travail dans laquelle la machine prend en charge les tâches routinières tandis que l'humain réalise les tâches de définition des buts, de formulation des hypothèses, de définition des critères et d'évaluation de solutions. Des recherches sur la coopération (ou collaboration²⁹) humain-machine se sont ensuite beaucoup développées à partir des années 80 pour développer des machines capables d'interagir comme un véritable partenaire, qui ne

substituent donc pas à l'humain, mais travaillent avec lui. Ces recherches ont été notamment suscitées par la volonté de sortir des approches substitutives et prothétiques de l'automatisation. De nombreux modèles qui précisent les caractéristiques que doit avoir un système pour développer une relation de coopération avec un humain ont été proposés (Christoffersen, & Woods, 2002 ; Clarke, & Smyth, 1993 ; Hoc, 2000, 2004 ; Klein, Woods, Bradshaw, Hoffman, & Feltovich, 2004 ; Millot, & Lemoine, 1998 ; Silverman, 1992 ; Terveen, 1995). Par exemple, selon Clarke & Smyth (1993), un système coopératif doit notamment être capable de reconnaître les buts de l'utilisateur ; de travailler avec lui à l'atteinte de buts de façon interactive ; de proposer des solutions alternatives, ou encore de soutenir la formation de nouvelles attitudes envers le système et la tâche. Selon, Klein *et al.* (2004), pour qu'un système « intelligent » soit véritablement collaboratif, il doit être en mesure de : comprendre l'activité et les intentions du ou des partenaires humains ; rendre ses actions, états et intentions reconnaissables/intelligibles ; s'engager dans la négociation des buts ; prédire les actions des partenaires humains ; participer à la gestion de l'attention, contrôler les coûts de coordination ; ou encore, se faire guider dans ses actions (régulation, contrôle). S'intéressant à la collaboration, Terveen (1995) distingue cinq caractéristiques centrales de cette relation : l'élaboration collective d'objectifs partagés ; la planification, la répartition des rôles et la coordination des actions ; l'élaboration et le maintien d'un contexte partagé ; la communication ; l'adaptation et l'apprentissage. Or, dans les approches de complémentarité que nous avons évoquées plus haut dans le contexte de la « nouvelle » IA, non seulement il y a une quasi-absence de référence à ces travaux antérieurs sur la coopération Humain-machine, mais ces notions de collaboration, de partenariat ou de symbiose sont aussi peu voire pas du tout définies comme si cela allait de soi. Par ailleurs, au regard des capacités que devraient avoir des machines capables de s'engager dans une *véritable* relation de coopération ou collaboration de type humaine, concevoir ces machines constitue un objectif particulièrement ardu et on peut même se demander s'il est possible de l'atteindre. Un système véritablement collaboratif suppose en effet qu'il soit doté de capacités complexes, sophistiquées qui sont encore largement hors de portée des systèmes d'IA à court et même à moyen terme. Ce problème avait déjà été identifié dans les années 90 (Salembier, 1994) et force est de constater qu'il reste entier malgré les avancées technologiques récentes. Tenant compte de la complexité de ce but, une position minimaliste comme celle suggérée par Hoc (2001) consiste à penser qu'il n'est pas raisonnable de « transférer toute la complexité de la coopération humaine à la coopération Homme-machine³⁰ » (p. 534). Mais, peut-on encore parler de coopération si celle-ci subit une réduction de sa complexité ? N'y a-t-il pas un risque à présenter des machines comme étant « coopératives » alors qu'elles ne disposent que de quelques capacités de cette nature ? Quel(s) niveau(x) de réduction serait acceptable pour considérer qu'une machine est capable de coopérer ? Ces questions restent ouvertes. Un autre problème est qu'il n'existe pas de définitions consensuelles de la coopération (ou collaboration), d'où, comme nous l'avons vu, la proposition de différents modèles de la coopération (ou collaboration) Humain-Machine. Une conséquence de cela est que les termes « coopération » ou « collaboration » renvoient en réalité à des modèles plus ou moins hétérogènes.

- 24 Le problème de la transférabilité de la coopération humaine à la relation Homme-Machine se pose également avec la notion de symbiose. Au sens biologique, une relation symbiotique implique une coévolution, une association émergente, intime et durable

entre deux organismes autonomes. Mais on peut se demander si cela a du sens de parler d'autonomie des machines, aussi intelligentes soient-elles. Comme le soutiennent Bradshaw, Hoffman, Woods et Johnson (2013), cette autonomie (de machines) ne serait-elle pas un « mythe »³¹ ? Si la notion de symbiose reste séduisante pour penser la relation Humain-Machine, son application reste ainsi encore largement programmatique. Un autre problème est que ces approches (coopérative, collaborative, symbiotique) « symétrisent » implicitement l'humain et les machines. Elles laissent supposer que les machines disposent des mêmes capacités que l'humain à collaborer ou coopérer. Or, entre ces deux entités, on peut distinguer au moins deux types d'asymétries fondamentales qui ne doivent pas être ignorées. Une première d'ordre « ontologique ». Par exemple, certaines approches de la collaboration humain-machine développées en IA se sont basées sur l'hypothèse d'une similarité fondamentale entre ces derniers du point de vue de l'action³² (Terveen, 1995). Ils sont modélisés comme des agents « rationnels » qui élaborent des plans pour réaliser des buts et infèrent les plans des partenaires³³. Or, cette symétrie a été largement remise en cause dans le cadre des critiques du cognitivisme. La seconde asymétrie est celle mise en évidence par Suchman (1987) concernant l'accès des machines au contexte de l'action.

25 Ainsi, ces termes de coopération, de collaboration ou de symbiose sont au mieux des métaphores. Bien qu'elles soient plus « modestes », les notions d'assistance, de conception en terme d'aide (Theureau, & Filippi, 1994), de complémentarité ou encore de « configuration » Humains-Machines (Suchman, & Weber, 2015) dans lesquelles les systèmes joueraient par exemple le rôle d'outils cognitifs ou d'instruments au sens de Woods *et al.* (1987) comme exposé plus haut, paraissent plus adéquates d'une part du fait de leur neutralité ontologique, et d'autre part, parce qu'ils ne font pas référence à des modèles de relation (coopération, collaboration, symbiose) qui ne cadrent pas (encore ?) avec l'état de l'art technologique. Une autre notion intéressante qui peut être mobilisée pour penser la relation entre Humain et IA tout en évitant les écueils soulevés par les concepts de coopération ou de symbiose est celle d'interdépendance que l'on retrouve dans certains modèles de la coopération (Castelfranchi, 1998 ; Schmidt, 2002). Ainsi, selon Johnson, Bradshaw, Feltovich, Hoffman, Jonker, van Riemsdijk *et al.* (2011), les machines « intelligentes » doivent être conçues en les dotant de capacités qui les rendent interdépendantes vis-à-vis des partenaires humains. Concevoir ces machines consisterait ainsi à orchestrer, configurer des relations dynamiques d'interdépendance (Johnson, *et al.*, 2011). Pour ces auteurs, c'est précisément pour sortir du « mythe » de l'autonomie que nous avons évoqué plus haut qu'il convient de penser la relation Humain-machine en termes d'interdépendance. Ces auteurs s'inscrivent également dans un cadre collaboratif, en particulier celui décrit par Klein *et al.* (2004). Pour ces auteurs, il est nécessaire de se placer dans ce cadre, car, en acquérant de plus en plus de capacités « cognitives », de perception, et d'interactions complexes, ces systèmes ne seront plus des outils, mais deviendront *inévitavelmente* des « partenaires ». Mais, outre la difficulté évoquée plus haut de concevoir des systèmes véritablement collaboratifs, on peut se demander si cette notion d'interdépendance ne comporte pas le risque de rendre l'humain totalement dépendant des machines et affecter son autonomie. Autrement dit, dans le contexte de développement de systèmes d'IA de plus en plus performants, il nous semble important de ne pas négliger les conséquences du degré d'interdépendance qui peut être créé.

26 Si la complémentarité entre Humains-IA dans ses différentes formes constitue une direction pertinente, moyennant les limites que nous avons soulignées à propos des

approches collaboratives ou coopératives, la nature de la nouvelle génération de systèmes d'IA (basés sur l'apprentissage machine) soulève cependant des questions concernant l'évolution de cette relation dans le temps. Une particularité de ces systèmes est que leur performance peut continuer d'évoluer s'ils sont régulièrement alimentés en données et entraînés. Comment dès lors envisager la relation Humain-IA dans le temps avec des systèmes qui vont progressivement s'améliorer au fur à mesure de leur apprentissage ? Faudra-t-il repenser régulièrement la répartition du travail entre humains et IA ? Que peut-il se passer dans des contextes où l'IA devient de plus en plus performante ? Cela peut-il induire une remise en cause de la complémentarité humains-IA au profit de cette dernière ? S'intéressant au domaine médical, Froomkin, Kerr, & Pineau (2019) estiment qu'une telle « dérive » est possible. Ils considèrent en effet que si les machines basées sur l'apprentissage machine s'avèrent plus performantes que les médecins pour réaliser des diagnostics, cela peut créer une pression légale et éthique à déléguer cette tâche à ces machines au détriment des médecins : « Avec le temps, un apprentissage machine efficace pourrait créer une pression légale et éthique forte pour déléguer le processus de diagnostic à la machine. » (p. 2) Autrement dit, au lieu de constituer une aide pour les médecins, ces machines risquent de les déposséder totalement de cet aspect de leur activité avec toutes les conséquences qui sont bien connues d'un tel phénomène (par exemple, substitution, perte d'expertise). Ce risque, qui n'est pas uniquement lié à la technologie, mais aussi au contexte social ou organisationnel dans lequel elle est déployée, avait déjà été identifié à propos des systèmes experts (Freyssenet, 1992, Terssac, 1994). Par exemple, de Terssac se demandait si ces systèmes ne deviendraient pas d'autant moins réfutables qu'ils démontreraient au fil du temps leur supériorité par rapport aux travailleurs. Le point soulevé par Froomkin *et al.* (2019) nous invite également à réfléchir à la question du développement des employés dans un contexte où les machines seront également capables de se « développer » grâce à l'apprentissage. Pour éviter le scénario qu'ils décrivent, ils insistent sur la nécessité de préserver le rôle des « médecins » dans la « boucle du diagnostic » et, dans cette perspective, proposent de revoir les règles qui encadrent la pratique médicale. Une autre piste de réflexion possible consiste à envisager des formes de « co-développement » Humain-IA dans lesquelles la complémentarité serait plus pérenne. L'humain apprend de l'IA, et inversement. Une réflexion plus large sur l'intégration des systèmes d'IA dans les situations de travail est ainsi nécessaire, car des facteurs sociaux et organisationnels peuvent affecter leurs usages et la place qu'ils peuvent prendre dans l'activité des salariés, non seulement à court terme, mais aussi à plus long terme. Plus généralement, comme l'ont déjà montré les études sur les systèmes experts, cette intégration et les transformations de l'activité qu'elle peut impliquer nécessitent une réflexion profonde sur l'organisation globale du travail et la manière dont elle peut évoluer.

- 27 Enfin, il est important de noter qu'à l'opposé de la complémentarité, notamment celle que nous défendons ici, c'est-à-dire envisager l'IA dans une perspective centrée sur l'humain, on voit apparaître des idées d'application de l'IA qui s'inscrivent dans une optique de contrôle de plus en plus fin de l'activité des opérateurs pour « optimiser la réalisation des tâches ». C'est par exemple le cas avec le brevet déposé par Amazon concernant un bracelet dont l'objectif est de guider de manière haptique (via des impulsions produites par ultrasons) les mouvements des opérateurs manutentionnaires vers les produits commandés qu'ils doivent récupérer dans les entrepôts où la société stocke ses marchandises. Ce type d'usage « tayloriste » de l'IA soulève de nombreuses

questions concernant d'une part la réduction de l'autonomie des employés en « contrôlant » une bonne part de leur activité et, d'autre part, leur surveillance dans la mesure où le bracelet constitue un moyen de vérification qu'ils sont bien actifs. Comme le notait de Terssac (1992) à propos des systèmes experts, les systèmes d'IA peuvent aussi être utilisés comme une source de contrôle dans les organisations.

4.2. Travailler avec des IA anthropomorphes

- 28 Bien que, ainsi que nous l'avons vu, la question du type de relation complémentaire entre Humains-Machine (collaboration, augmentation, coopération, partenariat, symbiose) qu'il faudrait mettre en place dans des situations de travail se pose depuis longtemps, il nous semble cependant que cette question en soulève d'autres qui concernent plus précisément les systèmes dotés de caractéristiques anthropomorphes ou qui s'en rapprochent. Pour illustrer ce point, un exemple intéressant est l'étude réalisée par Saupé et Mutlu (2015) dans trois entreprises sur l'interaction d'employés avec un cobot (appelé Baxter - voir Figure 2) développé par la société *rethinkrobotics*³⁴. Ce cobot, conçu sur la base de techniques d'IA, a la particularité d'être équipé d'un écran qui affiche des « yeux » qui sont censés simuler l'orientation de la machine selon ses actions. D'après les concepteurs, l'orientation du « regard » du robot est supposée servir d'indice qui peut aider l'humain qui travaille à côté de lui à interpréter et surtout anticiper ses actions.

Figure 1 : Baxter.

Figure 1: Baxter



- 29 Un premier point intéressant qui ressort de l'étude est que la relation à ce type de robots varie en fonction du rôle ou statut des employés. Du côté des managers, la relation était strictement instrumentale au sens où ils ne les envisageaient que comme

équipement industriel. La relation développée par les employés impliqués dans la production était différente. Ils ont en effet eu tendance à considérer leurs interactions avec ces robots comme semblables à des interactions avec des humains, à leur attribuer des intentions et une personnalité, et même à faire l'expérience d'émotions quand ils interagissaient avec ces machines. En outre, les « yeux » du robot induisaient un sentiment de sécurité dans la mesure où elles leur permettaient d'anticiper ses actions. Ses « yeux » ont également été interprétés comme un indice d'intelligence qui, ce faisant, influait sur le sentiment de confiance dans les comportements du robot³⁵. Certains des employés souhaitaient même que les capacités interactionnelles sociales de ces robots soient plus développées (par exemple, au plan conversationnel pour pouvoir leur parler comme à un collègue humain). L'étude semble donc montrer que le fait de les doter de caractéristiques anthropomorphes peut faciliter la coordination des actions et induire un sentiment de confiance. D'autres études réalisées dans un secteur industriel différent montrent que les opérateurs manifestent une préférence pour les cobots expressifs, c'est-à-dire qui manifestent des « indices sociaux » (*social cues*) comme le regard ou les hochements de tête (Elprama, El Makrini, Vanderborght, & Jacobs, 2016). Mais cette approche, qui consiste à anthropomorphiser les robots, peut être problématique dans la mesure où elle peut, par exemple, susciter de fortes attentes et des formes de sur-confiance qui ont été observées depuis longtemps dans l'interaction avec des automatismes. Toujours dans cette perspective anthropomorphique, il est intéressant de remarquer que les cobots sont de plus en plus lexicalement envisagés comme des « collègues » de travail (*co-workers*), non seulement dans les médias, mais également dans la littérature scientifique (par exemple, Andersen, Solund, & Hallam, 2014). Cette évolution est le signe d'une tendance plus générale qui consiste à « personnifier » les robots et à les doter de capacités d'interaction sociale. Pour ce qui concerne les situations de travail, cela pose plusieurs questions, par exemple : quelle relation les opérateurs vont développer avec des machines qui vont de plus en plus ressembler aux humains ? Les catégories classiques comme celles d'« instrument » ou « outil » restent-elles pertinentes pour qualifier ces machines ?

5. Le problème de « l'explicabilité » des systèmes d'IA

- 30 Si l'on considère que les systèmes d'IA, en particulier ceux basés sur l'apprentissage, vont être de plus en plus déployés dans les situations de travail, cette diffusion soulève la question de l'intelligibilité ou compréhension de leurs actions ou leur fonctionnement par les travailleurs qui interagiront avec eux. Ce problème n'est pas nouveau ; il s'est en effet posé avec les systèmes experts, et de façon plus générale, l'augmentation de l'automatisation dans les situations de travail³⁶. Plus précisément, il a en effet été constaté que ces systèmes généraient de l'opacité cognitive qui renvoie à la difficulté pour les travailleurs à élaborer une compréhension des systèmes, notamment du fait d'un manque d'informations sur leur état et leur fonctionnement. Des études ont montré que cette opacité crée de l'incertitude pour les utilisateurs, perturbe leurs activités et peut même dégrader leur travail (Falzon, 1989 ; Roth *et al.*, 1987). La machine devient une source de problèmes alors qu'elle est supposée être une aide, par exemple avec ce que Sarter, Woods, et Billings (1997) ont appelé « *automation surprises* », c'est-à-dire lorsque le système se comporte ou réagit de façon inattendue. Un autre problème est que cette opacité complique la reprise en main par l'humain de

la situation, notamment lorsqu'elle est dégradée. Qui plus est, elle peut dégrader la confiance de l'opérateur dans le système (Hoff, & Bashir, 2015). Or, la « transparence » du système semble être un élément important de la construction de cette confiance (French, Duenser, & Heathcote, 2018 ; Mercado, Rupp, Chen, Barnes, Barber, & Procci, 2016). Pour ce qui concerne plus spécifiquement les systèmes experts, de nombreuses études ont mis en évidence que la difficulté à interpréter leurs actions ou raisonnements était source de problèmes d'usage voire de rejet de ces systèmes (Clancey, 1983 ; Moore, & Swartout, 1988 ; Teach, & Shortliffe, 1981). Un grand nombre de recherches ont ainsi été menées sur la manière de rendre les actions et le fonctionnement des systèmes experts « explicables » ou interprétables (Bouri, Dieng, Kassel, & Safar, 1989 ; Clancey, 1983 ; Kass, & Finnin, 1988 ; Moore, & Swartout, 1988). Les travaux réalisés ont ainsi cherché à développer des modules spécifiques (sortes de programmes) capables d'expliquer ou tracer de façon intelligible pour l'utilisateur les raisonnements suivis par le système. Autrement dit, l'explication consiste à donner à l'utilisateur des éléments d'interprétation ou de compréhension des actions du système. Par exemple, le système MYCIN dont nous avons parlé plus haut a été doté d'un module complémentaire qui permet à l'utilisateur d'obtenir des explications du raisonnement suivi par le système, en lui indiquant principalement les connaissances et les règles (logiques ou probabilistes) mobilisées par ce dernier (van Melle, 1978). Autrement dit, le système explique « pourquoi » il est arrivé à une conclusion donnée à travers des interactions de type questions-réponses. Par exemple, si le système demande les résultats d'une analyse médicale concernant le taux de calcium dans le sang d'un patient pour poser un diagnostic, l'utilisateur peut lui demander pourquoi il pose cette question. Le système génère ensuite une réponse qui explique cette requête, en explicitant les règles et le raisonnement qui l'ont conduit à la formuler. Une caractéristique importante de ces systèmes experts est que les connaissances et les règles pouvaient être plus ou moins aisément exprimées en langage naturel, car ils étaient formalisés et programmés dans un code proche de ce langage (par exemple, utilisation de règles de type Si... alors). L'utilisation du langage naturel était ainsi censée rendre le raisonnement du système compréhensible par un utilisateur, quel que soit son niveau de compétences en informatique. Il convient cependant de noter que ce type d'explications avait certaines limites, par exemple le manque de justification des inférences réalisées par le système et la non-prise en compte des caractéristiques de l'utilisateur (par exemple, son degré d'expertise). Ces limites ont conduit des chercheurs à proposer d'autres modèles de l'explication, par exemple en l'envisageant comme une interaction structurée qui doit être adaptée à l'utilisateur (ses buts, ses connaissances, etc.) et de façon plus générale, au contexte (Swartout, & Moore, 1993). On retrouve cette idée d'adaptation de l'explication dans la notion de « transparence opérative » proposée par Rabardel (1995) et qui consiste à mettre en relation la transparence avec les besoins en informations de l'utilisateur, en fonction de ses buts, de ses compétences, etc. Autrement dit, l'explication doit être ajustée au contexte d'usage, incluant l'utilisateur. Une autre approche consiste à penser l'explication comme un processus coopératif qui implique une « négociation de sens » (Karsenty, & Brézillon, 1995). L'explicabilité a été également abordée dans le cadre des systèmes informatiques « sensibles » au contexte (*context-aware computing*) sous l'angle de la notion d'intelligibilité (Bellotti, & Edwards, 2001). Ces systèmes ont comme caractéristique de pouvoir adapter leurs actions sur la base de traitements d'éléments contextuels issus de capteurs placés dans l'environnement ou portés par des

utilisateurs (par exemple, l'activité en cours de l'utilisateur, sa présence ou son absence) ; ils impliquent là aussi l'utilisation de techniques d'IA et posent le même problème de compréhension de leurs actions que les systèmes experts. Comme le soulignent Bellotti et Edwards (2001), il est essentiel que ces systèmes soient capables d'indiquer (aux utilisateurs), de manière interprétable, les informations contextuelles dont ils disposent, comment ils les ont élaborées, et ce qu'ils font ou vont faire avec ces informations. C'est ce que les auteurs qualifient d'intelligibilité, qui de façon synthétique, correspond à la compréhension du fonctionnement du système.

- 31 Ce problème de l'opacité des machines se pose de manière beaucoup plus ardue dans le cas de l'apprentissage machine, en particulier les systèmes basés sur les réseaux de neurones. Selon les spécialistes du domaine, la compréhension du fonctionnement de ces systèmes est particulièrement difficile³⁷. C'est, par exemple, ce qu'explique Lipton (2018) :

« Bien que les procédures d'optimisation heuristique pour les réseaux de neurones soient manifestement puissantes, nous ne comprenons pas comment elles fonctionnent, et à l'heure actuelle, nous ne pouvons garantir a priori qu'elles fonctionneront sur de nouveaux problèmes³⁸. » (p. 5)

Pour être plus précis, ce problème concerne les modèles d'apprentissage machine au sens où nous l'avons indiqué plus haut, c'est-à-dire ce qui résulte de l'application de données d'apprentissage à un type d'algorithme. Cette difficulté de compréhension du fonctionnement des modèles d'apprentissage est particulièrement grande avec ceux qui reposent sur l'apprentissage profond qui est considéré comme la technique la plus performante, mais la moins compréhensible (Gunning, 2016). Selon Davis (cité dans Monroe, 2018), un autre spécialiste de ce domaine, la meilleure analogie dont nous disposons pour rendre compte du fonctionnement de systèmes basés sur cette technique est qu'ils développent une sorte « d'intuition » ou « d'instinct » (*gut instinct*). Une telle analogie permet de bien saisir l'ampleur du problème. Rahimi et Recht (2017), deux autres spécialistes de ces systèmes³⁹, vont même jusqu'à comparer la recherche en apprentissage automatique à de « l'alchimie » au sens où il est aujourd'hui possible de concevoir des algorithmes d'apprentissage qui atteignent des performances inédites sans que l'on soit en capacité de comprendre comment ils y parviennent⁴⁰. À travers cette métaphore, ils soulignent ainsi que la compréhension du fonctionnement de ces systèmes reste essentiellement empirique. Les modèles d'apprentissage qui reposent sur l'apprentissage profond sont ainsi considérés comme des « boîtes noires », « inscrutables » (Weld, & Bansal, 2019), en particulier lorsqu'ils mettent un grand nombre de couches de neurones, de nœuds et de coefficients qui calculent les liens d'activation entre les neurones. Ces difficultés soulèvent plusieurs questions, notamment : comment prouver à l'avance que ces systèmes feront bien ce qu'ils sont censés faire ? Comment expliquer, rendre intelligible ou encore interprétable les résultats qu'ils produisent ? De façon plus spécifique, l'intelligibilité des modèles basés sur l'apprentissage est posée comme essentielle pour au moins sept raisons (Weld, & Bansal, 2019). Premièrement, l'objectif d'une IA peut-être insuffisamment spécifié, mal défini voire (involontairement) délétère (par exemple, des études de systèmes qui calculent le risque de récidive criminelle ont montré que la focalisation sur la performance prédictive de ces systèmes a conduit à sous-estimer le fait que les données d'apprentissage utilisées sont entachées de biais sociaux qui ont comme conséquence de reproduire une forme de discrimination sociale qui se traduit par des prédictions plus favorables à certaines populations qu'à d'autres⁴¹). Deuxièmement, un système

peut établir de façon autonome des corrélations « inattendues » entre variables (*features*), qui (les corrélations) peuvent le conduire à produire des résultats erronés ; il importe donc de pouvoir les détecter⁴². Troisièmement, les performances d'un système peuvent se dégrader lorsqu'il est déployé dans des situations réelles, il est donc utile de comprendre les capacités de généralisation du système sur des données nouvelles. Quatrièmement, dans certains contextes, le contrôle d'un système par un utilisateur peut nécessiter que ce dernier comprenne pourquoi l'IA a produit un résultat donné, en particulier lorsqu'il est considéré comme insatisfaisant ou problématique (par exemple, un système qui effectue une recommandation de produits qui ne convient pas ou ne correspond pas au « profil » de l'utilisateur). Cinquièmement, le fait de disposer d'explications peut faciliter « l'acceptation » des décisions proposées par un système aux utilisateurs concernées (par exemple, un traitement médical). Sixièmement, l'intelligibilité d'un système peut aider à améliorer la connaissance de phénomènes ou de domaines particuliers (par exemple, la connaissance du jeu de go a été transformée par les performances du programme de jeu de go AlphaGo qui a « inventé » des coups inédits). Enfin, la difficulté à comprendre le fonctionnement de systèmes d'IA basés sur l'apprentissage génère une forte demande sociale à ce qu'ils soient transparents, interprétables, explicables pour des raisons légales ou éthiques⁴³. L'explication est ainsi posée comme un droit que devrait avoir tout citoyen (Goodman, & Flaxman, 2016). La commission européenne envisage même l'explication comme un principe éthique fondamental dans un récent rapport sur l'éthique de l'IA (« Lignes directrices en matière d'éthique pour une IA digne de confiance », 2018). Quelles formes doivent prendre les explications et comment prouver qu'elles sont correctes et fidèles au modèle reste cependant des questions largement ouvertes (Lipton, 2018 ; Rudin, 2019).

- 32 Cette question de l'explicabilité va constituer un défi majeur à mesure que les systèmes d'apprentissage machine se diffuseront progressivement dans les situations de travail. La difficulté à comprendre le fonctionnement de ces systèmes peut en effet avoir des conséquences problématiques dès lors qu'ils sont massivement déployés dans la société. Par exemple, Froomkin, Kerr et Pineau (2018) considèrent que le déploiement généralisé de systèmes basés sur l'apprentissage automatique dans le domaine médical pourrait amoindrir la qualité des traitements médicaux et poser le problème de la compréhension par les médecins des diagnostics générés par ces machines, et ce précisément en raison de l'opacité de ce type de système :

« Si nous en arrivons au point où la majorité des résultats cliniques recueillis dans les bases de données sont des diagnostics générés par l'apprentissage machine, il pourrait en résulter des scénarios de décisions futures qui ne sont pas facilement vérifiables ou compréhensibles par les médecins humains. Étant donné le fait bien documenté que les stratégies de traitement ne sont souvent pas aussi efficaces lorsqu'elles sont déployées dans la pratique clinique réelle qu'au stade de l'évaluation préliminaire, le manque de transparence introduit par les algorithmes de Machine Learning [ML] pourrait mener à une diminution de la qualité des soins. » (Froomkin, Kerr, & Pineau, 2019, p. 34).

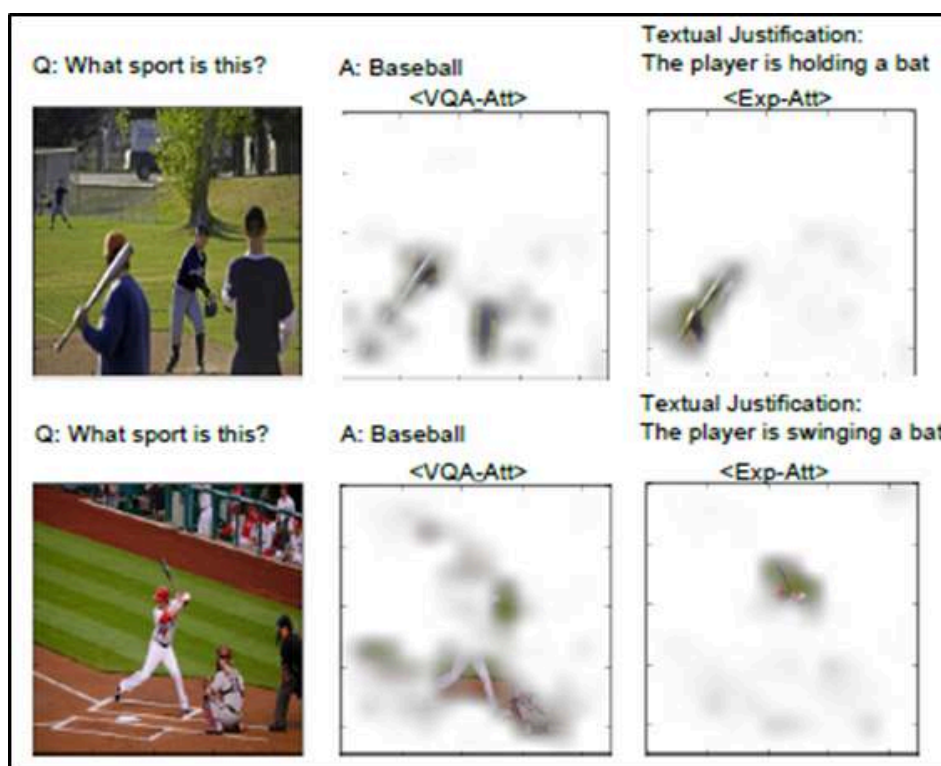
- 33 Concernant plus précisément les conséquences sur l'activité humaine, le déploiement de ces systèmes dans les situations de travail risque de faire apparaître les mêmes problèmes posés par l'opacité des machines, que nous avons évoqués plus haut, notamment à propos des systèmes experts.
- 34 La forte résurgence de ce problème de compréhension des systèmes d'IA a amené les chercheurs du domaine (de l'apprentissage machine) à proposer différentes approches

conceptuelles de ce problème, en s'appuyant sur les notions d'explicabilité, d'interprétabilité, d'intelligibilité, de transparence ou encore de justification. Par exemple, Doshi-Velez et Kim (2017) définissent l'interprétabilité comme la capacité à présenter ou expliquer quelque chose dans des termes compréhensibles par l'humain. Ils distinguent deux sortes d'interprétabilité : *globale*, qui porte sur la compréhension générale d'un modèle (d'apprentissage) ; *locale*, qui concerne l'explication d'une décision particulière produite par le modèle. Un point intéressant de leur approche est que, partant de cette définition, les auteurs (qui sont informaticiens) insistent sur l'importance de l'implication de l'humain dans l'évaluation de l'interprétabilité, dans le cadre d'applications concrètes avec des utilisateurs experts du domaine (par exemple, des médecins qui évaluent les explications données par un système de diagnostic médical) ou de tâches simplifiées expérimentales qui ne nécessitent pas une expertise particulière (par exemple, évaluer la capacité d'une explication à permettre à des humains de « prédire » une sortie d'un système à partir d'une entrée donnée). Pour Biran et Cotton (2017), un modèle d'apprentissage est interprétable si ses opérations sont compréhensibles par l'humain, notamment à travers des explications. À cette définition qui est plutôt vague de l'interprétabilité, ils ajoutent la notion de justification, qui, d'après eux, renvoie à la justesse d'une décision et qui n'exige pas d'explication précise du mécanisme qui l'a produit. Selon Weld et Bansal (2019) un modèle est intelligible si l'utilisateur humain est d'une part en capacité de prédire comment le changement d'une variable (par exemple, un petit accroissement de sa valeur) affectera la sortie du modèle, et, d'autre part, si ce changement produit réellement une modification de la réponse du système. Dans ce cas le modèle est dit interprétable. Miller (2019) parle de « transparence », qui, selon lui, renvoie à deux aspects : premièrement, l'explicabilité (ou interprétabilité) détermine dans quelle mesure l'humain est capable de comprendre les actions d'un système dans un contexte donné, deuxièmement, l'explication qui est donnée pour rendre compte d'une action. Outre cette diversité de définitions des notions d'explicabilité et d'interprétabilité, certains auteurs du domaine estiment nécessaire de différencier ces notions. Selon Lipton (2018), l'interprétabilité des systèmes d'IA à base d'apprentissage doit être analysée selon différentes dimensions qu'il convient de distinguer. Une première est la transparence qui renvoie à la compréhension du fonctionnement du modèle d'apprentissage. Trois types de transparence sont distingués : la « simulabilité » (*simulatability*) qui correspond à la capacité de l'humain à reproduire ou simuler dans un temps raisonnable le fonctionnement du système à partir des données et des paramètres du modèle (concrètement, effectuer les calculs qui produisent une sortie donnée), la décomposabilité (*decomposability*) qui signifie que chaque élément du système (les données qu'il prend en entrée, les paramètres et les calculs) est compréhensible de manière « intuitive » (descriptible en langage naturel), la transparence algorithmique (*algorithmic transparency*) qui concerne la capacité à prédire le fonctionnement de l'algorithme d'apprentissage avec de nouvelles données. La seconde dimension est l'interprétabilité post-hoc, qui, à la différence de la première, ne requiert pas une compréhension du fonctionnement interne modèle. Elle porte plus spécifiquement sur l'explication des sorties d'un système. Autrement dit, l'interprétabilité post-hoc correspond à l'explication. Contrairement à la première qui concerne surtout les spécialistes de l'IA ou concepteurs qui cherchent à comprendre leurs modèles, cette seconde dimension est directement en rapport avec ce qui nous intéresse ici dans la mesure où c'est ce type d'explications qui peut être utile ou

nécessaire pour des utilisateurs finaux dans le cadre d'activités de travail avec ce type de systèmes d'IA. Lipton distingue trois formes d'interprétations post hoc : les explications en langage naturel, sous forme de visualisations (du modèle ou des représentations construites par celui-ci), ou basées sur des exemples comparatifs (par exemple, dans un cadre médical, un système qui explique qu'il a classé une tumeur comme maligne, car il « trouve » qu'elle ressemble à d'autres tumeurs classées de cette façon). Selon Lipton, bien que leur interprétabilité du point de vue des trois premiers critères (simulabilité, décomposabilité et transparence algorithmique) soit particulièrement difficile, les réseaux de neurones profonds sont plus adaptés pour produire des interprétations post-hoc que d'autres modèles d'apprentissage. D'autres auteurs proposent de faire une distinction entre interprétation et explication. Montavon, Samek et Müller (2018) considèrent ainsi que l'interprétation consiste à traduire des concepts abstraits (par exemple, une classe de sorties) dans des formats ordinaires, c'est-à-dire compréhensibles par n'importe quel humain (par exemple, du texte ou des images). Une explication correspond à l'ensemble des éléments (*features*) représentés dans une forme interprétable qui ont contribué à une sortie (par exemple, une classification). Par exemple, pour un système qui classe des images, l'explication consistera en une carte de chaleur (*heatmap*) qui indique les pixels de l'image qui ont le plus influencé une classification réalisée par le système. L'explication consiste ainsi à mettre un résultat (une sortie) en relation avec des éléments qui ont contribué à ce résultat. Enfin, selon Gilpin, Bau, Yuan, Bajwa, Specter *et al.* (2018), l'explication doit être traitée selon deux aspects : l'interprétabilité (*interpretability*) qui correspond à la description du fonctionnement interne d'un système dans un langage compréhensible par l'humain, et la complétude (*completeness*) qui se réfère à la description précise et détaillée des opérations (par exemple, mathématiques) du système. D'après ces auteurs, le défi qui se pose est de produire des explications qui soient à la fois interprétables et complètes. Une façon de gérer ce défi consiste pour eux à trouver un compromis entre ces deux dimensions. De cette revue des approches de l'explicabilité, l'une des conclusions que l'on peut tirer est qu'il n'existe pas pour le moment de consensus sur la manière de définir l'interprétabilité ou l'explicabilité des systèmes d'apprentissage machine. On peut également noter une instabilité terminologique ; différentes notions sont utilisées de façon interchangeable ou différenciée. Une autre conclusion est que les travaux sur l'explication des systèmes experts (basés sur des règles et des connaissances symboliques) sont très rarement mentionnés dans la littérature actuelle sur l'apprentissage machine. Ce qui est particulièrement surprenant compte tenu du fait qu'il s'agit fondamentalement du même problème qui se pose dans les deux cas. Miller (2019) est l'un des rares qui évoque et suggère de s'appuyer sur ces travaux ; il propose également de s'appuyer sur les connaissances développées en sciences humaines et sociales sur l'explication. Un point notable au regard de ce qui nous intéresse dans cet article, à savoir l'interaction avec des systèmes d'IA dans le cadre du travail, est qu'il considère que l'explicabilité est avant tout un problème d'interaction Homme-Agent artificiel et souligne que l'explication n'est pas qu'une question de descriptions de relations entre des causes et des effets, mais qu'elle a des dimensions sociales, interactives, et contextuelles (à qui s'adresse-t-elle ? Pour quoi faire ? dans quelles circonstances ?). Ces dimensions, dont la mise en évidence n'est pas nouvelle, sont essentielles à prendre en compte dans la perspective de la conception de systèmes d'IA dont les actions requièrent d'être expliquées, c'est-à-dire rendues intelligibles pour les acteurs concernés.

- 35 À côté de ces travaux de conceptualisation, d'importants efforts de recherche sont actuellement menés sur les techniques permettant d'améliorer la compréhension du fonctionnement des systèmes d'apprentissage machine afin de rendre leurs sorties explicables⁴⁴ (par exemple, Montavon, Samek, & Müller, 2018 ; Park, Hendricks, Akata, Schiele, Darrell, & Rohrbach, 2018 ; Ribeiro, Singh, & Guestrin, 2016 ; Yosinski, Clune, Nguyen, Fuchs, & Lipson, 2015). En lien avec ce qui nous intéresse ici, l'interaction d'un utilisateur avec une IA, une partie de ces travaux développe des méthodes qui génèrent des explications intelligibles sous différentes formes, par exemple textuelles, graphiques ou multimodales, selon le domaine d'application du système. C'est ce que l'on peut voir dans l'exemple présenté dans la figure 2 ci-dessous qui concerne la catégorisation d'images. Dans cet exemple, comme dans d'autres, on retrouve une forme interactive de l'explication.

Figure 2 : Exemple d'une explication locale d'une sortie d'un système de classification d'images (Tiré de Park *et al.*, 2018). Ici, le système fournit une explication visuelle et textuelle de sa réponse (base-ball) à la question posée par l'utilisateur (de quel sport s'agit-il ?). Il indique les parties principales de l'image qui ont contribué à la prédiction (par exemple, le joueur tient une batte).
Figure 2. Example of a local explanation provided by an image classification system (from Park, Hendricks, Akata, Schiele, Darrell, & Rohrbach, 2018). Here, the system provides a textual and visual explanation of its response (baseball) to the question formulated by the user (what sport is this?). It indicates the main areas of the image that led to the prediction (e.g. the player is holding a bat)

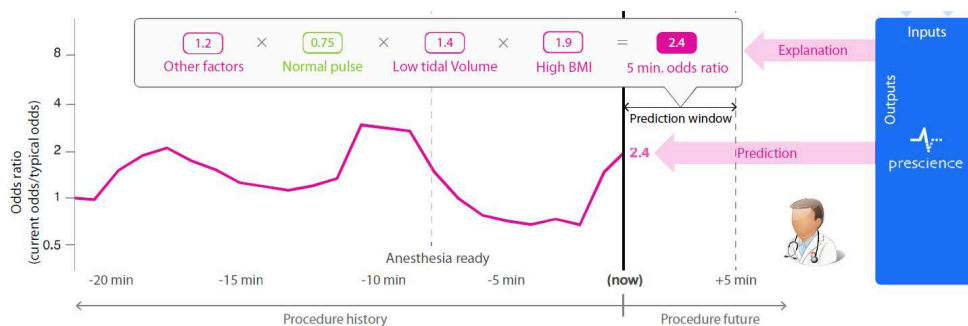


- 36 Le travail réalisé par Lundberg, Nair, Vavilala, Horibe, Eisses, Adams *et al.* (2018) constitue un autre exemple d'explication locale qui nous intéresse plus directement ici, car il montre comment les « prédictions » d'un système d'IA basé sur l'apprentissage peuvent être rendues intelligibles de façon ajustée à une activité de travail dans le domaine médical. Basé sur un modèle complexe d'apprentissage, le système en question établit des prédictions sur le risque d'hypoxémie (diminution anormale de la quantité d'oxygène contenue dans le sang) au cours d'une opération chirurgicale. Plus précisément, il fournit une valeur numérique qui correspond à un score de risque

d'hypoxémie pour un patient donné. Le calcul est réalisé sur la base d'une soixantaine de facteurs qui sont considérés par les anesthésistes comme importants à prendre en compte pour évaluer ce risque avant et au cours d'une opération (par exemple l'âge, l'indice de masse corporelle, le volume et le débit respiratoire, la fréquence cardiaque). L'explication consiste à indiquer à l'anesthésiste les facteurs principaux qui ont conduit le système à faire la prédiction à un instant donné, en précisant quels sont ceux qui augmentent le risque et ceux qui le diminuent (Figure 3). L'explication fournie se limite ainsi aux informations qui ont été considérées comme *directement pertinentes* pour les anesthésistes, compte tenu de la nature de leurs activités (elle a été déterminée avec eux). On retrouve ainsi dans cet exemple le principe selon lequel l'explication doit être adaptée au contexte d'usage du système (pour qui, pour quelle activité, dans quel but, etc.).

Figure 3 : Informations explicatives fournies par le système *Prescience*. Dans cet exemple, le système indique les principaux facteurs qui lui ont permis de calculer le risque d'hypoxémie à un instant t (2.4). Les facteurs qui majorent le risque sont en violet ; ceux qui le diminuent sont en vert. Source : Lundberg et al. (2018).

Figure 3. Explanations from the *Prescience* system. In this example, the system displays the most significant features that it used for calculating the risk of hypoxemia (2.4) at a given moment. The factors that increase this risk are in purple; those that decrease it are in green. Source: Lundberg et al. (2017)



- 37 Si les travaux en cours permettent d'avancer sur l'intelligibilité des modèles d'apprentissage, de nombreuses questions restent cependant ouvertes (Lavin, 2018). Premièrement, il n'existe pas de métriques précises permettant de dire qu'un modèle est plus ou moins interprétable qu'un autre. De façon plus générale, les critères d'interprétabilité/intelligibilité sont *sous-spécifiés*, c'est-à-dire pas assez bien définis pour être mesurables. En outre, il convient de noter qu'en pratique la conception de modèles explicables pose un problème de compromis entre intelligibilité et performance. En effet, plus un modèle d'apprentissage est performant moins il est interprétable et inversement. Ainsi, toutes choses étant égales par ailleurs, concevoir un modèle interprétable peut impliquer le développement d'un système moins performant.
- 38 Cette question de l'explicabilité ouvre des perspectives intéressantes pour l'ergonomie d'une part en termes de démarche qu'il faut mettre en œuvre pour déterminer les besoins potentiels en explication, et d'autre part, en termes de conception. Concernant le premier point, l'analyse de l'activité constitue un outil utile pour précisément déterminer les explications qu'il est nécessaire ou utile de donner aux travailleurs concernés. Le deuxième point porte sur les questions suivantes : comment concevoir les explications, c'est-à-dire comment les représenter ? Sous quelle(s) formes(s) ? Sur ce point, l'interface utilisateur constitue un enjeu central en termes de conception, car

c'est à travers elle que se matérialise l'explication. Une autre question qui peut être importante est de déterminer à quel moment elle doit être accessible. De façon générale, les travaux menés sur les systèmes experts constituent une ressource théorique et méthodologique importante pour aborder ces questions dans le cadre des systèmes basés sur l'apprentissage machine, car, comme dit plus haut, on retrouve bien le même type de problématique. Il faut cependant noter que la nature « non symbolique » des systèmes d'IA basés sur l'apprentissage machine soulève des questions de traduction et de représentation de leurs opérations à des fins explicatives qui sont beaucoup plus complexes. Comment en effet traduire en explications compréhensibles des opérations qui sont purement mathématiques ? Des recherches intéressantes dans ce domaine de la représentation sont actuellement menées (voir par exemple, Carter, & Nielsen, 2017).

- 39 Cet examen de la question de l'explicabilité serait incomplet sans indiquer l'existence d'un débat au sein même de la communauté des spécialistes de l'IA. Si la plupart d'entre eux ne remettent pas en question les systèmes de type « boîte noire » (ceux qui posent des problèmes d'explicabilité), des voix s'élèvent contre l'utilisation de ces systèmes. C'est notamment le cas de Rudin (2019) qui prône l'abandon pur et simple de ces systèmes, car elle estime d'une part que leur opacité soulève trop de problèmes techniques et sociaux, et d'autre part, que, contrairement à ce que l'on pense habituellement, ces systèmes ne sont pas nécessairement les plus performants. Elle ajoute et montre que dans le cas de ces systèmes, les explications fournies sont souvent incomplètes, voire même fausses. Pour toutes ces raisons, elle suggère qu'au lieu de chercher à développer des techniques « externes » d'explicabilité des systèmes de type boîte noire, il faut plutôt développer des systèmes qui sont « intrinsèquement » interprétables, c'est-à-dire basés sur des techniques d'apprentissage machine qui rendent possible de façon inhérente la compréhension de leur fonctionnement. Si l'on se place du point de vue de l'approche ergonomique de l'interaction Humain-machine, en particulier l'importance qu'elle donne à la conception de systèmes intelligibles pour les utilisateurs, ce point de vue paraît en phase avec cette approche. Mais, bien que les arguments de Rudin semblent convaincants, pour différentes raisons (enjeux industriels et académiques) il n'est pas du tout sûr que les systèmes de type « boîte noire » soient un jour abandonnés (au moins à court ou moyen terme). Dans ce contexte, il est important de garder à l'esprit que les explications du fonctionnement et des « sorties » de ces systèmes sont au mieux approximatives. Et dans tous les cas, comme nous l'avons vu, il est essentiel qu'elles soient contextualisées, c'est-à-dire adaptées dans leur forme, leur contenu et leur signification au contexte d'usage (pour qui, pour quoi faire, etc.).

Conclusion

- 40 Impulsée par les performances sans précédent réalisées par l'apprentissage machine dans une gamme de domaines qui ne cesse de s'élargir, une diffusion massive de l'IA sous différentes formes dans les situations de travail paraît aujourd'hui inéluctable. Les projets d'intégration de l'IA dans ces situations sont foisonnants et des expérimentations sont en cours⁴⁵, avec des objectifs variés (automatisation, optimisation de processus, réduction des coûts, amélioration de la performance, parfois dans une logique de réduction de la pénibilité du travail ou d'assistance aux employés,

mais aussi de contrôle de ces derniers dans d'autres cas, etc.). Partant de ce constat, l'intention principale qui a animé cet article était de s'interroger sur les questions et les enjeux soulevés par ces évolutions au regard des activités humaines et la relation Humain-Machine dans le cadre du travail. Une première conclusion de cet article est qu'une bonne partie de ces questions et enjeux est ancienne et qu'elle est abordée en ergonomie depuis longtemps. Tout d'abord, nous avons vu que c'est surtout la question des incidences de ces avancées sur l'emploi qui est débattue notamment en raison de la crainte d'une automatisation massive du travail dans de nombreux secteurs. Mais cette crainte, qui n'est pas nouvelle puisqu'elle réapparaît régulièrement depuis au moins la révolution industrielle, fait l'objet de controverses sur l'ampleur de cette automatisation. En outre, cette crainte laisse de côté le fait bien documenté depuis longtemps que l'automatisation quelle que soit sa nature se traduit surtout par une transformation des activités humaines. Le point important est de comprendre comment orienter cette transformation de façon pertinente pour les acteurs concernés. D'où l'importance des démarches de conception et d'intégration des technologies dans les environnements de travail. À l'opposé d'une approche substitutive de l'IA (remplacement de l'humain par celle-ci) qui constitue la toile de fond de cette crainte, nous avons vu qu'une approche mise en avant consiste à envisager l'Humain et l'IA en termes de complémentarité et non de remplacement du premier par le second. Dans ce cadre, différents modèles de complémentarité sont proposés, en particulier ceux de collaboration ou coopération, de partenariat ou encore de symbiose. Si ces modèles de relation Humain-machine, qui ne sont pas nouveaux, ont l'intérêt de permettre précisément de sortir de la vision substitutive et de plutôt chercher à tirer parti de ce que l'IA peut apporter aux travailleurs dans leurs activités en termes d'aide ou assistance, c'est-à-dire de s'inscrire dans une démarche centrée sur l'humain, nous avons vu qu'ils restent pour l'essentiel métaphoriques. L'état de l'art technologique actuel ne permet toujours pas de développer des systèmes véritablement coopératifs, collaboratifs, symbiotiques ou encore qui agissent comme des « partenaires ». Mais cela ne veut pas dire pour autant qu'il faille cesser de s'inspirer de ces modèles pour penser la relation Humain-IA. Ils resteront tout à fait pertinents à mesure que les capacités d'action et d'interaction des machines se développeront.

- 41 À côté de cette problématique récurrente portant sur le type de relation qu'il faudrait envisager entre humains et machine dans un cadre de complémentarité, certains systèmes d'IA soulèvent de nouvelles questions concernant cette relation. Comme nous l'avons vu avec l'exemple du robot Baxter, un point notable est que la dimension sociale de l'interaction est de plus en plus intégrée dans la conception de robots (les cobots) destinés au secteur industriel. Cette tendance suscite les questions suivantes : comment les humains vont travailler avec des systèmes (robots, etc.) qui ont de plus en plus de caractères anthropomorphes, c'est-à-dire incorporent de plus en plus de traits humains ? Quelles relations développeront-ils avec ces machines ? Jusqu'où faut-il aller dans ce « mimétisme anthropomorphique », au risque de susciter de fortes attentes qui ne sont pas en phase avec les capacités réelles des machines ? Cette question fait l'objet de vifs débats en robotique entre les chercheurs qui estiment qu'il faut éviter ce mimétisme, car il soulève de nombreux problèmes (attachement, isolement, confiance excessive dans la machine) et ceux qui estiment qu'il peut être utile dans certains contextes (Darling, 2017). Sur ce point, il nous semble qu'une position de principe qui s'abstrait du contexte d'usage n'est pas tenable. Doter les machines de caractères anthropomorphes parce qu'il est possible de le faire n'a pas non plus de sens. La

question centrale est en quoi cela peut-il être utile à l'activité ou à l'interaction humain-machine, tout en prenant en compte les risques soulignés ci-dessous.

- 42 Enfin, cette revue de questions sur l'IA a souligné un autre enjeu important, l'explicabilité des systèmes basés sur les techniques d'apprentissage machine, notamment celle de l'apprentissage profond. Bien qu'elle ne soit pas nouvelle, cet enjeu se pose avec beaucoup plus d'acuité, car, comme nous l'avons vu, la compréhension du fonctionnement des systèmes basés sur cette technique est particulièrement ardue. Comme l'ont montré les études passées sur les problèmes générés par l'opacité des systèmes experts et de façon plus générale des systèmes techniques complexes, cette question est importante à prendre en compte dans la conception et l'intégration des systèmes d'IA dans les situations de travail.
- 43 Pour conclure, pour faire face à ces enjeux et aux défis soulevés par les évolutions de l'IA, il paraît fondamental de maintenir et mobiliser une perspective centrée humain ou anthropocentrée (Rabardel, 1995), en mettant notamment l'accent sur les principes suivants (liste non exhaustive) : envisager les systèmes d'IA comme aides ou appuis cognitifs ou physiques à l'activité et qui s'articulent concevoir de façon fluide à celle-ci, considérer les possibilités et l'intérêt de l'automatisation ou délégation de tâches à l'aune de ce principe d'aide (que peut-elle apporter aux acteurs concernés ? Avec quelles conséquences sur leurs activités ?), veiller à ce que les systèmes ne deviennent pas une source additionnelle de tâches à réaliser ou d'obstacles à la réalisation du travail, ancrer la conception dans les réalités du travail actuel et la compréhension du travail futur en s'appuyant sur les méthodes existantes (par exemple, analyse de l'activité, simulations à partir de scénarios réalistes), mettre en œuvre une démarche de conception participative et itérative, mettre l'accent (en fonction de la situation) sur la nécessité de concevoir des systèmes dont les actions sont à toutes fins pratiques intelligibles ou explicables pour les acteurs concernés, faire en sorte que ces acteurs disposent d'une capacité de contrôle des systèmes (par exemple, valider les décisions qu'ils proposent, reprendre la main sur une tâche ou leur pilotage). Même s'ils varient dans la manière dont ils l'envisagent, c'est dans le sens de cette approche centrée humain qu'un certain nombre d'initiatives institutionnelles et de chercheurs (par exemple, Xu, 2019) appellent au développement d'une IA centrée humain⁴⁶. Compte tenu de la place centrale qu'occupe cette approche en ergonomie et des enjeux posés par l'IA, il y a là un champ de recherche et d'intervention immense pour les ergonomes.

BIBLIOGRAPHIE

Amalberti, R., & Deblon, F. (1992). Cognitive modelling of fighter aircraft process control: A step towards an intelligent on-board assistance system. *International Journal of Man-Machine Studies*, 36(5), 639-671.

Andersen, R. H., Solund, T., & Hallam, J. (2014). Definition and Initial Case-Based Evaluation of Hardware-Independent Robot Skills for Industrial Robotic Co-Workers. *41 st International Symposium on Robotics*. Munich, Germany, 2014, 1-7.

- Arntz, M., Gregory, T., & Zierahn, U. (2016). *The Risk of Automation for Jobs in OECD Countries: A Comparative Analysis*. Documents de travail de l'OCDE sur les questions sociales, l'emploi et les migrations. N° 189. Paris : Éditions OCDE. <https://doi.org/10.1787/5jlz9h56dvq7-en>.
- Bainbridge, L. (1983). Ironies of automation. *Automatica*, 19, 775-779.
- Bellotti, V., & Edwards, K. (2001). Intelligibility and accountability: human considerations in context-aware systems. *Hum.-Comput. Interact.*, 16(2), 193-212.
- Bencheikroun, T. H., Pavard, B., & Salembier, P. (1994). Design of cooperative systems in complex dynamic environments. In J. M. Hoc, C. Cacciabue & E. Hollnagell (Eds.), *Expertise and Technology - Cognition & Human-Computer Cooperation* (pp. 167-182). New Jersey : LEA.
- Bender, J., De Haan, J., & Bennett, D. (1995). Symbiotic approaches : Content and issues. In J. Bender, J. De Haan, & D. Bennett (Eds.), *The symbiosis of work and technology* (pp. 1-11). London : Taylor & Francis.
- Biran, O., & Cotton, C. (2017). Explanation and Justification in Machine Learning: A Survey. *IJCAI-17 Workshop on Explainable Artificial Intelligence (XAI)*.
- Boström, N. (2014) *Superintelligence: Paths, Dangers, Strategies*. Oxford : Oxford University Press.
- Bouri, M., Dieng, R., Kassel, G., & Safar, B. (1989). Vers des systèmes experts plus explicatifs. *Actes des 3^e journées nationales de PRC-GDR IA.*, Paris, 5-7 mars 340-355.
- Bradshaw, J.-M., Hoffman, R. R., Woods, D. D., & Johnson, M. (2013). The Seven Deadly Myths of "Autonomous Systems". *IEEE Intelligent Systems*, 28(3), 54-61.
- Brangier, E., Dufresne, A., & Hammes-Adelé, S. (2009). Approche symbiotique de la relation humain-technologie : perspectives pour l'ergonomie informatique. *Le Travail Humain*, 72(4), 333-353.
- Brooks, R.A. (1991). Intelligence without Representation. *Artificial Intelligence*, 47, 139-159.
- Brynjolfsson, E., & McAfee, A. (2012). *Race Against the Machines: How the Digital Revolution is Accelerating Innovation, Driving Productivity, and Irreversibly Transforming Employment and the Economy*. Digital Frontier Press.
- Cacciabue, P. C., Decortis, F., Drozdowicz, B., Masson, M., & Nordvik, J. P. (1992). COSIMO: A cognitive simulation model of human decision making and behavior in accident management of complex plants. *IEEE Trans. Systems, Man, and Cybernetics*, 22(5), 1058-1074
- Carr, N. (2017). *Remplacer l'humain : Critique de l'automatisation de la société*. Paris : L'Echappée.
- Carter, S., & Nielsen, M. (2017). Using Artificial Intelligence to Augment Human Intelligence, *Distill*.
- Castelfranchi, C. (1998). Modeling social action for AI agents. *Artificial Intelligence*, 10(1-2), 157-182.
- Christoffersen, K., & Woods, D.D. (2002). How to make automated system team players. In E. Salas (Ed.), *Advances in Human Performance and Cognitive Engineering Research*, vol2 (pp. 1-12). Amsterdam : Elsevier.
- Clancey, W.J. (1983). The epistemology of a rule-based expert system—a framework for explanation. *Artificial Intelligence*, 20(3), 215-251.
- Clarke, A.A., & Smyth, M.G.G. (1993). A co-operative computer based on the principles of human co-operation. *International Journal of Man-Machine Studies*, 38(1), 3-22.
- Collins, H. (2018). *Artificial Intelligence: Against Humanity's Surrender to Computers*. Polity Press.

- Crowston, K., & Bolici, F. (2019). Impact of machine learning on work. *Proceedings of the 52nd Hawaii International Conference on System Sciences*.
- Darling, K. (2017). "Who's Johnny?" Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy (March 23, 2015). In P. Lin, G. Bekey, K. Abney, & R. Jenkins (Eds.). *Robot Ethics 2.0*. (22 p.). Oxford : Oxford University Press.
- Daugherty, P., & Wilson, J. (2018). *Humans + Machine: Reimagining Work in the Age of AI*. Harvard Business Review Press.
- Davenport, T. H., & Kirby, K. (2016). Au-delà de l'automatisation. *Harvard Business Review*, Juin-juillet 2016, 45-53.
- Dekker, S.W.A., & Woods, D.D. (2002). MABA-MABA or Abracadabra? Progress on human-automation coordination. *Cognition, Technology & Work*. 4, 240-244.
- Dellermann, D., Calma, A., Lipusch, N., Weber, T., Weigel, S., & Ebel, P. (2019). The Future of Human-AI Collaboration: A Taxonomy of Design Knowledge for Hybrid Intelligence Systems. In *Hawaii International Conference on System Sciences (HICSS)*. Hawaii, USA.
- Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *CoRR, abs/1702.08608*, 2017.
- Dreyfus, H. (1972). *What computers can't do. The limits of artificial intelligence*. New-York : Harper & Row (traduction française, Flammarion, 1984, « Intelligence artificielle : Mythes et limites »).
- Dugdale, J., & Pavard, B. (2000). The Application of a Cognitive Engineering Methodology to Agent-based Simulation in the Social Sciences. In *Proceedings of Agent-Based Simulation Workshop*. May 2-3, Passau, Germany.
- Elprama, S., El Makrini, I., Vanderborght, B., & Jacobs, A. (2016). Acceptance of collaborative robots by factory workers: a pilot study on the role of social cues of anthropomorphic robots. *The 25th IEEE International Symposium on Robot and Human Interactive Communication*, New York.
- Engelbart, C. (1962). Augmenting Human Intellect: A conceptual framework. *SRI Summary Report AFOSR-322*.
- Falzon, P. (1989). Analyser l'activité pour l'assister. *Actes du 25^e Congrès de la SELF*, Lyon, 4-6 octobre.
- Ferber, J. (1995). *Les systèmes Multi-agents : vers une intelligence collective*. InterÉditions.
- Fitts, P. M. (Ed.). (1951). *Human engineering for an effective air navigation and traffic control system*. Washington DC : National Research Council.
- Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and autonomous systems*. 42(3-4), 143-166.
- Ford, M. (2016). *Rise of the Robots. Technology and the Threat of a Jobless Future*. New York : Basic Books
- French, B., Duenser, A., & Heathcote, A. (2018). *Trust in Automation - A Literature Review*. CSIRO Report EP184082. CSIRO, Australia.
- Frey, C.B., & M.A. Osborne (2013). *The Future of Employment: How Susceptible are Jobs to Computerization?* Working paper, Oxford : Oxford Martin School, University of Oxford.
- Freyssenet, M. (1992). Systèmes experts et division du travail. *Technologie, Idéologie, Pratiques*, X(2-4), 105-118.

- Froomkin, A. M., Kerr, I., & Pineau, J. (2019). When AIs Outperform Doctors: Confronting the Challenges of a Tort-Induced Over-Reliance on Machine Learning. *Arizona Law Review*, 61, 33.
- Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining Explanations: An Overview of Interpretability of Machine Learning. *DSAA 2018*, 80-89
- Goertzel, B. (2014). Artificial General Intelligence: Concept, State of the Art, and Future Prospects. *Journal of Artificial General Intelligence*, 5(1), 1-46.
- Goodman, B., & Flaxman, S. (2016). European union regulations on algorithmic decision-making and a “right to explanation”. *AI Magazine*, 38(3), 50-57.
- Griffith, D. (2007). Neo-symbiosis: A conceptual tool for system design. *HICSS, Proceedings of 40th Annual Hawaii International Conference on System Sciences (HICSS'07)*, p. 294 b.
- Griffith, D., & Greitzer, F.L. (2007). Neo-Symbiosis: The Next Stage in the Evolution of Human Information Interaction. *International Journal of Cognitive Informatics and Natural Intelligence*, 1, 39-52.
- Gunning, D. (2016). *Explainable Artificial Intelligence (XAI)*. <https://www.darpa.mil/program/explainable-artificial-intelligence>
- Haradji, Y., Guibourdenche, J., Reynaud, Q., Poizat, G., Sabouret, N., Sempé, F., Huraux, T., & Galbat, M. (2018). De la modélisation de l'activité humaine à la modélisation pour la simulation sociale : entre réalisme et fécondité technologique. *Activités*, 15(1). <https://doi.org/10.4000/activites.3106>
- Hoc, J.-M. (2000). La relation Homme-Machine en situation dynamique. *Revue d'Intelligence Artificielle*, 14(1-2/2000), 55-71.
- Hoc, J.M. (2001). Towards a cognitive approach to human-machine cooperation in dynamic situations. *International Journal of Human-Computer Studies*, 54, 509-540.
- Hoc, J. (2004). Vers une coopération homme-machine en situation dynamique. In P. Falzon (Ed.), *Ergonomie*, (pp. 269-283). Paris : Presses Universitaires de France.
- Hoff, K. A., & Bashir, M. (2015). Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors*, 57(3), 407-434.
- Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons*, 61(4), 577-586.
- Johnson, M., Bradshaw, J. M., Feltoich, P. J., Hoffman, R. R., Jonker, C., van Riemsdijk, B., & Sierhuis, M. (2011). Beyond Cooperative Robotics: The Central Role of Interdependence in Coactive Design. *IEEE: Intelligent Systems*, 26(3).
- Julia, L. (2019). *L'intelligence artificielle n'existe pas*. Paris : Éditions First.
- Kahneman, D. (2012). *Système 1 Système 2 - Les deux vitesses de la pensée*. Paris : Flammarion.
- Karppi, T., & Granata, Y. (2019). Non-artificial non-intelligence: Amazon's Alexa and the frictions of AI. *AI & Society*, 34(4), 867-876.
- Karsenty, L., & Brezillon, P. J. (1995). Cooperative problem solving and explanation. *Expert Systems With Applications*, 8(4), 445-462.
- Kass, R., & Finin, T. (1988). The Need for User Models in Generating Expert System Explanation. *Int. J. Expert Syst.*, 1(4), 345-375.

- Klein, G., Woods, D. D., Bradshaw, J. M., Hoffman, R. R., & Feltovich, P. J. (2004). Ten Challenges for Making Automation a “Team Player” in Joint Human-Agent Activity. *IEEE Intelligent Systems*, 19(6), 2004, 91-95.
- Kurenkov, A. (2015). A “Brief” History of Neural Nets and Deep Learning. *andreykurenkov.com*, December 24.
- Lavin, A. (2018). Interpreting AI Is More Than Black And White. <https://www.forbes.com/sites/alexanderlavin/2019/06/17/beyond-black-box-ai/> (Accès le 07/10/2019).
- Lecun, Y. (2019). *Quand la machine apprend. La révolution des neurones artificiels et de l'apprentissage profond*. Paris : Odile Jacob.
- Lee, K. M., Peng, W., Yan, C., & Jin, S. (2006). Can robots manifest personality?: An empirical test of personality recognition, social responses, and social presence in human-robot interaction. *Journal of Communication*, 56, 754-772.
- Lenat, D. B., Guha, R. V., Pittman, K., Pratt, D., & Shepherd, M. (1990). Cyc: toward programs with common sense. *Communication ACM* 33, 8 (August 1990), 30-49.
- Levesque, H. (2013). On Our Best Behaviour, *IJCAI-13*.
- Licklider, J.R. (1960). Man-Computer Symbiosis. *IRE Transactions on Human Factors in Electronics*, HFE-1, 4-11.
- Lipton, Z.C. (2018). The mythos of model interpretability. *acmqueue*, 16(3).
- Lundberg, S., Nair, B., Vavilala, M., Horibe, M., Eisses, M., Adams, T., Liston, D., Low, D., Newman, S., Kim, J., & Lee, S. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, 2, 749-760.
- McAfee, A., & Brynjolfsson, E. (2017). *Machine Platform Crowd: Harnessing our Digital Future*. New York : W.W. Norton and Company.
- McCarthy, J., Minsky, M., Rochester, N., & Shannon, C.E. (1955). *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*.
- McCulloch, W. S., & Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics*, 5, 115-133, 1943.
- Mercado, J.E., Rupp, M.A., Chen, J.Y., Barnes, M.J., Barber, D., & Procci, K. (2016). Intelligent agent transparency in human-agent teaming for multi-UxV management. *Human Factors*, 58(3), 401-415.
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267(2019), 1-38.
- Millot, P., & Lemoine, M.P. (1998). An attempt for generic concepts toward human-machine cooperation. *IEEE SMC*, San Diego, CA, octobre.
- Minsky, M., & Papert, S. (1969). *Perceptrons: An Introduction to Computational Geometry*. Cambridge MA : The MIT Press.
- Monroe, D. (2018). AI, explain yourself. *Communication ACM* 61, 11(October 2018), 11-13.
- Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73, 1-15.

- Moore, J. D., & Swartout, W. R. (1988). *Explanation in expert systems: A survey*. University of Southern California Marina Del Rey Information Sciences Institute. <http://www.dtic.mil/docs/citations/ADA206283>
- Newell, A., & Simon, H. A. (1976). *Computer Science as Empirical Inquiry: Symbols and Search*, *Communications of the ACM*, 19(3), 113-126.
- Nilsson, N. J. (2005). Human-Level Artificial Intelligence? Be Serious! *AI Magazine*, 26(4), 68-75.
- Norman, D. (2017). Design, business models, and human-technology teamwork: As automation and artificial intelligence technologies develop, we need to think less about human-machine interfaces and more about human-machine teamwork. *Research-Technology Management*, 60(1), 26-30.
- O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York, NY, USA : Crown Publishing Group.
- Parasuraman, R., Sheridan, Y., & Wickens, C. (2000). A Model for Types and Levels of Human Interaction with Automation. *IEEE Trans. Systems, Man and Cybernetics. Part A*, 30(3), 286-297.
- Park, D. H., Hendricks, L.-A., Akata, Z., Schiele, B. Darrell, T., & Rohrbach, M. (2018). Attentive Explanations: Justifying Decisions and Pointing to the Evidence. *Computer Vision and Pattern Recognition (CVPR)*, 2018.
- Peshkin M., & Colgate, J. E. (1999). Cobots. *Industrial Robot*, 26(5), 335-341.
- Rabardel, P. (1995). *Les hommes et les technologies ; approche cognitive des instruments contemporains*. Paris : Armand Colin.
- Rahimi, A., & Recht, B. (2017). *Reflections on Random Kitchens Sinks*. <http://www.argmin.net/2017/12/05/kitchen-sinks/>
- Ribeiro, M. T., Singh, & S., Guestrin, C. (2016). Why Should I Trust You? Explaining the Predictions of AnyClassifier. *arXiv:1602.04938 [cs.LG]*. <https://doi.org/10.1145/2939672.2939778>.
- Rosenblatt, R. (1957). *The perceptron, a perceiving and recognizing automaton (Project Para)*. Cornell Aeronautical Laboratory.
- Roth, E.M., Bennett, K.B., & Woods, D.D. (1987). Human interaction with an "intelligent" machine. *International Journal of Man-Machine Studies*, 27, 479-525.
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1, 206-215.
- Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.
- Salembier, P. (1994). Assistance coopérative aux activités complexes : l'exemple de la régulation du trafic aérien. In B. Pavard (Ed.), *Systèmes coopératifs : de la modélisation à la conception* (pp. 377-407). Toulouse : Octarès.
- Salembier, P., & Zouinar, M. (2004). Intelligibilité mutuelle et contexte partagé. *Activités*, 1(2). <https://doi.org/10.4000/activites.1243>
- Sarter, N. B., Woods, D. D., & Billings, C. E. (1997). Automation surprises. *Handbook of human factors and ergonomics*, 2, 1926-1943.
- Sauppé, A., & Mutlu, B. (2015). The Social Impact of a Robot Co- Worker in Industrial Settings. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15). ACM, New York, NY, USA. 3613-3622.

- Schmidt, K. (2002). Remarks on the complexity of cooperative work. *Revue d'Intelligence Artificielle*, 16(4-5), 443-483.
- Schmorrow, D.D., Stanney, K.M., & Reeves, L.M. (Eds.) (2006). *Foundations of augmented cognition* (2nd ed.). Arlington, VA : Strategic Analysis.
- Searle, J. R. (1980). Minds, Brains and programs, *The Behavioral and Brain Sciences*, vol. 3, Cambridge University Press.
- Seidel, S., Berente, N., Lindberg, A., Nickerson, J. V., & Lyytinen, K. (2019). Autonomous tools and design work: A triple-loop approach to human-machine learning. *Communications of the ACM*, 62(1), 50-57.
- Silverman, B. G. (1992). Human-Computer Collaboration. *Human-Computer Interaction*, 7(2), 165-196.
- Shapiro, S. (1992). *Encyclopedia of Artificial Intelligence* (2nd Edition). New York : Wiley.
- Sheridan, T.B., & Verplank, W. (1978). *Human and Computer Control of Undersea Teleoperators. Technical Report*. Boston : Man-Machine Systems Lab., Dept. of Mechanical Eng., Massachusetts Institute of Technology.
- Suchman, L. (1987). *Plans and situated actions: The Problem of Human-Machine Communication*. New York : Cambridge University Press.
- Suchman, L., & Weber, J. (2015). Human-Machine Autonomies. In N. Bhuta, S. Beck, R. Geiß, HY Liu, & C. Kreß (Eds.), *Autonomous Weapons Systems: Law, Ethics, Policy* (pp. 75-10). Cambridge, UK : Cambridge University Press.
- Swartout, W.R., & Moore, J. D. (1993). Explanation in second generation expert systems. *Second Generation Expert Systems*, 543-585.
- Teach, R. L., & Shortliffe, E. H. (1981). An analysis of physician attitudes regarding computer based clinical consultation systems. *Computers and Biomedical Research*, 14(6), 542-558.
- Terssac, G. de (1992) *Autonomie dans le travail*. Paris : PUF, coll. Sociologie d'Aujourd'hui.
- Terveen, L. G. (1995). Overview of human-computer collaboration. *Knowledge-Based Systems*, 8(2-3), 67-81.
- Theureau, J., & Filippi, G. (1994). Cours d'action et conception d'une situation d'aide à la coordination : le cas de la régulation du trafic du RER. *Sociologie du Travail*, n° XXXVI 4/94, 547-562.
- van Melle, V. (1978). The Structure of the MYCIN System. *International Journal of Man-Machine Studies*, 10(3), 313-322.
- Visetti, Y.M. (1991). Des systèmes experts aux systèmes à base de connaissances : à la recherche d'un nouveau schéma régulateur. *Intellectica*, 12, 221-279.
- Weiss, G. (Ed.) (1999). *Multiagent systems. A modern approach to distributed artificial intelligence*. MIT Press.
- Weld, D. S., & Bansal, G. (2019). The Challenge of Crafting Intelligible Intelligence. *Communications of the ACM*, June 2019, 62(6), 70-79.
- Winter, J.C.F. (de), & Dodou, D. (2014). Why the Fitts list has persisted throughout the history of function allocation. *Cognition, Technology & Work*, 16(1), 1-11.

Wisskirchen, G., Thibault Biacabe, B., Bormann, U., Muntz, A., Niehaus, G., Soler, G. J., & von Brauchitsch, B. (2017). *Artificial Intelligence and Robotics and Their Impact on the Workplace*. Rapport de l'IBA Global Employment Institute.

Woods, D. D., Roth, E. M., & Bennett, K. (1987). Explorations in joint human-machine cognitive systems. In W. Zachary & S. Robertson (Eds.), *Cognition, computing and cooperation*. Norwood NJ: Ablex.

Xu, W. (2019). Toward Human-Centered AI: A Perspective from Human-Computer Interaction. *Interactions*, July-August 2019.

Yosinski, J., Clune J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding Neural Networks Through Deep Visualization. *Deep Learning Workshop, International Conference on Machine Learning (ICML 2015)*.

Zouinar, M. (2000). *Analyse et modélisation des processus de construction et d'actualisation du contexte partagé*. Thèse de Doctorat, CNAM, Paris.

NOTES

1. Cette médiatisation est surtout alimentée par les inquiétudes spéculatives soulevées par l'IA comme la disparition des emplois sur laquelle nous reviendrons, la mise en péril de l'humanité ou le possible dépassement généralisé de l'humain par les machines dans le domaine cognitif (c'est l'hypothèse de la « singularité » qui, au passage, fait l'objet de nombreuses critiques).
2. Livres blancs, documents d'expertise, rapports institutionnels, livres de vulgarisation.
3. Dans la suite de cet article, ce terme sera utilisé pour qualifier les « sorties » d'un système d'IA, qui peuvent être une classification, une catégorisation, la prévision d'un événement futur ou la proposition d'une action. Ces « sorties » sont parfois qualifiées de « décisions ».
4. D'autres questions tout aussi importantes soulevées par la nouvelle génération de systèmes d'IA ne seront pas évoquées ici ou le seront de manière succincte faute de place. Il s'agit par exemple de celles de la confiance, la formation des salariés (quelles compétences doivent-ils développer pour travailler avec des systèmes d'IA apprenants ?), les démarches de conception et d'intégration de l'IA dans les organisations.
5. Pour des discussions plus récentes voir par exemple Karppi et Granata (2019). Julia (2019) estime même que l'IA « n'existe pas » car on ne sait pas ce qu'est l'intelligence. Comme d'autres auteurs (Carter & Nielsen, 2017), il préfère donner un autre sens à IA : Intelligence Augmentée. Compte tenu de ces débats, il est tentant d'éviter de parler d'Intelligence Artificielle mais étant donné son inscription historique dans la société, il est difficile de revenir en arrière. Nous garderons donc ici cette expression pour cette même raison, sachant qu'elle reste problématique.
6. Le débat sur la définition de l'intelligence reste donc largement ouvert.
7. "(...) every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it." McCarthy et al. (1955). Notons que cette perspective avait été déjà ouverte bien plutôt dans les années 40 par Alan Turing lorsqu'il s'est demandé dans un article désormais classique si les machines pouvaient penser, et dans lequel il proposait son fameux test éponyme (Test de Turing).
8. Développé par Rosenblatt (1957), il s'agit du premier modèle de réseau de neurones artificiels capable d'apprentissage. Minsky et Papert (1969) ont notamment montré que, dans sa version monocouche, le perceptron n'était pas en mesure de calculer la fonction booléenne XOR, c'est-à-dire le OU exclusif. Or le procédé algorithmique développé par Rosenblatt ne pouvait pas fonctionner avec plusieurs couches de « neurones » qui sont nécessaires pour faire ce type de calcul (Kurenkov, 2015).

9. Il convient de souligner que l'expression « systèmes experts » a fait l'objet d'un certain nombre de critiques qui se sont notamment traduits par l'abandon de cette expression au profit de celle de « systèmes à base de connaissances » (Visetti, 1991).
10. Voir aussi les critiques qui ont porté sur le statut de la « simulation » informatique de l'intelligence humaine (Visetti, 1991).
11. Développés notamment par Yann Lecun, ces réseaux s'inspirent de l'architecture des connexions du cortex visuel des mammifères.
12. Il faut noter que l'utilisation des techniques d'apprentissage actuelles peut nécessiter un travail humain qui peut être considérable : il faut rassembler les données d'apprentissage, les préparer, les étiqueter, sélectionner un modèle, choisir et régler ses paramètres, etc.
13. Contrairement aux robots industriels classiques, il s'agit de robots qui sont conçus pour interagir directement avec des humains dans une grande variété de contextes (à la maison, à l'école, à l'hôpital, etc.). Ces robots sont destinés à rendre différents types de service, comme assister les personnes âgées ou en situation de handicap, réaliser des tâches domestiques, jouer le rôle de compagnon ou de thérapeute pour accompagner des patients dans un cadre médical. Doter ces robots de capacités interactionnelles « sociales » constitue l'un des objectifs cruciaux de ce champ de la robotique dite « sociale » (Fong, Nourbakhsh, & Dautenhahn, 2003). Ces capacités sont considérées comme essentielles pour faciliter l'adoption des robots dans ces contextes (Lee, Peng, Yan, & Jin, 2006).
14. Une telle IA reste purement spéculative voire un mythe. Certains auteurs comme Boström prennent cependant au sérieux sa possible émergence.
15. Collins (2018) propose une distinction plus fine entre « IA faible » et IA « forte ». Il propose six niveaux d'IA : le premier correspond aux technologies actuelles, y compris les plus « simples » comme les thermostats qui sont capables de s'autoréguler, le deuxième se réfère aux « prothèses asymétriques » qui réalisent une activité typiquement humaine (par exemple, l'interaction vocale comme Siri) mais qui ne sont pas capables de réparer les dysfonctionnements humains (erreurs) alors que nous pouvons réparer les leurs (d'où l'asymétrie). Les niveaux suivants sont hypothétiques : le troisième est le domaine des prothèses symétriques ; le quatrième celui des machines qui reproduisent les processus internes de l'intelligence humaine ; le cinquième imagine des machines autonomes, ayant un corps de type humain et capables de former des « sociétés » ; le 6^e renvoie à des machines dotées d'une corporéité différente, capables de s'auto-améliorer au fil des générations et de former des « sociétés ».
16. "To pass the employment test, AI programs must be able to perform the jobs ordinarily performed by humans. Progress toward human-level AI could then be measured by the fraction of these jobs that can be acceptably performed by machines." (Nilsson, 2005)
17. Par exemple, dans le domaine juridique, l'informatisation des décisions de justice permet de développer des systèmes d'IA qui calculent les « chances » de gagner un procès, en fonction de la juridiction, du magistrat et du cabinet d'avocat.
18. <http://reports.weforum.org/future-of-jobs-2018/preface/>
19. Voir le site suivant qui répertorie les maladies pour lesquelles l'IA obtient de meilleurs scores de diagnostic que les médecins depuis 2016 : <https://spectrum.ieee.org/static/ai-vs-doctors>
20. Les auteurs reconnaissent que les algorithmes peuvent comporter des biais et doivent donc faire l'objet d'analyse.
21. Cette approche consistant à s'appuyer sur cette distinction entre système 1 et 2 a été mobilisée par d'autres auteurs pour définir le rôle de la technologie vis-à-vis de l'humain dans le cadre d'activités de travail. C'est le cas de Griffith et Greitzer (2007) selon qui la technologie informatique doit surtout soutenir le système 2 pour le rendre plus efficient puisqu'il est lent. Mais il ne s'agit pas pour les auteurs de s'inscrire dans une logique de substitution (transfert des fonctions du système 2 aux machines) mais d'augmentation de l'humain, en particulier de ces capacités de raisonnement.

22. Le sens donné par ces auteurs à cette notion est différent de celui de l'approche instrumentale (Rabardel, 1995). Elle est utilisée au sens « d'outil cognitif » qui se caractérise notamment par le fait qu'il reste sous le contrôle de l'utilisateur : "Rather than deploying machine power in the form of a prosthesis--a replacement or remedy for perceived human deficiencies, cognitive tools can be conceptualized as instruments--a means for effecting something in the hands of a competent practitioner." (Roth *et al.*, 1987, p. 503)
23. Pour cette raison, Dekker et Woods (2000) l'ont qualifié de méthode MABA-MABA (« Men-Are-Better-At / Machines-Are-Better-At »).
24. Pour dépasser ces limites, d'autres modèles d'automatisation ont été proposés (voir par exemple, Parasuraman, Sheridan, & Wickens, 2000 ; Sheridan, & Verplank, 1978).
25. On retrouve également cette approche dans le courant de la robotique collaborative ou cobotique (Peshkin, & Colgate, 1999).
26. https://www.nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf
27. Il existe aujourd'hui un champ de recherche entièrement consacré à l'Augmentation de la Cognition Humaine (Schmorrow, Stanney, & Reeves, (2006). Ce champ a été notamment initié par la DARPA, qui est l'agence du département de la Défense des Etats-Unis, chargée de la recherche et du développement des nouvelles technologies destinées aux militaires. Notons également que cette vision augmentative a beaucoup influencé des acteurs importants du développement de l'informatique « personnelle » comme Alan Kay et Steve Jobs.
28. La notion de symbiose entre humains et machines a été reprise par certains auteurs (par exemple, Bender, De Haan, & Bennett, 1995 ; Brangier, Dufresne, & Hammes-Adelé, 2009 ; Griffith, 2007), et mobilisée plus récemment par Jarrahi (2018) dans le contexte de l'IA.
29. Ces deux notions sont souvent utilisées de manière similaire ou interchangeable. Dans la suite, nous garderons cette équivalence même s'il est possible de les distinguer.
30. "However, we think that it is not reasonable, at any price, to transfer all the complexity of human human cooperation to human machine cooperation. The human machine couple is indeed a chimera and the current limits of the machine do not enable us to make such a transfer." (p. 534)
31. Voir aussi la critique très intéressante de Suchman et Weber (2015) qui proposent de repenser l'autonomie des machines en termes de « configurations » humains-machines.
32. Cette approche s'inscrivait dans le mouvement du cognitivisme.
33. Prenant en compte ce problème de la symétrisation ontologique au regard de la notion de symbiose, Griffith (2007) a proposé une approche, qu'il appelle Néo-symbiose, dans laquelle l'humain a une « position supérieure » (super ordinate) du fait qu'il possède une conscience. Pour cet auteur, cette caractéristique permet ainsi de souligner l'asymétrie « ontologique » entre l'humain et machine qui ne possède pas une telle conscience.
34. Cette société a été fondée par Rodney Brooks, dont nous avons déjà parlé à propos de « l'intelligence sans représentations ».
35. Notons que ces phénomènes d'anthropomorphisme dans la relation aux machines ont été observés dans d'autres contextes de travail où les robots n'avaient pas une apparence humaine (Darling, 2017)
36. D'autres types de programmes d'IA posent également ce problème d'intelligibilité, c'est le cas des systèmes de planification automatique (Weld, & Bansal, 2019)
37. Il convient de noter que ce n'est pas le cas de tous les algorithmes d'apprentissage. Il existe toute une classe d'algorithmes dont il est possible de comprendre le fonctionnement.
38. "While the heuristic optimization procedures for neural networks are demonstrably powerful, we don't understand how they work, and at present cannot guarantee a priori that they will work on new problems."
39. <http://www.argmin.net/2017/12/05/kitchen-sinks/>

40. Ce positionnement a ouvert un débat épistémologique et éthique intéressant au sein de la communauté des spécialistes de l'apprentissage automatique entre ceux qui estiment que la recherche d'une explication théorique rigoureuse du fonctionnement de ces algorithmes n'est pas utile (l'approche empirique est suffisante) et ceux qui pensent qu'elle est nécessaire, notamment du fait du déploiement des systèmes basés sur ces algorithmes dans la société. Certains pensent même qu'il faut renoncer à comprendre, et prendre acte de la complexité insondable de ces systèmes.

41. Une étude sur l'utilisation du système de prédiction criminelle COMPAS au Etats-Unis a mis en lumière cette conséquence. Voir : <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>

42. Un exemple emblématique est celui d'un système qui établit qu'un asthmatique a peu de risque de décéder d'une pneumonie, ce qui est erroné. Cette erreur vient du fait que l'apprentissage du système s'est appuyé sur des données d'où il a tiré la conclusion suivante : des patients asthmatiques atteints d'une pneumonie et bénéficiant d'un soin intensif ont moins de risque de décéder de cette maladie que les autres.

43. Cette demande est également liée à la découverte de biais sociaux (raciaux, de genre, etc.) dans les systèmes basés sur l'apprentissage automatique (O'Neill, 2016).

44. Aux États-Unis, la DARPA dont nous avons déjà parlé, a même lancé un programme de recherche appelé « XAI » (*eXplainable AI*) spécifiquement dédié à ce problème.

45. Il reste à voir dans quelle mesure elles se traduiront par un véritable déploiement dans des situations réelles.

46. Par exemple, l'institut de « l'IA centrée humain » (*Human-Centered Artificial Intelligence*) créé par l'université de Stanford ou encore.

RÉSUMÉS

Depuis quelques années, l'Intelligence Artificielle (IA) connaît un regain d'intérêt sans précédent grâce à d'importantes avancées technologiques, notamment dans le domaine de l'apprentissage machine (*machine learning*), qui étendent les capacités des ordinateurs et accroissent leurs performances dans un grand nombre de domaines (traitement du langage, compréhension de la parole, reconnaissance d'images, robotique, etc.). Ces avancées ouvrent de vastes perspectives en termes d'innovation technologique et d'automatisation dans les situations de travail. Cet article s'intéresse aux questions et enjeux soulevés par ces évolutions concernant l'activité humaine au travail. Il montre que la majorité des questions soulevées ne sont pas nouvelles. Il s'agit en particulier des questions concernant les incidences de l'automatisation sur le travail, et la manière d'envisager la répartition du travail et la relation entre Humain et IA. Certaines questions qui ne sont pas nouvelles se posent avec plus d'acuité, c'est le cas d'une part de « l'explicabilité » des systèmes d'IA basés sur l'apprentissage et, d'autre part, des conséquences à plus long terme de cet apprentissage sur l'activité humaine. Enfin, de nouvelles questions émergent, comme celles qui concernent plus particulièrement le travail avec des machines intelligentes qui exhibent des caractères anthropomorphes.

Over recent years we have seen an unprecedented revival of interest in Artificial Intelligence (AI) due to major technological advances, particularly in the field of machine learning, which extend the capabilities of computers and increase their performance in a large number of domains

(language processing, speech understanding, image recognition, robotics, etc.). These advances have opened up vast opportunities in terms of technological innovation and automation in work situations. This article focuses on the questions and issues raised by these developments concerning human activity at work. The main conclusions of this article are, first of all, that the majority of the issues raised are not new. These include issues concerning the consequences of automation on work, and how to approach the division of tasks and the relationship between AI and Human. Secondly, some of the questions that are not new are becoming more challenging, such as the explainability of AI systems. Finally, new questions are emerging, such as those relating more particularly to work with “intelligent” machines that exhibit anthropomorphic features.

INDEX

Keywords : artificial intelligence, machine learning, automation, activity, work, human-machine relationships, explainability

Mots-clés : intelligence artificielle, automatisation, activité, travail, relation humain-machine

AUTEUR

MOUSTAFA ZOUINAR

Orange labs et CNAM, 40-48 avenue de la République, 92320 Chatillon, France.

moustafa.zouinar@orange.com