Organisation des Nations Unies pour l'éducation, la science et la culture

Commission mondiale d'éthique des connaissances scientifiques et des technologies (COMEST)

SHS/COMEST/EXTWG-ETHICS-AI/2019/1
Paris, le 26 février 2019
Original anglais

ÉTUDE PRÉLIMINAIRE SUR L'ÉTHIQUE DE L'INTELLIGENCE ARTIFICIELLE

En s'appuyant sur les travaux de la COMEST concernant l'éthique de la robotique (2017) et les implications éthiques de l'Internet des objets, un Groupe de travail élargi de la COMEST sur l'éthique de l'intelligence artificielle a rédigé la présente étude préliminaire.

Le présent document ne prétend pas être exhaustif et ne représente pas nécessairement les opinions des États membres de l'UNESCO.

ÉTUDE PRÉLIMINAIRE SUR L'ÉTHIQUE DE L'INTELLIGENCE ARTIFICIELLE

TABLE DES MATIÈRES

ÉTUDE PRÉLIMINAIRE SUR L'ÉTHIQUE DE L'INTELLIGENCE ARTIFICIELLE

INTRODUCTION

1. Le monde est confronté à un rapide essor de l'« intelligence artificielle » (IA). Grâce aux progrès accomplis dans ce domaine, on dispose maintenant de machines capables d'apprendre et d'effectuer des tâches cognitives autrefois réservées aux êtres humains. Cette évolution technologique aura certainement des implications sociales et culturelles importantes. S'agissant d'une technologie cognitive, les implications de l'IA sont intimement liées aux domaines centraux de: éducation, science, culture et communication. Les algorithmes jouent aujourd'hui un rôle crucial dans la sélection de l'information et des nouvelles que les individus consultent, la musique qu'ils écoutent et les décisions qu'ils prennent. Les systèmes d'IA sont une source croissante d'information pour les médecins, les scientifiques et les magistrats. Dans le domaine de la recherche scientifique, l'IA sert désormais à analyser et interpréter les données. De plus, le remplacement progressif du travail humain par les technologies intelligentes exige de la part de la main d'œuvre de nouvelles formes de résilience et de flexibilité. Des intellectuels influents, comme Stephen Hawking, ont même exprimé la crainte que l'IA ne devienne une menace pour la survie de l'humanité, dans la mesure où elle pourrait prendre le contrôle de bien des aspects de nos vies quotidiennes et de l'organisation de la société.

2. Le terme d'« intelligence artificielle » a été lancé dans les années 1950 pour qualifier des machines capables de réaliser plus que des tâches routinières. Avec l'augmentation de la puissance de calcul, le terme a été appliqué à des machines capables d'apprendre. Bien qu'il n'existe pas une définition unique de l'IA, il est communément admis que les machines basées sur l'IA, ou sur l'« informatique cognitive », ont la capacité potentielle d'imiter, voire de dépasser des capacités cognitives humaines comme la détection, l'interaction linguistique, le raisonnement et l'analyse, la résolution de problèmes et même la créativité. En outre, ces « machines intelligentes » peuvent démontrer des facultés d'apprentissage comparables à celles des humains, avec des mécanismes d'autorelation et d'autocorrection, grâce à des algorithmes d'« apprentissage machine » ou « automatique », voire même d'« apprentissage profond », utilisant des « réseaux neuronaux » imitant le fonctionnement du cerveau humain.

3. Récemment, de grandes firmes technologiques multinationales, dans de nombreuses régions du monde, ont commencé à investir massivement dans l'utilisation de l'IA dans leurs produits. La puissance de calcul a atteint un niveau tel qu'il permet l'exécution d'algorithmes d'une grande complexité et le traitement des « big data », les gigantesques données de masse pouvant servir à l'apprentissage automatique. Ces entreprises ont accès à une puissance de calcul quasi illimitée, ainsi qu'aux données collectées auprès de milliards d'individus pour « nourrir » leurs systèmes d'IA en intrants d'apprentissage. De plus, par le biais de leurs produits, l'IA gagne rapidement en influence dans la vie quotidienne des individus et dans des secteurs professionnels comme la santé, l'éducation, la recherche scientifique, les communications, les transports, la sécurité et l'art.

4. Cette profonde influence de l'IA suscite des inquiétudes qui pourraient ébranler la confiance des individus dans ces technologies. Elles vont de la crainte de la criminalité, de l'escroquerie et du vol d'identité au harcèlement ou aux agressions sexuelles, des discours de haine et de la discrimination à la désinformation, et, plus généralement, de la question de la transparence des algorithmes à celle de la confiance que l'on peut avoir dans les systèmes d'IA. Dans la mesure où nombre de ces problèmes ne peuvent être

résolus par la seule réglementation, l'UNESCO a proposé une gouvernance multiparties prenantes comme un moyen optimal d'associer les différents acteurs à la formulation et à la mise en œuvre de normes, d'une éthique et d'une politique, et d'autonomiser les utilisateurs.

5. Du fait de ses profondes incidences sociales, beaucoup d'organisations et de gouvernements s'inquiètent des implications éthiques de l'IA. La Commission européenne a créé un Groupe d'experts de haut niveau sur l'IA composé de représentants des milieux universitaires, de la société civile et des entreprises, et une Alliance européenne pour l'IA, large plate-forme ouverte de discussion sur tous les aspects du développement de l'IA et de ses incidences. Le Groupe européen d'éthique des sciences et des nouvelles technologies a émis une *Déclaration sur l'IA, la robotique et les systèmes « autonomes »* (EGE, 2018). La Commission européenne a publié une *Communication sur l'intelligence artificielle pour l'Europe* (CE, 2018) et le Conseil de l'Europe a produit divers rapports sur l'IA et créé un Comité d'experts chargé de travailler sur *les dimensions des droits de l'homme du traitement automatisé des données et des différentes formes d'intelligence artificielle*. L'organisation IEEE a lancé une Initiative mondiale sur l'éthique des systèmes autonomes et intelligents. L'OCDE a lancé un projet intitulé « Vers le numérique », qui vise à aider les décideurs dans tous les domaines stratégiques concernés à mieux comprendre la révolution numérique en marche dans les différents secteurs de l'économie et l'ensemble de la société. L'OCDE a également nommé un groupe d'experts (l'AIGO), chargé de fournir des orientations sur les principes de cadrage de l'intelligence artificielle dans la société. L'UIT et l'OMS ont créé un groupe de discussion sur « l'intelligence artificielle au service de la santé ». De nombreux pays ont aussi entamé une réflexion sur leurs orientations éthiques et politiques à l'égard de l'IA, comme le rapport Villani en France (Villani *et al.*, 2018), le rapport de la Chambre des lords au Royaume-Uni (House of Lords, 2017), et le rapport du Bureau exécutif du Président des États-Unis (2016).

6. L'UNESCO a une perspective unique à apporter à ce débat. L'IA a des implications pour les domaines de travail centraux de l'UNESCO. Aussi, en plus des nombreux cadres et lignes directrices d'ordre éthique qui sont actuellement élaborés par les gouvernements, les entreprises et les organisations de la société civile, l'UNESCO peut apporter une approche pluridisciplinaire, universelle et holistique au développement de l'IA au service du genre humain, du développement durable et de la paix.

7. À cet égard, il existe plusieurs cadres et initiatives sur lesquels s'appuyer. Citons, en premier lieu, le cadre des *droits de l'homme*, à la base de la Déclaration de principes de Genève du Sommet mondial sur la société de l'information (SMSI), qui stipule que « l'utilisation des TIC et la création de contenus devrait respecter les droits de l'homme et les libertés fondamentales d'autrui, notamment la vie privée ainsi que la liberté d'opinion, de conscience et de religion, conformément aux instruments internationaux pertinents » (SMSI, 2003). Le SMSI (2005) propose une approche multiparties prenantes appelant à une coopération efficace entre toutes les parties prenantes – gouvernements, secteur privé, société civile, organisations internationales et communautés techniques et universitaires. Dans le cadre du processus de suivi du SMSI, l'UNESCO a adopté cette approche multiparties prenantes et se charge de mettre en œuvre les grandes orientations concernant l'accès (C3), le télé-enseignement (C7), la diversité culturelle (C8), les médias (C9) et les dimensions éthiques de la société de l'information (C10).

8. En second lieu, on peut s'appuyer sur le cadre concernant *l'universalité de l'Internet* et sur les *principes ROAM* connexes, tels qu'approuvés en 2015 par la 38e Conférence générale (UNESCO, 2015b). Ces principes – Respect des droits humains, Ouverture, Accessibilité, Multiples acteurs – sont le fruit de l'étude « Des clés pour la promotion de

sociétés du savoir inclusives » de l'UNESCO réalisée pour la 38e Conférence générale (UNESCO, 2015a). Dans le document final de la conférence, intitulé « Interconnecter les ensembles », l'UNESCO s'engage à promouvoir une réflexion, des activités de recherche et un dialogue public éthiques, fondés sur les droits humains, concernant les implications des technologies nouvelles et émergentes et leurs effets potentiels sur la société. En outre, à sa 18e session, le Conseil intergouvernemental du Programme Information pour tous (PIPT) de l'UNESCO a examiné et approuvé le Code d'éthique pour la société de l'information, élaboré par le Groupe de travail du PIPT sur l'éthique de l'information.

9. Dans son examen des implications éthiques de l'IA, la présente étude expliquera d'abord ce qu'est l'intelligence artificielle, comment elle fonctionne et en quoi elle diffère des autres technologies. La deuxième section étudiera les aspects éthiques de l'IA, en partant des domaines de compétence de l'UNESCO que sont l'éducation, la science, la culture et la communication, ainsi que les dimensions éthiques et mondiales de la paix, de la diversité culturelle, de l'égalité des genres et de la durabilité. Cet examen devrait être perçu comme une exploration, non comme une analyse exhaustive, de questions qui vont de la diversité culturelle à la confiance dans la science, de la créativité artistique à la réflexion critique, et d'une prise de décisions basée sur l'IA au rôle de l'intelligence artificielle dans les pays en développement. La troisième section de cette étude préliminaire dégagera les dimensions centrales que toute réflexion éthique digne de ce nom devrait adopter du point de vue de l'UNESCO.

I. QU'EST-CE QUE l'IA?

I.1. Définition

10. L'idée d'une « intelligence artificielle » (IA), ou d'êtres, de machines ou d'outils « créés artificiellement » et « intelligents », parsème l'histoire humaine. On en trouve diverses formes dans les religions, les mythologies, la littérature et les traditions philosophiques en Occident comme ailleurs. En tant que tels, ces exemples témoignent de l'éternelle curiosité du genre humain pour de telles entités, et bien qu'exprimée sous des apparences culturellement variées, il s'agit d'une curiosité partagée et transculturelle. Aujourd'hui, la fascination que suscite l'IA – y compris ses dimensions éthiques – est amplifiée par son développement et ses applications dans le réel.

11. Tout examen des implications éthiques de l'IA exige d'en préciser les sens possibles. Le terme a été forgé en 1955 par John McCarthy, Marvin L. Minsky, Nathaniel Rochester et Claude E. Shannon. Ces derniers proposaient de mener une « étude de l'intelligence artificielle (...) en partant de l'hypothèse selon laquelle tous les aspects de l'apprentissage ou tout autre trait de l'intelligence peuvent en principe être si précisément décrits qu'on peut construire une machine pour les simuler » (McCarthy *et al.*, 2006 [1955], p. 12). Au fur et à mesure du développement et de la diversification du domaine, au cours des décennies suivantes, on a vu croître le nombre des significations données à « l'IA » et on ne dispose pas actuellement d'une définition universellement reconnue. Les différentes définitions de l'IA renvoient à différentes approches disciplinaires comme l'informatique, l'ingénierie électrique, la robotique, la psychologie et la philosophie.

12. Malgré le nombre et la diversité des définitions de l'IA, un certain consensus existe, au niveau le plus général, sur l'idée d'en distinguer deux aspects : l'un que l'on qualifie le plus souvent de « théorique » ou de « scientifique », et l'autre, de « pragmatique » ou de « technologique ».

13. Parler d'IA « théorique » ou « scientifique » consiste à « utiliser les concepts et les modèles de l'IA pour tenter de répondre à des questions sur les êtres humains et d'autres êtres vivants » (Boden, 2016, p. 2). L'IA « théorique » ou « scientifique » est ainsi naturellement liée à des disciplines comme la philosophie, la logique, la linguistique, la psychologie et les sciences cognitives. Elle soulève des questions telles que : qu'entend-on par « intelligence » et comment distinguer l'intelligence « naturelle » de l'intelligence « artificielle » ? La pensée a-t-elle nécessairement besoin d'un langage symbolique ? Peut-on créer une « IA forte » (une intelligence *authentique* de la même espèce et aussi générale que l'intelligence humaine), par opposition à une « IA faible » (qui ne fait qu'*imiter* l'intelligence humaine et ne peut exécuter qu'un nombre limité d'instructions étroitement définies) ? Bien qu'il s'agisse de questions d'ordre théorique ou scientifique, elles contiennent des dimensions métaphysiques ou spirituelles (par exemple, sur le caractère unique de l'humanité ou sur le libre arbitre) qui ont elles-mêmes des implications éthiques indirectes, mais néanmoins sérieuses.

14. L'IA « pragmatique » ou « technologique » intéresse l'ingénierie. Elle s'appuie sur les diverses branches de l'IA – les exemples classiques en sont le traitement automatique du langage naturel, la représentation des connaissances, le raisonnement automatisé, l'apprentissage machine, l'apprentissage profond, la vision par ordinateur et la robotique (Russell et Norvig, 2016, p. 2-3) – afin de créer des machines ou des programmes capables d'exécuter indépendamment des tâches qui exigeraient autrement une intelligence et une intervention humaines. L'IA « pragmatique » ou « technologique » a connu un succès remarquable en combinaison avec les TIC (technologies de l'information et de la communication). Les innovations en matière d'IA sont utilisées aujourd'hui dans de nombreux domaines de la vie moderne, comme les transports, la médecine, la communication, l'éducation, la science, la finance, le droit, le domaine militaire, le commerce, les services au consommateur et les loisirs. Toutes ces innovations posent des problèmes éthiques, qui vont de la disparition des emplois traditionnels à une déshumanisation générale des relations humaines et de la société dans son ensemble, en passant par les dommages physiques ou psychologiques qu'elles pourraient causer aux êtres humains. À l'heure actuelle, aucun système d'IA ne peut être considéré comme un agent intelligent généraliste, capable de fonctionner efficacement dans un large éventail d'environnements, ce qui est une faculté propre à l'intelligence humaine.

15. Une des particularités de l'IA a trait à son « étrangeté » pour nous humains, au sens où cette intelligence fonctionne d'une façon qui nous semble déroutante et mystérieuse. Cette « étrangeté » tient surtout à ce qu'on pourrait appeler son « activité sans conscience ». Une IA de très haut niveau comme AlphaGo ou Watson peut exécuter des tâches impressionnantes sans reconnaître ce qu'elle fait. AlphaGo a battu plusieurs maîtres de go sans même savoir qu'il jouait à un jeu humain appelé go. Watson a répondu instantanément à des questions diaboliques, que la plupart des humains ont du mal à même comprendre dans le temps imparti. Pourtant, Watson ne « répond » pas au sens où l'entendent les humains : il « calcule » les probabilités des diverses réponses possibles à partir de l'analyse automatisée de la base de données dont il dispose. AlphaGo et Watson fonctionnent brillamment sans être conscients de ce qu'ils font.

16. L'incertitude concernant l'authenticité du « jeu » d'AlphaGo et des « réponses » de Watson soulève à coup sûr d'importantes questions philosophiques. Toutefois, plus crucial au niveau éthique est le fait que nous autres humains ne sommes pas accoutumés à ce type d'intelligence. Lorsque nous sommes confrontés à des chefs-d'œuvre artistiques, littéraires ou scientifiques, nous percevons naturellement l'intelligence « consciente » qui se trouve derrière : nous reconnaissons la personnalité exceptionnelle

de Beethoven derrière sa 9e Symphonie, et l'extraordinaire capacité de recherche derrière le théorème d'incomplétude de Gödel. Le simple fait que nous ne devrions pas appliquer ce principe de base familier aux performances brillantes d'une IA de pointe quand nous interagissons avec elle pose des défis sociaux et éthiques majeurs. Étant accoutumés aux interactions émotionnelles et sociales avec des agents au comportement intelligent, nous avons naturellement le même type d'interaction avec ces « IA très performantes sans conscience » que sont les robots dits « affectifs » ou « sociaux », tel qu'un « assistant personnel intelligent » (Alexa, Siri, Assistant Google). Au stade actuel de la technologie, l'IA de haut niveau sans conscience est incapable de répondre correctement aux attentes émotionnelles et sociales complexes des agents humains, alors que son comportement extérieur couplé à l'imagination humaine pourrait susciter en nous l'espoir « irréaliste » d'interagir authentiquement avec elle. Il est important que nous nous souvenions que l'apparente « émotivité » de l'IA est avant tout le fruit de notre imagination. Il y a un consensus général sur le fait que les systèmes artificiellement intelligents ne sont pas conscients au sens de l'expérience humaine, même s'ils sont capables de répondre à des questions concernant le contexte de leurs actions. Il est important de ne pas assimiler expérience et intelligence, bien que certains experts aient suggéré que les récents progrès de l'IA puissent aussi être une raison de réexaminer l'importance de cette expérience ou de cette conscience dans le fait d'être humain. Si l'expérience est au cœur de l'humanité, les considérations éthiques doivent veiller à ce que son rôle ne soit ni écarté ni affaibli, mais au contraire protégé et renforcé par l'utilisation de l'IA. Cela dit, il est possible que notre expérience avec l'IA de haut niveau sans conscience puisse néanmoins avoir une influence sur nos interactions avec les humains ordinaires conscients.

I.2. Comment l'IA fonctionne-t-elle ?

17. Pour pouvoir exécuter les tâches d'un cerveau humain, une machine intelligente doit être capable de percevoir son environnement et de recueillir des données de façon dynamique, de les traiter rapidement et d'apporter des réponses – à partir de son « expérience » passée, de principes préétablis de prise de décisions et de son anticipation de l'avenir. Mais la technologie de l'IA est une TIC classique : elle est basée sur la collecte ou l'acquisition de données, sur leur stockage, leur traitement et leur restitution. Les caractéristiques exceptionnelles des machines cognitives sont dues aux quantités, qui sont transformées en qualités. La technologie de l'IA repose sur les éléments suivants :

(a) *des données dynamiques :* le système doit être exposé à des environnements changeants et à l'ensemble des données acquises par différents capteurs, pour pouvoir les classer et les stocker, et être capable de les traiter rapidement ;

(b) *un traitement rapide :* les machines cognitives doivent réagir rapidement. L'IA doit donc posséder des moyens de calcul et de communication fiables, rapides et puissants ;

(c) *des principes de prise de décisions :* la prise de décisions de l'IA est basée sur des algorithmes d'apprentissage machine. Par conséquent, sa réponse à une instruction donnée dépend de son « expérience », c'est-à-dire, des données auxquelles elle a été exposée. Les algorithmes permettant aux machines cognitives de prendre des décisions sont basés sur quelques principes généraux auxquels obéit l'algorithme et qu'il s'efforce d'optimiser, compte tenu des données qui lui ont été fournies.

La capacité actuelle d'intégrer efficacement l'acquisition de données dynamiques et les algorithmes d'apprentissage automatique pour une prise de décisions rapide permet de créer des « machines cognitives ».

I.3. Qu'est-ce qui distingue l'IA des autres technologies ?

18. La plupart des technologies du 20e siècle reposent sur la modélisation. Autrement dit, les chercheurs étudient la nature et proposent un modèle scientifique pour la décrire, et la technologie progresse à partir de ces modèles. La compréhension de la propagation des ondes électromagnétiques est par exemple à l'origine de la communication sans fil. Par contre, la modélisation du cerveau humain est une entreprise qui semble encore bien loin de nous permettre de créer une machine cognitive à partir d'un modèle. C'est pourquoi l'IA s'appuie sur une autre approche : une approche par les données.

19. L'approche par les données est au cœur de l'*apprentissage automatique*, qui s'appuie généralement sur des « réseaux de neurones artificiels » (RNA). Les RNA sont formés de séries de nœuds de conception similaire aux neurones du cerveau, connectés en couches successives. Les nœuds de la couche d'entrée reçoivent les informations produites par l'environnement, une transformation non linéaire étant ensuite appliquée à chaque nœud. Ces systèmes « apprennent » à exécuter des instructions en étudiant des exemples (données étiquetées), généralement sans qu'on leur ait fourni des règles ou des modèles de tâche précise. L'apprentissage profond, enfin, s'appuie sur plusieurs couches de RNA, qui permettent à la machine de reconnaître des concepts complexes comme un visage ou un corps humain, de comprendre une conversation et de classer toutes sortes d'images.

20. Le principal obstacle, pour que l'IA ait des facultés semblables à celle des humains, est sa scalabilité. Le fonctionnement de ces machines dépend des données auxquelles elles sont exposées, et ne sera optimal que si elles ont un accès illimité aux données pertinentes. Il peut y avoir des limites techniques à cet accès, mais la façon dont les données sont choisies et classées est aussi un problème socioculturel (Crawford, 2017). La classification est une opération culturelle et un produit de l'histoire, et peut biaiser les décisions prises par l'algorithme. En exposant la même machine à des séries variées de données, on peut réduire ce biais, mais non le supprimer (Executive Office of the President, 2016). Il convient donc de souligner que, pour respecter l'article 27 de la Déclaration universelle des droits de l'homme – toute personne a le droit de participer aux bienfaits qui résultent du progrès scientifique – et garantir la diversité des ensembles de données fournis à l'IA, il est important de promouvoir le renforcement des capacités des États, en termes à la fois de compétences humaines et d'infrastructures.

21. La technologie de l'IA a muri sous l'impulsion de firmes multinationales que ne retiennent pas les contraintes locales et nationales. De plus, pour garantir la vélocité de traitement et la fiabilité des systèmes, le traitement informatique est géographiquement dispersé et l'emplacement d'une machine d'IA n'est pas défini par l'endroit où elle fonctionne. En pratique, l'IA s'appuie sur la technologie en nuage, qui permet d'installer les unités de stockage et de traitement à n'importe quel endroit. La technologie de l'IA présente les caractéristiques suivantes :

(a) bien que nombre de ses applications se trouvent dans la sphère publique, la technologie de l'IA est élaborée et dirigée par des sociétés multinationales, qui relèvent pour la plupart du secteur privé et ont moins d'obligations envers le bien public ;

(b) l'IA n'est pas confinée dans une localisation physique, d'où la difficulté de réglementer cette technologie aux niveaux national et international ;

(c) la technologie est fondée sur l'accessibilité aux données à la fois personnelles et publiques ;

(d) les technologies de l'IA ne sont pas neutres, elles sont au contraire intrinsèquement biaisées à cause des données qui les entraînent, et des choix opérés durant cet entraînement ;

(e) les décisions prises par l'IA et les machines cognitives ne peuvent être totalement prévisibles ni explicables. Au lieu de fonctionner mécaniquement ou de façon déterministe, le logiciel d'IA apprend à l'aide de données dynamiques à mesure qu'il se développe et intègre l'expérience du monde réel à sa prise de décisions.

II. CONSIDÉRATIONS ÉTHIQUES

22. L'intelligence artificielle a des implications sociales et culturelles importantes. Comme beaucoup d'autres technologies de l'information, l'IA soulève des questions concernant la liberté d'expression, la vie privée et la surveillance, la propriété des données, la partialité et la discrimination, la manipulation de l'information et la confiance, les relations de pouvoir et l'impact environnemental lié à sa consommation d'énergie. En outre, l'IA crée aussi de nouveaux défis, qui ont trait à son interaction avec les capacités cognitives humaines. Les systèmes basés sur l'IA ont des implications pour la compréhension et l'expertise humaines. Les algorithmes des médias sociaux et des sites d'information peuvent contribuer à la diffusion de fausses informations et avoir des répercussions néfastes sur le sens que nous donnons aux « faits » et à la « vérité », ainsi que sur le débat et l'engagement politiques. L'apprentissage automatique peut inculquer et exacerber les préjugés, entraînant de l'inégalité et de l'exclusion et menaçant la diversité culturelle. L'envergure et le pouvoir que confère la technologie de l'IA accentuent l'asymétrie entre les individus, les groupes et les pays, y compris la « fracture numérique » au sein des pays et entre eux. Cette fracture peut être encore exacerbée par le manque d'accès à des éléments fondamentaux comme les algorithmes d'entraînement et de classement, les données nécessaires pour entraîner et évaluer les algorithmes, les ressources humaines exigées pour coder, installer le logiciel et préparer les données, ainsi que les moyens informatiques nécessaires au stockage et au traitement des données.

23. L'intelligence artificielle exige donc une analyse attentive. Du point de vue de l'UNESCO, les principales questions éthiques posées par l'intelligence artificielle sont ses implications pour la culture et la diversité culturelle, l'éducation, la connaissance scientifique et la communication et l'information. En outre, compte tenu de l'orientation mondiale de l'UNESCO, les thèmes éthiques de portée planétaire que sont la paix, la durabilité, l'égalité des genres et les défis spécifiques de l'Afrique, méritent aussi une attention particulière.

II.1. Éducation

24. L'intelligence artificielle remet en question à bien des égards le rôle de l'éducation dans la société. D'abord, l'IA oblige à repenser le rôle sociétal de l'éducation. Le déplacement de la main d'œuvre causé par certaines formes d'IA impose, entre autres mesures, le recyclage professionnel des employés, ainsi qu'une nouvelle approche de formulation des qualifications visées par les programmes éducatifs. En outre, dans un

monde dominé par l'IA, l'éducation devrait donner aux citoyens les moyens nécessaires pour acquérir de nouvelles formes de pensée critique, y compris leur « sensibilisation aux algorithmes » et la capacité de réfléchir à l'impact de l'IA sur l'information, la connaissance et la prise de décisions. Un deuxième domaine de questionnement éthique concernant l'IA et l'éducation a trait à son rôle dans le processus éducatif lui-même, en tant qu'élément des environnements d'apprentissage numérique, de la robotique éducative et des systèmes d'« analyse de l'apprentissage », qui tous exigent d'être élaborés et mis en œuvre de façon responsable. Enfin, les ingénieurs et les concepteurs de logiciels devraient être éduqués à assurer une conception et une mise en œuvre responsables de l'IA.

II.1.1 Le rôle sociétal de l'éducation

25. Une des principales préoccupations sociétales concernant l'IA est le déplacement de la main d'œuvre. La rapidité des changements induits par l'IA crée des défis sans précédent (Illanes *et al.*, 2018). Il faudra, dans le proche avenir, reconvertir un nombre important de travailleurs, et l'IA aura aussi une profonde incidence sur les carrières que devront emprunter les étudiants. Selon le sondage en panel mené par McKinsey en 2017, « les chefs d'entreprise considèrent de plus en plus comme une priorité urgente l'investissement dans le recyclage et l'amélioration des compétences des personnels existants » (Illanes *et al.*, 2018).

26. L'IA incite donc les sociétés à repenser l'éducation et ses rôles sociaux. L'éducation formelle traditionnelle dispensée par les universités pourrait se révéler insuffisante face à l'essor des économies numérisées et des applications de l'IA. Jusqu'à présent, le modèle éducatif standard visait généralement l'acquisition d'un ensemble de « connaissances clés » (Oppenheimer, 2018) et mettait l'accent sur des apprentissages formels comme la lecture, l'écriture et le calcul. Au XXI^{e} siècle, l'information et la connaissance sont devenus omniprésents, exigeant non seulement une « maîtrise des données » permettant aux élèves de lire, d'analyser et de gérer efficacement cette information, mais aussi une « maîtrise de l'IA » autorisant une réflexion critique sur la part prise par les systèmes informatiques intelligents dans la reconnaissance des besoins d'information et la sélection, l'interprétation, le stockage et la représentation des données.

27. De plus, dans un marché du travail en constante évolution, le système éducatif ne peut plus viser à former les individus pour une profession particulière. L'éducation devrait leur permettre d'être adaptables et résilients, préparés pour un monde dans lequel les technologies créent un marché de la main d'œuvre dynamique et où les employés doivent se rescolariser en permanence. Les idées actuelles concernant « l'apprentissage tout au long de la vie » devront peut-être être étendues à un modèle d'éducation continue, comprenant l'élaboration d'autres types de cycles d'enseignement et de diplômes.

II.1.2 L'IA dans l'enseignement et l'apprentissage

28. En procurant des cours gratuits et de qualité et d'autres ressources d'enseignement par le biais de l'Internet, les ressources éducatives libres (REL) ont été un ajout important au paysage de l'apprentissage. Les REL offrent un moyen inégalé d'améliorer l'éducation à travers le monde, mais ce potentiel est encore loin d'être pleinement réalisé, comme l'attestent les faibles taux d'achèvement des cours en ligne ouverts à tous (MOOC). L'ampleur et la diversité des ressources disponibles ont créé deux problèmes. Il y a d'abord la difficulté de trouver la bonne ressource, que ce soit pour l'apprenant individuel ou pour un enseignant désireux de la réutiliser comme support d'enseignement. Il en découle une seconde difficulté, qui est la réduction de la diversité, quand certaines

ressources l'emportent en popularité sur des contenus plus pertinents, mais plus difficiles d'accès.

29. Un exemple en est le projet « X5GON » d'Horizon 2020 (visant à créer un réseau mondial de REL par-delà les différences de modalités, de cultures, de langues, de thèmes et de sites : https://www.x5gon.org/). Ce projet, financé par l'Union européenne, élabore des méthodes d'intelligence artificielle permettant aux apprenants et aux enseignants de repérer les ressources dont ils ont besoin pour atteindre leurs objectifs d'apprentissage, en fonction de leur situation particulière. Ainsi, on pourra orienter un enseignant africain vers des exposés faisant appel aux savoirs locaux et autochtones et adaptés au contexte culturel et local qui est le sien, mais aussi permettre à un apprenant d'une autre partie du monde curieux de comprendre les défis spécifiques de l'Afrique de trouver un contenu africain pertinent pouvant être traduit de la langue locale.

30. L'IA peut ainsi potentiellement s'attaquer aux deux difficultés citées plus haut. Elle surmonte la première en aidant à identifier les ressources les mieux à même de répondre aux besoins de l'apprenant ou de l'enseignant grâce à une modélisation de leurs intérêts et de leurs objectifs, tout en exploitant une représentation enrichie des vastes réservoirs de REL disponibles dans le monde. En adaptant les recommandations aux besoins individuels de l'apprenant ou de l'enseignant, elle surmonte aussi la seconde, car ces recommandations ne renverront plus par défaut à la ressource la plus populaire sur un sujet donné. Et elle a aussi la possibilité de relier entre eux les apprenants de différentes cultures, ce qui améliorera le partage interculturel des idées et donc favorisera la compréhension et le respect mutuels.

II.1.3 Éduquer les ingénieurs de l'IA

31. Le développement des technologies du futur appartient aux experts techniques. Traditionnellement, les ingénieurs sont formés pour élaborer des produits dotés de performances optimales pour un minimum de ressources (énergie, spectre, encombrement, poids, etc.), en fonction de contraintes extérieures données. Au cours des dernières décennies, l'éthique des technologies a développé différents moyens de faire une place à la réflexion éthique, à la responsabilité et au raisonnement dans le processus de conception. S'agissant de l'IA, l'expression « conception conforme à l'éthique » (EAD, ethically aligned design) a été forgée pour désigner un processus de conception incorporant explicitement des valeurs humaines (IEEE, 2018).

32. Il est capital d'appliquer l'EAD à l'IA ainsi qu'à d'autres systèmes autonomes et intelligents, car cela permet de poser les questions éthiques à un moment où cette technologie peut encore être adaptée. Le principe de « *privacy by design* », ou de respect de la vie privée dès la conception, en est un bon exemple. On violera moins la vie privée si on ne stocke pas toutes les données, mais seulement celles qui sont nécessaires pour une tâche spécifique. C'est le cas du comptage de foule, autrement dit du comptage d'individus dans une foule à partir de clichés photographiques. En ce cas, si le cliché est prétraité de façon à n'extraire que le contour des silhouettes, les personnes ne pourront pas être reconnues et l'algorithme de comptage fonctionnera efficacement sans violer la vie privée. De même, les concepteurs de l'IA peuvent examiner d'autres problèmes éthiques tels que la prévention des biais algorithmiques ou la traçabilité, afin de limiter les possibilités d'usage abusif de la technologie, et d'améliorer l'explicabilité des décisions algorithmiques.

33. Actuellement, la formation des ingénieurs dans le monde repose pour l'essentiel sur des enseignements scientifiques et technologiques sans lien intrinsèque avec une analyse des valeurs humaines expressément vouées à améliorer le bien-être humain et

la qualité de l'environnement. Il est essentiel de changer cela pour former les futurs ingénieurs et informaticiens à une conception éthique des systèmes d'IA. Cela exige une prise de conscience explicite des implications et des conséquences sociétales et éthiques potentielles de la technologie en développement, et des risques d'utilisation abusive. L'IEEE (organisation mondiale regroupant plus de 400 000 ingénieurs électriques) sensibilise déjà à cette question dans le cadre de son initiative mondiale sur l'éthique des systèmes autonomes et intelligents (https://ethicsinaction.ieee.org/). Cette question consiste aussi à s'engager activement en faveur de l'intégration de la question du genre ainsi que de la diversité sociale et culturelle des ingénieurs, et à promouvoir une prise en compte holistique des implications sociétales et éthiques de la conception d'un système d'IA. Les occasions de dialogue entre les ingénieurs et le public devraient être encouragées pour favoriser la communication concernant les besoins et points de vue de la société et le déroulement quotidien du travail et des activités de recherche des ingénieurs.

II.2. L'IA et la connaissance scientifique

34. Dans le domaine des pratiques scientifiques, l'IA devrait avoir des implications profondes. Dans les sciences naturelles et sociales, tout comme les sciences du vivant et les sciences environnementales, elle remet fondamentalement en question nos conceptions de la compréhension et de l'explication scientifiques. Cela rejaillit également sur la manière dont nous appliquons la connaissance scientifique dans les contextes sociaux.

II.2.1 L'IA et l'explication scientifique

35. Du fait de la puissance croissante des formes d'apprentissage automatique et d'apprentissage profond, l'IA bouscule l'idée que nous nous faisons d'une explication scientifique satisfaisante et de ce que nous pouvons naturellement attendre d'une théorie scientifique efficace. Selon la conception traditionnelle de la science – ce qu'on a appelé le modèle « déductif-nomologique » – une bonne explication scientifique est capable de prévoir avec exactitude un phénomène donné en s'appuyant sur des lois, des théories et des observations scientifiques. Par exemple, on peut légitimement dire qu'on ne peut expliquer comment la Lune tourne autour de la Terre selon la mécanique newtonienne que lorsque nous sommes capables de l'appliquer de façon déductive pour prévoir l'orbite lunaire. Ces prévisions s'appuient habituellement sur la compréhension causale, ou sur une compréhension unifiante de phénomènes apparemment disparates.

36. L'IA, par contre, peut fidèlement produire des prévisions d'une exactitude impressionnante à partir d'ensembles de données sans nous en fournir aucune explication causale ou unifiante. Ses algorithmes n'utilisent pas les mêmes concepts sémantiques que les humains pour parvenir à la compréhension scientifique d'un phénomène. Cet écart entre des prévisions efficaces d'une part, et une compréhension scientifique satisfaisante d'autre part, jouera certainement un rôle clé dans les pratiques scientifiques, ainsi que dans les prises de décisions basées sur l'IA.

37. Cela pourrait avoir des répercussions sur la confiance qu'inspire la science, qui s'est toujours appuyée sur une méthode scientifique expliquant les différents phénomènes de manière systématique et transparente, en formulant ses prévisions de manière rationnelle et fondée sur des faits. L'efficacité apparente avec laquelle les algorithmes d'apprentissage automatique parviennent à des résultats comparables sans s'appuyer sur un tel modèle justifié scientifiquement pourrait avoir des répercussions négatives sur la perception et l'évaluation publiques de la science et de la recherche scientifique.

38. De plus, la recherche montre que la qualité de l'apprentissage automatique est fortement dépendante des données disponibles utilisées pour entraîner les algorithmes. Mais dans la mesure où la plupart des applications de l'IA sont élaborées par des entreprises privées, la transparence concernant ces données n'est pas toujours suffisante, contrairement à la méthode scientifique traditionnelle qui garantit la validité des résultats en exigeant leur reproductibilité, autrement dit la possibilité de les reproduire en répétant les mêmes expériences.

II.2.2 L'IA, les sciences du vivant et la santé

39. Dans les sciences du vivant et en particulier en médecine, l'évolution des technologies de l'IA a significativement transformé le paysage de la santé et de la bioéthique au fil des ans. Elles peuvent avoir des effets positifs, tels qu'une plus grande précision de la chirurgie robot-assistée, ou une meilleure prise en charge des enfants autistes, mais soulèvent des questions éthiques, telles que leur coût dans un contexte de rareté des ressources dont dispose le système de santé, ou la transparence dont elles devraient s'accompagner pour que soit respectée l'autonomie des patients.

40. Du point de vue individuel, l'IA procure au public non averti une nouvelle façon d'aborder la santé et les questions médicales. La consultation des sites Internet et la multiplication des applications d'autodiagnostic sur le téléphone mobile permettent à tout un chacun de poser un diagnostic sans recourir à un professionnel de santé. Cela pourrait avoir des implications pour l'autorité médicale et l'acceptation de l'automédication, laquelle présente aussi des dangers. Cela modifie aussi la relation entre le médecin et son patient, d'où la nécessité d'instaurer une forme de réglementation, sans pour autant nuire à l'innovation et à l'autonomie.

41. Les technologies de l'IA pourraient libérer du temps que les personnels de santé pourraient consacrer à leurs patients, par exemple en facilitant la saisie de données et les tâches administratives, mais elles pourraient aussi remplacer les éléments holistiques et humains des soins. La technologie bien connue Watson for Oncology d'IBM a révolutionné le traitement du cancer, mais soulève aussi d'importantes questions concernant le caractère et les attentes de l'expertise et de la formation médicales, ainsi que les responsabilités des médecins utilisateurs. Des questions similaires se posent face au développement des « chatbots » ou assistants virtuels pour personnes en demande de soutien et de conseils psychologiques, des applications de détection précoce des épisodes de troubles psychiatriques, ou des systèmes d'IA utilisés pour poser un diagnostic psychiatrique à partir d'informations personnelles collectées sur les réseaux sociaux et l'Internet – ce qui a aussi, évidemment, d'importantes implications pour le respect de la vie privée. En outre, s'agissant des personnes âgées, on voit apparaître des technologies à base d'IA comme les robots sociaux, qui peuvent être utiles du point de vue médical pour des patients atteints de démence, par exemple, mais soulèvent aussi le problème de la réduction des soins humains et de l'isolement social qui en découle.

42. L'IA confère aussi une nouvelle dimension au débat en cours opposant « amélioration de l'être humain » et « thérapie ». On assiste à des tentatives d'intégration de l'IA dans le cerveau humain au moyen d'une « interface neuronale » : un treillis qui grandirait avec le cerveau, servirait d'interface directe cerveau-ordinateur et circulerait dans les veines et les artères de l'organisme hôte (Hinchliffe, 2018). Cette évolution technologique a des implications importantes sur la question de la nature humaine et du fonctionnement « normal » de l'être humain.

II.2.3 L'IA et les sciences de l'environnement

43. L'IA peut apporter sa contribution aux sciences de l'environnement à travers une série d'applications. Elle peut être utilisée pour traiter et interpréter des données intéressant l'écologie, la biologie des systèmes, la bioinformatique et la recherche spatiale et climatique, améliorant ainsi la compréhension scientifique des processus et des mécanismes. L'amélioration du recyclage, la gestion et la réhabilitation environnementales et une consommation d'énergie plus efficace peuvent avoir des effets positifs directs sur l'environnement. Dans l'agriculture, l'IA peut apporter une amélioration de la production agricole (grâce, par exemple, à l'automatisation de la fertilisation et de l'irrigation) et du bien-être animal, et réduire les risques induits par les maladies, les organismes nuisibles et les mauvaises conditions météorologiques. D'un autre côté, l'IA pourrait amener les humains à percevoir la nature différemment, dans le bon sens, en lui faisant prendre conscience de sa beauté ou de son autonomie, ou dans le mauvais sens, en accentuant l'« instrumentalisation » de la nature ou le clivage entre les êtres humains et les animaux ou l'environnement.

44. Pour toutes ses applications, les avantages potentiels doivent être mis en balance avec l'impact environnemental du cycle complet de production de l'IA et des technologies de l'information (TI). Cela va de l'extraction des terres rares et autres matières premières à l'énergie nécessaire pour produire et faire fonctionner les machines et aux déchets produits durant leur production et à la fin de leurs cycles de vie. Davantage d'IA a de fortes chances d'ajouter aux inquiétudes grandissantes suscitées par les volumes croissants de déchets électroniques et la pression sur les terres rares générée par l'industrie informatique. Outre leurs impacts sur l'environnement et la santé, les déchets électroniques ont d'importantes répercussions sociopolitiques, liées notamment à leur exportation vers les pays en développement et les populations vulnérables (Heacock *et al.*, 2015).

45. La gestion des risques de catastrophes est un domaine où l'IA peut apporter sa contribution en matière de prévention et de réaction face aux catastrophes environnementales comme les tsunamis, les séismes, les tornades et les ouragans. Un exemple concret en est le géoserveur G-WADI (Information sur l'eau et le développement) de l'UNESCO, dont l'objectif est d'informer la planification et la gestion d'urgence des risques hydrologiques, comme les inondations, les sécheresses et les événements météorologiques extrêmes. Son système de soutien PERSIANN (*Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks*) est un algorithme d'extraction de précipitations à partir d'images satellite, qui fournit des informations en temps quasi réel. L'algorithme de son système de classification des nuages (accessible sur : http://hydis.eng.uci.edu/) a été optimisé pour observer les précipitations extrêmes, notamment à très haute résolution spatiale, et il est largement utilisé dans le monde pour suivre la trajectoire des tempêtes. Il offre aussi une application mobile iRain (http://fr.unesco.org/news/irain-new-mobile-app-promote-citizen-science-and-support-water-management) pour laquelle la production participative permet d'associer des scientifiques amateurs à la collecte de données.

46. Il est intéressant de noter que des entreprises privées ont elles-mêmes récemment contribué à la gestion des catastrophes. Un exemple en est le projet de prévision des inondations assisté par l'IA de Google (https://www.blog.google/products/search/helping-keep-people-safe-ai-enabled-flood-forecasting/). À cet égard, le développement de technologies de l'IA potentiellement utiles à la gestion des catastrophes devrait être encouragé.

II.2.4 L'IA et les sciences sociales

47. Généralement parlant, la recherche en sciences sociales a pour objet d'identifier la structure causale des interactions personnelles et sociales. Comme la plupart des phénomènes sociaux sont influencés à plusieurs titres par une série de facteurs causaux, les chercheurs en sciences sociales s'appuient traditionnellement sur l'analyse statistique des données empiriques pertinentes pour déterminer les principaux facteurs causaux et l'ampleur de leurs effets. Ce faisant, il leur est essentiel de distinguer les corrélations purement statistiques des relations de causalité réelles. Il est clair que l'IA pourrait aider les chercheurs en sciences sociales à s'orienter parmi les énormes masses de données disponibles afin de dégager des mécanismes causaux plausibles, et de vérifier leur validité. Mais l'IA pourrait aussi « surajuster » les données, et proposer des « pseudo »-relations de causalité totalement inexistantes. Cette possibilité pourrait susciter des controverses sociales, notamment lorsque les relations causales proposées sont éthiquement sensibles, comme les suggestions de différences intellectuelles entre les races. Là encore, nous ne devrions pas accepter automatiquement les « conclusions » de l'IA sans une évaluation humaine.

II.2.5 La prise de décisions basée sur l'IA

48. Les techniques d'IA peuvent avoir un impact considérable dans un large éventail de domaines, qui vont des professions juridiques et des systèmes judiciaires au soutien à la prise de décisions des organes publics législatifs et administratifs. Elles peuvent par exemple aider les avocats à améliorer leur efficacité et leur précision dans le domaine du conseil juridique et du contentieux, ce qui serait bénéfique aux avocats, à leurs clients et à l'ensemble de la société. Les logiciels actuellement à la disposition des magistrats peuvent être complétés et améliorés par des outils d'IA qui les aident dans leurs décisions (CEPEJ, 2018).

49. Une question essentielle concernant ce type d'usages porte sur la nature et l'interprétation des résultats des algorithmes, qui ne sont pas toujours intelligibles pour les humains[1]. Cette question peut être étendue au plus large domaine de la prise de décisions fondée sur les données. Capable d'analyser, de traiter et de catégoriser de vastes quantités de données de nature très variée et susceptibles d'évoluer rapidement, une intelligence artificielle est considérée comme capable de proposer – et, si elle y est autorisée, de prendre – des décisions dans des situations complexes. Les exemples d'usages examinés dans le présent rapport couvrent la gestion environnementale, la prévision et la gestion des catastrophes, l'anticipation des troubles sociaux et la planification d'opérations militaires.

50. La validité d'une décision assistée par l'IA n'en devrait pas moins être considérée avec prudence. Une telle décision n'est pas nécessairement juste, équitable, exacte ou appropriée. Elle est exposée aux inexactitudes, à des résultats discriminatoires, à des biais acquis ou introduits et aux limites du processus d'apprentissage. Un être humain a

---

[1] Comme l'indique K.D. Ashley : « Dans la mesure où un algorithme d'apprentissage automatique apprend des règles à partir de régularités statistiques pouvant surprendre les humains, ces règles ne leur sembleront pas nécessairement raisonnables. […] Bien que les règles induites par la machine puissent conduire à des prévisions fiables, elles ne se réfèrent pas à l'expertise humaine et peuvent ne pas être intelligibles pour les humains comme le seraient des règles construites manuellement par un expert. Les règles […] déduites par l'algorithme n'étant pas nécessairement le reflet d'une connaissance ou d'une expertise légale explicite, elles peuvent ne pas répondre aux critères du raisonnable pour un expert humain » (Ashley, 2017, p. 111).

non seulement une « vision du monde » bien plus large, mais il ou elle a aussi une connaissance tacite supérieure à celle de l'IA dans les situations critiques et complexes, comme les décisions à prendre sur un théâtre d'opérations. Idéalement, la décision serait celle qu'un humain prendrait si il ou elle avait pu traiter la masse considérable de données dans des délais raisonnables. Mais les humains ont d'autres capacités et prennent leurs décisions selon des architectures de prise de décisions fondamentalement différentes, y compris une sensibilité aux biais potentiels.

51. Il est peu probable que l'IA ait un jour – au moins dans le proche avenir – la capacité de gérer des données ambiguës et évoluant rapidement, ou d'interpréter et d'exécuter ce qu'auraient été les intentions humaines si l'homme avait pu gérer des données complexes et diversifiées. Même lorsqu'un être humain « au courant » est là pour modérer une décision machine, cela peut ne pas être suffisant pour produire une « bonne » décision : dans la mesure où l'IA cognitive ne prend pas ses décisions de la même façon que l'être humain, celui-ci ne pourrait pas disposer des connaissances et des informations nécessaires pour décider si les actions guidées par les données répondent aux intentions humaines. En outre, le comportement stochastique de l'IA cognitive, ajouté à l'incapacité consécutive de l'être humain de savoir pourquoi ce choix particulier a été fait par le système, signifie que ce choix aura moins de chances d'inspirer la confiance.

52. Une illustration, incitant à la prudence, de certains des problèmes posés par l'utilisation de l'IA pour assister la prise de décisions dans les contextes sociaux est fournie par l'Allegheny Family Screening Tool (AFST), un modèle prédictif utilisé pour prévoir les actes de négligence et de maltraitance des enfants à Allegheny, en Pennsylvanie (voir https://www.alleghenycountyanalytics.us/wp-content/uploads/2017/07/AFST-Frequently-Asked-Questions.pdf). Cet outil a été adopté avec la certitude que les données permettraient une prise de décisions objective et impartiale qui viendrait au secours d'administrations publiques aux ressources limitées. Le service qui l'a mis en œuvre était sans doute bien intentionné. Mais des recherches récentes ont montré que l'AFST a des conséquences néfastes pour la population qu'il espérait servir (Eubanks, 2018b, p. 190 ; Eubanks, 2018a). Il suréchantillonne les pauvres et utilise des extrapolations pour comprendre et prévoir les actes de maltraitance infantile d'une façon qui ne peut que désavantager les familles laborieuses pauvres. La discrimination structurelle déjà subie par les pauvres s'en trouve exacerbée, avec des effets disproportionnellement négatifs sur des communautés vulnérables.

53. Dans certains contextes, demander à l'IA de décider (que ce soit de manière totalement autonome ou humainement assistée) peut même être considéré comme un pacte avec le diable : pour profiter de la vitesse et des immenses capacités d'ingestion et de catégorisation de données de l'IA, il va en effet nous falloir renoncer à la possibilité d'influencer ces décisions. Elles peuvent aussi avoir de graves conséquences, notamment dans les situations de conflit.

II.3. Culture et diversité culturelle

54. L'IA devrait avoir des implications substantielles dans le domaine de la culture et des expressions artistiques. Bien qu'on n'en soit qu'aux balbutiements, on voit se profiler les premiers exemples de collaboration artistique entre des algorithmes intelligents et la créativité humaine, ce qui pourrait poser un jour de sérieux problèmes pour les droits des artistes, les industries culturelles et créatives (ICC) et l'avenir du patrimoine. En même temps, le rôle des algorithmes dans les médias en ligne diffusés en continu et dans la

traduction automatique aura certainement des implications pour la diversité culturelle et les langues.

II.3.1 Créativité

55. L'intelligence artificielle est de plus en plus connectée à la créativité et aux pratiques artistiques humaines, du logiciel d'« auto-accord » qui corrige automatiquement les fausses notes des chanteurs, aux algorithmes qui aident à créer des œuvres d'art visuelles, composent de la musique ou écrivent des romans et des poèmes. La créativité, comprise comme la capacité de produire des contenus nouveaux et originaux en faisant appel à l'imagination ou à l'invention, joue un rôle central dans les sociétés ouvertes, inclusives et pluralistes. Pour cette raison, l'impact de l'IA sur la créativité humaine mérite un examen attentif. Si l'IA s'avère un puissant outil de création, elle soulève d'importantes questions sur l'avenir de l'art, les droits et la rémunération des artistes et l'intégrité de la chaîne de valeurs de la création.

56. L'exemple du « nouveau Rembrandt » – la réalisation d'une nouvelle toile du maître par une intelligence artificielle et une imprimante 3D – en est une bonne illustration (Microsoft Europe, 2016). Des œuvres d'art de ce type obligent à redéfinir ce qu'est un « auteur », pour pouvoir rendre justice au travail de création à la fois du « véritable » auteur et des algorithmes et technologies qui ont produit l'œuvre d'art elle-même. Cela soulève une autre question : que faire lorsque l'IA est capable de créer elle-même une œuvre d'art ? Lorsqu'un auteur humain est remplacé par des machines et des algorithmes, peut-on envisager d'attribuer des droits d'auteur ? Un algorithme peut-il et devrait-il être reconnu comme un auteur, et jouir des mêmes droits qu'un artiste ?

57. Bien que l'IA soit clairement capable de produire des œuvres « originales », c'est toujours à des êtres humains que l'on doit la mise au point des technologies de l'IA et des algorithmes, et souvent aussi la création des œuvres dont s'inspire l'art produit par l'IA. De ce point de vue, on peut considérer l'IA comme une nouvelle technique artistique, donnant lieu à une nouvelle forme d'art. Si nous voulons préserver l'idée que les créations de l'IA ont un auteur, il faut procéder à l'analyse des différents auteurs qui se trouvent « derrière » chaque œuvre d'art, et des relations qu'ils ont entre eux. En conséquence, nous devons élaborer de nouveaux cadres pour distinguer la piraterie et le plagiat de l'originalité et de la créativité et reconnaître la valeur du travail créatif humain dans nos interactions avec l'IA. Nous avons besoin de ces cadres pour éviter l'exploitation délibérée du travail et de la créativité des êtres vivants, et garantir une rémunération et une reconnaissance adéquates des artistes, l'intégrité de la chaîne de valeurs culturelle, et la capacité du secteur de la culture à fournir des emplois décents.

II.3.2 Diversité culturelle

58. L'IA est également étroitement liée à la diversité culturelle. Bien qu'elle soit susceptible d'avoir des effets positifs sur les industries culturelles et créatives, les artistes et les entrepreneurs culturels ne disposent pas tous des compétences et des moyens pour utiliser les technologies basées sur l'IA dans la création et la distribution de leurs œuvres. La logique commerciale des grandes plates-formes pourrait aboutir à une concentration accrue de l'offre, des données et des revenus de la culture entre les mains d'un petit nombre d'acteurs, avec des répercussions potentiellement négatives sur la diversité des expressions culturelles plus généralement, y compris le risque de créer une nouvelle fracture de la création, et une marginalisation croissante des pays en développement.

59. À mesure que ces plates-formes deviennent le principal moyen d'accéder aux œuvres d'art, il est crucial de veiller à leur diversité et à un accès équitable des artistes de tous genres et de toutes origines à ces plates-formes. Dans ce contexte, les artistes des pays en développement méritent une attention particulière. Les artistes et les entrepreneurs culturels devraient avoir accès à la formation, aux moyens financiers, aux infrastructures et aux équipements nécessaires pour participer à cette nouvelle sphère et à ce nouveau marché culturels.

60. En outre, les algorithmes utilisés par les sociétés de diffusion médiatique en continu, comme Spotify et Netflix, ont une influence majeure sur le choix de musique et de films auxquels le public a accès. Dans la mesure où ces plates-formes non seulement mettent les œuvres d'art à disposition, mais où elles *suggèrent* aussi à quelles œuvres leurs utilisateurs peuvent accéder, il est important que leurs algorithmes soient conçus de façon à ne pas privilégier certaines œuvres au détriment d'autres en limitant leurs suggestions aux œuvres d'un genre particulier les plus en vue, ou aux choix les plus populaires de leurs utilisateurs et de l'entourage de ces derniers. D'autres institutions ont exprimé des préoccupations similaires (ARCEP, 2018). La transparence et la responsabilité de ces algorithmes sont essentielles pour assurer l'accès à des expressions culturelles diverses et une participation active à la vie culturelle.

61. En ce qui concerne le patrimoine culturel, l'IA peut également jouer un rôle important. L'IA peut être utilisée, par exemple, pour surveiller et analyser les changements subis par les sites du patrimoine, en lien avec la pression du développement, le changement climatique, les catastrophes naturelles et les conflits armés. Elle peut aussi être utilisée pour surveiller le trafic illicite d'objets culturels et la destruction de biens culturels et appuyer la collecte de données nécessaires aux efforts de réhabilitation et de reconstruction.

II.3.3 Langues

62. Dans un monde qui se mondialise à vive allure, la traduction automatisée des langues est appelée à jouer un rôle croissant. L'IA aura par suite un impact substantiel sur les langues et l'expression humaine, dans toutes les dimensions de la vie. Cette réalité oblige à adopter une attitude prudente vis-à-vis des langues « naturelles » (par opposition aux langues artificielles et au code informatique) et de leur diversité. Car la langue, après tout, est le fondement de l'identité humaine, de la cohésion sociale, de l'éducation et du développement humain. Depuis sa création, l'UNESCO reconnaît l'importance des langues pour promouvoir l'accès à une éducation de qualité, bâtir des sociétés du savoir inclusives et transmettre le patrimoine et les expressions culturelles (UNESCO, 2002).

63. Un élément central de la relation complexe entre l'IA et les langues est le rôle intermédiaire des « langages formels » (c'est-à-dire des langages formés de mots composés au moyen d'un alphabet). Les technologies de l'IA ont souvent besoin que les mots et les phrases exprimés dans une des nombreuses langues naturelles parlées dans le monde soient traduits dans des langages formels pouvant être traités par les ordinateurs. La traduction de nombreuses langues naturelles en langage formel n'est pas un processus neutre, toute traduction d'une langue naturelle en langage formel entraînant une « perte » de sens, car il est impossible de formaliser entièrement toutes les spécificités et les idiosyncrasies des langues.

64. Un deuxième élément est la traduction entre langues naturelles, qui se fait par le biais de ces langages formels. Il y a plusieurs difficultés inhérentes à la traduction automatique ; les mots peuvent avoir des sens différents dans les différentes langues, et il peut ne pas exister de correspondance linguistique ou conceptuelle entre les langues.

En ce cas, la traduction est ardue, sinon techniquement impossible. En outre, les connotations contextuelles et culturelles des mots et expressions ne sont pas toujours pleinement traduisibles. Bien qu'elle se soit beaucoup améliorée ces dernières années, au moins dans les langues les plus parlées, la traduction automatique est souvent trop incertaine pour être utilisée, par exemple, dans des domaines techniques où la précision lexicale et conceptuelle est cruciale, ou dans l'expression culturelle et la littérature.

65. Ces deux aspects de la traduction automatique ont d'importantes implications, non seulement en ce qui concerne la qualité de la traduction et le risque de confusion entre les langues, mais aussi la diversité linguistique. Il est hautement probable que la traduction automatique, au moins dans un premier temps, sera d'abord développée pour les principales langues mondiales, notamment l'anglais. Cette technologie exige de grands ensembles de données compilées dans les traductions effectuées par l'homme. De tels ensembles de données sont souvent indisponibles en quantités suffisantes pour des langues moins parlées. Dans le même temps, cette technologie peut aussi jouer un rôle positif, en permettant aux individus de s'exprimer dans des langues moins répandues.

66. Un processus analogue est déjà en cours à la radio. Si les chaînes de radio commerciales produisent surtout du contenu dans les langues les plus parlées, renforçant ainsi les cultures incarnées par les langues dominantes, les diffuseurs communautaires génèrent souvent du contenu dans les langues locales, favorisant ainsi le pluralisme et la diversité des médias. Comme le souligne le manuel de l'UNESCO consacré aux médias communautaires : « [Les médias communautaires] sont présents dans toutes les régions du monde, car les mouvements sociaux et les organisations communautaires ont cherché un moyen d'exprimer dans leur langue leurs interrogations, leurs préoccupations et leurs cultures » (UNESCO, 2013, p. 7). Les médias de masse peuvent donc effectivement contribuer à la préservation des langues et de la diversité culturelle.

67. De même, la traduction automatique est déjà utilisée comme un instrument de promotion de la diversité et de protection des langues autochtones. Ainsi, en Australie, un chercheur du *Centre d'excellence pour la dynamique des langues* de l'ARC a enregistré près de 50 000 heures d'expression orale. Pour traiter ces enregistrements, les linguistes devaient sélectionner de courts fragments pouvant contenir des éléments grammaticaux et lexicaux clés, et à cette fin écouter les enregistrements et les transcrire. Sans l'IA, il leur aurait fallu pour cela 2 millions d'heures. Cet usage de l'IA leur a déjà permis de modéliser 12 langues autochtones parlées en Australie, dont le *kunwok,* le *kriol,* le *mangarayi,* le *nakkara,* le *pitjantjatjara,* le *warlpiri* et le *wubuy* (O'Brien, 2018).

68. Ces exemples prouvent que l'IA, comme toute technologie, devrait être développée et utilisée de manière à ne pas menacer la diversité culturelle mais au contraire à la préserver. Si nous voulons préserver le multilinguisme et l'interopérabilité des langues, il faut prévoir des moyens techniques et financiers adéquats pour ce faire (Palfrey et Gasser, 2012 ; Santosuosso et Malerba, 2015).

II.4. Communication et information

69. L'intelligence artificielle joue un rôle croissant dans le traitement, la structuration et la transmission de l'information. Le journalisme automatisé et la diffusion algorithmique de nouvelles sur les médias sociaux ne sont que des exemples parmi d'autres de cette évolution, qui pose des questions concernant l'accès à l'information, la désinformation, la discrimination, la liberté d'expression, le respect de la vie privée et la formation aux médias et à l'information. Dans le même temps, une attention doit être prêtée aux nouvelles fractures numériques entre les pays et au sein des différents groupes sociaux.

II.4.1 Désinformation

70. L'IA peut faciliter la libre circulation de l'information et l'activité journalistique, mais elle peut également servir à la diffusion de fausses informations, parfois qualifiées du terme anglais de « fake news » (bobards). Les exemples récents, comme l'affaire Cambridge Analytica, ont montré que des algorithmes conçus pour éviter les partis pris politiques humains, au moment de décider quel contenu sera mis en avant sur les médias sociaux, peuvent être utilisés pour favoriser délibérément la diffusion, auprès de certains groupes cibles, de contenus forgés, manipulateurs et porteurs de discorde. Dans certains cas, ces contenus peuvent inclure des informations frauduleusement présentées comme des nouvelles, ainsi que des contenus utilisés comme propagande à caractère émotionnel.

71. Cela peut avoir des effets néfastes sur les normes de la discussion courtoise et éclairée, sur la confiance sociale et le débat public, et même sur l'exercice de la démocratie. L'existence d'opinions différentes et parfois polarisées est un trait habituel des sociétés ouvertes et démocratiques qui offrent un espace public libre et ouvert. Or, les algorithmes des médias sociaux peuvent exacerber la polarisation des opinions en intensifiant et amplifiant les contenus émotionnels grâce aux boutons « j'aime », « partager », « retweeter », au complètement automatique des champs de recherche et à d'autres formes de recommandations et d'incitations, aboutissant aux bien nommées « bulles de filtrage » ou « chambres d'écho », au lieu de fournir une infrastructure de discussion et de débat. Les individus partageant la même « bulle » peuvent être exposés à des contenus d'information filtrés, et, en retour, l'espace public ouvert peut devenir caractérisé par des groupes d'opinion de plus en plus homogénéisés, qui sont en même temps de plus en plus polarisés les uns par rapport aux autres.

72. Bien que certaines grosses entreprises de médias sociaux commencent à reconnaître le problème et la nécessité d'y remédier par une approche multiparties prenantes, associant à la fois la société civile et les autorités régulatrices des États, les solutions semblent encore incertaines. Un des moyens de les explorer est d'utiliser le cadre ROAM de l'UNESCO (Respect des droits humains, Ouverture, Accessibilité pour tous, Multiples acteurs) pour identifier de façon systématique où des améliorations peuvent être apportées et quels liens elles ont avec l'ensemble des principes en jeu.

73. Parfois, la modération du contenu peut se justifier, précisément comme un moyen d'éviter la diffusion de fausses nouvelles et de contenus incitant à la violence, à la haine et à la discrimination, et de prévenir une communication personnelle agressive. Le filtrage peut être effectué par des humains, mais il est souvent assisté, voire même automatisé, grâce à des algorithmes d'IA. La difficulté en ce cas est de ne pas se contenter d'identifier le contenu incriminé, mais d'éviter aussi que le filtre soit trop inclusif, ce qui l'exposerait à des accusations de censure et de restriction automatisées appliquées à une expression légitime. La réponse à la désinformation et au « discours de haine » devrait s'appuyer sur les normes internationales de la liberté d'expression et être conformes aux conventions et déclarations des Nations Unies dans ce domaine (Article 19, 2018a).

II.4.2 Journalisme de données et journalisme automatisé

74. L'apparition récente d'une IA aux fonctionnalités puissantes a des implications pour le journalisme, et ce, de plusieurs manières. D'une part, la possibilité croissante d'utiliser les données et les outils informatiques dans la recherche journalistique peut favoriser le travail des journalistes. De l'autre, l'IA pourrait aussi s'arroger certaines de leurs tâches. Les tâches routinières, pour lesquelles on dispose de quantités de données « issues de la pratique », sont les premières à pouvoir être imitées par l'IA, et la routine constitue

effectivement une part substantielle du travail journalistique : recueillir et sélectionner les données pertinentes, résumer les résultats et les décrire dans des termes clairs. L'IA accomplit déjà des travaux de rédaction relativement simples, à format déterminé, dans des domaines qui doivent être constamment tenus à jour, comme l'actualité des marchés ou les résultats sportifs. Cette avancée ne présente pas que des mauvais côtés : elle peut aussi libérer les journalistes pour leur permettre de se consacrer à des travaux supérieurs d'interprétation, d'analyse, de vérification et de présentation de l'information.

75. La rédaction automatisée des nouvelles sans intervention ou supervision humaine est une réalité qui est souvent cachée au lecteur. Dès 2006, certains services d'information (par exemple Thomson Financial) ont annoncé l'utilisation d'ordinateurs pour produire des articles basés sur des données, en vue de transmettre rapidement des informations à leurs utilisateurs. En 2014, Wibbitz (Israël) a obtenu le Grand prix du Forum UNESCO/Netexplo, en proposant une application qui permet aux chaînes d'information de créer facilement une vidéo à partir de contenus textuels trouvés sur l'Internet, présentant un résumé de leurs idées maîtresses. Depuis peu, plusieurs grands médias utilisent le « robot-journalisme » : Le Monde, Press Association et Xinhua, pour ne citer qu'eux, ont informé se servir d'algorithmes de production de langage naturel pour couvrir différents sujets journalistiques.

76. La production et la diffusion de contenu médiatique délèguent de plus en plus le pouvoir d'analyse et de décision à des algorithmes sophistiqués. Dans des proportions croissantes, les organisations des médias ont recours à des algorithmes pour analyser les préférences et les modèles de consommation des utilisateurs (personnalisation). Appliqués au journalisme, les algorithmes sont cette fois sollicités pour analyser des communautés géographiques spécifiques en quête de variables démographiques, sociales et politiques, en vue de produire l'information qui aura le plus de pertinence pour ces communautés, comme les prévisions météorologiques ou les résultats sportifs. Cette pratique peut potentiellement soutenir le journalisme et les journaux locaux. L'IA contribuera ainsi à renforcer les modèles économiques du journalisme.

77. En même temps, le journalisme basé sur l'IA soulève des questions de responsabilité, de transparence et de droits d'auteur. La responsabilité peut être un problème lorsqu'il s'avère laborieux d'imputer une faute s'agissant de diffusion d'information basée sur des algorithmes, par exemple en cas de diffamation. La transparence et la crédibilité posent problème lorsque les consommateurs ne réalisent pas ou ne peuvent réaliser que le contenu est généré par une machine, quelles en sont les sources et dans quelle mesure l'information est vérifiée ou même fausse – le débat actuel à propos des « deep fakes » ou vidéos truquées en est un exemple extrême. Quant à la question des droits d'auteur, elle ne va pas tarder à se poser, puisque le contenu généré par l'IA dépend de moins en moins des intrants humains, raison pour laquelle certains préconisent d'attribuer une forme de responsabilité en matière de droit d'auteur aux algorithmes eux-mêmes.

78. Pour relever ces défis, beaucoup sont d'avis que les journalistes et les rédacteurs en chef devraient se rapprocher des constructeurs d'algorithmes. Le lancement récent d'une plate-forme Open Source par Quartz AI Studio, un projet américain destiné à aider les journalistes à se faire seconder dans diverses tâches par l'apprentissage automatique, en est un exemple.

II.5. L'IA au service de la paix et de la sécurité

79. Conformément à la mission et au mandat de l'UNESCO consistant à promouvoir et construire la paix, la présente étude souhaite également examiner le rôle de l'intelligence artificielle dans les questions touchant à la consolidation de la paix et à la sécurité. Le fait que cela inclue la possibilité d'un usage militaire de l'IA n'affaiblit en rien son engagement en faveur de la paix.

80. L'IA, fait-on valoir, est capable d'analyser, de traiter et de catégoriser de très grandes quantités de données évoluant rapidement et de nature extrêmement variée (Payne, 2018 ; Roff, 2018 ; Gupta, 2018). Parmi les données « *objectives* », on peut citer les images satellitaires et autres imageries de surveillance, les signaux et l'intelligence électronique, tandis que les données « *subjectives* » comprendraient les rapports, documents, fils d'actualité, participations aux médias sociaux et données politiques et sociologiques. L'IA est présentée comme capable de catégoriser toute cette masse de données pour identifier les menaces externes et internes, découvrir les objectifs et les stratégies des acteurs, interpréter les intentions complexes et pluridimensionnelles sous-tendant leurs activités, et définir les stratégies à même de prévenir ou de contrer leurs actions.

81. Un tel outil d'intelligence situationnelle pourrait se révéler un instrument puissant de prévention et de résolution des conflits (Spiegeleire *et al.*, 2017). Il pourrait fournir des indications sur les facteurs à l'origine d'une entreprise humaine et leurs résultantes, avec une application possible à la déradicalisation. Une « intelligence prospective » à base d'apprentissage automatique pourrait prévoir la montée des troubles sociaux et de l'instabilité sociétale, et suggérer des moyens de prévention. Des informations plus poussées sur les facteurs de conflit pourraient inciter des déclencheurs de conflit potentiels à renoncer à mettre leurs mauvaises intentions à exécution. Nous pourrions être capables de détecter les pathologies sociales à un stade précoce, d'identifier quelles actions pourraient désamorcer une situation potentiellement dangereuse, ou de trouver des voies d'apaisement efficaces afin de déjouer des tentatives d'attiser les ardeurs sectaires. Au niveau sociétal, en repérant et en nous aidant à comprendre les dynamiques qui renforcent ou affaiblissent la résilience sociétale, l'IA pourrait nous conduire vers une société plus résiliente, et nous aider à nous diriger vers un monde paisible et débarrassé des conflits.

82. Du côté négatif, *l'IA transformera la nature et la pratique des conflits*, ce qui aura des répercussions sociétales dépassant de loin les questions purement militaires (Payne, 2018 ; Spiegeleire *et al.*, 2017). Non seulement elle transformera le recours à la force explosive en améliorant l'efficacité du déploiement des systèmes d'armes, mais l'IA promet aussi d'accroître de façon importante la vitesse et la précision dans tous les domaines, de la logistique militaire, du renseignement et de l'intelligence situationnelle à la planification des combats et à l'exécution ou aux opérations. Le système d'IA lui-même pourrait être chargé de formuler ses propres suggestions d'actions à mener : il pourrait créer un ensemble d'instructions destinées à exploiter les faiblesses de l'ennemi qu'il aura évaluées grâce à ses propres analyses, ou, parce qu'il aura trouvé le mode de fonctionnement de l'ennemi ou de l'insurgé, concevoir des parades contre un acte d'agression prédit. Il pourrait aussi mener son propre exercice de simulation d'un conflit pour sonder les réponses probables à des actions particulières.

83. La vitesse à laquelle ces outils de planification pourraient fonctionner augmenterait la capacité d'agir dans les situations évoluant rapidement. On peut envisager, par exemple, le développement d'une réponse algorithmique à une attaque coordonnée

grâce, par exemple, à des nuées de drones ou d'autres moyens militaires sans pilote comme les missiles assaillants. La vélocité de la réponse assistée par l'IA peut être perçue comme une incitation à en faire usage, et elle pourrait donc avoir un effet potentiellement déstabilisant. Ou des effets réellement désastreux, comme le montrent les exemples passés d'avertissements lancés par des machines auxquels, heureusement, le commandement humain n'a pas donné réponse. Il n'en reste pas moins qu'un État qui ne s'engagerait pas dans cette voie d'une réponse pilotée par l'IA serait fortement désavantagé, ce qui encourage la prolifération de cette capacité.

84. Il se peut qu'une machine de prise de décisions assistée par l'IA, par exemple une arme totalement autonome, décide par elle-même d'attaquer et de tuer et passe à l'action sans intervention humaine. L'idée d'une entité *non humaine* dotée d'un tel pouvoir d'intervention pourrait modifier radicalement notre approche de la politique au niveau le plus large. De plus, la proximité entre les usages militaires potentiels de l'IA et ses applications civiles (la facilité d'armer) signifie qu'il ne s'agit pas d'une catégorie clairement délimitée, ce qui complique à la fois l'éthique et la réglementation de son développement et de son application.

85. Bien qu'on puisse considérer l'IA simplement comme une nouvelle révolution en matière militaire, qui permet aux forces armées de faire la même chose avec les mêmes outils, son réel potentiel « révolutionnaire » (Payne, 2018 ; Spiegeleire *et al.*, 2017) réside peut-être dans la *transformation du concept de « force armée »* en une force dont les armes sont plus subtiles que les engins explosifs. Le pouvoir de l'IA dans les conflits tient non seulement à l'amélioration des technologies physiques, mais aussi à sa redéfinition de la « force armée ».

86. On y assiste déjà dans le cyberespace, où l'IA apporte à la fois des capacités de défense et d'attaque. Grâce aux techniques de recherche de motifs, d'apprentissage profond et d'observation des déviations par rapport à l'activité normale, il devient possible de détecter les vulnérabilités logicielles, puis d'en faire une arme en empêchant à l'ennemi de se défendre. Les réseaux neuronaux profonds pourraient détecter et prévenir les intrusions. Pour être efficace, la cyberdéfense devra être rapide et par conséquent avoir un haut degré d'autonomie.

87. La propagande est une autre arme que l'IA a renforcée. La facilité avec laquelle on peut truquer des voix, des images et des informations, et les propager auprès de publics choisis, menace l'ingénierie sociale et la formation de l'opinion publique, concourant à sa déformation. Par nature, l'IA facilite les mensonges convaincants et favorise les falsifications. Le risque de saper notre confiance dans l'intégrité de l'information qui en résulte augmente les possibilités d'une erreur d'appréciation, tant tactique que stratégique, des intentions d'un adversaire présumé.

88. L'IA offre aussi des capacités de sabotage économique et de perturbation des infrastructures essentielles. En plaçant la guerre radioélectronique en mode cognitif, l'IA pourrait jouer un rôle déterminant pour perturber l'accès au spectre électromagnétique. Des systèmes utilisant l'apprentissage automatique, les algorithmes « intelligents » et le filtrage adaptatif des signaux, sont déjà commercialisés.

89. Enfin, en ce qui concerne la sécurité *interne* des États, l'utilisation de l'analyse d'ensembles de données et de la reconnaissance faciale implique une nouvelle relation entre la société et les institutions chargées de la protéger. Cela a évidemment d'importantes conséquences éthiques.

II.6. L’IA et l’égalité des genres

90. Les systèmes d’IA ont d’importantes implications pour l’égalité des genres, dans la mesure où ils peuvent refléter les préjugés sociétaux existants, avec le pouvoir de les exacerber. La plupart de ces systèmes sont construits à partir de séries de données reflétant le monde réel, et celui-ci peut être biaisé, inéquitable et discriminatoire (Marda, 2018). Récemment, on s’est aperçu du caractère sexiste d’un outil de recrutement utilisé par Amazon, qui privilégiait les candidats de sexe masculin pour les emplois techniques (Reuters, 2018). Ces systèmes peuvent être dangereux, non seulement parce qu’ils perpétuent les inégalités de genre dans la société, mais aussi parce qu’ils incorporent ces inégalités de manière opaque, tout en étant considérés comme « objectifs » et « rigoureux » (O’Neil, 2018).

91. Ces inégalités proviennent essentiellement des modalités d’apprentissage des machines. Dans la mesure où l’apprentissage automatique dépend des données fournies, une attention particulière est nécessaire pour promouvoir des données sensibles à l’égalité des genres, ainsi que la collecte de données ventilées par genre. Dans le cas de l’outil de recrutement d’Amazon, le biais provenait de ce que l’outil apprenait à partir des candidats antérieurs – majoritairement de sexe masculin – et qu’il avait donc « appris » qu’il fallait préférer les postulants aux postulantes (Short, 2018). L’attention prêtée aux données biaisées devrait donc aider à réduire l’angle mort dans la manière dont les systèmes d’IA peuvent être adaptés et conçus pour les hommes et pour les femmes. En outre, l’application de données ventilées par genre aux fonctions analytiques de l’IA offre une occasion de mieux saisir les questions de genre auxquelles nous sommes actuellement confrontés.

92. Il est important de noter que les inégalités de genre interviennent dès les premiers stades de la conceptualisation et de la conception des systèmes d’IA. Dans les domaines techniques, la disparité de genre est patente et bien connue (Hicks, 2018), qu’il s’agisse d’inégalité dans les salaires ou les promotions (Brinded, 2017). C’est ce qu’on a appelé le « tuyau percé » : la baisse de 40 % de la participation des femmes aux secteurs de la technologie et de l’ingénierie entre le moment où elles décrochent leur diplôme et celui où elles deviennent des cadres dans ces domaines (Wheeler, 2018). Le faible pourcentage de femmes dans la main d’œuvre de l’IA – et dans le développement des capacités informatiques en général – fait que les femmes n’ont pas une voix égale au chapitre dans les processus de prise de décisions concernant la conception et le développement des systèmes d’IA. Par suite, nous risquons de construire ces technologies uniquement pour certains segments de la population (Crawford, 2016).

93. De plus, les biais que les individus véhiculent dans leur vie de tous les jours peuvent se trouver reflétés et même amplifiés par le développement et l’utilisation des systèmes d’IA. Le « genrage » des assistants numériques, par exemple, pourrait renforcer l’image des femmes comme étant subordonnées et obéissantes. De fait, ce sont des voix de femmes qui sont régulièrement choisies pour les bots d’assistance personnelle, qui remplissent essentiellement des fonctions de service aux consommateurs, alors que la majorité des bots dans des services professionnels comme les secteurs juridiques et financiers, par exemple, sont codés comme des voix d’hommes. Cela a des répercussions au niveau de l’éducation pour ce qui est de notre compréhension des compétences « masculines » par opposition aux compétences « féminines » et de notre définition de l’autorité par rapport aux positions subalternes. En outre, dans les systèmes d’IA, la notion de « genre » se résume souvent à un simple choix entre masculin et féminin. Celui-ci

ignore et écarte clairement les individus transgenres, au risque d'exercer une discrimination humiliante à leur encontre (Costanza-Chock, 2018).

II.7. L'Afrique et les défis de l'IA

94. On assiste en Afrique, comme dans d'autres régions en développement, à une accélération de l'utilisation des technologies de l'information et de l'IA. La nouvelle économie numérique émergente suscite d'importants défis et opportunités pour les sociétés créatives africaines.

95. Concrètement, en terme de connectivité des infrastructures, l'Afrique est profondément déficitaire et très en retard par rapport aux autres régions en développement ; sur le plan des connexions nationales, des liens régionaux et de l'accès continu à l'électricité, son handicap est lourd. Les services d'infrastructure se paient très cher, même si un nombre croissant d'Africains – y compris dans les bidonvilles urbains – possèdent un téléphone mobile.

96. Les problèmes de développement des pays africains sont nombreux. Le cadre des droits de l'homme et les Objectifs de développement durable (ODD) sont un moyen cohérent d'orienter le développement de l'IA. Par suite, comment partager les technologies et les connaissances relatives à l'IA et les orienter vers les priorités que les pays en développement ont eux-mêmes définies ? Priorités qui portent sur des défis tels que les infrastructures, les compétences, les connaissances, les capacités de recherche et la disponibilité des données locales, comme souligné lors du Forum de l'UNESCO sur l'intelligence artificielle en Afrique qui s'est tenu à l'Université Mohammed VI polytechnique à Benguérir (Maroc) les 12 et 13 décembre 2018.

97. Les femmes jouent un rôle crucial. Actrices économiques d'un grand dynamisme en Afrique, les femmes exercent la majorité des activités agricoles, détiennent un tiers des entreprises et représenteraient, dans certains pays, jusqu'à 70 % des employés. Elles sont les principaux leviers de l'économie domestique et du bien-être familial, et jouent un rôle directeur totalement indispensable au sein de leurs communautés et nations respectives. En plaçant l'égalité des genres au centre de sa stratégie de promotion du développement en Afrique, la Banque africaine de développement reconnaît le rôle fondamental de la parité des genres pour parvenir à une croissance inclusive et bâtir des sociétés résilientes. L'accès à l'éducation, à la maîtrise de l'IA et, plus globalement, aux technologies de l'information et de la communication (TIC) sont des éléments clés pour autonomiser les femmes et éviter leur marginalisation.

98. Avec une attention particulière pour la recherche scientifique, la science, la technologie, l'ingénierie et les mathématiques, ainsi que pour une éducation à la citoyenneté fondée sur les valeurs, les droits et les devoirs, l'IA devrait être intégrée au politiques et aux stratégies nationales de développement en s'appuyant sur les cultures, les valeurs et les connaissances endogènes pour développer les économies africaines.

III. INSTRUMENT À CARACTÈRE NORMATIF

III.1. Déclaration ou recommandation ?

99. Le Groupe de travail a examiné attentivement deux des outils normatifs de l'UNESCO – la déclaration et la recommandation –, qui sont en lien avec les analyses contenues dans les deux premières sections de la présente étude préliminaire concernant l'éthique de l'IA. Le Groupe de travail s'est également appuyé sur l'expérience antérieure

du COMEST, qui est à l'origine de la Déclaration de principes éthiques en rapport avec les changements climatiques de 2017 et a participé à la révision de la Recommandation concernant la science et les chercheurs scientifiques de 2017. Le Groupe de travail a évalué les avantages et les inconvénients de chacun de ces deux outils normatifs.

100. Concernant la proposition de Déclaration sur l'éthique de l'intelligence artificielle, le Groupe de travail a noté la très récente augmentation du nombre des déclarations de principes éthiques concernant l'IA en 2018. La *Déclaration de Montréal pour un développement responsable de l'IA* (Université de Montréal, 2018), la *Déclaration de Toronto : protection du droit à l'égalité et à la non-discrimination dans le domaine de l'apprentissage automatique* (Amnesty International et Access Now, 2018), et la Déclaration du *Future of Life Institute* concernant les *Principes d'Asilomar sur l'IA* (*Future of Life Institute*, 2017) sont le fruit d'initiatives différentes et sont soutenues par des organisations variées (universités, gouvernements, associations professionnelles, entreprises, ONG). Il convient d'ajouter à cette série de déclarations plusieurs propositions éthiques telles que les *Lignes directrices concernant l'éthique d'une IA digne de confiance* du Groupe d'experts de haut niveau sur l'IA de la Commission européenne, qui prend appui sur les droits de l'homme, et le deuxième document de l'IEEE (actuellement objet de consultations) qui prône une conception conforme à l'éthique (*Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*), et qui s'adresse aux ingénieurs et vise à intégrer des valeurs dans les systèmes intelligents autonomes. Toutes ces initiatives sont positives dans la mesure où elles lancent le débat sur l'éthique de l'IA à différents niveaux.

101. Le Groupe de travail n'en a pas moins conclu qu'il existe une forte hétéronomie dans les principes et dans la mise en œuvre des valeurs promues par l'une et l'autre. Cette hétéronomie est la conséquence à la fois de la définition de l'IA qui a été choisie, et des objectifs qui sont poursuivis : gouvernance, éducation des ingénieurs, politiques publiques. La question est la suivante : une Déclaration de l'UNESCO sur l'éthique de l'IA permettrait-elle à cette hétéronomie d'être fédérée en un petit nombre de principes directeurs qui apporteraient une réponse globale aux questions éthiques soulevées par l'IA, ainsi qu'aux préoccupations spécifiques de l'UNESCO dans les domaines de l'éducation, de la culture, de la science et de la communication ? Le Groupe de travail estime cela possible, avec le risque, toutefois, qu'au cours du processus conduisant à cette déclaration, les États membres ne s'accordent essentiellement que sur quelques principes généraux, abstraits et non contraignants, puisqu'il s'agit d'une déclaration. Dans une telle perspective, une Déclaration de l'UNESCO sur l'éthique de l'IA apporterait-elle de la valeur ajoutée par rapport à d'autres déclarations et initiatives en cours ? Il est douteux qu'un tel instrument s'établisse immédiatement comme une référence internationale, dans un contexte de concurrence entre cadres éthiques, à un moment où les technologies sont naissantes et où leurs utilisations ne sont pas encore stabilisées.

102. Le Groupe de travail a donc examiné si une recommandation ne constituerait pas un outil plus approprié dans la situation actuelle. Au niveau international, au niveau européen et dans le contexte politique national de plusieurs pays, on s'oriente vers des formes similaires de réglementation en ce qui concerne l'économie numérique, qui prennent aussi en compte les relations entre les deux principales puissances numériques : les États-Unis et la Chine. Les critiques croissantes concernant l'absence de transparence, les biais ou le comportement des grandes entreprises, et la méfiance grandissante du public face aux cyberattaques, créent un nouveau climat politique qui n'est pas sans incidence sur le développement de l'IA. Le mouvement de réglementation numérique, initié par l'Union européenne sur la protection des données personnelles

pourrait donc être étendu au niveau international dans des domaines émergents comme l'IA. Toutefois, à ce niveau, les outils en sont encore à leurs premiers stades d'élaboration, même si la stratégie poursuivie par l'OCDE, à travers son Groupe d'experts sur l'intelligence artificielle (AIGO), met l'accent sur la responsabilité, la sécurité, la transparence, la protection et la reddition des comptes :

> L'OCDE soutient les gouvernements par l'analyse des politiques, le dialogue et l'identification des meilleures pratiques. Nous déployons de gros efforts de cartographie des impacts économiques et sociaux des technologies de l'IA, de leurs applications et de leurs implications politiques. Il s'agit notamment d'améliorer la mesure de l'IA et de ses impacts, et d'éclairer sur des questions politiques importantes comme l'évolution des marchés de la main d'œuvre et les compétences de l'ère numérique, le respect de la vie privée, la reddition des comptes sur les décisions habilitées par l'IA, et les questions de responsabilité, de sécurité et de sûreté que soulève l'intelligence artificielle (OCDE, 2019).

103. Les priorités de l'OCDE en matière de politiques publiques ont surtout trait à la gouvernance de l'IA et aux bonnes pratiques dans ce domaine. Il semble ici que l'approche de l'UNESCO pourrait être complémentaire à celle de l'OCDE au niveau international, mais en se concentrant sur des aspects généralement négligés comme la culture, l'éducation, la science et communication. Ces dimensions ont une incidence directe sur la vie quotidienne des individus et des populations et sur leurs aspirations individuelles et collectives. L'approche de l'UNESCO concernant une Recommandation sur l'éthique de l'IA serait présentée comme une alternative complémentaire à une vision de la gouvernance économique. Le Groupe de travail estime donc qu'une recommandation, bien qu'exigeant plus de temps et d'énergie qu'une déclaration, permettrait à l'UNESCO de se distinguer non seulement quant à son contenu éthique, mais aussi du fait des propositions spécifiques adressées à ses États membres. Un des buts est de donner aux États les moyens et capacités nécessaires pour intervenir dans les principaux domaines impactés par le développement de l'IA, comme la culture, l'éducation, la science et la communication.

104. Cette recommandation devrait avoir deux dimensions. La première est l'affirmation d'une série de principes fondamentaux pour une éthique de l'IA. La deuxième est la formulation de propositions spécifiques visant à aider les États à suivre la mise en œuvre, et réglementer, les usages de l'IA dans les domaines relevant du mandat de l'UNESCO au moyen du mécanisme de présentation de rapports de la recommandation, et à identifier des outils d'évaluation éthique pour l'examen régulier de leurs politiques destinées à guider le développement de l'IA. À cet égard, l'UNESCO serait particulièrement bien placée pour apporter une perspective pluridisciplinaire, ainsi qu'une plate-forme universelle pour l'élaboration d'une Recommandation sur l'éthique de l'IA. En particulier, l'UNESCO serait capable de réunir à la fois les pays développés et les pays en développement, différents points de vue culturels et moraux ainsi qu'un éventail varié de parties prenantes des sphères publiques et privées au sein d'un processus réellement international d'élaboration d'un ensemble complet de principes et de propositions concernant l'éthique de l'IA.

105. La section suivante présente quelques-unes de ces propositions.

III.2. Suggestions concernant un instrument à caractère normatif

106. Sur la base de cette analyse des implications potentielles de l'intelligence artificielle pour la société, le Groupe de travail aimerait suggérer une série d'éléments qui pourraient

être inclus dans une éventuelle Recommandation sur l'éthique de l'IA. Ces suggestions incarnent la perspective mondiale de l'UNESCO, ainsi que ses domaines de compétence.

107. En premier lieu, le Groupe de travail aimerait suggérer une série de principes généraux concernant le développement, la mise en œuvre et l'utilisation de l'IA. Ces principes sont les suivants :

(a) **droits humains** : l'IA devrait être développée et mise en œuvre conformément aux normes internationales relatives aux droits de l'homme ;

(b) **inclusivité** : l'IA devrait être inclusive, et s'efforcer d'éviter les biais, de favoriser la diversité et d'empêcher l'apparition d'une nouvelle fracture numérique ;

(c) **épanouissement** : l'IA devrait être développée en vue d'améliorer la qualité de la vie ;

(d) **autonomie** : l'IA devrait respecter l'autonomie humaine en exigeant un contrôle humain permanent ;

(e) **explicabilité** : l'IA devrait être explicable, et permettre d'obtenir une idée précise de son fonctionnement ;

(f) **transparence** : les données utilisées pour alimenter les systèmes d'IA devraient être transparentes ;

(g) **sensibilisation et initiation** : il convient d'informer les citoyens en les sensibilisant aux algorithmes et en leur procurant une compréhension élémentaire du fonctionnement de l'IA ;

(h) **responsabilité** : les concepteurs et les entreprises devraient prendre l'éthique en considération lorsqu'ils conçoivent un système intelligent autonome ;

(i) **reddition des comptes** : des dispositifs devraient être élaborés de façon à obliger à rendre des comptes sur les décisions pilotées par l'IA et le comportement des systèmes d'IA ;

(j) **démocratie** : l'IA devrait être élaborée, mise en œuvre et utilisée conformément aux principes démocratiques ;

(k) **bonne gouvernance** : les gouvernements devraient présenter des rapports réguliers sur leur utilisation de l'IA dans la police, le renseignement et la sécurité ;

(l) **durabilité** : pour toutes les applications de l'IA, les avantages potentiels doivent être mis en balance avec l'impact environnemental de l'intégralité du cycle de production de l'IA et de la TI.

108. Plus particulièrement, le Groupe de travail aimerait souligner certaines questions éthiques centrales concernant les thèmes spécifiques de l'UNESCO :

(a) **éducation** : l'IA exige que l'éducation encourage la maîtrise de l'IA, la réflexion critique, la résilience sur le marché du travail et l'enseignement de l'éthique aux ingénieurs ;

(b) **science** : l'IA exige une introduction responsable dans la pratique scientifique, et dans la prise de décisions basée sur les systèmes d'IA, qui requiert une évaluation et un contrôle humains, et doit éviter d'exacerber les inégalités structurelles ;

(c) **culture** : l'IA devrait favoriser la diversité culturelle, l'inclusivité et l'épanouissement des êtres humains, en évitant de creuser la fracture numérique. L'approche multilingue devrait être encouragée ;

(d) **communication et information** : l'IA devrait renforcer la liberté d'expression, l'accès universel à l'information, la qualité du journalisme, et des médias libres, indépendants et pluralistes, tout en évitant la diffusion de fausses informations. Il convient de promouvoir une gouvernance multiparties prenantes ;

(e) **paix** : afin de contribuer à la paix, l'IA pourrait être utilisée pour obtenir des informations sur les facteurs de conflit, et son activité ne devrait jamais échapper au contrôle humain ;

(f) **Afrique** : l'IA devrait être intégrée dans les politiques et les stratégies nationales de développement, en prenant appui sur les cultures, les valeurs et les connaissances endogènes pour développer les économies africaines ;

(g) **genre** : les préjugés de genre devraient être évités dans l'élaboration des algorithmes, dans les ensembles de données utilisées pour les entraîner, et dans leur utilisation pour la prise de décisions ;

(h) **environnement** : l'IA devrait être développée de façon durable en prenant en considération la totalité du cycle de production de l'IA et des TI. L'IA peut être utilisée dans la surveillance environnementale et la gestion des risques, ainsi que dans la prévention et l'atténuation des crises environnementales.

BIBLIOGRAPHIE

AI Now. 2016. *The AI Now Report: The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term*. New York, The White House and the New York University's Information Law Institute. Available at: https://ainowinstitute.org/AI_Now_2016_Report.pdf.

Ajunwa, I., Crawford, K., and Schultz, J. 2017. Limitless Worker Surveillance. *California Law Review.* No. 735, pp. 101-142.

Allen, G. and Chan, T. 2017. Artificial Intelligence and National Security. *Harvard Kennedy School, Belfer Center for Science and International Affairs.* Online. Available at: https://www.belfercenter.org/publication/artificial-intelligence-and-national-security.

Amnesty International and Access Now. 2018. *The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems.* Toronto, RightsCon 2018. Available at: https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf.

ARCEP (Autorité de régulation des communications électroniques et des postes). 2018. *Smartphones, tablets, voice assistants... Devices, the weak link in achieving an open Internet*. Paris, ARCEP. Available at: https://www.arcep.fr/uploads/tx_gspublication/rapport-terminaux-fev2018-ENG.pdf.

Article 19. 2018a. *Free speech concerns amid the « fake news » fad*. Online. Available at: https://www.article19.org/resources/free-speech-concerns-amid-fake-news-fad/

Article 19. 2018b. *Privacy and Freedom of Expression in the Age of Artificial Intelligence*. Online. Online. Available at: https://www.article19.org/wp-content/uploads/2018/04/Privacy-and-Freedom-of-Expression-In-the-Age-of-Artificial-Intelligence-1.pdf.

Ashley, K.D. 2017. Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age. Cambridge, Cambridge University Press.

Boden, M.A. 2016. *AI: Its Nature and Future*. Oxford, Oxford University Press.

Brinded, L. 2017. « Robots are going to turbo charge one of society's biggest problems », *QUARTZ* (28 December 2017). Online. Available: https://qz.com/1167017/robots-automation-and-ai-in-the-workplace-will-widen-pay-gap-for-women-and-minorities/.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B. and Anderson, H. 2018. *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation*. Available at: https://arxiv.org/ftp/arxiv/papers/1802/1802.07228.pdf.

Bunnin, N. and Yu, J. 2008. *The Blackwell dictionary of western philosophy*. John Wiley & Sons.

Butterfield, A., Ngondi, G.E. and Kerr, A. eds. 2016. *A dictionary of Computer Science*. Oxford, Oxford University Press.

Costanza-Chock, S. 2018. « Design justice, AI, and escape from the matrix of domination », *Journal of Design and Science.* Online. Available at: https://jods.mitpress.mit.edu/pub/costanza-chock.

Crawford, K. 2016. « Artificial Intelligence's White Guy Problem », *The New York Times* (Opinion, 25 June 2016). Online. Available at: https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html.

Crawford, K. 2017. « The Trouble with Bias », NIPS 2017 Keynote. Available at: https://www.youtube.com/watch?v=fMym_BKWQzk.

Cummings, M. L., Roff, H. M., Cukier, K., Patakilas, J. and Bryce, H. 2018. *Artificial Intelligence and International Affairs: Disruption Anticipated.* Chatham House Report. Available at: https://www.chathamhouse.org/sites/default/files/publications/research/2018-06-14-artificial-intelligence-international-affairs-cummings-roff-cukier-parakilas-bryce.pdf.

Brookfield Institute and Policy Innovation Hub (Ontario). 2018. *Policymakers: Understanding the Shift*. Online. Available at: https://brookfieldinstitute.ca/wp-content/uploads/Brookfield-Institute_-The-AI-Shift.pdf.

Eubanks, V. 2018a. « A Child Abuse Prediction Model Fails Poor Families », *WIRED*. Online. Available at: https://www.wired.com/story/excerpt-from-automating-inequality/.

Eubanks, V. 2018b. *Automating Inequality: How high tech tools profile, police, and punish the poor*. New York, St. Martin's Press.

European Commission (EC). 2018. *Artificial Intelligence for Europe.* Communication from the Commission to the European Parliament, the European council, the Council, the European Economic and Social Committee and the Committee of the Regions. Brussels, European Commission. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0237&from=EN.

European Commission for the Efficiency of Justice (CEPEJ). 2018. *European ethical Charter on the use of Artificial Intelligence in judicial systems and their environment.* Strasbourg, CEPEJ. Available at: https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c.

European Group on Ethics in Science and New Technologies (EGE). 2018. *Statement on AI, Robotics, and Autonomous System*. Brussels, European Commission. Available at: https://publications.europa.eu/en/publication-detail/-/publication/dfebe62e-4ce9-11e8-be1d-01aa75ed71a1/language-en/format-PDF/source-78120382.

Executive Office of the President (USA). 2016. *Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights*. Washington, D.C., Executive Office of the President. Available at: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.

Frankish, K. and Ramsey, W.M. eds. 2014. *The Cambridge handbook of artificial intelligence*. Cambridge, Cambridge University Press.

Future of Life Institute. 2017. *Asilomar AI Principles.* Cambridge, Future of Life Institute. Available at: https://futureoflife.org/ai-principles/?cn-reloaded=1.

Gupta, D.K. 2018. « Military Applications of Artificial Intelligence », *Indian Defence Review* (22 March 2019). Online. Available at: http://www.indiandefencereview.com/military-applications-of-artificial-intelligence/.

Heacock, M., Kelly, C.B., Asante, K.A., Birnbaum, L.S., Bergman, Å.L., Bruné, M.N., Buka, I., Carpenter, D.O., Chen, A., Huo, X. and Kamel, M. 2015. « E-waste and harm to vulnerable populations: a growing global problem », *Environmental health perspectives*, Vol. 124, No. 5, pp. 550-555.

Hicks, M. 2018. « Why tech's gender problem is nothing new », *The Guardian* (12 October 2018). Online. Available at: https://amp.theguardian.com/technology/2018/oct/11/tech-gender-problem-amazon-facebook-bias-women?__twitter_impression=true.

Hinchliffe, T. 2018. « Medicine or poison? On the ethics of AI implants in humans », *The Sociable*. Online. Available at: https://sociable.co/technology/ethics-ai-implants-humans/.

House of Lords. 2017. *AI in the UK: ready, willing and able?* London, House of Lords Select Committee on Artificial Intelligence. Available at: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf.

Illanes, P., Lund, S., Mourshed, M., Rutherford, S. and Tyreman, M. 2018. *Retraining and reskilling workers in the age of automation*. Online, McKinsey Global Institute. Available at: https://www.mckinsey.com/featured-insights/future-of-work/retraining-and-reskilling-workers-in-the-age-of-automation.

Institute of Electrical and Electronic Engineers (IEEE). 2018. *Ethically Aligned Design – Version 2 for Public Discussion*. New Jersey, The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Available at: https://ethicsinaction.ieee.org/.

Laplante, P.A. 2005. *Comprehensive dictionary of electrical engineering*. Boca Raton, CRC Press.

Latonero, M. 2018. *Governing Artificial Intelligence: Upholding Human Rights & Dignity*. Data & Society. Available at: https://datasociety.net/wp-content/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf.

Marda, V. 2018. « Artificial Intelligence Policy in India: A Framework for Engaging the Limits of Data-Driven Decision-Making », *Philosophical Transactions A: Mathematical, Physical and Engineering Sciences*. Online. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3240384.

Matias, Y. 2018. Keeping people safe with AI-enabled flood forecasting. *The Keyword* (24 September 2018). Online. Available at: https://www.blog.google/products/search/helping-keep-people-safe-ai-enabled-flood-forecasting/.

Matsumoto, D.E. 2009. *The Cambridge dictionary of psychology.* Cambridge, Cambridge University Press.

McCarthy, J., Minsky, M. L., Rochester, N., Shannon, C. E. 2006 [1955]. « A proposal for the Dartmouth Summer Research Project on Artificial Intelligence », *AI Magazine*, vol. 27, no. 4, pp.12-14.

Microsoft Europe. 2016. « The Next Rembrandt », *Microsoft News Centre Europe*. Online. Available at: https://news.microsoft.com/europe/features/next-rembrandt/.

National Science and Technology Council (USA). 2016. *The National Artificial Intelligence Research and Development Strategic Plan*. Washington, D.C., National Science and Technology Council. Available at: https://www.nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf.

O'Brien, A. 2018. « How AI is helping preserve Indigenous languages », *SBS News*. Online. Available at: https://www.sbs.com.au/news/how-ai-is-helping-preserve-indigenous-languages.

O'Neil, C. 2018. « Amazon's Gender-Biased Algorithm Is Not Alone », *Bloomberg Opinion* (16 October 2018). Online. Available at: https://www.bloomberg.com/opinion/articles/2018-10-16/amazon-s-gender-biased-algorithm-is-not-alone.

OECD. 2019. *Going Digital.* Paris, OECD. Available at: http://www.oecd.org/going-digital/ai/.

Oppenheimer, A. 2018. *¡Sálvese quien pueda!: El futuro del trabajo en la era de la automatización*. New York, Vintage Espanol.

Palfrey, J.G. and Gasser, U. 2012. *Interop: The Promise and Perils of Highly Interconnected Systems*. New York, Basic Books.

Payne, K. 2018. « Artificial Intelligence: A Revolution in Strategic Affairs? », *Survival*, Vol. 60, No. 5, pp. 7-32.

Peiser, J. 2019. « The Rise of the Robot Reporter », *The New York Times* (5 February 2019). Online. Available at: https://www.nytimes.com/2019/02/05/business/media/artificial-intelligence-journalism-robots.html.

Reuters. 2018. « Amazon ditched AI recruiting tool that favored men for technical jobs », *The Guardian* (11 October 2018). Online. Available at: https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine.

Roff, H.M. 2018. 2018. « COMPASS: a new AI-driven situational awareness tool for the Pentagon? », *Bulletin of the Atomic Scientists* (10 May 2018). Online. Available at: https://thebulletin.org/2018/05/compass-a-new-ai-driven-situational-awareness-tool-for-the-pentagon/.

Rosenberg, J.M. 1986. *Dictionary of artificial intelligence and robotics*. New York, John Wiley & Sons.

Russell, S.J. and Norvig, P. 2016. *Artificial Intelligence: A Modern Approach*, 3rd ed. Harlow, Pearson.

Santosuosso, A. and Malerba, A. 2015. « Legal Interoperability As a Comprehensive Concept in Transnational Law », *Law, Innovation and Technology*, Vol. 5, No. 1, pp. 51-73.

Short, E. 2018. « It turns out Amazon's AI hiring tool discriminated against women », *Siliconrepublic* (11 October 2018). Online. Available at: https://www.siliconrepublic.com/careers/amazon-ai-hiring-tool-women-discrimination.

Spiegeleire, S. De, Maas, M. and Sweijs, T. 2017. *Artificial Intelligence and the Future of Defence*. The Hague, The Hague Centre for Strategic Studies.

UNI Global Union. 2016. Top 10 principles for ethical artificial intelligence. Switzerland, UNI Global Union. Available at: http://www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf.

UNICEF. 2017. *Children in a Digital World*. New York UNICEF. Available at: https://www.unicef.org/publications/files/SOWC_2017_ENG_WEB.pdf.

United Nations Educational, Scientific and Cultural Organization (UNESCO). 2002. *UNESCO Universal Declaration on Cultural Diversity: a vision, a conceptual platform, a pool of ideas for implementation, a new paradigm*. Paris, UNESCO. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000127162.

UNESCO. 2013. *Community Media: A Good Practice Handbook*. Paris, UNESCO. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000215097.

UNESCO. 2015a. *Keystones to foster inclusive knowledge societies: access to information and knowledge, freedom of expression, privacy and ethics on a global internet.* Paris, UNESCO. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000232563.

UNESCO. 2015b. *Outcome document of the « CONNECTing the Dots: Options for Future Action » Conference*. Paris, UNESCO. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000234090.

University of Montreal. 2018. *Montreal Declaration for a Responsible Development of AI*. Montreal, University of Montreal. Available at: https://www.montrealdeclaration-responsibleai.com/.

Vernon, D. 2014. *Artificial cognitive systems: A primer*. Cambridge, MIT Press.

Villani, C., Schoenauer, M., Bonnet, Y., Berthet, C., Cornut, A.-C., Levin, F. and Rondepierre, B. 2018. *For A Meaningful Artificial Intelligence: Towards a French and European Strategy*. Paris. Available at: https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf.

Wheeler, T. 2018. « Leaving at Lightspeed : the number of senior women in tech is decreasing », *OECD Forum* (23 March 2018). Online. Available: https://www.oecd-forum.org/users/91062-tarah-wheeler/posts/31567-leaving-at-lightspeed-the-number-of-senior-women-in-tech-is-decreasing.

World Summit on the Information Society (WSIS). 2003. *Declaration of Principles. Building the Information Society: A global challenge in the new Millenium*. Geneva, WSIS. Available at: http://www.itu.int/net/wsis/docs/geneva/official/dop.html.

WSIS. 2005. *Tunis Agenda for the Information Society*. Tunis, WSIS. Available at: http://www.itu.int/net/wsis/docs2/tunis/off/6rev1.html.

ANNEXE : COMPOSITION DU GROUPE DE TRAVAIL ÉLARGI DE LA COMEST SUR L'ÉTHIQUE ET L'IA

1. **M Peter-Paul VERBEEK (co-Coordinateur)**
   Professeur de philosophie de la technologie à l'Université de Twente, Pays-Bas
   Membre de la COMEST (2016-2019)

2. **Mme Marie-Hélène PARIZEAU (co-Coordinatrice)**
   Professeur, Faculté de philosophie, Université Laval, Québec, Canada
   Membre de la COMEST (2012-2019)
   Présidente (2016-2019) et Vice-Présidente (2014-2015) de la COMEST

3. **M. Tomislav BRACANOVIĆ**
   Associé de recherche, Institut de philosophie, Zagreb, Croatie
   Membre de la COMEST (2014-2021)
   Rapporteur de la COMEST (2018-2019)

4. **M. John FINNEY**
   Professeur émérite de physique, Département de physique et d'astronomie, Londres, Royaume-Uni
   Coordinateur du Groupe de travail sur l'éthique scientifique, Conférence Pugwash sur la science et les problèmes internationaux
   Membre *ex-officio* de la COMEST

5. **M. Javier JUAREZ MOJICA**
   Commissaire, Conseil de l'Institut fédéral des télécommunications de Mexico, Mexique
   Membre du Groupe d'experts sur l'IA de l'OCDE (AIGO)
   Membre de la COMEST (2018-2021)

6. **M. Mark LATONERO**
   Directeur de recherche, Données et droits de l'homme, Data & Society, États-Unis d'Amérique

7. **Mme Vidushi MARDA**
   Responsable du programme numérique pour ARTICLE 19 (Mme Marda réside en Inde)

8. **Mme Hagit MESSER-YARON**
   Professeur d'ingénierie électrique et ancienne Vice-Présidente pour la recherche et le développement, Université de Tel Aviv, Tel Aviv, Israël
   Membre, Comité exécutif, Initiative mondiale de l'IEEE sur l'éthique des systèmes autonomes et intelligents
   Membre de la COMEST (2016-2019)

9. **M. Luka OMLADIC**
   Maître de conférence, Université de Ljubljana, Ljubljana, Slovénie
   Membre de la COMEST (2012-2019)

10. **Mme Deborah OUGHTON**
Professeur et Directrice de recherche, Centre de la radioactivité environnementale, Université norvégienne des sciences de la vie
Membre de la COMEST (2014-2021)

11. **M. Amedeo SANTOSUOSSO**
Fondateur et Directeur scientifique, Centre européen du droit, de la science et des nouvelles technologies (ECLT), Université de Pavie, Pavie, Italie,
Président, Première chambre, Cour d'appel de Milan, Italie
Membre de la COMEST (2018-2021)

12. **M. Abdoulaye SENE**
Sociologue de l'environnement, Coordinateur pour « l'éthique, la gouvernance et la responsabilité environnementale et sociale », Institut des sciences environnementales, Université Cheikh Anta Diop, Dakar, Sénégal
Membre de la COMEST (2012-2019)
Vice-Président de la COMEST (2016-2019)

13. **M. John SHAWE-TAYLOR**
Chaire UNESCO en intelligence artificielle, *University College* de Londres et Président de la Fondation *Knowledge 4 All,* Royaume-Uni de Grande-Bretagne et d'Irlande du Nord

14. **M. Davide STORTI**
Spécialiste de programme, Section pour l'application des TIC dans l'éducation, la science et la culture, Secteur de la communication et de l'information, UNESCO

15. **M. Sang Wook YI**
Professeur de philosophie, Université Hanyang, Séoul, République de Corée
Membre de la COMEST (2018-2021)