unesco

Mila

ANGLES MORTS DE LA GOUVERNANCE DE L'IA

SOUS LA DIRECTION DE
Benjamin Prud'homme
Catherine Régis
and Golnoosh Farnadi

ANGLES MORTS DE LA GOUVERNANCE DE L'IA

Publié en 2023 par l'Organisation des Nations Unies pour l'éducation, la science et la culture

(UNESCO), 7, place de Fontenoy, 75007 Paris, France et par Mila – Institut québécois d'intelligence artificielle, 6666, rue Saint-Urbain, Montréal, Québec, Canada H2S 3H1.

ISBN: 978-92-3-200285-3

Éditeurs et éditrices : Benjamin Prud'homme, Catherine Régis, Golnoosh Farnadi, Vanessa Dreier, Sasha Rubel (2021), Charline d'Oultremont.

Coordination : Amanda Leal de Lima Alves

Conception graphique : Alphatek

Traduction professionnelle : Martine Sénecal, Daly-Dallaire Services de traduction

Conception de la page couverture : Frédérick Gélinas

BREF RÉSUMÉ

18 propositions sélectionnées offrant une approche pluraliste, informée et critique de la gouvernance de l'IA

NAVIGUER L'IMPRÉVISIBLE: RÉFLEXIONS SUR LA GOUVERNANCE DE L'IA

Au cours de la prochaine décennie, l'intelligence artificielle (IA) continuera d'avoir une influence et un effet considérables sur nos sociétés. Alors que ces avancées scientifiques et technologiques se succèdent à un rythme soutenu, il est essentiel de poursuivre et accélérer la conversation mondiale et inclusive sur leur développement et leur gouvernance.

C'est dans ce contexte que Mila et l'UNESCO unissent leurs forces pour mener un travail collectif pour cerner des questions importantes et novatrices quant à la gouvernance de l'IA. La présente publication est une compilation de 18 chapitres sélectionnés à la suite d'un appel à contributions lancé mondialement en 2021. Les articles présentés traversent les frontières disciplinaires et géographiques en plus d'inclure différentes perspectives, incluant celles d'universitaires, de membres de la société civile et d'innovateurs et innovatrices. Cette publication vise à influencer la conversation sur l'IA pour que celle-ci porte moins sur ce que nous savons déjà que sur ce qui échappe à notre regard : les angles morts de l'IA. Les sujets abordés sont donc variés et incluent notamment la perspective du droit international et des droits des peuples autochtones, les audits des systèmes d'IA, l'alignement de ces technologies avec les objectifs de développement durable des Nations Unies et la centralisation du pouvoir décisionnel que permet cette technologie de rupture.

Les membres des milieux politiques, de la recherche et de la société civile engagés dans cette conversation mondiale sur l'IA bénéficieront des perspectives présentées pour faire face à l'immense tâche qui leur incombe : assurer un développement de l'IA qui soit éthique, inclusif et conforme aux droits humains.

unesco

« Les guerres prenant naissance dans l'esprit des hommes et des femmes, c'est dans l'esprit des hommes et des femmes que doivent être élevées les défenses de la paix »

À PROPOS DES ORGANISATIONS

Les applications d'intelligence artificielle (IA) continuent d'élargir les possibilités permettant d'atteindre les objectifs de développement durable et L'UNESCO cherche à tirer parti de ces opportunités dans ses champs de compétence en éducation, en sciences, en culture, en communications et en information. L'UNESCO mène des réflexions sur des préoccupations pressantes découlant du développement rapide de l'IA et ce, du point de vue des droits de la personne et de l'éthique.

Ces préoccupations sont diverses ; parmi elles figurent le rôle de l'IA dans l'avenir de l'éducation, dans les défis omniprésents de la désinformation et des discours haineux en ligne, dans l'utilisation de l'IA pour atteindre les objectifs de développement durable et autonomiser les pays du Sud, et dans la promotion de l'égalité des sexes dans le secteur de l'IA et la lutte contre les biais algorithmiques.

Mila a pour mission d'être un pôle mondial pour les avancées scientifiques qui inspirent l'innovation et l'essor de l'IA au bénéfice de tous. tes. Mila est situé à Montréal et réunit plus de 1000 chercheur.e.s universitaires qui poursuivent des travaux de recherche en IA dans une grande variété de sous-domaines et de champs d'application, incluant pour soutenir la lutte aux changements climatiques, la préservation de la biodiversité, l'accélération des découvertes en santé et la pleine réalisation des droits humains. Les travaux menés à Mila sont guidé.e.s par notre engagement à étudier et à développer des cadres appuyant l'avancement et la mise en œuvre d'une IA responsable.

En tant que chef de file mondial dans le domaine, Mila contribue également aux efforts nationaux et internationaux visant à favoriser un dialogue social et une mobilisation sur des questions d'éthique, de responsabilité et de justice sociale en lien avec l'IA et sa gouvernance.

Réalisé avec le soutien financier de

Québec

REMERCIEMENTS

L'équipe éditoriale souhaite remercier les personnes de l'UNESCO suivantes pour avoir soutenu la réalisation de cette publication : Marielza Oliveira, Guy Berger, Prateek Sibal, Cédric Wachholz et Jacinth Chia.

L'équipe éditoriale tient à remercier le gouvernement du Québec pour son soutien financier dans la réalisation de cet ouvrage, en particulier le ministère des Relations internationales et de la Francophonie et les Fonds de recherche du Québec.

L'équipe éditoriale désire aussi exprimer sa reconnaissance envers ceux et celles qui ont partagé leur temps, leurs connaissances et leur travail en répondant à l'appel à contributions dans le but d'explorer des approches créatives, novatrices et ambitieuses en lien avec la gouvernance de l'intelligence artificielle. Nous remercions également tous les auteurs et autrices qui ont pris le temps de partager leurs idées et dont les voix ont été essentielles à la réalisation de cette publication.

L'équipe éditoriale souhaite remercier les personnes qui ont contribué à la mise en forme, à la traduction, à la révision et à la conception de la publication : Martine Sénecal, Daly-Dallaire Services de traduction, Marie Zumstein, Noah Oder, Laura Gagliano, Frédérick Gélinas et Alphatek.

Enfin, l'équipe éditoriale tient à remercier chaleureusement Amanda Leal de Lima Alves, dont le travail de coordination inlassable a permis la réalisation de cette publication.

AVANT-PROPOS DE L'UNESCO

L'UNESCO est l'agence spécialisée des Nations Unies chargée de bâtir la paix grâce à la coopération internationale dans les domaines de l'éducation, des sciences, de la culture, de l'information et de la communication. Dans le cadre de son mandat, l'UNESCO s'efforce d'approfondir les connaissances, de favoriser la collaboration mondiale et d'offrir des conseils en matière de politiques sur des questions clés liées à l'innovation et à la transformation numériques. Cette publication s'inscrit donc dans cette expertise et se veut la dernière offre d'une série de ressources de connaissances de pointe alors que l'UNESCO continue d'exercer ses nombreuses fonctions, notamment en tant que laboratoire d'idées.

Au fur et à mesure que s'étendent le développement et l'utilisation de l'IA, celle-ci continue de défier ce que nous pensions jadis être possible. Il y a tout juste 80 ans, les informaticiens tentaient de trouver une façon pour permettre aux ordinateurs d'effectuer des tâches simples, comme l'enregistrement de commandes. Aujourd'hui, les applications de l'IA se sont étendues au traitement du langage naturel, aux processus judiciaires, aux véhicules autonomes et à la cartographie des maladies, pour n'en citer que quelques-unes. L'IA offre de nouvelles possibilités pour atteindre les objectifs de développement durable des Nations Unies, y compris dans les domaines de compétence de l'UNESCO en matière d'éducation, de sciences naturelles, sociales et humaines, de culture, et de communication et d'information. Cependant, même si l'IA recèle un grand potentiel pour favoriser le développement durable, la complexité et le rythme de son développement posent un défi non seulement en matière de gouvernance de l'IA, mais aussi de protection et de promotion des droits humains.

Grâce à son rôle de premier plan en coopération internationale, l'UNESCO a orienté la réflexion mondiale autour des préoccupations urgentes liées à l'IA. Une partie de cet effort a conduit à l'adoption en 2021, par les 193 États membres de l'UNESCO, de la *Recommandation sur l'éthique de l'intelligence artificielle*, le premier outil normatif mondial dans ce domaine.

Avec la compréhension commune que l'IA doit avant tout être centrée sur la personne, l'UNESCO reconnaît aussi les défis auxquels font face plusieurs pays dans le développement et la gouvernance de l'IA. À ce titre, cette publication, issue d'une collaboration entre l'UNESCO et Mila – Institut québécois d'intelligence artificielle, est une contribution importante aux réflexions sur les défis de gouvernance et les écarts en matière de capacités humaines et institutionnelles auxquels font face les pays pour assurer une utilisation digne de confiance et responsable de l'IA.

Par exemple, l'initiative de l'UNESCO pour la formation des juges a permis de mettre au point un ensemble complet d'outils pour former les acteurs du judiciaire à l'application de l'IA et à son impact dans l'administration de la justice. En outre, nous dialoguons avec des fonctionnaires, qui sont des parties prenantes importantes dans le développement des politiques en matière d'IA, afin d'en arriver à une compréhension des possibilités et des défis liés à l'IA dans leur travail. L'UNESCO a aussi effectué des recherches sur des questions liées au genre et à l'IA, notamment sur les écarts de compétences numériques et les préjugés liés au genre dans les algorithmes d'IA ou sur l'impact de l'IA sur les femmes dans le milieu professionnel.

Je suis convaincu que cette publication contribuera à fournir aux membres du milieu politique et de la société civile les perspectives critiques nécessaires pour veiller à ce que le développement de l'IA atteigne son plein potentiel dans le respect des libertés et droits fondamentaux. J'espère que les lecteurs et lectrices trouveront dans chacun des chapitres les réponses que nous cherchons ainsi que les questions que nous devons poser pour que les technologies de l'IA ne laissent personne de côté.

Nous espérons donc que cette publication aidera à renforcer la contribution essentielle que les technologies numériques, et plus particulièrement l'IA, peuvent apporter à la promotion de sociétés inclusives et pacifiques, lorsque ces technologies sont appliquées dans une approche fondée sur les droits humains, et à la mise en place d'une IA digne de confiance et responsable.

L'UNESCO remercie Mila et tous ceux et celles qui ont contribué à la réalisation de cette publication. Nous vous souhaitons une lecture inspirante et seront ravis de dialoguer avec vous.

TAWFIK JELASSI
Sous-directeur général pour la Communication et l'information, UNESCO

AVANT-PROPOS DE MILA – INSTITUT QUÉBÉCOIS D'INTELLIGENCE ARTIFICIELLE

Depuis sa création, Mila s'efforce d'atteindre les plus hauts niveaux de leadership scientifique dans le domaine de l'intelligence artificielle en plaçant le développement responsable et éthique de l'IA au cœur de sa mission. Notre collaboration avec l'UNESCO témoigne de notre engagement à démocratiser ces connaissances et à contribuer aux efforts de coopération internationale dans ce domaine.

L'innovation technologique a déjà une influence et un effet marqués dans toutes les sphères de nos vies. Les progrès réalisés par l'IA dans des domaines comme la santé, l'agriculture et les sciences climatiques offrent d'immenses possibilités qui étaient jusqu'à récemment inimaginables. Cependant, l'IA pose également d'importants risques ; c'est pourquoi une énergie inlassable doit être consacrée au développement de systèmes d'IA responsables et socialement bénéfiques. Cela signifie des systèmes d'IA qui respectent et soutiennent la pleine réalisation des droits humains, l'État de droit, les principes énoncés dans la *Recommandation sur l'éthique de l'intelligence artificielle* de l'UNESCO ainsi que la réalisation des objectifs de développement durable (ODD). Il s'agit d'un programme collectif ambitieux, mais essentiel. Et si des percées scientifiques sont à la fois prévisibles et souhaitables, c'est à la gouvernance de l'IA qui revient le rôle primordial de déterminer comment ces avancées peuvent être encadrées pour assurer leur caractère bénéfique et équitable à l'échelle mondiale.

Cet engagement à contribuer au développement d'une IA responsable est profondément ancré dans la communauté Mila. Dès 2018, Mila menait, avec d'autres, la *Déclaration de Montréal pour un développement responsable de l'intelligence artificielle* qui visait à orienter le développement éthique de l'IA en formulant des principes clés ayant une forte légitimité démocratique. En 2020, dans le cadre du processus menant à l'adoption de la *Recommandation sur l'éthique de l'intelligence artificielle* de l'UNESCO, Mila a co-mené le *Dialogue inclusif sur l'éthique de l'IA*. Plusieurs membres de la communauté Mila mènent aussi des recherches à l'intersection de l'IA et du développement durable, de la santé, de l'équité, de l'éthique et de la gouvernance. Enfin, Mila dirige des projets appliqués dans plusieurs domaines afin de tirer parti de la puissance de l'IA pour soutenir la réalisation des ODD. Il s'agit entre autres de projets mobilisant l'IA pour lutter contre la crise climatique, soutenir la prévention de la traite des personnes et de l'esclavage moderne, contribuer à l'élaboration de politiques d'IA inclusives ainsi que pour détecter les préjugés et biais liés au genre dans les textes écrits.

Cette nouvelle collaboration avec l'UNESCO est un exemple additionnel de la conviction qui anime la communauté Mila. Les perspectives des auteurs et autrices, issus de divers horizons disciplinaires, géographiques et professionnels, y convergent pour amplifier la portée de ces réflexions. Les chapitres explorent, par exemple, les implications du développement de l'IA pour les communautés autochtones et les personnes LGBTI, le besoin urgent d'égalité entre les sexes dans les écosystèmes de l'IA, la réduction des inégalités par l'accès aux connaissances liées à l'IA, ainsi que les moyens de s'assurer que l'IA soutienne la réalisation de l'*Agenda 2030 pour le développement durable.*

Enfin, nous sommes fiers et reconnaissants de présenter cette publication aux côtés de l'UNESCO, qui a joué un rôle de premier plan sur l'éthique de l'IA avec l'adoption du premier instrument normatif mondial dans ce domaine. Nous remercions les contributeurs et contributrices pour leur engagement, et espérons que les membres des milieux de la politique, de la société civile et de la recherche pourront désormais s'y intéresser.

VALÉRIE PISANO
Présidente et chef de la direction, Mila – Institut québécois d'intelligence artificielle

TABLE DES MATIÈRES

INTRODUCTION

L'intelligence artificielle fait désormais partie de notre vie quotidienne. Elle est utilisée dans un large éventail de domaines comme la santé, les transports et la cybersécurité, influençant notre façon de communiquer, de travailler et d'apprendre. L'IA offre de grandes possibilités, mais pose également des risques qui étaient encore imprévus il y a seulement quelques décennies. Sa gouvernance est devenue une priorité mondiale, mobilisant les universités, les gouvernements, la société civile et les organisations internationales. Alors que le développement de l'IA continue de s'accélérer, il est à prévoir que ses effets sur nos sociétés s'amplifieront. Dans ce contexte, des conversations mondiales et inclusives sont essentielles pour nous aider à faire la lumière sur ces défis et à imaginer de nouvelles façons de les comprendre puis de les relever.

C'est pourquoi Mila et l'UNESCO ont uni leurs forces dans le but de favoriser un échange multipartite sur les questions importantes qui doivent être abordées pour appuyer le développement responsable de l'IA. Pour assurer l'inclusion de diverses perspectives, nous avons publié un appel mondial à contributions. Cette publication est une compilation de 18 chapitres sélectionnés, lesquels dépassent les frontières disciplinaires, culturelles et géographiques. Ils présentent les points de vue d'un large éventail d'acteurs dont des membres des milieux universitaire, politique et de la société civile. L'objectif était d'élargir la conversation pour porter notre attention non pas sur ce que nous comprenons déjà, mais plutôt sur ce que nous ne savons pas encore : les angles morts de l'intelligence artificielle. En d'autres termes, cette publication vise à offrir un espace où se côtoient des points de vue divergents, nouveaux et nuancés sur la gouvernance de l'IA telle qu'elle est vécue et comprise par les nombreux acteurs qui contribuent à son développement et à sa mise en œuvre.

Ce livre offre des réflexions et des propositions à l'égard de nombreux sujets importants. On y trouve notamment des chapitres sur les risques et les possibilités de l'IA pour les droits des peuples autochtones (Toki et Phillips), sur les effets potentiels de ces développements technologiques sur les communautés LGBTI (Horner), sur les moyens par lesquels la gouvernance de l'IA peut garantir la prise en compte de diverses voix, y compris celles des pays du Sud (Png) et sur les tensions qui peuvent surgir entre différents principes éthiques lorsque l'IA est mobilisée en période de pandémie (Voarino et Régis). D'autres chapitres présentent des comptes rendus stimulants des effets de l'IA sur les discussions législatives et l'élaboration des politiques (Clement-Jones ; i4Policy), sur les possibilités d'utiliser l'IA pour soutenir le développement durable (Letouzé *et al.* ; Shawe-Taylor *et al.*) et sur l'innovation inclusive (Future Earth ; Dieng). Les contributions nous invitent aussi à explorer les défis et les possibilités que présentent l'équité et l'éthique au sein de l'industrie de l'IA (Farnadi *et al.*) et au rôle de l'IA en matière d'éducation (Mishra *et al.*). Enfin, certaines des plus grandes voix mondiales du domaine proposent leurs réflexions sur la manière dont les écosystèmes d'IA pourraient mieux soutenir l'innovation à des fins socialement bénéfiques (Bengio *et al.*), sur les effets importants et parfois dévastateurs des préjugés dans les données (Crawford), sur la centralisation du pouvoir décisionnel (Ng et Brynjolfsson) et sur la nécessité de repenser les audits des systèmes d'IA par des tiers (Algorithmic Justice League).

Avec cette publication, nous espérons offrir des perspectives fructueuses pour nous aider à faire face à l'immense tâche qui nous attend, soit celle de façonner le développement de l'IA de manière à ce que personne ne soit laissé de côté. Il faut donc travailler à la mise en place de systèmes d'IA centrés sur l'humain, inclusifs, éthiques et durables, ainsi que respectueux des droits humains et de l'État de droit. Cette publication est notre humble contribution à cet effort. Elle n'est en aucun cas exhaustive, et un grand nombre d'autres initiatives et conversations seront nécessaires pour que le monde puisse tirer parti des possibilités qu'offre l'IA, tout en endiguant les risques qu'elle pose. Nous espérons que les réflexions stimuleront des discussions sur certains des défis les plus urgents en matière de gouvernance de l'IA et fourniront des idées nouvelles pour soutenir le développement responsable de l'intelligence artificielle.

CHANGEMENT DE L'EXTÉRIEUR : VERS DES AUDITS DE TIERCE PARTIE FIABLES EN MATIÈRE D'IA

INIOLUWA DEBORAH RAJI

Candidate au doctorat en sciences informatiques à l'UC Berkeley et chercheuse à l'Algorithmic Justice League.

SASHA COSTANZA-CHOCK

Directrice de recherche et conception à l'Algorithmic Justice League.

JOY BUOLAMWINI

Fondatrice et directrice exécutive à l'Algorithmic Justice League.

Au nom de l'Algorithmic Justice League

ODD 9 - Industrie, innovation et infrastructure

ODD 10 - Inégalités réduites

ODD 11 - Villes et communautés durables

ODD 16 - Paix, justice et institutions efficaces

ODD 17 - Partenariats pour la réalisation des objectifs

CHANGEMENT DE L'EXTÉRIEUR : VERS DES AUDITS DE TIERCE PARTIE FIABLES EN MATIÈRE D'IA

RÉSUMÉ

Lorsque des systèmes d'intelligence artificielle (IA) causent des préjudices, il est important d'établir les parties prenantes en cause et de les tenir responsables. Les audits en matière d'IA sont un mécanisme de responsabilisation gagnant récemment en popularité, et les auditeurs et auditrices de l'IA font partie d'un écosystème en croissance. L'« audit en matière d'IA », désigne dans ce chapitre le processus par lequel une partie évalue un système ou un produit d'IA selon un ensemble particulier de critères, puis fournit des conclusions et des recommandations. Ces audits servent à déterminer si les systèmes d'IA répondent aux objectifs de rendement ou encore s'ils soulèvent des préoccupations relativement à des préjugés et autres préjudices, à la protection des données et à la confidentialité, à la transparence et la responsabilité, au respect des normes et des obligations réglementaires, aux pratiques de travail ou aux impacts environnementaux. Dans le domaine de l'IA ont cours des audits de première partie (internes), de seconde partie (menés en sous-traitance) ou de tierce partie (externes, menés de manière entièrement indépendante). De tierces parties, chercheuses ou chercheurs indépendants, journalistes d'enquête, organismes défendant les intérêts de la communauté, cabinets d'avocats et organismes de réglementation, ont mené bon nombre des audits en matière d'IA les plus percutants à ce jour. Or, malgré l'importance de leur rôle dans la responsabilisation de l'IA, ces tierces parties ont été largement négligées dans les politiques.

Dans ce chapitre, nous proposons sept interventions clés pour renforcer la capacité des tierces parties à examiner les systèmes d'IA : la protection juridique quant à l'accès des tierces parties, la certification des auditeurs et auditrices, l'élaboration de normes visant les produits d'IA, le signalement des incidents préjudiciables, la divulgation publique obligatoire de l'usage de systèmes d'IA, le passage de cadres axés sur les préjugés de l'IA à des cadres axés sur les préjudices de l'IA et la mise au point de mécanismes de responsabilisation garantissant des interventions par audit appropriées.

> En relevant ces « chaînons manquants », nous espérons faire progresser la réglementation qui protège et soutient la capacité des profanes, telles les tierces parties et d'autres parties prenantes, à examiner les systèmes d'IA. Nous estimons que des audits de tierce partie valables aideront à protéger les droits fondamentaux des communautés les plus susceptibles d'être lésées par l'usage de systèmes d'IA.

INTRODUCTION

Les systèmes d'intelligence artificielle (IA) sont trop souvent conçus et utilisés de façon telle qu'ils reproduisent des formes existantes d'inégalité systémique et qu'ils causent de véritables torts, en particulier à des groupes marginalisés. Les préjudices causés par les systèmes d'IA sont de plus en plus manifestes. Ils sont désormais bien recensés dans les publications de recherche, et présents dans la culture populaire ainsi que dans les discussions et les propositions qui émergent relativement à l'élaboration des politiques. Malgré la sensibilisation croissante du public quant à des préjudices potentiels ou réels, ces problèmes restent cependant difficiles à cerner, à évaluer et, en fin de compte, à résoudre.

Nous écrivons ce chapitre à titre de chercheuses de l'Algorithmic Justice League, une organisation qui a pour mission de susciter une prise de conscience quant aux répercussions de l'IA, de fournir à cet égard des résultats de recherches empiriques, de faire entendre des groupes particulièrement touchés par les effets de l'IA, et d'inciter les chercheurs et chercheuses, les décideurs et décideuses ainsi que les gens de l'industrie à réduire les préjugés de l'IA et à atténuer les préjudices qu'elle cause.

Dans ce chapitre, nous nous concentrons sur un mécanisme de responsabilisation qui a selon nous été sous-estimé et sous-utilisé à ce jour : les audits en matière d'IA. Parmi les nombreux mécanismes pouvant assurer une responsabilisation dans le domaine de l'IA, les audits ont la capacité exemplaire d'aiguiser la conscience publique et de mener à des rappels de produits effectifs, à des mesures réglementaires ainsi qu'au règlement fructueux de litiges. Cependant, plusieurs des parties prenantes externes qui effectuent des audits de tierce partie et s'efforcent de protéger des communautés contre d'éventuels préjudices causés par l'IA sont elles-mêmes vulnérables aux représailles de puissantes entreprises technologiques. Ces auditeurs et auditrices formant un groupe diversifié, provenant du milieu universitaire de la recherche, du secteur privé, d'organisations non gouvernementales, de cabinets d'avocats, d'organismes de réglementation ou d'autres organismes du secteur public, sont souvent laissés à eux-mêmes quand ils se demandent comment procéder. Ils et elles doivent concevoir et exécuter des audits sans beaucoup de conseils, de soutien ou de protection de la part des décideurs et décideuses.

Nous présentons donc dans les pages qui suivent certaines interventions nécessaires en matière de politiques pour permettre et affermir les audits de tierce partie de systèmes d'IA. À l'heure actuelle, le rôle de ces tierces parties dans l'écosystème plus large de la responsabilité en matière d'IA fait l'objet d'une considération limitée dans les politiques publiques. Les décideurs et décideuses ont un important rôle à jouer pour assurer la protection et le soutien continus de ceux et celles qui choisissent de jouer ce rôle essentiel. Dans ce chapitre, nous commençons par définir les termes de base, puis décrivons sept interventions stratégiques déterminantes qui sont nécessaires à la tenue d'audits de tierce partie efficaces. Ces interventions concernent :

1) la protection juridique des tierces parties menant des audits en IA ;

2) la certification des auditeurs et auditrices ;

3) l'élaboration de normes visant les produits d'IA ;

4) le signalement des incidents dans lesquels l'IA a causé des préjudices ;

5) la divulgation publique obligatoire de l'usage de systèmes d'IA ;

6) le passage de cadres axés sur les préjugés de l'IA à des cadres axés sur les préjudices de l'IA ; et

7) la mise au point de mécanismes de responsabilisation visant à s'assurer que les résultats des audits mènent à des changements.

Nous espérons que ce travail fournira un point de départ à une discussion fort nécessaire sur les interventions qui visent à soutenir les tierces parties menant des audits, discussion qui s'inscrit dans le contexte élargi d'une politique de responsabilisation de l'IA.

CONTEXTE

Pour commencer, nous résumerons brièvement plusieurs concepts clés afin de préciser ce que nous entendons par audits en matière d'IA. Ensuite, nous décrirons l'écosystème émergent dont font partie les auditeurs et auditrices de l'IA en distinguant les audits de première, de seconde et de tierce partie puis en résumant quelques-unes des politiques qui façonnent le paysage actuel. Nous mettrons en évidence certaines des avenues prometteuses que les décideurs et décideuses ont explorées jusqu'à présent, pour finalement souligner les chaînons manquants dans l'écosystème de responsabilité de l'IA.

Dans le discours public ainsi que dans les milieux politiques, on a tendance à utiliser les expressions « intelligence artificielle », « systèmes de prise de décisions automatisée », « systèmes algorithmiques » et « apprentissage automatique » de manière quelque peu interchangeable et fâcheusement vague. Dans ce chapitre, nous utilisons « systèmes d'IA » en tant que générique et faisons occasionnellement référence à d'autres expressions quand la précision est nécessaire (Richardson, 2021). Nous utiliserons « systèmes d'IA » pour désigner une gamme de systèmes sociotechniques qui automatisent entièrement ou partiellement des processus mettant en cause le traitement de l'information et la reconnaissance de formes. La plupart des systèmes d'IA en usage sont conçus au moyen de techniques d'apprentissage automatique et donc fortement influencés par les données d'entraînement qui façonnent leurs extrants. Ils visent à imiter ou à automatiser certains processus cognitifs même si, dans la pratique, la plupart sont mis en œuvre pour exécuter une tâche précise tels que la classification, le classement ou l'identification. Nous reconnaissons que l'expression « système d'IA » est imprécise, mais nous avons choisi d'employer le vocabulaire le plus courant dans les discussions actuelles sur les politiques afin d'ancrer nos recommandations dans ce contexte.

Qu'entendons-nous par audits en matière d'IA ?

En dehors de l'industrie de l'IA, l'audit s'est progressivement imposé dans plusieurs domaines comme un mécanisme de responsabilisation courant. Contrairement à d'autres formes d'assurance contre le risque, telles les études d'impact ou les listes de contrôle, l'audit a tendance à faire appel à la précision lorsqu'un système est évalué par rapport à une norme connue. En tant qu'évaluation *post hoc* du système, il peut fournir des constats explicites nets quant aux limites et aux risques. En déterminant si une organisation ou un produit est conforme aux exigences ou non, les audits contribuent à déterminer s'il faut sélectionner un fournisseur, si un produit est prêt pour son lancement ou si un autre doit être

rappelé. Les audits constituent un instrument essentiel que des groupes peuvent éventuellement utiliser pour critiquer et influencer les personnes ayant le pouvoir décisionnel sur des systèmes d'IA qui ont une incidence sur eux (Wieringa, 2020).

Par « audit en matière d'IA », nous entendons dans ce chapitre un processus par lequel une partie évalue un système ou un produit d'IA selon un ensemble particulier de critères, puis fournit des conclusions et des recommandations à l'organisation auditée, au public ou à une autre entité, par exemple à un organisme de réglementation ou à la cour en tant qu'élément de preuve dans une action judiciaire. Ces audits servent à déterminer si les systèmes d'IA répondent ou non aux attentes, que ce soit relativement à des objectifs de rendement déclarés (par exemple, la précision de la prédiction ou de la classification) ou à d'autres préoccupations telles que : les préjugés et la discrimination (divergence de rendement entre divers groupes de personnes), la protection des données, la confidentialité, la sécurité et le consentement, la transparence, l'explicabilité et la responsabilité, le respect des normes, des principes éthiques et des obligations légales et réglementaires ou encore les pratiques de travail, la consommation d'énergie ou les impacts environnementaux.

Idéalement, les audits des systèmes d'IA, comme ceux qui ont cours dans d'autres domaines, devraient être menés par des parties officiellement reconnues par un organisme de certification crédible (mais il faut minutieusement définir les processus de certification afin d'éviter que l'industrie les accapare ou que les chercheuses et chercheurs indépendants en soient exclus). Les audits devraient également être menés de manière à répondre à des attentes précises, généralement énoncées sous forme de normes clairement définies et largement reconnues. Il faudrait de plus encourager les auditeurs et auditrices à exposer leurs conclusions (sans révéler de renseignements personnels) et les protéger de diverses manières des réactions hostiles de certaines sociétés devant leurs conclusions. Comme nous en discutons longuement plus loin dans le chapitre, la protection dont bénéficient habituellement les tierces parties dans d'autres domaines semble encore absente en IA.

Écosystème émergent des audits de première, de seconde et de tierce partie en IA

À mesure que l'on comprend les préjugés de l'IA et les préjudices qu'ils causent, les audits en matière d'IA se popularisent. Il existe trois grandes catégories d'audits, de première partie, de seconde partie et de tierce partie, selon qui les exécute. Ceux et celles qui mènent un audit de première partie sont employés par l'organisation qui conçoit le système d'IA. Ils veillent en interne à ce que celui-ci se conforme aux objectifs de rendement définis par la direction. De nombreuses entreprises technologiques de premier plan ont formé ou sont en train de mettre en place des équipes internes d'« éthique de l'IA » qui font en effet figure de première partie. Mentionnons à titre d'exemple l'équipe Facebook's Society and AI Lab (SAIL), l'équipe Fairness, Accountability, Transparency, and Ethics (FATE) de Microsoft, l'équipe ML Ethics, Transparency and Accountability (META) de Twitter, le groupe Justice by Design de Paypal ainsi que les équipes Ethical AI et Responsible Innovation de Google.

Aux commandes d'audits de seconde partie se trouvent des consultants et consultantes ou des collaborateurs et collaboratrices généralement embauchés en sous-traitance par les entreprises qui souhaitent confier la tâche d'audit à des spécialistes qualifiés ou bénéficier d'une nouvelle perspective. Ils et elles évoluent dans le secteur en pleine croissance des fournisseurs de services d'audit en matière d'IA. Ces fournisseurs sont aussi bien de petites entreprises émergentes ou sociétés de conseil (telles que ORCAA et Parity) que des équipes travaillant pour de grands cabinets-conseils (tels que Deloitte, McKinsey et Accenture) et proposant un examen des produits d'IA d'autres sociétés sous l'angle éthique, juridique ou technique. Certaines équipes ayant commencé à mener des audits de première partie en IA au sein de grandes entreprises technologiques ont ensuite proposé d'agir en tant que seconde partie pour d'autres sociétés, notamment des équipes d'audit de Google (Simonite, 2020) et d'IBM (IBM, 2021). Les auditeurs et auditrices agissant en tant que première ou seconde partie entretiennent une relation contractuelle avec l'organisation ciblée par l'audit.

Ceux et celles qui mènent des audits de tierce partie n'entretiennent en revanche aucune relation contractuelle avec l'organisation ciblée. Puisqu'ils et elles sont étrangers à celle-ci, ils et elles ont souvent un accès limité au système audité, ce qui limite certaines techniques d'audit. Ils et elles conservent cependant une indépendance totale, ce qui leur permet de poser des questions épineuses sur les extrants du système ou de présenter des conclusions négatives.

Les audits de tierce partie sont le travail de personnes extérieures à l'organisation ciblée par l'audit. Ces parties prenantes externes scrutent les organisations et les systèmes audités. Il s'agit notamment de chercheuses et chercheurs indépendants, d'équipes de journalistes d'enquête (travaillant par exemple au sein de The Markup ou de l'Associated Press), d'organisations de la société civile, de cabinets d'avocats et, dans certains cas, d'organismes de réglementation. Ils et elles mènent des vérifications externes indépendantes sur les préjudices causés par les systèmes d'IA sans être liés par une obligation contractuelle envers l'entité qui conçoit, vend ou exploite le système d'IA en question. La relation entre les tierces parties et les organisations ciblées par un audit peut être conflictuelle par moments, mais ce n'est pas toujours le cas.

Nous remarquons également l'émergence d'un certain nombre de centres universitaires et de groupes de recherche qui se penchent sur les aspects sociaux, techniques et juridiques associés aux préjudices causés par des algorithmes. Aux États-Unis, il y a notamment l'AI Now Institute de l'Université de New York et le Center for Critical Internet Inquiry de l'Université de Californie à Los Angeles. Le Royaume-Uni n'est pas en reste avec l'Institute for Ethics in AI de l'Université d'Oxford, l'Ada Lovelace Institute et bien d'autres. Certains de ces centres de recherche effectuent des audits formels de systèmes d'IA, en tant que seconde partie, travaillant pour des entreprises en vertu d'un contrat signé, ou à titre de tierce partie.

Effet des audits de tierce partie

L'indépendance, la crédibilité et l'intégrité des auditeurs et auditrices sont depuis longtemps source de préoccupation dans d'autres domaines, notamment financier (AICPA, 2017) et environnemental (Gunningham, 1993) ou liés à la sécurité ou à la salubrité des aliments (Lytton et McAllister, 2014). Les personnes ou organisations qui mènent des audits de première ou de seconde partie sont soumises aux conditions fixées par les organisations ciblées. Seules les tierces parties sont libres d'agir indépendamment des demandes formulées par celles-ci. Ainsi, puisqu'elles sont libres d'obligations contractuelles ou de conflits d'intérêts, elles peuvent poser des questions épineuses que d'autres ne poseraient pas. S'il le faut, elles peuvent agir à l'encontre des préférences de l'organisation ciblée pour la tenir responsable.

De tierces parties ont mené bon nombre des audits en matière d'IA des plus percutants à ce jour. Par exemple, les journalistes d'enquête de ProPublica ont démontré la présence de préjugés raciaux dans l'évaluation du risque de récidive, ce qui a fait réagir le fournisseur du système en cause et a plus largement remis en question le recours aux outils d'évaluation du risque dans le système judiciaire (Angwin *et al.*, 2016). L'enquête de ProPublica a également exposé la manière dont Facebook avait permis un ciblage publicitaire discriminatoire en matière d'emploi, de vente et de location (Gillum et Tobin, 2019), une histoire qui a conduit à un règlement de cinq millions de dollars et à des modifications des systèmes publicitaires de cette plateforme (Spinks, 2019). Des enquêtes ultérieures de l'organisation The Markup ont néanmoins révélé que le problème persistait (Keegan, 2021). The Markup a également montré comment les algorithmes liés à l'admission dans des écoles de la ville de New York reproduisent une ségrégation raciale (Lecher et Varner, 2021) et comment les personnes racisées essuient deux fois plus de refus de la part des sociétés de financement hypothécaire que les personnes blanches (Martinez et Carollo, 2021). Les chercheuses indépendantes Joy Buolamwini, Inioluwa Deborah Raji et Timnit Gebru ont fait ressortir des disparités quant au degré de précision des technologies de reconnaissance faciale vendues par d'importants fournisseurs de ce monde relativement

au genre de la personne et à la couleur de sa peau (Buolamwini et Gebru, 2018), un constat qui a conduit IBM, Amazon et Microsoft à instaurer un moratoire pour une durée indéterminée sur la vente de tels outils de reconnaissance faciale à la police (Raji et Buolamwini, 2019). Ce constat a également fondé une plainte de l'American Civil Liberties Union (ACLU) contre la police de Détroit après l'arrestation injustifiée d'un homme noir, Robert Williams, en raison d'un mauvais appariement opéré par la technologie de reconnaissance faciale (Hill, 2020). Environ à la même période, le National Institute of Standards and Technology a audité l'algorithme de 189 logiciels de 99 sociétés conceptrices pour constater que la plupart d'entre eux fonctionnaient considérablement moins bien pour les personnes racisées que pour les personnes blanches (Grother *et al.*, 2019). De manière semblable, le cabinet d'avocats Foxglove a poussé le gouvernement britannique à revoir sa position sur l'attribution de notes à des étudiants et étudiantes par un algorithme, par suite d'une analyse ayant montré que cet algorithme désavantageait les personnes à faible revenu (Foxglove, 2020).

Ces exemples montrent clairement que les audits de tierce partie constituent des interventions importantes en matière de responsabilisation. De tels audits des systèmes d'IA ont porté une attention soutenue et croissante aux préjugés, aux préjudices, à l'équité et à la responsabilité en ce qui a trait aux systèmes d'IA. Nous pensons cependant que les politiques actuelles relatives à l'IA ont fortement sous-estimé ou même ignoré l'importance des audits de tierce partie dans le développement équitable et imputable des systèmes d'IA.

Politiques qui façonnent le paysage des audits en matière d'IA

L'orientation des politiques qui régissent et façonnent la mise au point de systèmes d'IA évolue rapidement. Bien que l'Algorithmic Justice League n'exerce pas de pression quant à des projets de loi en particulier, notre organisation suit de près les changements législatifs qui surviennent, en particulier aux États-Unis et au sein de l'Union européenne. Nous nous réjouissons de la présentation de divers projets de règlements visant à régir des pans de l'IA qui n'ont encore fait l'objet d'aucun contrôle, même si nous pensons que les audits de tierce partie n'ont pas reçu une attention suffisante.

Récemment, des organismes de réglementation comme l'Information Commissioner's Office (ICO) du Royaume-Uni, soit le principal organe qui réglemente la protection des données au pays, sont entrés dans la mêlée avec des directives sur la façon dont les entreprises et les organismes gouvernementaux pouvaient auditer leurs systèmes. En réaction au Règlement général sur la protection des données (RGPD), l'ICO a publié des directives pour aider les entreprises à voir comment elles devraient rendre compte de la gestion des données et des algorithmes (Kazim et Koshiyama, 2020). Les directives de l'ICO représentent un pas dans la bonne direction même si, à notre avis, elles mettent trop l'accent sur le cadre relatif à la protection des données, qui accapare la discussion politique au sein de l'Union européenne. Elles font du contrôle des renseignements personnels le principal levier de commande des systèmes d'IA, ce qui ne devrait pas toujours être le cas. Évaluer que les données sont traitées de manière responsable nécessite un accès direct à la cible d'un audit. Ainsi, les interventions du Royaume-Uni et de l'Union européenne tendent vers la formulation de conseils sur la façon dont les auditeurs et auditrices *internes,* ou premières parties, des entreprises privées pourraient influer sur la prise de décisions des équipes d'ingénierie ou de produits concernant la collecte, le stockage et l'utilisation de données. Bien que les audits de première ou de seconde partie s'avèrent utiles en IA, ils ne remplacent pas l'examen minutieux externe d'une tierce partie. De telles interventions réglementaires peuvent certes conduire à des pratiques d'ingénierie responsables, mais elles n'aident guère des observateurs et observatrices de l'extérieur à faire valoir leurs préoccupations particulières. L'*Algorithmic Accountability Act*, une loi adoptée en 2019 dans le contexte d'une revue de la législation aux États-Unis, souligne la nécessité de mener en interne des études d'impact relatives aux algorithmes semblables, à certains égards, aux études d'impact relatives à la protection des données mentionnées dans le RGPD, mais ayant sans doute une portée élargie étant donné qu'on ne s'intéresse pas qu'aux mesures de protection des données (United States Congress, 2019). Il s'agit encore une fois d'un pas dans la bonne

direction, mais cette loi se concentre malheureusement sur l'élaboration de telles études en interne plutôt que d'exiger la vérification d'une tierce partie. De plus, bien qu'elle propose que les entreprises fournissent des détails sur leurs systèmes aux organismes de réglementation, la loi de 2019 n'exige pas la divulgation publique des études d'impact relatives aux algorithmes (MacCarthy, 2019). En outre, celles-ci s'assimilent davantage à des outils internes de réflexion ouverte qu'à des évaluations strictes de la conformité. Alors que plusieurs propositions d'études d'impact soutiennent le besoin d'une participation communautaire accrue (Metcalf *et al.*, 2021), la mise en place effective de telles analyses ne tient généralement pas compte d'un large éventail de points de vue et ressemble en fait à un audit interne de première partie (Selbst, 2021).

Aux États-Unis, d'autres initiatives récentes, telles que l'adoption de l'*Algorithmic Justice and Online Platform Transparency Act of 2021* (Markey, 2021) et de l'*Automated Decision Systems Accountability Act of 2021* (Cuevas, 2020), montrent une prise de conscience accrue quant au rôle des organismes fédéraux en tant que tierces parties menant des audits, en particulier à celui de la Federal Trade Commission. Il s'agit d'une évolution positive vers la surveillance externe des systèmes d'IA, même si les détails pratiques restent ambigus. Dans l'Union européenne, les récentes politiques de responsabilisation relatives à l'IA visant les entreprises de médias sociaux ont également adopté le langage de l'audit. En particulier, le *Online Harms Bill* du Royaume-Uni et le *Digital Services Act* de la Commission européenne mentionnent explicitement la nécessité d'un examen indépendant des systèmes d'IA employés (United Kingdom Government, 2020 ; European Commission, 2022). De telles interventions restent cependant limitées. Par exemple, les auditrices et auditeurs indépendants décrits à l'article 28 du *Digital Services Act* sont en réalité des secondes parties, à savoir des consultantes ou consultants embauchés par l'organisme ciblé pour exécuter le mandat d'examen. L'article 31 décrit davantage une participation de tierces parties, mais en restreint la définition aux chercheurs et chercheuses universitaires ainsi qu'aux organismes de réglementation, excluant par omission d'autres éventuelles tierces parties comme les journalistes d'enquête, les cabinets d'avocats ou les organisations de la société civile. Le *Digital Services Act* mentionne effectivement la nécessité de contrôler les auditeurs et auditrices, mais ne prévoit ni mécanisme de certification clair des auditeurs et auditrices ni normes précises en fonction desquelles les audits des systèmes d'IA devraient être menés.

Ainsi, malgré de récentes initiatives encourageantes quant aux politiques de responsabilisation en matière d'IA, nous sommes encore loin d'un écosystème qui permet la participation effective de tierces parties à des audits en IA. Les normes et le cadre réglementaire dont nous avons besoin pour garantir que les tierces parties ont la certification, la protection et le soutien nécessaires pour jouer leur rôle ne sont pas en place. Pour assurer l'équité et la responsabilité dans le déploiement des systèmes d'IA, les communautés les plus susceptibles d'en subir des inconvénients doivent être mieux représentées dans le processus d'audit, d'examen ou d'évaluation. Les tierces parties pouvant jouer ce rôle doivent être certifiées et soutenues par un ensemble de politiques qui garantissent leur indépendance, leur intégrité et leur efficacité. Dans le reste de ce chapitre, nous présentons sept chaînons manquants de la politique en matière d'IA qui, selon nous, sont nécessaires pour aider les tierces parties à effectuer leur travail d'audit.

CHAÎNONS MANQUANTS

De multiples interventions sont nécessaires pour faire en sorte que les tierces parties jouent leur rôle dans les audits en matière d'IA. Malgré de récentes initiatives, nous n'avons encore vu aucune action législative, même sous forme de proposition, qui satisfait au critère de base d'un contrôle efficace. Nous proposons ici sept interventions clés qui renforceront à notre avis le discours actuel relativement aux politiques :

1. **Protection juridique quant à l'accès des tierces parties.** Une fois leur écosystème solidifié, les auditrices et auditeurs certifiés doivent être en mesure d'effectuer leur travail. Nous avons besoin d'instruments de politiques qui offrent aux tierces parties l'**accès protégé à l'information** dont elles ont besoin pour mener des évaluations indépendantes des systèmes d'IA, selon leurs différentes priorités et préoccupations.

2. **Certification et formation des auditeurs et auditrices.** Un **processus formel de certification** des auditeurs et auditrices de première, de seconde et de tierce partie est une condition préalable pour soutenir l'accès, garantir l'intégrité des parties et assurer la qualité de l'audit. Les auditeurs et auditrices de systèmes d'IA doivent être évalués par un organisme de certification qui s'assure qu'ils adhèrent à des attentes nationales ou internationales inclusives en matière de conduite et de compétence. Cela dit, nous faisons une mise en garde : il faut prévoir les processus de certification de manière à éviter que l'industrie les accapare ou que les chercheuses et chercheurs indépendants en soient exclus.

3. **Élaboration de normes.** Des **normes claires et largement reconnues quant aux produits d'IA,** qui traduisent des attentes de haut niveau relativement à ces systèmes et à leur utilisation, doivent être établies. Des normes clairement définies et élaborées selon un processus transparent sont une condition préalable à des audits valables en matière d'IA.

4. **Suivi des incidents préjudiciables.** Aucun système d'IA n'est parfait. Il est nécessaire de **signaler et de suivre les incidents dans lesquels l'IA cause des préjudices** afin de faire en sorte que les personnes touchées par les systèmes d'IA partagent leur expérience et leurs préoccupations à cet égard. Normaliser le suivi des incidents préjudiciables implique une véritable compréhension des problèmes, l'amélioration des systèmes par les fournisseurs et les exploitants, une meilleure surveillance des organismes de réglementation, des actions judiciaires, si nécessaire, et une sensibilisation accrue du public par rapport à de tels incidents, par exemple par une meilleure couverture dans la presse.

5. **Généralisation des avis d'utilisation.** Une politique de responsabilisation en matière d'IA devrait inclure la **divulgation publique obligatoire quant à l'usage de tout système d'IA** susceptible de causer des préjudices. Les organismes publics, en particulier, doivent être tenus d'informer le public lorsqu'ils achètent, pilotent ou déploient des systèmes d'IA. La divulgation publique permet aux tierces parties de repérer des cibles d'audit. Elle représente de plus une exigence de base pour obtenir le consentement éclairé des personnes qui utiliseront ces systèmes d'IA ou seront touchées par leurs effets.

6. **Recadrage pour aller au-delà des préjugés de l'IA.** Une politique en matière d'IA devrait aborder un large éventail de préjudices causés par l'IA plutôt que de se concentrer uniquement sur des mesures techniques liées à l'exactitude et aux préjugés. Elle nécessite également des **définitions valables** de tous les termes clés et la reconnaissance des multiples formes de préjudices.

7. **Mise en place de mécanismes de responsabilisation post-audit.** Les audits de tierce partie ne sont en fin de compte utiles que si plusieurs mécanismes mènent à la résolution des problèmes qu'ils mettent à jour. Une politique relative à l'IA devrait inclure divers **outils de mise en application** pour garantir que, par suite d'un audit, les entreprises en divulguent les principales conclusions, qu'elles cherchent à se conformer aux normes et à la loi, qu'elles apportent des améliorations et qu'elles réparent les préjudices.

Dans les sections suivantes, nous préciserons brièvement chacune de nos sept recommandations.

1. Accès et protection des tierces parties

Une fois leur écosystème solidifié, les auditrices et auditeurs certifiés doivent être en mesure d'effectuer leur travail. Nous avons besoin d'instruments de politiques qui offrent aux tierces parties un accès protégé aux données dont elles ont besoin pour mener des évaluations indépendantes des systèmes d'IA, selon leurs différentes priorités et préoccupations.

Les audits de tierce partie sont des mesures de responsabilisation nécessaires tout au long du cycle de vie des systèmes d'IA. Les tierces parties peuvent mettre en lumière des problèmes imprévus, sous-estimés ou ignorés par ceux et celles qui mettent au point, achètent, déploient ou entretiennent des systèmes d'IA. Elles peuvent également attirer l'attention sur diverses répercussions que subissent des parties prenantes marginalisées dont on tient souvent peu compte. Comme elles n'entretiennent aucune relation contractuelle avec l'organisation ciblée par l'audit, elles sont moins susceptibles d'être influencées par les préférences, les attentes ou les priorités de cette organisation. De plus, elles ont tendance à représenter un plus large éventail de points de vue que les parties prenantes internes et jettent donc un nouveau regard critique sur des questions clés qui pourraient autrement demeurer sous-estimées ou même ne pas être soulevées.

Un audit de tierce partie est celui que réalise un organisme d'audit indépendant du fournisseur et du client, et est exempt de tout conflit d'intérêts (ASQ, n.d.). Les audits de tierce partie jouent un rôle unique dans la responsabilisation relative aux algorithmes[1]. Les audits de première et de seconde partie, bien que potentiellement utiles, ne suffisent pas à assurer la mise au point de systèmes d'IA équitables et responsables. Ces types d'audits présentent des limites à certains égards. Par exemple, les audits de première ou de seconde partie ont souvent lieu seulement en réaction à un problème, après que celui-ci a été soulevé par les organismes de réglementation ou le public. Les auditrices et auditeurs agissant à titre de première partie divulguent rarement leurs conclusions au-delà de certaines équipes de leur entreprise. La diffusion publique des conclusions d'audits de seconde partie est généralement limitée par des accords de non-divulgation. La publication des conclusions d'audits de première ou de seconde partie en matière d'IA est particulièrement improbable lorsque la démarche révèle des problèmes importants quant aux préjugés de l'IA ou aux préjudices causés par celle-ci. Bien que ces types d'audits offrent parfois un accès illimité aux systèmes d'IA et qu'ils constituent donc un outil utile pour leur développement ou l'évaluation préalable à leur déploiement, ils ne sont pas à l'abri de l'influence de l'organisation ciblée, ce qui accentue la nécessité d'obtenir la perspective de tierces parties (Raji *et al.*, 2020).

Une politique en matière d'IA devrait veiller à ce que les tierces parties certifiées soient en mesure de recueillir les données dont elles ont besoin pour évaluer les systèmes d'IA sans craindre qu'on leur en bloque l'accès et surtout sans redouter que la cible de l'audit leur fasse subir des représailles juridiques. En effet, malgré une récente décision judiciaire[2] prévoyant une exception pour les chercheurs et chercheuses menant des études sur la discrimination algorithmique (Weigner, 2020), de tierces parties établies aux États-Unis ont déclaré à l'Algorithmic Justice League qu'elles craignaient des poursuites en vertu du *Computer Fraud and Abuse Act* lorsqu'elles risquent d'enfreindre les conditions d'utilisation d'un produit dans le cadre de l'audit qu'elles mènent (United States, 1984).

1. Les audits de tierce partie sont parfois être appelés audits « indépendants » ou « externes ». À notre avis, il est capital qu'ils soient réalisés par des personnes ou des organisations totalement indépendantes de la cible de l'audit. Il importe de faire la distinction entre les audits de tierce partie et ceux de seconde partie. Quand une entreprise qui conçoit ou utilise l'IA embauche un consultant, il s'agit d'un audit de seconde partie.

2. Pour de l'information sur la cause Sandvig c. Barr, intentée par l'ACLU, consulter le https://www.aclu.org/press-releases/federal-court-rules-big-data-discrimination-studies-do-not-violate-federal-anti.

En matière de politiques, nous recommandons donc les interventions suivantes :

- **Assurer la protection juridique des tierces parties.** Actuellement, les lois sur la fraude informatique, les abus et la cybersécurité (comme la *Computer Fraud and Abuse Act* aux États-Unis) peuvent rendre les tierces parties ou les parties remettant des faits en question vulnérables à des poursuites judiciaires relatives à la récupération de données qui sont importantes pour mener l'audit. De telles lois devraient toutes prévoir des exceptions pour les journalistes et les chercheurs et chercheuses universitaires.
- **Exiger des audits de tierce partie des systèmes d'IA** étant conçus, achetés ou employés par tout organisme gouvernemental ou bénéficiaire de fonds fédéraux.
- **Fournir un accès aux données aux tierces parties approuvées.** Les tierces parties ne sont pas rémunérées par l'organisation auditée et n'y sont pas liées contractuellement. Par conséquent, elles tendent à ne bénéficier que d'un accès limité à la cible de l'audit. Cela se manifeste souvent par un manque d'accès aux données et au code, ainsi que par un accès restreint à la documentation ou à des discussions avec les responsables de la conception des systèmes à propos des motifs qui sous-tendent leur prise de décisions. Nous avons observé que de nombreuses politiques incluent des directives relatives aux audits internes. Celles-ci sont potentiellement utiles, mais nous devons également constater des exigences faisant en sorte que les tierces parties approuvées bénéficient de l'accès nécessaire pour examiner avec succès les produits avant et après leur déploiement.
- **Soutenir et permettre les audits de tierce partie des systèmes d'IA.** Nous proposons que, plutôt que de se concentrer uniquement sur des directives internes en matière de responsabilité, les organismes de réglementation considèrent aussi leur rôle potentiel dans le soutien et l'habilitation des tierces parties. De récentes mesures visent à soutenir les parties s'intéressant à la réglementation ainsi que les universitaires, par exemple l'article 31 du *Digital Services Act* soumis au parlement européen (Ponce, 2020), l'exigence d'un examen de la Federal Trade Commission prévue dans l'Algorithmic Justice and Online Platform Transparency Act of 2021 (Markey, 2021) et les exigences énoncées dans le *California Automated Decision Systems Accountability Act* quant à la production, par des organismes étatiques, de rapports relatifs à la responsabilisation des systèmes de prise de décisions automatisée (Chau, 2020).
- **Renforcer les mesures de protection à l'égard des lanceurs et lanceuses d'alertes,** tant du secteur privé que du secteur public, qui se préoccupent de technologies et dénoncent des pratiques, des produits ou des services liés aux algorithmes violant des normes, des obligations légales ou réglementaires, les droits civils ou la législation en matière de droits humains.

2. Certification des auditeurs et auditrices

Un processus formel de certification des auditeurs et auditrices de première, de seconde et de tierce partie améliorerait la qualité et la fiabilité des audits en matière d'IA. Les auditeurs et auditrices de systèmes d'IA doivent être évalués par des organismes de certification qui s'assurent qu'ils respectent les normes internationales et qu'ils et elles ont la formation et la qualification adéquates. Cela dit, nous faisons une mise en garde : il faut prévoir les processus de certification de manière à éviter que l'industrie les accapare ou que les chercheuses et chercheurs indépendants en soient exclus.

Des auditeurs et auditrices de systèmes d'IA sont au cœur d'une industrie émergente, mais il existe peu de mécanismes de contrôle pour s'assurer qu'ils et elles sont à la fois qualifiés et vraiment indépendants des organisations ciblées par les audits. La politique actuelle en matière d'IA n'a pas réussi à imposer une surveillance publique ou une certification de ces auditeurs et auditrices. Les fournisseurs de systèmes d'IA sont donc actuellement en mesure d'embaucher n'importe qui pour effectuer un travail qu'ils appellent « audit », pour ensuite déclarer que leurs systèmes ont été audités. Notamment, de nombreux cabinets annoncent désormais qu'ils acceptent de passer de lucratifs contrats d'audit

de seconde partie avec des organismes gouvernementaux ou des entreprises du secteur privé. L'Algorithmic Justice League cartographie actuellement le paysage des auditeurs et auditrices de l'IA, et nous avons repéré 65 entités qui prétendent mener des audits de première, de seconde ou de tierce partie (Algorithmic Justice League, 2021).

Si une personne souhaite fournir des services médicaux, elle doit être autorisée à le faire par un ordre ou conseil des médecins. Si une autre veut pratiquer le droit, elle doit réussir l'examen du barreau de chaque endroit où elle veut travailler. Bien qu'il existe de multiples façons d'organiser la certification, nous pensons que celle qui vise les auditeurs et auditrices en IA est une pièce importante du casse-tête. Cette certification peut structurer le processus de vérification pour renforcer la confiance dans les systèmes d'IA, élargir l'accès accordé aux parties certifiées et garantir une qualité uniforme des audits.

Nous estimons également que la certification doit reposer sur des définitions et des normes claires indiquant ce qu'est un audit valable. En l'absence de telle mesure, la porte reste ouverte à de multiples scénarios dans lesquels les audits ne remplissent pas la fonction qu'ils devraient en matière d'IA. Les auditeurs et auditrices de l'IA peuvent fournir d'excellents services, mais également être plus ou moins compétents sur le plan technique, ou plus ou moins conscients des principaux facteurs sociaux, historiques, culturels ou contextuels. De nombreuses parties menant des audits en IA se concentrent uniquement ou principalement sur les aspects techniques d'une « égalité algorithmique » plutôt que de viser l'objectif plus noble de l'équité ou de réduire les préjudices que cause l'IA tout au long de son cycle de vie. Les audits désincarnés qui se concentrent uniquement sur le rendement d'un modèle, sans tenir compte également des contextes, des produits, de leurs exploitants et des communautés qui interagissent avec les systèmes du monde réel, ne peuvent traiter de manière adéquate les menaces et les dangers émergents. En outre, la qualité des données d'audit importe. Si les auditeurs et auditrices n'ont accès qu'à une partie des données n'étant pas représentatives du système examiné, ils et elles pourraient formuler des conclusions beaucoup trop optimistes ou pessimistes.

De plus, sans une certification indépendante, des parties bien intentionnées, frauduleuses ou opportunistes peuvent mener des audits « minces » ou faibles relativement aux algorithmes. Certaines mesures incitent fortement des entreprises à chercher des auditeurs ou auditrices qui poseront leur « sceau d'approbation » même quand les systèmes d'IA entraînent des préjudices concrets, ce qui s'appelle un blanchiment éthique (Bietti, 2020). Il existe également des obstacles à la surveillance d'un système après l'apposition d'un sceau d'approbation, ce qui constitue un problème parce que l'usage et le contexte évoluent continuellement et que peuvent survenir d'autres préjudices.

Ajoutons qu'il faut faire preuve de vigilance dans la mise en œuvre de la certification. Son processus doit être transparent et protégé de la mainmise des entreprises. Dans les pires cas, la certification pourrait servir à exclure indûment des organisations de la société civile qui seraient pourtant les mieux placées pour défendre les intérêts des groupes susceptibles de subir des préjudices associés à l'IA. Il faut examiner attentivement les mesures formelles de certification pour s'assurer qu'il y a une diversité de tierces parties menant des audits en IA.

En matière de politiques, nous recommandons donc les interventions suivantes :

- **Éduquer et former les auditeurs et auditrices.** Étant donné que des systèmes d'IA se déploient à l'échelle mondiale et dans tous les secteurs et qu'on emploie l'expression « auditeur ou auditrice en IA » à toutes les sauces, nous considérons que le groupe actuel de personnes et d'organisations ayant une expérience pratique de l'audit en matière d'IA est assez limité. Il faut un effort concerté pour former les parties et coordonner le secteur des audits en matière d'IA. Nous pensons que les décideurs et décideuses politiques devraient envisager un soutien à l'éducation et à la formation des parties menant des audits dans le domaine, et plus particulièrement des tierces parties. Une telle mesure renforcerait l'écosystème, y compris des organisations locales capables de mener des audits et de certifier que les systèmes d'IA respectent les normes convenues, se conforment aux obligations

réglementaires locales et nationales, et n'enfreignent pas la loi (notamment la législation sur les droits humains). Nous encourageons également l'élaboration de normes permettant aux auditeurs et auditrices d'évaluer les systèmes d'IA par rapport aux principes de l'entreprise (par exemple, les principes éthiques internes liés à l'IA), à des normes techniques internationales, ou encore aux attentes ou demandes de la communauté.

- **Investir dans des outils, modèles et procédures d'audit en matière d'IA.** Cet investissement aiderait à normaliser les pratiques dans le domaine de l'audit et à regrouper les attentes. La normalisation visant des outils et des processus particuliers améliorerait également l'éducation, la formation et le contrôle.
- **Évaluer l'indépendance des auditeurs et auditrices.** Ceux et celles qui participent à des audits de seconde ou de tierce partie peuvent se présenter à tort comme indépendants et ainsi camoufler leur affiliation à une entreprise ou le fait qu'ils et elles reçoivent un financement (Addalla et Abdalla, 2021), et donc éviter de divulguer des conflits d'intérêts. La certification devrait comprendre un examen des éventuels conflits d'intérêts pouvant entraver la capacité des auditeurs et auditrices à obtenir des mandats dans un contexte donné ou pour une cible en particulier.

Les processus formels de formation, de contrôle et de certification des auditeurs et auditrices forment une composante nécessaire d'une future politique en matière d'IA[3]. S'ils sont bien organisés, les organismes de certification reconnus qui évaluent les parties pouvant auditer des systèmes d'IA représentent un outil clé vers des systèmes équitables et responsables. Une politique relative à l'IA qui favorise la certification des parties garantirait que les systèmes d'IA sont conformes aux normes de l'industrie, aux obligations légales et réglementaires locales, nationales et internationales, et à la *Déclaration universelle des droits de l'homme*.

3. Normes

Une politique liée à l'IA doit soutenir l'élaboration de normes claires et largement reconnues quant aux produits et processus d'IA, qui traduisent des attentes de haut niveau relativement à ces systèmes et à leur utilisation. Des normes clairement définies et élaborées selon un processus transparent sont une condition préalable à des audits valables en matière d'IA.

Dans toute évaluation, une norme est nécessaire pour comparer le rendement réel d'un système à un rendement attendu ou idéal. Dans les audits en matière d'IA en particulier, l'auditeur ou auditrice doit pouvoir compter sur des normes clairement définies afin d'examiner si le système ou le produit répond aux attentes et de tenir les parties prenantes en cause responsables s'il s'en écarte dangereusement. Les normes jouent donc un rôle crucial dans l'établissement des exigences et des indicateurs de rendement nécessaires à des audits fiables. Parfois, ce sont des organismes gouvernementaux qui déterminent et appliquent ces normes, énoncées formellement dans des lois ou une réglementation. D'autres fois, ce sont des acteurs de l'industrie qui les élaborent, en négociant et en cherchant un consensus.

L'IA étant malheureusement un domaine peu réglementé et offrant peu d'occasions de fédérer l'industrie, les normes largement acceptées quant aux systèmes, produits et processus de l'IA restent rudimentaires ou inexistantes (Mittelstadt, 2019). Sur la scène internationale, de telles normes demeurent peu élaborées, mais apparaissent lentement. Des normes d'ingénierie de l'IA voient le jour au sein de plusieurs organismes de normalisation nationaux et internationaux, tels que, notamment,

3. Pour avoir un aperçu de l'approche de l'Organisation internationale de normalisation (ISO) en matière d'évaluation de la conformité, consulter le https://www.iso.org/fr/conformity-assessment.html.

le National Institute of Standards and Technology[4] des États-Unis (Cochrane, 1966), l'Organisation internationale de normalisation (ISO)[5] et l'Institute of Electrical and Electronics Engineers (IEEE)[6] (Shahriari et Shahriari, 2017). En l'absence de consensus sur les normes, les entreprises technologiques ont décidé d'énoncer publiquement des principes liés à l'IA (Jobin *et al.*, 2019) ou de prévoir en interne des critères relatifs à son déploiement (Raji *et al.*, 2020). Elles exposent ainsi leurs propres attentes éthiques concernant les systèmes d'IA. Au mieux, il s'agit de principes et de critères de haut niveau, difficiles à mettre en application et représentant des mesures volontaires d'autorégulation (Bietti, 2020). Fait important, ces déclarations de principes ne découlent pas nécessairement d'une réelle compréhension des préjudices et sont rarement élaborées en consultation ou en collaboration avec les communautés les plus touchées par ces préjudices (Metcalf et Moss, 2019). Les chercheurs et chercheuses universitaires ainsi que les organisations de la société civile ont également mis au point, séparément, des cadres qui traitent leurs préoccupations et leurs attentes en ce qui a trait aux systèmes sociotechniques, mais ces cadres restent aussi pour la plupart de haut niveau et difficiles à mettre en application par les équipes d'ingénierie (Krafft *et al.*, 2021).

Certaines propositions liées à la responsabilisation de l'IA se sont concentrées sur la certification des produits (IEEE Standards Association, 2019). Celle-ci s'avère utile dans certains contextes, mais soyons avertis : toute approche qui utilise les normes comme une « liste de contrôle » avant le déploiement présente d'évidentes limites. La conformité aux normes doit être considérée comme un seuil ou un point de départ pour attester le rendement minimal du produit, et non comme la finalité. Les normes et les points de référence servent à définir des attentes en matière de rendement de même qu'à répertorier et à préciser des préoccupations potentielles, et elles s'apparentent à l'expression d'une forme idéalisée des systèmes d'IA.

Voici nos recommandations à l'égard des normes :

- **Adopter des normes servant de guides, et non de listes de contrôle.** Idéalement, les normes doivent être suffisamment souples pour s'adapter à l'attitude du public et à la compréhension qu'il en a, qui évoluent. Elles guident éventuellement les audits, les études d'impact, les rapports d'incident et d'autres formes d'évaluation, mais ne doivent pas dicter automatiquement les conditions de déploiement des produits d'IA.
- **Élaborer des normes visant les processus, et pas que les résultats.** En plus des normes servant à évaluer des résultats (notamment le taux de précision de la prédiction ou de la classification), il est crucial d'adopter des normes axées sur les processus de mise au point des produits d'IA qui définissent les attentes à cet égard. Les normes liées aux processus doivent ainsi prévoir les meilleures pratiques d'ingénierie quant à la collecte et à l'utilisation consensuelles des données, des exigences à respecter pour la documentation (Gebru *et al.*, 2021 ; Mitchell *et al.*, 2019 ; Raji et Yang, 2019), des critères à atteindre avant le déploiement, des processus de signalement des incidents et d'intervention pour y réagir, et d'autres mesures d'évaluation. Par exemple, une technologie de reconnaissance faciale qui obtient de bons résultats au *Face Recognition Vendor Test* ne limite pas nécessairement les atteintes à la vie privée lors de la collecte de données (Learned-Miller *et al.*, 2020). Seules des normes prévoyant la manière de recueillir, de distribuer et d'utiliser des données préviennent de tels préjudices[7].

4. Cet organisme publie de l'information à jour sur les avancées liées à l'évaluation des systèmes d'IA : https://www.nist.gov/news-events/events/2021/06/ai-measurement-and-evaluation-workshop.
5. L'ISO présente les mises à jour concernant la norme ISO/IEC JTC 1/SC 42 sur l'IA au https://www.iso.org/fr/committee/6794475.html.
6. L'IEEE publie de l'information à jour sur l'IA au https://standards.ieee.org/initiatives/artificial-intelligence-systems/index.html.
7. De l'information détaillé sur la norme ISO-IEC 19794-5 sur les formats d'échange des données biométriques est accessible au https://www.iso.org/fr/standard/38749.html.

- **Mettre au point des normes visant une conformité réglementaire, et pas que certains éléments techniques.** Certains de ceux et celles qui mènent des audits de seconde partie en IA prétendent être en mesure d'évaluer le rendement d'un système par rapport à des points de référence techniques ou à sa conformité aux lois et à la réglementation[8]. Nous estimons que les parties autorisées à auditer les systèmes d'IA devraient être aptes à évaluer tant leur conformité aux normes techniques que le respect des lois locales, nationales et internationales, y compris de la *Déclaration universelle des droits de l'homme*. Elles devraient également être en mesure d'évaluer les systèmes d'IA en fonction des principes énoncés par l'entreprise (tels les principes éthiques internes par rapport à l'IA) et des propositions formulées par les organismes qui défendent les intérêts de la société civile en ce qui a trait aux préjudices potentiels et réels subis par des communautés vulnérables. De plus, de telles demandes devraient être soumises à l'examen d'une tierce partie, et ceux et celles qui mènent des audits de seconde partie de cette nature devraient être tenus de respecter des normes claires définies par les organismes de certification.
- **Favoriser des normes évolutives après consultation, notamment des personnes susceptibles de subir des préjudices.** Enfin, l'élaboration de normes est un processus toujours inachevé qui doit constamment s'adapter à des contextes en évolution. Alors que des organismes de normalisation tels que l'ISO ou l'IEEE veillent généralement aux normes d'ingénierie, nous sommes préoccupés par le fait que de puissantes entités du secteur privé ou d'États-nations s'emparent de la normalisation et l'édulcorent. Nous aimerions que les processus de normalisation incluent la consultation d'organisations représentant les personnes les plus susceptibles d'être lésées par les systèmes d'IA plutôt que de se limiter aux chercheurs et chercheuses entretenant des liens avec l'industrie (Veale, 2020). Nous recommandons donc l'élaboration indépendante de normes souples quant aux produits d'IA afin de définir des critères de base à appliquer aux produits et aux processus.

4. Suivi des incidents préjudiciables

Aucun système d'IA n'est parfait. Nous devons signaler et suivre les incidents dans lesquels l'IA cause des préjudices afin de faire en sorte que les personnes touchées par les systèmes d'IA partagent leur expérience et leurs préoccupations à cet égard. Normaliser le suivi des incidents préjudiciables implique une véritable compréhension des problèmes, un souci d'amélioration des systèmes par les fournisseurs et les exploitants, une meilleure surveillance des organismes de réglementation, des actions judiciaires, si nécessaire, et une sensibilisation accrue du public par rapport à de tels incidents, par exemple par une meilleure couverture dans la presse.

Le suivi des incidents est une pratique qui a déjà cours dans certains secteurs, comme la sécurité de l'information (Kenway et François, 2021). Bien réagir aux incidents comprend plusieurs activités clés, notamment la découverte de ceux-ci (en d'autres termes, apprendre qu'un incident s'est produit), le rapport et le suivi (documenter l'incident et diffuser l'information), la vérification (confirmer que l'incident a été causé par le système en question ou qu'il pourrait se reproduire), le classement selon un échelon (signaler le degré de gravité ou d'urgence de l'incident), l'atténuation (modifier le système de sorte qu'il ne continue pas à causer des préjudices, idéalement par une analyse des causes profondes plutôt que par un correctif superficiel), la réparation (prendre des mesures pour s'assurer que toute personne lésée sent qu'on a reconnu le préjudice qu'elle a subi et qu'on l'a traité puis, dans certains cas, l'indemniser), et la divulgation (faire part du problème aux parties prenantes indiquées, y compris les autres acteurs de l'industrie, les organismes de réglementation et le public).

8. Voir par exemple Parity AI (https://www.getparity.ai) et Credo AI (https://www.credo.ai).

S'il est bien organisé, le suivi des incidents incite les personnes qui ont subi des préjudices causés par les systèmes d'IA (ou celles qui les défendent) à décrire leur expérience d'une manière informative, ce qui servira à remettre en question des systèmes problématiques, à les améliorer (dans certains cas, cesser de les utiliser), et à demander réparation. Recueillir systématiquement des rapports relatifs aux incidents préjudiciables constitue une étape essentielle pour mieux comprendre les risques associés au déploiement des systèmes d'IA et pour garantir l'atténuation et la réparation des préjudices. Actuellement, il n'existe toutefois aucune proposition de politique, obligation, norme, ni système fonctionnel quant au suivi des incidents liés aux préjudices causés par l'IA.

En matière de politiques, nous recommandons donc les interventions suivantes :

- **Prévoir des normes et des bases de données quant au suivi des incidents préjudiciables liés à l'IA.** Les Nations Unies, les organismes publics nationaux et les organismes de réglementation devraient collaborer pour concevoir et gérer des bases de données qui recensent les préjudices causés par l'IA afin de documenter les cas connus où des systèmes d'IA ont enfreint les lois existantes ou ont nui à des personnes, puis d'en discuter. De tels systèmes de suivi des incidents doivent être adaptés aux besoins de chaque administration. Idéalement, ils suivraient des normes convenues en ce qui a trait à la classification des incidents et à l'évaluation de leur degré de gravité, entre autres critères. Une base de données internationale sur les incidents liés à l'IA gérée par les Nations Unies établirait idéalement la norme et pourrait également être exploitée dans des projets nationaux. Nous y voyons une mesure similaire à l'établissement de rapports d'incident diffusés à l'échelle de l'industrie dans d'autres domaines, par exemple en cybersécurité.
- **Exiger le signalement et le suivi des incidents préjudiciables.** En plus d'établir des normes quant au signalement d'incidents, les décideurs et décideuses politiques devraient exiger des fournisseurs et des exploitants de systèmes d'IA qu'ils emploient certains mécanismes pour signaler les préjudices, les abus, les répercussions diverses, les défaillances du système et d'autres incidents. En ce qui concerne les systèmes à haut risque en particulier, les organismes de réglementation devraient exiger des fournisseurs qu'ils publient régulièrement un résumé des rapports d'incident dévoilant la fréquence et la gravité des incidents ainsi que les mesures prises pour atténuer le problème. Des mécanismes de signalement et de suivi des incidents accessibles au public favorisent la responsabilisation en combinant des pressions exercées de toutes parts, que ce soit par les organismes de réglementation, des journalistes et le grand public, ou en raison de l'existence de recours collectifs ou de la concurrence du secteur privé. Une base de données publique et des normes partagées quant au suivi et au signalement des incidents garantiront que les divers acteurs de l'écosystème se penchent sur les problèmes repérés, partagent des connaissances sur des problèmes communs à tous les secteurs et renforcent la confiance en exposant les systèmes d'IA à un examen externe minutieux.

5. Avis d'utilisation

Une politique de responsabilisation en matière d'IA devrait inclure la divulgation publique obligatoire quant à l'usage de tout système d'IA susceptible de causer des préjudices. Les organismes publics, en particulier, doivent être tenus d'informer le public lorsqu'ils achètent, pilotent ou déploient des systèmes d'IA. La divulgation publique permet aux tierces parties de repérer des cibles d'audit. Elle représente de plus une exigence de base pour obtenir le consentement éclairé des personnes qui utiliseront ces systèmes d'IA ou seront touchées par leurs effets.

Souvent, les personnes qui subissent directement un préjudice découlant des systèmes d'IA ne connaissent pas les façons dont un produit ou un outil peut le causer. Elles reconnaissent le tort, ses répercussions et ce que cela implique dans leur vie quotidienne, mais peinent à déterminer comment un produit en particulier a causé le problème. Dans certaines situations où l'IA cause des préjudices, les gens découvrent l'existence du système en question après avoir remarqué le nom d'un fournisseur ou une interface utilisateur, ou après avoir été informés par une figure institutionnelle comme

un représentant ou une représentante des forces de l'ordre ou d'un centre d'aide juridique. Pour renverser cette tendance, nous pensons qu'une politique relative à l'IA doit commencer à imposer une divulgation publique quant à l'usage des systèmes d'IA.

La divulgation publique est l'un des chaînons les plus faibles en ce qui concerne la réglementation actuelle. Le public a le droit d'obtenir différentes formes de divulgation relativement aux systèmes d'IA. Les établissements qui utilisent un outil d'IA doivent publier de l'information sur le fait qu'ils emploient cet outil, les raisons et la manière dont celui-ci a été acquis, son fonctionnement et les éventuels préjudices qu'il aurait pu causer. Nous devons adopter des normes et des lois qui veillent à ce que les personnes soient informées de l'usage de systèmes d'IA, qu'elles connaissent les options d'exclusion, et qu'elles sachent faire appel de décisions et signaler des préjudices (Brennan Center for Justice, 2017). Si un système d'IA nécessite des garde-fous pour être employé de manière sûre et efficace, il faut également le mentionner.

La divulgation publique relative à l'usage de systèmes d'IA améliore la capacité des personnes à comprendre ce qui se passe et permet également aux tierces parties de repérer des cibles d'audits. Bien entendu, le degré de divulgation varie en fonction de plusieurs facteurs, notamment du niveau de risque et de la gravité des préjudices éventuellement causés par le système d'IA. Les obligations en matière de divulgation et de transparence pourraient être très différentes pour les organisations publiques et les organisations privées.

En matière de politiques, nous recommandons donc les interventions suivantes :

- **Mettre en place un avis d'intention quant à la conception ou au déploiement de systèmes d'IA.** Lorsque des organismes publics et des institutions ont l'intention de concevoir ou de déployer des systèmes d'IA, le public doit être averti et consulté dès le début du processus, selon un degré d'urgence qui varie en fonction des risques et de la gravité des préjudices qui pourraient en résulter. De plus, il importe de divulguer de l'information suffisante et de fournir un accès aux organismes de réglementation, au public ou aux tierces parties certifiées pour permettre une évaluation valable.
- **Émettre un avis d'utilisation.** Dans certains cas, les organismes de réglementation peuvent exiger d'entités publiques et privées qu'elles divulguent l'usage d'un système d'IA. Par exemple, dans le système judiciaire, chaque prévenu faisant l'objet d'une évaluation du risque de récidive devrait recevoir une notification à cet égard. Dans le secteur privé, il est possible, dans certains contextes, d'exiger que chaque personne postulant à un emploi soit informée si sa candidature est traitée par un outil de sélection intelligent, et la loi en vigueur peut contenir des dispositions qui permettent à certaines d'entre elles (voire à n'importe laquelle) de se retirer du processus, à la manière des dispositions relatives aux accommodements raisonnables contenues dans l'*Americans with Disabilities Act*. L'avis d'utilisation peut également s'étendre à un avis de collecte de données semblable à celui que recommande le *Règlement général sur la protection des données*, par exemple aux notifications liées au consentement quant à la collecte d'information effectuée au cours de la visite de sites Web ou aux avis indiquant la présence de caméras de surveillance.
- **Fournir une explication quant à l'usage et à la justification de l'organisation.** Les personnes doivent savoir *quand* et *comment* il y a emploi de systèmes d'IA dans les produits et services publics et privés. Cette information doit présenter les capacités et les limites du système d'une manière facile à comprendre. De plus, les organisations doivent justifier *pourquoi* elles utilisent le système d'IA.
- **Demander un consentement.** Dans de nombreux contextes, tant publics que privés, la politique en matière d'IA peut explicitement exiger le consentement de l'utilisateur ou l'utilisatrice en ce qui a trait à la collecte, à l'utilisation des données et à la participation aux processus de prise de décisions automatisée. Pour la collecte des données, les organismes de réglementation peuvent exiger une option d'inclusion (*opt-in*) plutôt que d'exclusion (*opt-out*). À titre d'exemple, Facebook a fait face à un important recours collectif après le moissonnage de données biométriques sans consentement

éclairé (Singer et Isaac, 2020). La collecte de renseignements personnels, de données biométriques et d'autres données de nature délicate de même que l'utilisation de cette information pour concevoir en aval des modèles d'apprentissage automatique devraient nécessiter un consentement explicite.

- **Prévoir, dans le processus d'approvisionnement, des mécanismes qui garantissent le recours à des systèmes d'IA équitables et responsables.** Chaque fois qu'un organisme gouvernemental lance un processus d'acquisition d'un système d'IA, il devrait respecter des obligations précises en matière de divulgation et de consultation publiques. Dans les administrations fédérales comme celle des États-Unis, les gouvernements des États et des localités qui reçoivent un financement du palier fédéral devraient également respecter de telles obligations. Un processus public solide et transparent devrait mener à l'élaboration d'exigences en matière d'approvisionnement. Celles-ci pourraient comprendre la diffusion d'un avis public quant à l'intention de déployer un système d'IA, lequel comprendrait une justification du projet et prévoirait une période de commentaires et la tenue d'audiences, une obligation de rendre des comptes (par exemple, par un rappel des caractéristiques du produit (Mitchell *et al.*, 2019), des études d'impacts des algorithmes[9] (McKelvey et MacDonald, 2019) ou la production de fiches techniques (Gebru *et al.*, 2021), la divulgation des principales conclusions des audits et des études d'impact et les exigences en matière de consentement, d'option d'exclusion, de recours, de rapports d'incident et de résolution des problèmes.
- **Prescrire des obligations strictes en matière d'équité et de responsabilité pour le financement public lié au développement de systèmes d'IA.** Les partenariats public-privé, les contrats du secteur privé, les subventions de recherche accordées aux établissements universitaires et les fonds reçus de l'administration fédérale par les gouvernements des États et des localités pour mettre au point des systèmes d'IA, y compris (mais pas seulement) des systèmes de prise de décisions automatisée, des produits équipés d'IA ou des modèles à usage général, devraient tous être soumis à de strictes obligations en matière d'équité et de responsabilité. Comme pour l'approvisionnement, ces obligations peuvent comprendre de la documentation quant au rendement, la divulgation obligatoire en vue d'informer les parties prenantes en cause, des études d'impact avant et après le déploiement, etc.
- **Exiger que la recherche liée au développement de systèmes d'IA financée par des fonds publics recueille des données permettant diverses études d'impact.** Les établissements universitaires et les organismes de financement public tels que (aux États-Unis) la National Science Foundation (NSF), les National Institutes of Health (NIH) ou la Defense Advanced Research Projects Agency (DARPA), notamment, devraient faire en sorte de mieux cerner les limites, les risques et les préjudices associés aux systèmes d'IA en exigeant que les recherches financées par l'État recueillent des renseignements démographiques ou d'autres données catégorielles pertinentes à diverses études d'impact. Ils devraient aussi documenter la recherche de données ainsi que l'étiquetage et l'interprétation des données recueillies.
- **Mettre en place des mécanismes visant à améliorer la divulgation par le secteur privé de leur usage de systèmes d'IA.** Alors que les organismes publics ont en général l'obligation de divulguer de l'information, les entreprises privées n'en ont pas souvent. Par exemple, alors qu'une régie publique du logement peut être contrainte, à la suite d'une demande d'accès à l'information, d'indiquer qu'elle utilise un système d'IA pour trier des locataires, le ou la propriétaire d'une résidence privée peut faire usage d'un tel système sans le dire à quiconque. Les locataires potentiels qui craignent de se voir refuser un logement en raison de leur genre, de leur origine ou d'un handicap peuvent ne jamais savoir qu'un système d'IA a été employé dans leur sélection.

9. Voir l'outil d'évaluation du gouvernement du Canada au https://www.canada.ca/fr/gouvernement/systeme/gouvernement-numerique/innovations-gouvernementales-numeriques/utilisation-responsable-ai/evaluation-incidence-algorithmique.html.

6. Recadrage : au-delà des préjugés de l'IA

Une politique en matière d'IA devrait aborder un large éventail de préjudices causés par l'IA plutôt que de se concentrer uniquement sur des mesures techniques liées à l'exactitude et aux préjugés. Elle nécessite également des définitions valables de tous les termes clés et la reconnaissance des multiples formes de préjudices.

À l'Algorithmic Justice League, nous rencontrons régulièrement des organisations communautaires et des personnes qui subissent directement des préjudices causés par les systèmes d'IA. Ces personnes n'ont pas tendance à nous parler de taux de précision ou de « préjugé algorithmique ». Elles disent s'inquiéter pour le loyer, et se demandent comment payer l'épicerie ou si leur enfant réussit bien à l'école… En plus de tout cela, elles s'inquiètent de ce qui se passerait si un système de reconnaissance faciale défectueux, installé à l'entrée de leur immeuble sans leur consentement, les empêchait d'y entrer tard le soir (Bellafante, 2019).

Les systèmes d'IA peuvent causer des préjudices de plusieurs manières : taux de précision variables dans la prédiction ou la classification de divers groupes de personnes, déploiement de systèmes sans avertissement ni consentement, opacité dans la façon dont le système prend des décisions, absence d'une option de retrait ou de recours permettant de contester les décisions prises par l'IA, systèmes dysfonctionnels qui ne répondent pas aux attentes annoncées et systèmes d'IA qui sont conçus et déployés de manière telle qu'ils entravent la vie privée ou la sécurité des personnes.

Voici nos recommandations à l'égard d'un recadrage :

- **Faire basculer les cadres stratégiques en matière d'IA pour délaisser l'optique axée sur les « préjugés » et s'ouvrir à une réflexion sur les préjudices causés par l'IA.** Il est nécessaire de réorienter ainsi les discussions pour faire en sorte que les audits de tierce partie répondent à un large éventail de préoccupations en fonction des priorités de la communauté que les auditeurs et auditrices pourraient représenter. Ce cadre élargi se prête également bien aux techniques de validation externes, et permet à l'audit de sonder le système et de critiquer des enjeux qui vont au-delà des préjugés.
- **Recentrer les politiques de responsabilisation de l'IA sur les répercussions des systèmes plutôt que sur leur précision.** Un système d'IA « exempt de préjugés » peut rester un système « inéquitable ». Le cadre qui se limite aux « préjugés » a tendance à mettre l'accent sur un mécanisme plutôt que sur ses effets. Pourtant, un système qui aurait été « libéré de préjugés » pour respecter un critère précis lié à une tâche en particulier, mais dont l'emploi entraîne par ailleurs des répercussions négatives au sein d'une certaine population, est toujours nocif. Bien qu'il soit essentiel de déterminer l'origine ou la cause profonde des mauvaises décisions prises par des systèmes d'IA, nous devons également comprendre l'importance d'analyser les effets qui se produisent en aval. Notre objectif n'est pas seulement d'« éliminer les préjugés » d'un système selon des normes techniques, mais aussi d'améliorer l'expérience des personnes qui en subissent des effets, selon leurs propres critères. Tout préjudice, même s'il semble minime, mérite d'être signalé. La situation n'a pas à s'aggraver pour que des personnes sentent un effet négatif. Quiconque subit des préjudices causés par un système d'IA devrait pouvoir exprimer son inquiétude.
- **Mettre l'accent sur les préjudices pour fonder les actions en justice.** Les préjugés des modèles ne justifient pas nécessairement des poursuites judiciaires, mais documenter les répercussions et les préjudices qu'ils causent en aval contribue à une meilleure stratégie juridique. Mettre l'accent sur les préjudices favorise les actions en justice, notamment les recours collectifs, qui sont d'importants mécanismes de réparation. Bien que la législation contre la discrimination ait dominé les discussions sur la responsabilité civile et les questions d'équité, de nombreuses définitions formelles de l'équité au sein des systèmes d'IA restent incompatibles avec les notions juridiques entourant la discrimination (Xiang et Raji, 2019). En outre, nous devons discuter plus amplement de droit de la responsabilité civile délictuelle, de responsabilité associée aux produits, de négligence, de protection des consommateurs,

et d'autres concepts juridiques servant à protéger les personnes contre des effets néfastes. Si les tierces parties menant des audits en IA vont au-delà des préjugés pour se concentrer sur les préjudices, elles peuvent contribuer à des actions judiciaires qui amélioreront l'expérience de ceux et celles qui subissent actuellement des répercussions négatives provoquées par des systèmes d'IA non responsables.

7. Mécanismes de responsabilisation post-audit

Les audits de tierce partie ne sont en fin de compte utiles que si plusieurs mécanismes mènent à la résolution des problèmes qu'ils mettent à jour. Une politique relative à l'IA devrait inclure divers outils de mise en application pour garantir que, par suite d'un audit, les entreprises en divulguent les principales conclusions, qu'elles apportent des améliorations en conséquence, qu'elles cherchent à se conformer aux normes et à la loi, et qu'elles réparent les préjudices.

Responsabilisation est le maître mot. Nous devons tenir responsables ceux et celles qui ont le pouvoir de mettre au point, de déployer et d'utiliser des systèmes d'IA quant à l'incidence de ceux-ci sur les personnes et les communautés vulnérables. En nous efforçant de réduire les préjudices, nous visons d'abord à améliorer l'existence des personnes directement touchées par des systèmes d'IA non responsables, et nous nous détournons de mesures et d'objectifs arbitraires qui ne seraient pas absolument pertinents dans l'expérience que font les personnes de l'IA. La responsabilisation requiert non seulement de mener des audits, mais également d'agir en fonction de leurs conclusions. Celles-ci doivent entraîner des changements concrets dans la vie des personnes en cause, qu'elles visent le retrait ou une révision des produits d'IA qui posent problème.

En matière de politiques, nous recommandons donc les interventions suivantes :

- **Mettre en place une surveillance continue des systèmes d'IA.** Il importe de surveiller les systèmes d'IA, en particulier leurs divers effets sur les populations marginalisées. À l'instar du secteur de la cybersécurité, dans lequel le Secure Development Lifecycle a fait évoluer les normes, nous devons opérer un virage dans le domaine de l'IA pour assurer l'équité et la responsabilisation tout au long du cycle de vie d'un projet. Cela comprend la conception, la planification, la collecte de données, le développement de modèles, les tests, et le déploiement d'un système d'IA puis son suivi. Les audits ponctuels ne suffisent pas à garantir l'équité et la responsabilisation. Bien que certaines lois récentes imposent un audit préalable au déploiement et des études d'impact, nous pensons qu'il est important de prévoir des mécanismes d'évaluation continue auxquels les tierces parties, notamment, auraient recours. Un système d'IA qui fonctionne bien en laboratoire peut malgré tout causer des préjudices lorsqu'on le déploie, et même avoir des effets contraires à la loi. Diverses répercussions et des préjudices peuvent survenir après la mise en place de systèmes d'IA et pour différentes raisons, y compris la complexité de systèmes qui évoluent au fil du temps ou encore des changements dans la façon de configurer ou d'utiliser l'outil en question.
- **Exiger des plans pour faire suite à l'audit et atténuer les préjudices.** En plus d'audits normalisés, d'une surveillance continue et de la production de rapports d'incidents, nous devrions également exiger des fournisseurs et des exploitants de systèmes d'IA qu'ils élaborent et mettent en œuvre des plans d'atténuation des préjudices prévoyant la manière de réagir à des préjudices pouvant survenir à tout moment. En matière de cybersécurité, il est désormais courant pour les entreprises d'avoir des équipes d'intervention en cas d'incident ainsi que des instruments facilitant le signalement des incidents, la vérification, le classement selon un échelon et la résolution. Une telle pratique doit également devenir la norme en ce qui concerne les systèmes d'IA.

- **Requérir la divulgation publique des principales conclusions de l'audit.** Bien que les fournisseurs de systèmes d'IA invoquent des objections légitimes quant à la divulgation publique d'information sur leurs produits, que ce soit en raison d'enjeux liés au secret industriel ou du désir de protéger les renseignements personnels des utilisateurs et utilisatrices, nous pensons que les têtes dirigeantes en matière d'IA devraient rendre obligatoire la divulgation publique des principales conclusions des audits. Les données sur le rendement évalué par rapport à des normes et à des points de référence connus, y compris celles tirées d'audits de première, de seconde et de tierce partie, doivent être accessibles au public. Les exigences d'une politique en matière de divulgation publique des principales conclusions transformeraient considérablement la responsabilisation.
- **Prévoir la réparation des préjudices.** Enfin, les fournisseurs et les exploitants de systèmes d'IA doivent se charger de prendre en considération les conclusions et les recommandations formulées par des auditrices ou auditeurs certifiés, les rapports d'incidents et d'autres lacunes ou préjudices, et adopter des mesures correctives. Ils devraient être tenus d'apporter des améliorations en conséquence, de viser la conformité aux normes et à la législation, et de réparer les préjudices qu'aurait révélés un audit.

CONCLUSION

Dans ce chapitre, nous avons décrit une série de « chaînons manquants » qu'il serait, selon nous, nécessaire d'aborder dans les discussions actuelles au sujet d'une politique en matière d'IA pour permettre la tenue d'audits de tierce partie fiables et efficaces. En matière de politiques, de multiples interventions doivent soutenir, protéger et encourager les audits de tierce partie, qui représentent un outil clé de surveillance. Les décideurs et décideuses politiques peuvent adopter des mesures précises pour favoriser la participation d'auditeurs et auditrices externes à la conception d'une IA équitable et responsable, peu importe le domaine d'application. Nous avons proposé une série d'interventions que nous considérons comme nécessaires et nous pensons qu'il faut les inscrire au cœur des discussions actuelles en matière de politiques. L'audit de tierce partie doit devenir une composante centrale des futures propositions de politiques relatives à l'IA.

Nous avons décrit sept interventions qui nous semblent cruciales pour soutenir les tierces parties menant des audits en IA. Elles concernent : 1) la protection juridique quant à l'accès des tierces parties, 2) la certification des auditeurs et auditrices, 3) l'élaboration de normes visant les produits d'IA, 4) le signalement des incidents préjudiciables, 5) la divulgation publique obligatoire de l'usage de systèmes d'IA, 6) le passage de cadres axés sur les préjugés de l'IA à des cadres axés sur les préjudices de l'IA et 7) la mise au point de mécanismes de responsabilisation garantissant des interventions appropriées quand des audits révèlent que des systèmes d'IA s'écartent des normes ou enfreignent la législation locale, nationale ou internationale qui s'applique, y compris la *Déclaration universelle des droits de l'homme*.

Ces premiers sujets de préoccupation ne constituent, parmi un vaste ensemble de mesures de responsabilisation dans le domaine de l'IA, que le début d'un nécessaire réexamen du rôle des auditeurs et auditrices agissant à titre de tierce partie. Nous souhaitons que nos propositions suscitent des discussions constructives et des actions de la part des responsables des politiques en vue de rendre les systèmes d'IA plus équitables et imputables qu'ils ne le sont actuellement.

RÉFÉRENCES

Abdalla, M. et Abdalla, M. 2021. The Grey Hoodie Project : Big tobacco, big tech, and the threat on academic integrity. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 287-297. https://dl.acm.org/doi/abs/10.1145/3461702.3462563

AICPA. 2017. *Audit and Accounting Guide Depository and Lending Institutions : Banks and Savings Institutions, Credit Unions, Finance Companies, and Mortgage Companies*. Hoboken, NJ : John Wiley & Sons.

Algorithmic Justice League. 2021. *AI Audits Landscape Mapping 2021 (public)*. Google Docs. https://docs.google.com/spreadsheets/d/17MP8s0PxTluEt1YOV4kWeBz2SpEqk7VunyZVSWDGA54/edit?usp=embed_facebook.

Angwin, J., Larson, J., Mattu, S. et Kirchner, L. 2016. *Machine bias*. ProPublica. 23 mai. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

ASQ. n.d. *What is an Audit ? Types of Audits & Auditing Certification*. ASQ. https://asq.org/quality-resources/auditing.

Bellafante, G., 2019. *The landlord wants facial recognition in its rent-stabilized buildings. Why ?* New York Times. 28 mars. https://www.nytimes.com/2019/03/28/nyregion/rent-stabilized-buildings-facial-recognition.html

Bietti, E. 2020. From ethics washing to ethics bashing : a view on tech ethics from within moral philosophy. Dans *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 210-219. https://dl.acm.org/doi/abs/10.1145/3351095.3372860

Brennan Center for Justice. 2017. *The Public Oversight of Surveillance Technology (POST) Act : A Resource Page*. Brennan Center for Justice. https://www.brennancenter.org/our-work/research-reports/public-oversight-surveillance-technology-post-act-resource-page

Buolamwini, J. et Gebru, T. 2018. Gender shades : Intersectional accuracy disparities in commercial gender classification. Dans *Conference on Fairness, Accountability and Transparency*, pp. 77-91. https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Chau. 2020. Public contracts : automated decision systems, Assemblée legislative de la Californie, États Unis, No. AB-13. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202120220AB13.

Cochrane, R.C. 1966. *Measures for progress : A history of the National Bureau of Standards*, vol. 13. National Bureau of Standards, Department of Commerce, US.

Cuevas, E. 2020. Chau introduces automated decision systems accountability Act of 2021. Communiqué de presse, « Ed Chau Assembly District 49 ». 8 décembre 2020. https://web.archive.org/web/20210616125930/

European Commission. 2022. Digital services act package. https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package

Foxglove. 2020. *We put a stop to the A Level grading algorithm !* London, Foxglove. Nouvelles sur le site Web. 17 août. https://www.foxglove.org.uk/2020/08/17/we-put-a-stop-to-the-a-level-grading-algorithm/

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H. et Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, vol. 64, nº 12, pp. 86-92. https://arxiv.org/abs/1803.09010.

Gillum, J. et Tobin, A., 2019. Facebook won't let employers, landlords or lenders discriminate in ads anymore. Propublica. 19 mars. https://www.propublica.org/article/facebook-ads-discrimination-settlement-housing-employment-credit

Grother, P., Ngan, M. et Hanaoka, K. 2019. *Face Recognition Vendor Test (FVRT) : Part 3, Demographic Effects*. Gaithersburg, Md., National Institute of Standards and Technology.

Gunningham, N. 1993. Environmental auditing : Who audits the auditors ? *Environmental and Planning Law Journal*, vol. 10, p. 229.

Hill, K. 2020. Wrongfully accused by an algorithm. The New York Times. 24 juin. https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html

IBM. 2021. Trustworthy AI is human-centered. Site Wb d'IBM. https://www.ibm.com/watson/trustworthy-ai

IEEE. 2019. IEEE Standard for Safety Levels with Respect to Human Exposure to Electric, Magnetic, and Electromagnetic Fields, 0 Hz to 300 Ghz. *IEEE Access*, vol. 7, pp. 171346-171356.

Jobin, A., Ienca, M. et Vayena, E. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, vol. 1, nº 9, pp. 389-399. https://www.nature.com/articles/s42256-019-0088-2

Kazim, E. et Koshiyama, A. 2020. *A review of the ICO's draft guidance on the AI auditing framework*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3599226

Keegan, J. 2021. *Facebook got rid of racial ad categories. Or did it ?* The Markup, 9 juillet. https://themarkup.org/citizen-browser/2021/07/09/facebook-got-rid-of-racial-ad-categories-or-did-it

Kenway, J., et François, C. 2021. *Bug Bounties for Algorithmic Harms ? Lessons from Cybersecurity Vulnerability Disclosure for Algorithmic Harms Discovery, Disclosure, and Redress*. Washington, DC : Algorithmic Justice League. https://www.ajl.org/bugs

Krafft, P.M., Young, M., Katell, M., Lee, J.E., Narayan, S., Epstein, M., Dailey, D., Herman, B., Tam, A., Guetler, V. et Bintz, C. 2021. An action-oriented AI policy toolkit for technology audits by community advocates and activists. Dans *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 772-781. https://dl.acm.org/doi/10.1145/3442188.3445938

Learned-Miller, E., Ordónez, V., Morgenster, J. et Buolamwini, J. 2020. *Facial recognition technologies in the wild : A call for a federal office*. Algorithmic Justice League. https://assets.website-files.com/5e027ca188c99e3515b404b7/5ed1145952bc185203f3d009_FRTsFederalOfficeMay2020.pdf

Lecher, C. et Varner, M. 2021. How we investigated NYC high school admissions. The Markup, 26 mai. https://themarkup.org/show-your-work/2021/05/26/how-we-investigated-nyc-high-school-admissions

Lytton, T.D. et McAllister, L.K. 2014. Oversight in private food safety auditing : Addressing auditor conflict of interest. *Wisconsin Law Review*, nº 6/2014, pp. 289-337.

MacCarthy, M. 2019. An examination of the *Algorithmic Accountability Act of 2019*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3615731

Markey, Ed. 2021. Senator Markey, Rep. Matsui Introduce Legislation To Combat Harmful Algorithms And Create New Online Transparency Regime. Communiqué de presse, 27 mai. https://www.markey.senate.gov/news/press-releases/senator-markey-rep-matsui-introduce-legislation-to-combat-harmful-algorithms-and-create-new-online-transparency-regime

Martinez, E. et Carollo, M. 2021. Dozens of mortgage lenders showed significant disparities. Here are the worst. The Markup, 25 août. https://themarkup.org/denied/2021/08/25/dozens-of-mortgage-lenders-showed-significant-disparities-here-are-the-worst

McKelvey, F. et MacDonald, M. 2019. Artificial intelligence policy innovations at the Canadian Federal Government. *Canadian Journal of Communication*, vol. 44, nº, 2, pp. 43-50. https://www.canada.ca/fr/gouvernement/systeme/gouvernement-numerique/innovations-gouvernementales-numeriques/utilisation-responsable-ai/evaluation-incidence-algorithmique.html

Metcalf, J. et Moss, E. 2019. Owning ethics : Corporate logics, Silicon Valley, and the institutionalization of ethics. *Social Research : An International Quarterly*, vol. 86, nº 2, pp. 449-476.

Metcalf, J., Moss, E., Watkins, E. A., Singh, R. et Elish, M.C. 2021. Algorithmic impact assessments and accountability : The co-construction of impacts. Dans *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 735-746. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3736261

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D. et Gebru, T. 2019. Model cards for model reporting. Dans *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220-229. https://dl.acm.org/doi/10.1145/3287560.3287596

Mittelstadt, B. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, vol. 1, nº 11, pp. 501-507. https://www.nature.com/articles/s42256-019-0114-4

Open Government Partnership. 2021. *Algorithmic Accountability for the Public Sector*. https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector

Ponce, A. 2020. *The Digital Services Act Package : Reflections on the EU Commission's Policy Options.* ETUI Research Paper-Policy Brief No. 12/2020. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3699389

Raji, I.D. et Buolamwini, J. 2019. Actionable auditing : Investigating the impact of publicly naming biased performance results of commercial ai products. Dans *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 429-435. https://www.media.mit.edu/publications/actionable-auditing-investigating-the-impact-of-publicly-naming-biased-performance-results-of-commercial-ai-products/

Raji, I.D. et Yang, J. 2019. About ML : Annotation and benchmarking on understanding and transparency of machine learning lifecycles. 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada. https://arxiv.org/pdf/1912.06166.pdf

Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D. et Barnes, P. 2020. Closing the AI accountability gap : Defining an end-to-end framework for internal algorithmic auditing. Dans *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 33-44. https://arxiv.org/abs/2001.00973

Richardson, R. 2021. Defining and demystifying automated decision systems. *Maryland Law Review*, à paraître. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3811708

Selbst, A. D. 2021. An institutional view of algorithmic impact assessments. *Harvard Journal of Law & Technology*, vol. 35, nº 10, pp. 117-191.

Shahriari, K. et Shahriari, M. 2017. IEEE standard review—Ethically aligned design : A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. Dans *2017 IEEE Canada International Humanitarian Technology Conference (IHTC)*, pp. 197-201. https://ieeexplore.ieee.org/document/8058187

Simonite, T. 2020. Google offers to help others with the tricky ethics of AI. *Wired*, 28 août. https://www.wired.com/story/google-help-others-tricky-ethics-ai

Singer, N. et Isaac, M. 2020. Facebook to pay $550 million to settle facial recognition suit. *New York Times*, 29 janvier. https://www.nytimes.com/2020/01/29/technology/facebook-privacy-lawsuit-earnings.html

Spinks, C. N. 2019. Contemporary housing discrimination : Facebook, targeted advertising, and the Fair Housing Act. *Houston Law Review*, vol. 57, nº 4 https://houstonlawreview.org/article/12762-contemporary-housing-discrimination-facebook-targeted-advertising-and-the-fair-housing-act

Team, A. P. 2021. *Artificial Intelligence Measurement and Evaluation at the National Institute of Standards and Technology*. Washington, D.C., National Institute of Standards and Technology. https://www.nist.gov/news-events/events/2021/06/ai-measurement-and-evaluation-workshop

United Kingdom Government. 2020. *Draft Online Safety Bill*. London, Department for Digital, Culture, Media and Sport. https://www.gov.uk/government/publications/draft-online-safety-bill

United States. 1984. *Computer Fraud and Abuse Act*, 18 U.S.C., § 1030. https://uscode.house.gov/view.xhtml?req=(title:18%20section:1030%20edition:prelim)

United States Congress, 2019. *H.R.2231 – 116th Congress (2019-2020) : Algorithmic Accountability Act of 2019*, April 11. https://www.congress.gov/bill/116th-congress/house-bill/2231

Veale, M. 2020. A critical take on the policy recommendations of the EU high-level expert group on artificial intelligence. *European Journal of Risk Regulation*, vol. 11, nº 1, pp. 1-10. https://www.cambridge.org/core/journals/european-journal-of-risk-regulation/article/abs/critical-take-on-the-policy-recommendations-of-the-eu-highlevel-expert-group-on-artificial-intelligence/FF6FF91A0C140E58B4B527C68E0C5321

Weiger, C., Smith, K.C., Cohen, J.E., Dredze, M. et Moran, M. B. 2020. How internet contracts impact research : Content analysis of terms of service on consumer product websites. *Public Health and Surveillance*, vol. 6, nº 4, pp. 1-15. https://publichealth.jmir.org/2020/4/e23579/

Wieringa, M. 2020. What to account for when accounting for algorithms : A systematic literature review on algorithmic accountability. Dans *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 1-18. https://dl.acm.org/doi/abs/10.1145/3351095.3372833

Xiang, A. et Raji, I.D. 2019. On the legal compatibility of fairness definitions. https://arxiv.org/abs/1912.00761

LE SECTEUR DE L'IA DU POINT DE VUE DE L'ÉTHIQUE

GOLNOOSH FARNADI

Professeure adjointe d'apprentissage automatique à HEC Montréal et professeure associée à l'Université de Montréal. Membre académique principal de MILA – Institut québécois d'intelligence artificielle et titulaire de chaire en IA Canada-CIFAR.

AMANDA LEA DE LIMA ALVES

Assistante de recherche au FATE – Fairness, Accountability, Transparency, and Ethics Lab, membre de l'IA pour l'humanité à MILA – Institut québécois d'intelligence artificielle. Elle détient une maîtrise en sciene politique de l'Université de Montréal.

REBECCA SALGANIK

M. Sc. à MILA – Institut québécois d'intelligence artificielle et à l'Université de Montréal. Membre de FATE – Fairness, Accountability, Transparency, and Ethics Lab dirigé par professeure Golnoosh Farnadi.

ODD 9 - Industrie, innovation et infrastructure

ODD 10 - Inégalités réduites

ODD 11 - Villes et communautés durables

ODD 12 - Consommation et production responsables

ODD 16 - Paix, justice et institutions efficaces

ODD 17 - Partenariats pour la réalisation des objectifs

LE SECTEUR DE L'IA DU POINT DE VUE DE L'ÉTHIQUE

RÉSUMÉ

De nos jours, tous les secteurs d'activité font partie du secteur de l'intelligence artificielle (IA). L'agriculture, les soins de santé, les finances, l'éducation et les arts, entre autres, utilisent non seulement des modèles d'apprentissage automatique et d'IA tout au long de leur chaîne d'approvisionnement, mais chacun et chacune d'entre nous interagit avec l'ensemble de l'écosystème de leurs algorithmes, peut-être même sans nous en rendre compte, et ce, pour le meilleur et pour le pire et sans une compréhension des principes qui guident le fonctionnement de ces modèles. Dans cet article, nous tentons de faire la lumière sur la manière dont le secteur traite l'éthique en IA, soulignant les failles qui ont permis l'émergence d'iniquités et explorant des voies possibles pour l'avenir. Nous avons discuté avec sept chercheurs et chercheuses en éthique de l'IA issus du milieu universitaire et du secteur pour dévoiler le scénario qui prévaut actuellement en éthique dans le secteur : les défis, les possibilités et les parties prenantes. Nous explorons trois des enjeux les plus importants soulevés par les experts et expertes et proposons une analyse de trois avenues possibles pour une IA éthique.

Puisque les modèles d'IA jouent un rôle de plus en plus important dans nos vies, changer la façon dont ils fonctionnent deviendra de plus en plus difficile. Plusieurs rapports font état de comportements discriminatoires, démontrant à quel point les préjugés peuvent devenir profondément intégrés dans les technologies. La culture « d'agir vite et de casser des choses » a engendré de nombreuses conséquences sociétales négatives. Nous décrivons de quelle manière un manque d'indicateurs d'équité unifiés, un manque de diversité et un manque de normes éthiques ont mené à la tempête parfaite d'irresponsabilité et d'inaction. Ensuite, nous explorons de quelle manière une participation plus large peut lever le voile sur le secteur des technologies et démocratiser la discussion concernant l'équité en IA. Nous nous penchons aussi sur la manière dont la sensibilisation et l'élargissement de l'accès à une éducation en IA fondée sur l'éthique peuvent

> opérer un changement important dans la façon dont nous approchons la conception et la mise en œuvre des technologies. Finalement, nous soutenons que la démocratisation de la participation et l'instauration de règles de jeu équitables pour les discussions doivent être suivies de politiques concrètes, de règlements et de réformes organisationnelles.

INTRODUCTION

Au cours des dernières années, la mise en œuvre d'algorithmes d'intelligence artificielle (IA), et les modèles d'apprentissage automatique qu'ils utilisent pour comprendre le monde, a engendré un immense mouvement technologique qui s'est infiltré dans toutes les facettes de nos vies. L'agriculture, les soins de santé, les finances, l'éducation et les arts, entre autres, ont embrassé l'utilisation d'algorithmes de prise de décisions. Chaque jour, nous interagissons avec un écosystème d'algorithmes, non seulement à titre d'utilisateurs actifs et d'utilisatrices actives, mais aussi de manière passive, sans même nous en rendre compte. En raison de cette utilisation, les complexités de l'IA touchent maintenant directement la qualité de vie des gens.

Or, même si nous avons intégré ces algorithmes dans nos économies mondiales, une série de questions cruciales demeurent sans réponse. Qui supervise les impacts potentiels du développement et de la mise en œuvre de l'IA dans nos secteurs importants ? De quelle manière garantissons-nous que les effets négatifs de ces systèmes ne surpassent pas les avantages potentiels ? De quelle manière sont comblées les lacunes entre le développement de l'IA, sa mise en œuvre et ses résultats dans le secteur des technologies ?

L'urgence de ces questions s'accroît exponentiellement avec la vitesse à laquelle les nouvelles technologies sont introduites dans nos vies. Il existe des limites inhérentes à la compréhension de systèmes créés pour réfléchir mathématiquement à un monde qui ne peut être entièrement décrit par des règles mathématiques. Sans un cadre de surveillance rigoureux, ces systèmes, une fois mis en œuvre dans un aussi vaste répertoire d'applications, sont susceptibles de produire des résultats qui n'ont pas été prévus durant le processus de conception. Dans des situations où nous accordons un pouvoir immense à ces modèles, sans par ailleurs en assurer une surveillance adéquate, leurs erreurs peuvent entraîner de réelles conséquences désastreuses. Malheureusement, il s'agit d'un aspect de l'IA qui est souvent ignoré.

Récemment, les coûts éthiques réels de certains systèmes d'IA ont été mis en lumière. Au cours des dernières années, les experts et expertes ont découvert des cas dans lesquels les modèles perpétuaient et même consolidaient les formes de discrimination qui sont présentes au sein de notre société (Angwin *et al.*, 2016, Spielkamp, 2017, Yong, 2018, Grind *et al.*, 2019). La constatation que leur discrimination algorithmique ciblait spécifiquement des groupes démographiques couramment marginalisés est encore plus préoccupante. Le tollé qui a suivi ces découvertes a favorisé l'émergence d'un nouveau domaine : l'équité algorithmique en apprentissage automatique.

Intuitivement, le fondement de l'équité algorithmique est de veiller à ce que les modèles d'apprentissage automatique ne discriminent pas des groupes de personnes en particulier. Dans notre discours courant, l'équité algorithmique est souvent traitée comme un exercice spécialisé, technique ou universitaire. Or, l'équité est un sujet qui nous concerne tous et toutes et non pas seulement les codeurs et codeuses. En fait, les valeurs et les prémisses de l'équité algorithmique trouvent leur origine dans des concepts

telles l'égalité et la non-discrimination. Nous pourrions, bien entendu, nous demander pourquoi nous avons besoin d'un tel mouvement alors qu'il existe plusieurs concepts sociaux et juridiques bien établis qui, à bien des égards, guident les pratiques non discriminatoires. Cependant, lorsqu'il est question d'applications d'IA, il semble exister un paradoxe en ce qui concerne la responsabilité en matière de pratiques injustes. Alors que dans certaines situations une personne peut être tenue responsable d'une discrimination, ceux et celles ayant conçu un modèle d'IA perpétuant un comportement discriminatoire ne font pas face aux mêmes conséquences. Cette incohérence a créé une situation où, à l'insu de tous et toutes, les systèmes d'IA peuvent perpétuer une injustice à grande échelle auprès d'un groupe de personnes au sein de notre société. À cet égard, l'équité algorithmique n'est pas une question technique, mais bien une question socioculturelle.

Bien que des efforts didactiques soient une première étape importante pour éliminer les préoccupations en matière d'équité, ces efforts ne sont pas suffisants. Comment combler le fossé qui existe entre le domaine de l'équité algorithmique, qui est à la fois nouveau et prometteur, les pratiques du secteur qui façonnent les produits et services alimentés par l'IA et nous tous et toutes, les consommateurs et consommatrices, les entreprises, la société civile et les gouvernements ? Les besoins de qui l'IA devrait-elle servir ?

Le domaine de l'IA en est encore à ses premiers pas. Mais, comme ces modèles jouent un rôle plus important dans nos vies, changer la façon dont ils fonctionnent sera de plus en plus difficile. L'opacité des pipelines d'IA et leur enchevêtrement avec des structures de mises en œuvre font en sorte qu'il est souvent difficile de cerner exactement où les choses se passent mal. Ceci s'explique en raison du fait que la discrimination peut se produire partout dans le processus d'IA : dans les données, dans le modèle et, même, dans le résultat. Afin d'opérer des changements concrets, nous devons lever le voile sur le fonctionnement interne du secteur de l'IA et comprendre les défis que comporte la création de produits éthiques.

D'une certaine façon, la seule issue possible est de continuer. En préparation de ce chapitre, nous avons communiqué avec des chercheurs et chercheuses en IA issus du milieu universitaire et du secteur possédant des antécédents diversifiés, mais partageant tous et toutes notre engagement à faire progresser un mouvement de sensibilisation éthique au sein du secteur de l'IA. Leurs perspectives ne tiennent pas seulement compte de leur expertise, mais aussi de la voix de groupes sous-représentés en IA, que ce soit en raison de leur genre, de leur race ou de l'endroit géographique où ils vivent. Nous avions préparé des questions concernant leurs perspectives sur l'IA éthique dans le secteur des technologies, abordant les aspects d'équité et de diversité, les avantages et désavantages de l'IA et les possibles voies à suivre pour mettre en œuvre des cadres réglementaires. Leurs réponses ont été recueillies par écrit ou lors d'entretiens vidéo. Lors de la rédaction de ce chapitre, nous avons choisi les éléments clés de nos discussions avec ces experts et expertes du domaine, portant une attention particulière sur les lacunes importantes qui mettent en péril le développement d'une IA éthique. En outre, nous présentons leurs points de vue judicieux dans trois domaines d'engagement pouvant inspirer une meilleure éthique en IA dans l'avenir au sein du secteur. Notre objectif ultime est d'ouvrir la boîte noire du secteur de l'IA et de mettre en lumière les discussions nécessaires en ce qui concerne les pratiques actuelles en éthique de l'IA.

Participants et participantes

Margaret Mitchell : Margaret Mitchell est chercheuse et travaille en IA éthique. Elle se penche actuellement sur les tenants et aboutissants du développement éthique de l'IA dans le secteur des technologies. Auteure de plus de 50 articles sur la génération automatique de textes en langage naturel, les technologies d'aide, la vision par ordinateur et l'éthique en IA, elle détient plusieurs brevets dans le domaine de la génération de conversations et l'analyse des sentiments. Elle a travaillé chez Google AI à titre de scientifique de recherche où elle a fondé et codirigé le groupe d'IA éthique de Google qui

se penchait sur la recherche fondamentale en éthique de l'IA et sur l'opérationnalisation interne de l'IA éthique chez Google. Avant de se joindre à Google, elle était chercheuse chez Microsoft Research et ses travaux portaient plus particulièrement sur la génération de langage à partir de la vision par ordinateur. Elle a fait un stage postdoctoral à Johns Hopkins où elle s'intéressait à la modélisation bayésienne et à l'extraction d'informations. Elle détient un Ph. D. en informatique de l'Université d'Aberdeen ainsi qu'une maîtrise en linguistique informatique de l'Université de Washington. De 2005 à 2012, durant ses études supérieures, elle a aussi travaillé dans les domaines de l'apprentissage automatique, des troubles neurologiques et des technologies d'aide à l'Oregon Health and Science University. Elle a dirigé plusieurs ateliers et initiatives à l'intersection de la diversité, de l'inclusion, de l'informatique et de l'éthique. Ses travaux lui ont mérité des prix d'Ash Carter, secrétaire à la défense, et de la American Foundation for the Blind et ont été mis en œuvre par plusieurs sociétés technologiques.

Rumman Chowdhury : La passion de Dre Rumman Chowdhury se situe à l'intersection de l'IA et de l'humanité. Pionnière dans le domaine de l'éthique algorithmique appliquée, elle crée des solutions sociotechniques de pointe pour une IA éthique, explicable et transparente. Elle est actuellement directrice de l'équipe META (l'équipe chargée de l'éthique, de la transparence et de la responsabilité en apprentissage automatique) chez Twitter où elle dirige une équipe de chercheurs, chercheuses, ingénieurs appliqués et ingénieures appliquées qui cernent les préjudices algorithmiques sur la plateforme et les atténue. Elle a, auparavant, été cheffe de la direction et fondatrice de Parity, une compagnie fournissant une plateforme d'audit algorithmique pour les entreprises. Elle a aussi agi à titre de responsable mondiale de l'IA responsable chez Accenture Applied Intelligence, dirigeant la conception de Fairness Tool, le premier outil algorithmique pour cerner et atténuer les préjugés dans les systèmes d'IA dans le secteur. Dre Chowdhury a coécrit un article sur les influences et les impacts de cet outil qui a été publié dans le Harvard Business Review.

Francisco Marmolejo-Cossío : Francisco Marmolejo-Cossío est postdoctorant à la School of Engineering and Applied Sciences (SEAS) de l'Université Harvard et chercheur chez Input Output Hong Kong (IOHK). Auparavant, il a été bousier en perfectionnement de carrière en informatique au Balliol College de l'Université d'Oxford. Il a obtenu un Ph. D. en informatique théorique sous la supervision de Paul Goldberg et un baccalauréat ès arts en mathématiques à l'Université Harvard avec une mineure en neuroscience en 2012. Il est également coorganisateur de l'initiative de recherche Mechanism Design for Social Good (MD4SG).

Arisa Ema : Arisa Ema est professeure agrégée à l'Université de Tokyo et chercheuse invitée au RIKEN Center for Advanced Intelligence Project au Japon. Elle s'intéresse aux études en science et technologie et, plus particulièrement, aux avantages et risques de l'IA et, pour mener ses travaux, a mis en place un groupe de recherche interdisciplinaire. Elle est membre du comité d'éthique de la Japanese Society for Artificial Intelligence qui a publié, en 2017, des lignes directrices sur l'IA éthique. Elle siège également au conseil d'administration de la Japan Deep Learning Association (JDLA) et préside le groupe d'étude sur la gouvernance de l'IA. Elle est aussi membre du Council for Social Principles of Human-centric AI du Bureau du cabinet qui, en 2019, a publié le document intitulé « Social Principles of Human-Centric AI ».

Vidushi Marda : Vidushi Marda, chercheuse et avocate indienne, étudie l'impact sociétal des systèmes d'IA. Elle travaille actuellement en tant qu'agente de programme principale chez ARTICLE 19, une organisation mondiale de droits humains, où elle dirige les travaux de recherche et la mobilisation sur les implications en matière de droits fondamentaux de l'apprentissage automatique. Elle est membre du groupe d'experts et d'expertes sur la gouvernance des données et de l'IA de l'initiative Global Pulse des Nations Unies et siège au comité directeur chez RealML. Dans son travail, elle interagit avec des communautés techniques, de politiques, universitaires et de défense des intérêts et a notamment été citée par la Cour suprême de l'Inde dans le cadre d'une décision fondatrice sur le droit à la vie privée, par le comité spécial sur l'intelligence artificielle de la Chambre des lords du Royaume-Uni et par la rapporteuse spéciale sur la liberté d'opinion et d'expression des Nations Unies.

Joanna Shields : Vétérane du secteur technologique, Joanna Shields a aidé à bâtir certaines des sociétés à forte croissance chefs de file, dont Google, Aol et Facebook, et a mené plusieurs entreprises en démarrage vers la réussite. Elle est actuellement cheffe de la direction de BenevolentAI, une société pharmaceutique de pointe qui mise sur l'apprentissage automatique et l'IA pour développer des médicaments plus efficaces. Elle préside la Plénière du groupe d'experts multipartite et copréside le comité directeur du Partenariat mondial sur l'intelligence artificielle, qui est soutenu par l'OCDE, et a déjà agi à titre de ministre de la Sécurité d'Internet et sous-secrétaire d'État du Royaume-Uni, d'ambassadrice pour les secteurs numériques du Royaume-Uni, de conseillère spéciale en économie numérique du premier ministre, de présidente du conseil et cheffe de la direction de TechCityUK et d'administratrice non dirigeante du London Stock Exchange Group. En 2014, elle a fondé WePROTECT.org, une alliance mondiale œuvrant à la protection des enfants contre la violence et l'exploitation en ligne. Aussi, en 2014, elle s'est vu remettre l'Ordre de l'Empire britannique du Royaume-Uni pour ses services au secteur numérique et ses services bénévoles aux jeunes et la Chambre des lords l'a nommée Paire à vie.

Ulrich Aïvodji : Ulrich Aïvodji est professeur adjoint d'informatique au Département de génie logiciel et des technologies de l'information de l'ÉTS Montréal. Ses travaux de recherche portent sur la sécurité informatique, la protection de la vie privée, l'optimisation combinatoire et l'apprentissage automatique. Il se concentre actuellement sur divers aspects de l'apprentissage automatique digne de confiance, comme l'équité, l'apprentissage automatique préservant la vie privée et l'explicabilité. Avant d'occuper son poste actuel, il était postdoctorant à l'UQAM et travaillait avec Sébastien Gambs sur l'éthique et la protection de la vie privée en apprentissage automatique. Il a obtenu son Ph. D. en informatique de l'Université Toulouse III sous la supervision de Marie-José Huguet et de Marc-Olivier Killijian. Durant son Ph. D., il a été affilié au LAAS-CNRS à titre de membre des groupes de recherche TSF et ROC et a travaillé sur des technologies respectueuses de la vie privée pour le covoiturage.

POURQUOI SE SOUCIER DE L'ÉQUITÉ AU SEIN DU SECTEUR DE L'IA ?

Les pièges courants de l'IA peuvent être illustrés par le tristement célèbre modèle COMPAS (Angwin *et al.*, 2016). COMPAS, acronyme de Correctional Offender Management Profiling for Alternative Sanctions, est un outil qui a été conçu par une entreprise privée dans le but de prédire le risque de récidive d'une personne accusée d'une infraction pénale. Ce modèle a été utilisé en tant que preuve dans le système judiciaire américain lors d'enquête sur remise en liberté. En 2016, ProPublica, un organe de presse spécialisé dans le journalisme d'enquête, a publié un article dans lequel ses journalistes alléguaient que le modèle était plus susceptible de prédire un risque de récidive plus élevé parmi les personnes accusées noires (Larson, 2016). En fait, après un examen approfondi des chiffres, ils ont constaté que les personnes accusées noires étaient près de deux fois plus susceptibles que les personnes accusées caucasiennes de se faire étiqueter comme étant à haut risque, sans, dans les faits, récidiver. En termes simples, ceci suggérait que le modèle avait commencé à perpétuer des stéréotypes raciaux préjudiciables et faux au sujet des groupes de citoyens et citoyennes les plus susceptibles de commettre des crimes.

Qu'est-ce que cela signifie et pourquoi cela s'est-il produit ? Lorsque la nouvelle a été diffusée, l'entreprise responsable du logiciel COMPAS, Northpointe, a publié une déclaration informant le public que, compte tenu de leurs notions mathématiques de l'équité, leur outil n'était pas partial (Dietrich, 2016). Et, dans les faits, après avoir examiné leurs notions mathématiques des préjugés, les personnes responsables de l'audit ont confirmé que compte tenu de ces notions, le modèle faisait des prédictions non partiales.

Cette situation complexe illustre parfaitement bien les lacunes en matière d'éthique dans le secteur de l'IA que nous discuterons dans ce chapitre, soit 1) le manque de définitions de l'équité, 2) le manque de diversité et 3) le manque de normes éthiques.

Il est important de noter que la raison principale qui a mené à cette situation était le manque de cohérence dans ce qui définit l'équité. D'abord, nous explorerons comment il est possible qu'un algorithme puisse être qualifié « d'équitable » alors qu'il perpétue des comportements discriminatoires envers un groupe démographique. Nous expliquerons aussi comment un manque de diversité dans le secteur des technologies peut exacerber le manque de surveillance des conséquences algorithmiques. Ensuite, nous explorerons comment le manque de normes éthiques, le manque de règlements et l'opacité algorithmique permettent aux propriétaires de systèmes d'intelligence artificielle de définir unilatéralement ce que signifie l'équité en ce qui a trait à leurs produits. Nous analyserons comment le manque de participation plus large aux discussions réglementaires et le manque général de sensibilisation à l'égard de l'équité algorithmique créent un vide qui permet à des modèles discriminatoires de continuer d'exister, sans surveillance. Finalement, nous passerons en revue ces points, tenant compte de nos conversations approfondies avec ces experts et expertes de l'IA éthique, et terminerons par une discussion sur des orientations d'améliorations potentielles.

1. Le manque de définitions de l'équité

La question suivante est au cœur même de l'atteinte de l'équité : qu'est-ce exactement l'équité et comment pouvons-nous la définir dans les modèles d'IA et d'apprentissage automatique ?

Intuitivement, l'équité consiste à veiller à ce que les prédictions d'un système algorithmique ne discriminent pas de manière contraire à l'éthique un certain groupe ou une personne. Or, du point de vue algorithmique, l'équité doit souvent être définie en termes mathématiques, ce qui comprend une notion de vrais positifs, de faux positifs, de vrais négatifs et de faux négatifs. Par exemple, prédire qu'une personne commettra un crime quand, en réalité, elle ne le fait pas est un faux positif. Cependant, prédire qu'une personne n'est pas atteinte d'un cancer lorsque, dans les faits, elle l'est est un faux négatif. L'une des difficultés dans l'établissement d'une définition de l'équité est que ce mot possède différentes définitions, souvent contradictoires, qui dépendent profondément du problème considéré. En 2018, les universitaires dans le domaine de l'équité algorithmique avaient répertorié 21 définitions uniques de l'équité, la plupart d'entre elles étant entièrement incompatibles les unes par rapport aux autres (Verma, 2018).

> Les lignes directrices en matière d'éthique recommandent souvent que les systèmes d'IA intègrent une liste de propriétés de fiabilité, comme l'équité, la sécurité, la protection de la vie privée, l'explicabilité et la transparence. Toutefois, notre compréhension de la manière dont interagissent ces propriétés en est encore à un stade embryonnaire. Mettre en œuvre ces technologies sans comprendre ces interactions est une pure fiction qui causera plus de tort que de bien.
>
> – Ulrich Aïvodji

Encore plus surprenant est le fait qu'aucune de ces notions mathématiques de l'équité ne peut réellement saisir l'essence d'une expérience « équitable ». Comme Rumman Chowdhury l'explique dans sa discussion sur les audits d'équité, il en est ainsi, parce que du point de vue de l'utilisateur ou de l'utilisatrice, l'équité n'est pas simplement un objectif mathématique, mais bien une expérience. Il ne s'agit pas uniquement d'évaluer objectivement si un ensemble d'algorithmes est équitable, mais plutôt de tenir compte de tous les aspects qui contribuent à l'expérience d'un utilisateur ou d'une utilisatrice qui interagit avec la technologie.

> Pour une expérience utilisateur ou utilisatrice en particulier, plusieurs algorithmes fonctionnent simultanément. […] C'est l'une des choses qui manque vraiment [dans la] conversation réglementaire. Nous traitons un algorithme et disons « il s'agit de l'algorithme et l'algorithme doit être audité ». [Mais,] il n'y a pas d'algorithme Twitter. Twitter utilise plusieurs algorithmes, mais [vous n'avez uniquement] qu'une expérience Twitter.
>
> – Rumman Chowdhury

Comme l'explique Rumman Chowdhury, même si nous procédions à un audit formel d'un seul algorithme, les interactions des utilisateurs ou utilisatrices avec les modèles d'IA ont plusieurs facettes. Ils et elles interagissent avec un écosystème de modèles, et non avec un seul. Ce problème ne fait qu'exacerber la fluidité de la définition de l'équité. Que se passe-t-il si nous permettons de faux positifs dans un modèle et des faux négatifs dans un autre ? Qu'est-ce que cela signifiera pour l'utilisateur ou l'utilisatrice ?

Francisco Marmolejo-Cossío souligne que nos besoins en matière d'équité sont en constante évolution, précisément parce qu'ils sont si dépendants du contexte. Ceci crée une situation dans laquelle la recherche et la mise en œuvre d'objectifs sont constamment modifiées pour ressembler à chaque besoin unique en matière d'équité. Alors que nous découvrons une panoplie de contextes de plus en plus larges dans lesquels les notions d'équité doivent être définies, la liste des définitions de l'équité s'allonge pour tenir compte de notions de moralité plus complexes.

> [L'équité] est une pratique grandement interdisciplinaire. Il ne s'agit pas simplement de l'aspect technique provenant des STIM ainsi que des techniques algorithmiques et d'optimisation que nous utilisons, il s'agit aussi du contexte sociétal qui est intégré dans les caractéristiques précises ou dans les intrants dont nous disposons.
>
> – Francisco Marmolejo-Cossío

Nos experts et expertes ont soulevé un point crucial à savoir la différence profonde entre des indicateurs d'équité et les indicateurs standards actuellement utilisés dans la formation des modèles d'IA. Margaret Mitchell soulève cette distinction et explique ses conséquences :

> L'une des choses que nous devons faire est de concevoir des protocoles standards pour l'évaluation. Ce qui est actuellement à la fine pointe consiste encore à utiliser des indicateurs que d'autres personnes ont définis, soit ceux que l'on voit couramment et par défaut dans la littérature, comme le score F1. Et se concentrer uniquement à faire grimper ce nombre. Dans l'espace de la sécurité et de l'équité, ce que plusieurs d'entre nous tentent de faire [est de dire] « non, en fait, il faut décomposer la performance du modèle. Il est nécessaire d'évaluer comment le modèle se comporte dans plusieurs différents contextes en utilisant plusieurs différents indicateurs ». Et, actuellement, ce n'est pas ce que nous faisons. La communauté de l'apprentissage automatique ne l'a pas encore réalisé. L'évaluation doit tenir compte du contexte sociétal, ce qui consiste donc à se demander qui est le plus susceptible d'être lésé et comment le système est le plus susceptible d'être mal utilisé et [...] de faire des erreurs problématiques.
>
> – Margaret Mitchell

Ce que met en lumière Margaret Mitchell est les méthodes d'évaluation actuellement utilisées pour entraîner les systèmes d'IA. Dans le contexte actuel du secteur, un modèle est souvent évalué selon sa capacité à optimiser une quelconque fonction mathématique complexe qui doit représenter la satisfaction de l'utilisateur ou de l'utilisatrice. Un système de recommandation de vidéos, par exemple, apprendrait à prédire avec exactitude la probabilité d'un utilisateur ou d'une utilisatrice à cliquer sur une vidéo. Cependant, cette évaluation ne tient aucunement compte des implications sociétales pour une personne regardant cette vidéo. Dans les faits, si l'utilisateur ou l'utilisatrice retourne ensuite à une vidéo, le modèle pourrait être récompensé deux fois, même si la vidéo présente du contenu violent ou perturbant qui pourrait avoir des conséquences plus tard dans la vie de cet utilisateur ou de cette utilisatrice. Margaret Mitchell souligne l'importance de créer des méthodes d'évaluation qui tiennent compte du contexte sociétal et plaide en faveur de la prise en compte de nouveaux indicateurs qui intègrent l'évaluation des risques associés à divers résultats. Selon elle, bien que les chercheurs et chercheuses universitaires s'intéressent aux objectifs d'équité et que ces derniers soient discutés dans ce milieu, ils sont rarement intégrés dans les modèles du secteur. Ceci mène à de sérieux problèmes puisque si un modèle n'a jamais été testé pour son équité, il est impossible de prouver qu'il est inéquitable.

Plus préoccupant encore est le fait que, parfois, ces modèles finissent par se comporter de manière que les concepteurs et conceptrices n'avaient pas prévu. En 2018, par exemple, Amazon a mis en œuvre un outil d'embauche fondé sur l'IA pour régler les déséquilibres de genre dans son codage de la main-d'œuvre (Dastin, 2018). Or, on a rapidement découvert que le modèle sélectionnait de manière disproportionnée des candidats plutôt que des candidates. Même si les concepteurs et conceptrices avaient expressément pour intention d'éviter ce problème, ils et elles avaient entraîné ce modèle en utilisant un ensemble de données alimenté par des données de C.V. provenant largement d'hommes et le modèle avait adopté un C.V. d'homme typique comme règle d'or pour l'embauche. Concrètement, ceci signifiait que tout C.V. ayant des attributs différents, comme le fait d'avoir fréquenté une université pour femmes seulement, était classé plus bas (Dastin, 2018). Par conséquent, si aucun changement n'avait été apporté à sa fonction d'évaluation dans le but de tenir compte de ce déséquilibre et de l'éliminer, le modèle aurait perpétué le déséquilibre auquel il avait été exposé.

> Il est très difficile pour une personne d'obtenir réparation si elle croit être victime de préjugé ou de discrimination, parce que ce qu'elle demande est : « Eh bien, prouvez qu'un biais algorithmique existe dans le système. » Lorsque vous n'êtes pas un expert ou une experte de la science des données, vous ne savez pas comment ça fonctionne.
>
> – Rumman Chowdhury

Comme Rumman Chowdhury l'explique, le manque de définitions de l'équité et la faible participation du secteur dans le domaine de l'équité privent les utilisateurs et utilisatrices de remettre en question les pratiques discriminatoires d'un modèle d'IA. À l'avenir, cette situation ne fera qu'exacerber l'iniquité des structures qui jouent un rôle de plus en plus important dans nos vies quotidiennes.

2. Le manque de diversité

À l'origine de l'équité dans tout algorithme sont les personnes qui le conçoivent. Comme le soutient Arisa Ema, « *qui* en discute ne peut être étranger à ce *qui* est discuté ». Nous pouvons établir un lien entre les notions de diversité d'équipe et des résultats d'équité algorithmique. Intuitivement, compte tenu de la complexité associée à la création d'un algorithme entièrement équitable, il est clair que sans l'emploi d'un bassin diversifié de penseurs et de penseuses, une entreprise a peu de chance d'y arriver. Dans le meilleur scénario, nous pourrions espérer que les équipes d'ingénierie sont constituées d'un bassin de travailleuses et de travailleurs interdisciplinaires et multidimensionnels qui sont profondément conscients et conscientes des besoins et des préoccupations des utilisateurs et utilisatrices. Mais, comme nous le démontrons dans cette section, ce n'est pas le cas. Nous expliquons, dans cette discussion, de quelle manière un manque de diversité crée un environnement dans lequel des pratiques d'IA contraires à l'éthique peuvent prendre le dessus.

Rumman Chowdhury décrit comment une réelle équité commence au tout début de la conception d'un algorithme. Bien que la question de l'équité soit souvent externalisée à des « experts et expertes de l'équité », elle explique pourquoi cela n'accroît pas nécessairement les normes éthiques.

> Et le problème de [séparer les fonctions responsables de superviser l'IA, comme la protection de la vie privée et la sécurité] est que vous vous retrouvez avec une résistance active, une résistance passive et une inaction générale. Ou, comme nous l'observons, les gens pensent que cela accroît votre charge de travail parce qu'ils ne veulent pas le faire. [...] L'une des raisons pour lesquelles nous avons choisi [d'adopter ce comportement] est [parce que] le ou la responsable du modèle est ciblé comme la personne possédant la plus grande expertise au sujet du modèle. C'est elle qui l'a construit. C'est elle qui le connaît le plus. [Il y a] cette idée d'externalisation morale, soit « n'êtes-vous pas des gens d'éthique ? Ne devriez-vous donc pas le faire, comme, allez rendre mon modèle éthique ? ». Mais je ne dispose pas du personnel pour aller dans votre modèle, faire ces changements, etc., et, ensuite, m'adresser à vous pour obtenir votre approbation. C'est un échec. [...]
>
> – Rumman Chowdhury

Si l'éthique n'est considérée qu'une fois le produit conçu, il y a beaucoup moins de chances que des changements fondamentaux soient effectués. Séparer les fonctions de conception de produits et d'éthique crée une division artificielle entre la conception du produit et ses résultats en matière d'équité. Par ailleurs, cela nuit à la collaboration entre les créateurs et créatrices et les auditeurs et auditrices et place le fardeau de l'équité sur les personnes à qui l'on donne souvent moins de pouvoir organisationnel. Rumman Chowdhury soutient que l'évaluation de l'équité d'un algorithme devrait être une préoccupation tout au long de sa conception et de son développement et non pas simplement externalisée à une autre équipe une fois que le travail a déjà été fait. En d'autres mots, il est nécessaire de prendre du recul et de réévaluer la valeur de l'interdisciplinarité, tant pour la société, puisqu'elle garantirait une plus grande supervision tout au long du processus de création de technologies, que pour le secteur, puisqu'elle réduirait les coûts à long terme.

La nécessité d'une main-d'œuvre consciente de l'équité fait en sorte que les décisions d'embauche sont une étape cruciale en matière d'IA éthique. Sans personnel conscient de l'équité qui peut s'approprier le modèle d'un point de vue éthique, les entreprises sont forcées d'externaliser leurs évaluations éthiques.

> Une diversité au sein des équipes est cruciale, non seulement au sein de celles qui traitent d'équité, mais bien au sein du secteur en général. Nous construisons des produits et fournissons des services qui changent la manière dont les gens vivent leur vie et, en général, ces produits sont destinés aux personnes partout sur la planète. Les équipes qui construisent ces produits doivent donc refléter le public auquel sont destinés leurs outils.
>
> – Rumman Chowdhury

Comme l'expliquent Rumman Chowdhury et Francisco Marmolejo-Cossío, la diversité est la première étape pour mettre en place une équipe axée sur l'équité. C'est pour cette raison qu'ils décrivent la diversité dans l'embauche comme un élément crucial pour assurer l'équité algorithmique.

> L'ensemble [du] processus de va-et-vient qui permet de peaufiner les modèles en intégrant une plus grande réalité terrain est très important, et ce, même pour ceux et celles d'entre nous qui pensent de manière un peu plus théorique. Dans ce contexte, la diversité des antécédents est très importante. Et, bien entendu, cette diversité se manifeste dans plusieurs aspects. La diversité socio-économique, la diversité géographique, la diversité des genres sont aussi des enjeux extrêmement importants. En tant qu'homme aidant à diriger certains de ces groupes, je trouve qu'un des moyens qui s'avère le plus efficace pour nous est d'essayer de maintenir une équipe de direction très diversifiée.
>
> – Francisco Marmolejo-Cossío

C'est ici que le lien entre la diversité et ses effets sur l'équité ressort. Les entreprises qui créent les modèles d'IA doivent comprendre les conséquences de leur travail. Donc, comme Rumman Chowdhury et Francisco Marmolejo-Cossío l'ont souligné, les équipes qui construisent ces produits doivent, au moins à un certain degré, refléter le public à qui les outils sont destinés. Ainsi, l'écart entre les créateurs et créatrices d'une technologie et ceux et celles qui évaluent les impacts de la manière dont elle est mise en œuvre devient une lacune critique qui est directement liée à un manque de diversité au sein des équipes d'IA.

Au cours des dernières années, le secteur des technologies a commencé à réagir aux appels de la société en faveur de l'embauche de travailleurs et travailleuses issus de plus larges horizons. Mais, malheureusement, la mise en place d'équipes diversifiées se bute à plusieurs obstacles, plus particulièrement dans le cas de postes hautement spécialisés. Comme le mentionne Margaret Mitchell, même lorsque les entreprises veulent embaucher des personnes issues de groupes de minorités, l'environnement technologique peut ne pas leur permettre de s'épanouir dans leur rôle.

> En général, la diversité en technologie est horrible, en partie parce que, bien que les entreprises puissent (en quelque sorte) comprendre ce qu'est la diversité, elles ne semblent pas comprendre ce qu'est l'inclusion ou comment elle fonctionne. Ceci signifie donc que, même si les entreprises sont en mesure d'*embaucher* des personnes ayant des caractéristiques qui sont sous-représentées dans le secteur des technologies, elles peinent à les retenir. Pour ceux et celles qui possèdent des caractéristiques sous-représentées, la diversité sans l'inclusion est synonyme de torture professionnelle.
>
> – Margaret Mitchell

Ceci s'ajoute au fait que les personnes occupant des postes élevés sont les seules à prendre des décisions de conception cruciales. À elle seule, l'embauche d'un bassin diversifié de nouveaux ingénieurs et de nouvelles ingénieures ne changera pas le fait que les postes les plus élevés sont majoritairement occupés par des personnes qui ne sont pas issues de la diversité.

> Pire encore, les efforts en matière de diversité tendent à se concentrer sur des personnes que les dirigeants et dirigeantes considèrent comme étant « sous » eux et elles, comme de nouveaux ingénieurs ou de nouvelles ingénieures. Puisque les personnes occupant des échelons plus bas ont peu à dire dans la définition de la culture, contrairement à celles qui occupent les postes les plus élevés et qui tendent à posséder les caractéristiques dominantes dans le secteur des technologies, l'entreprise, dans son ensemble, devient alors structurellement raciste, sexiste, etc., repoussant ceux et celles qui ne sont pas alignés sur les attentes et les normes des comportements inhérents à la culture qui, quant à elles, sont définies du haut vers le bas.
>
> – Margaret Mitchell

Dans sa discussion sur la promotion d'une IA digne de confiance au sein des entreprises, Margaret Mitchell établit un lien très étroit entre l'inclusion et la diversité d'un milieu de travail et des résultats d'équité algorithmique.

> L'un des plus grands obstacles au fonctionnement d'une IA éthique au sein des entreprises est le fait qu'il n'existe pas un flux d'informations du bas vers le haut. Plus vous occupez un poste élevé, plus on vous accorde uniquement le pouvoir d'une personne possédant une expertise. Parallèlement, si vous occupez un poste plus bas, il est impossible pour vous de dire quoi que ce soit à quiconque. Personne d'autre ne vous écoutera ou ne se souciera de vous parce que l'entreprise a déclaré qu'une personne était l'expert ou l'experte. Et, souvent, les personnes qui occupent les postes les plus élevés semblent penser qu'elles sont les expertes. Vous savez, si ces personnes font partie de cette entreprise depuis si longtemps, c'est un peu comme une pensée de groupe. La possibilité d'un flux d'informations du bas vers le haut, des experts et expertes de l'entreprise à la direction, est nécessaire. Or, en ce moment, ce n'est pas le cas.
>
> – Margaret Mitchell

Francisco Marmolejo-Cossío soulève aussi la façon dont la diversité et l'équité sont interreliées. Il soutient qu'un manque de diversité au sein d'une équipe d'IA est l'une des raisons pouvant expliquer pourquoi elles ne cherchent pas de contributions des personnes sur le terrain, ce qui est crucial pour tester les exigences en matière d'équité des applications d'IA.

> Les points de vue des personnes qui travaillent sur le terrain sont cruciaux. Ils permettent de déterminer si la mise en œuvre d'une manière particulière d'un élément passerait totalement à côté de certaines réalités terrain. Je trouve aussi que, dans certains groupes dans lesquels j'ai travaillé dans le passé, cette considération est absente soit simplement parce qu'elle n'est pas présente en raison de problèmes systémiques ou d'un manque de diversité, mais aussi, parfois, par commodité,

où une certaine commodité mathématique entre en quelque sorte en ligne de compte quand on travaille avec les modèles.

– Francisco Marmolejo-Cossío

Vidushi Marda plaide en faveur d'une réforme plus importante en ce qui a trait aux structures de pouvoir au sein des entreprises technologiques qui s'étendrait bien au-delà des pratiques d'embauche aux échelons plus bas. Elle souligne de quelle manière les dynamiques de pouvoir peuvent avoir des effets qui se font ressentir dans les discussions sur l'équité et la diversité.

Bien qu'une diversité au sein des équipes soit un solide atout pour comprendre de manière globale l'impact de l'IA dans les sociétés, une prise de conscience beaucoup plus fondamentale est nécessaire au sein du secteur de l'IA éthique, puisque ce sont actuellement les personnes au pouvoir qui décident des normes, de la manière dont elles sont satisfaites et du moment où elles sont satisfaisantes. Ceci représente un mauvais alignement des mesures incitatives et des efforts visant à empêcher, au minimum, une réglementation et une responsabilité concrètes.

– Vidushi Marda

Bien que la diversité et l'équité soient intimement liées, il devient clair de la discussion avec Vidushi Marda que l'une ne peut être traitée comme un substitut à l'autre. Après tout, l'embauche de femmes ne peut assurer à elle seule que les algorithmes qu'elles créent ne discrimineront pas les femmes. Comme l'explique Margaret Mitchell, un élément crucial dans l'embauche d'une main-d'œuvre diversifiée est de veiller à ce que cette diversité soit présente dans l'ensemble de la hiérarchie de l'entreprise. Le point que soulève Vidushi Marda s'appuie sur cette idée pour confirmer que, sans une prise de conscience élargie de la part des structures de pouvoir présentes au sein des organisations, une réelle réforme des pratiques d'IA contraires à l'éthique d'une entreprise est impossible.

Malheureusement, l'éveil des acteurs du secteur aux idées de diversité n'a pas mené à un examen plus approfondi des systèmes sous-jacents qui la crée. Au contraire, les pratiques d'embauche sont, en quelque sorte, devenues des couteaux à double tranchant dans les discussions sur l'équité. Pour en tenir compte, les entreprises technologiques se sont, au cours des dernières années, adaptées et embauchent plus de personnes issues de minorités. Or, s'appuyant sur ce que Margaret Mitchell dit en ce qui a trait à la nécessité d'un « flux d'informations du bas vers le haut », l'embauche de nouveaux membres du personnel ne suffit pas pour changer la culture existante d'une organisation. En opérant des changements en surface, les entreprises finissent donc souvent par éviter les appels en faveur de changements structuraux plus larges et plus radicaux en ce qui a trait aux disparités de pouvoir inhérentes à leur structure organisationnelle.

Il existe actuellement un consensus social quant à l'importance d'aborder la question de l'équité en IA. Parallèlement, cela a peut-être rendu plus difficile le fait de remarquer nos préjugés inconscients. [...] la reconnaissance croissante de l'importance de l'éthique en IA est éclipsée par les conventions et les préjugés inconscients de la communauté elle-même. *Qui* discute [d'éthique en IA] ne peut être étranger à ce *qui* est discuté. « Des principes aux pratiques » est l'un des thèmes centraux du débat récent sur l'IA. Pour être conscients et conscientes de tels préjugés inconscients, il est essentiel de non seulement établir des principes comme l'équité, mais aussi de mettre en place les mécanismes de gouvernance appropriés pour mettre en pratique ces principes.

– Arisa Ema

Ceci a un impact incroyable sur l'équité des algorithmes étant produits, puisque les personnes qui pourraient mieux comprendre les conséquences d'un modèle sont exclues de la conversation, tandis que les personnes qui définissent la rhétorique sur l'équité sont celles qui perpétuent une telle exclusion. Comme le dit Joanna Shields :

> Aujourd'hui, les innovateurs, les innovatrices, les créateurs et les créatrices ne représentent pas nécessairement la population générale. Ce manque de diversité nuit considérablement à la manière dont sont conçus, développés et mis en œuvre les produits d'IA et d'apprentissage automatique. Au fil des ans, nous avons été à même de voir l'IA répliquer les déséquilibres de pouvoir historiques, menant [par exemple] à des services de reconnaissance d'images effectuant des classifications choquantes de personnes issues de minorités et même à des services de reconnaissance faciale de pointe identifiant incorrectement des personnes à la peau plus foncée.
>
> – Joanna Shields

Ultimement, nous constatons que le manque de diversité dans la conception et la mise en œuvre stratégique de l'IA peut empêcher la réalisation des prémisses de l'IA éthique et de ses effets.

3. Le manque de normes éthiques

Les enjeux dont nous avons parlé commencent à s'aggraver. L'absence de définitions universelles de l'équité, jumelée à l'homogénéité des dirigeants et dirigeantes du secteur, a créé un système dans lequel les organisations technologiques décident de la manière dont est construite et mise en œuvre l'IA essentiellement sans contribution extérieure. Comme nos experts et expertes l'expliqueront, le manque de normes d'éthique largement convenues en IA a créé un vide de pouvoir important dans lequel les acteurs du secteur définissent et renforcent leurs propres normes éthiques[10]. Comme le mentionnent les personnes interviewées, les normes qui existent sont, quant à elles, créées pour servir les fins de chaque entreprise. Enfin, nous aborderons aussi de quelle manière la culture actuelle de chaque secteur technologique peut exacerber ces enjeux en raison de son approche de mise en œuvre de l'IA qui consiste « à agir vite et à casser des choses ».

Comme nous l'avons présenté dans notre première section, sans l'établissement d'indicateurs de référence, nous sommes incapables d'évaluer l'équité actuelle des modèles d'IA qui sont mis en œuvre par le secteur technologique. Margaret Mitchell souligne ce fait dans sa discussion du rôle des valeurs dans la définition des actions d'une entreprise. Vidushi Marda soulève aussi l'importance pour une entreprise de disposer d'une forme organisée de normes éthiques qui gouvernent ses comportements.

> Afin de régler les désaccords fondés sur les valeurs, il est important, dès le départ, d'avoir déjà défini et convenu des valeurs de base. Lorsqu'une entreprise, une organisation ou une équipe, etc., est créée, ces créateurs et créatrices apportent avec eux et elles des valeurs implicites qui ont une incidence sur ses décisions. Rendez ces valeurs explicites et mettez-les à jour au fur et à mesure que des personnes d'horizons différents sont prises en compte. C'est cet ensemble de valeurs, les « principes d'une organisation », qui contribue à définir ce qui doit être fait.
>
> – Margaret Mitchell

> Dans le meilleur scénario, les normes éthiques sont construites selon les exigences minimales établies par les droits humains internationaux (l'ensemble de principes le plus universel à notre disposition et qui jouit d'une compréhension et d'un fondement juridique communs dans tous les pays). Deuxièmement, elles sont construites et prises en compte parallèlement avec une prise

10. Le 24 novembre 2021, peu de temps après la rédaction de ce chapitre, l'UNESCO a publié sa Recommandation sur l'éthique de l'IA. Il s'agit d'une première étape importante. Pour obtenir plus d'informations, consultez le https://unesdoc.unesco.org/ark:/48223/pf0000379920.page=14.

> en compte fondamentale des incitatifs institutionnels, structurels et historiques qui, à l'heure actuelle, sous-tendent les organisations et les entreprises.
>
> – Vidushi Marda

Or, l'établissement de valeurs d'entreprise n'est pas suffisant si aucune structure n'est en place pour mettre en application leur importance. Comme l'explique Vidushi Marda, bien qu'il soit courant pour des entreprises du secteur technologique d'établir certains principes directeurs, sans une structure de mise en application, elles ne sont pas tenues responsables d'y adhérer.

> Les normes éthiques jouent un rôle important pour situer les relations publiques d'une organisation. Établir une compréhension de ce que l'entreprise ou l'organisation souhaite idéalement réaliser permet aux parties prenantes de reconnaître les possibilités qui sont acceptables ou non. Toutefois, dans sa forme actuelle, cela ne met pas en place des mécanismes de responsabilité ni des obligations de transparence et n'aborde pas les questions cruciales du pouvoir.
>
> – Vidushi Marda

Vidushi Marda explique ensuite que ces « principes directeurs », tels qu'ils sont établis par chaque entreprise, ne sont pas susceptibles de créer un code de conduite qui peut être respecté dans le secteur plus large.

> Il est important de reconnaître d'emblée que « l'éthique » et « l'IA éthique » revêtent différentes significations pour différentes personnes au sein d'une même organisation et entre les organisations. En pratique, ceci se traduit par le fait que l'on attribue le sérieux d'une norme que l'on s'efforce d'atteindre à de vagues conditions, et ce, sans une compréhension commune des conditions elles-mêmes. Même au sein d'une organisation et d'une entreprise, cela signifie qu'il y a peu ou pas de coordination sur quand et comment sont atteints ou ignorés les normes éthiques.
>
> – Vidushi Marda

Arisa Ema soulève la question de la complexité de la chaîne d'approvisionnement de l'IA et de quelle manière toute gouvernance ou approche réglementaire doit tenir compte des dynamiques complexes des différents niveaux d'acteurs privés. Ses propos sont très pertinents et soulignent que le secteur n'est pas homogène. Hormis les grandes sociétés technologiques mondiales et celles qui offrent des produits directement aux consommateurs et consommatrices (commerces électroniques de détail), il y a plusieurs plus petits acteurs locaux qui sont impliqués. Il est important, par exemple, de considérer l'incidence de la gouvernance et des règlements sur les entreprises en démarrage. Les fournisseurs de service figurent parmi les plus petits acteurs qui lient les différentes entreprises tout au long de la chaîne d'approvisionnement jusqu'à ce que la technologie atteigne les consommateurs et consommatrices (« B2B2C »). Elle met en garde contre la taille considérable des chaînes d'approvisionnement et les défis qu'elles posent en matière de responsabilité.

> En général, la gouvernance de l'IA concerne l'établissement de principes visant à assurer la sécurité et la fiabilité de l'IA au sein d'une entreprise ou d'une organisation et la mise en œuvre de mesures de contrôle pour le développement et l'utilisation. Elle découle aussi de la crainte des entreprises que leurs produits et services ne soient pas acceptés par la société, si elles n'abordent pas les défis que pose l'IA. C'est pour cette raison que le débat sur l'éthique de l'IA est maintenant reconnu comme un « problème directement lié à la stratégie de gestion » des entreprises. Par conséquent, les entreprises mondiales (en particulier) ont créé des comités d'éthique individuels. Cependant, il est souvent techniquement et économiquement difficile pour les petites et moyennes entreprises ainsi que pour les entreprises en démarrage disposant de ressources limitées de mettre en œuvre une gouvernance semblable. Au cours des dernières années, les gouvernements locaux et nationaux ont commencé à exiger une gouvernance des données et de l'IA comme condition d'achat de service,

> ce qui devrait servir de mesure incitative encourageant les entreprises à renforcer leur gouvernance de l'IA. Mais pour les entreprises en démarrage, les grandes exigences en matière d'éthique et de gouvernance de l'IA deviennent aussi un obstacle économique. Il s'agit d'une conséquence ironique, étant donné que l'idée de l'éthique en IA est sous-tendue par des principes qui partagent une vision d'une société qui reconnaît l'inclusivité au sein de la diversité. [...] [Or,] dans une longue chaîne d'approvisionnement, les principes du développement et de l'utilisation de l'IA ne sont pas toujours partagés par les entreprises en aval. Dans le cas d'un accident ou d'un incident dans une entreprise en aval, la mesure dans laquelle la responsabilité peut être retracée à l'entreprise en amont devient difficile à établir. Par conséquent, il est difficile pour une seule entreprise ou organisation de se prémunir contre tous les risques [...].
>
> – Arisa Ema

Réfléchissant aux dangers potentiels en matière d'éthique au sein du secteur de l'IA, Joanna Shields souligne de quelle manière la concentration du pouvoir au sein du secteur technologique crée une situation dans laquelle les personnes définissant ce qu'est l'équité pour chaque organisation n'ont aucun intérêt pour une réforme de l'équité. En outre, Ulrich Aïvodji met en garde contre le fait que les lignes directrices en matière d'éthique sont souvent si larges qu'elles n'ont pas de conséquences concrètes pour les personnes chargées de les appliquer. Cette faille apparemment involontaire pourrait en fait être stratégique pour les entités non conformes.

> Le fait que le pouvoir d'influencer l'utilisation de l'IA et son impact soit concentré entre les mains de quelques personnes est un réel danger. Dans une large mesure, il s'agit de la réalité actuelle des produits, des applications et des services omniprésents que nous utilisons tous les jours. [...] J'étais aux premières loges de la première vague de la révolution numérique, une foire d'empoigne entrepreneuriale dans laquelle les géants technologiques vivaient selon la devise « agir vite et casser des choses ». Leur objectif inébranlable était de croître par tous les moyens nécessaires pour dominer ces secteurs émergents. Il n'existait aucun cadre ou plan international pour gérer la technologie qui progressait dans le secteur privé. Nous sommes donc maintenant forcés de nous attaquer aux conséquences imprévues qui menacent notre vie privée. Ces préjugés classent injustement les personnes et limitent leurs possibilités, en plus de répandre de la désinformation et du contenu illégal.
>
> – Joanna Shields

> Les « lignes directrices en matière d'éthique » actuelles sont présentées comme une liste de recommandations que les entreprises sont « encouragées » à suivre. Une telle approche présente un risque évident de blanchiment éthique, soit un exercice de communication dans lequel des entités malhonnêtes projettent la fausse impression qu'elles respectent une recommandation particulière alors que cela peut ne pas être le cas.
>
> – Ulrich Aïvodji

Ensemble, nos experts et expertes soulignent un conflit d'intérêts dans le fait de dépendre uniquement sur l'autoréglementation pour assurer des progrès en matière d'IA digne de confiance. Ils et elles expliquent que la mise en application réelle de l'équité dans les algorithmes est un processus profondément complexe qui nécessite un engagement financier et un engagement philosophique sincère. Plus important encore, ils et elles rappellent que la mise en œuvre de l'équité à ce moment-ci de la croissance du secteur nécessiterait de s'écarter sérieusement du statu quo, ce qui pourrait ne pas être financièrement viable.

> Si des normes éthiques sont établies avec la compréhension qu'elles doivent faciliter l'adoption plus rapide et sans friction des systèmes d'IA, elles ne sont clairement pas destinées à être le mécanisme de responsabilité que nous envisageons. [...] Il est aussi essentiel de comprendre que ceux et celles qui plaident en faveur d'une IA éthique sont également ceux et celles qui détiennent le pouvoir, soit

> les GAFAM, les acteurs qui adhèrent aux idées du solutionnisme technologique. Même des initiatives éthiques bien intentionnées sont contraintes à des réalités organisationnelles et structurales qui privilégient la rapidité plutôt que l'examen et le déploiement plutôt que la délibération.
>
> – Vidushi Marda

Comme l'expliquent nos experts et expertes, si les seuls acteurs définissant les normes éthiques sont ceux qui, dans le développement de l'IA, possèdent des intérêts intrinsèquement orientés vers le profit, il y a peu de garanties que l'éthique, la diversité et l'inclusion seront sérieusement prises en compte. En ce sens, si nous souhaitons éviter de perpétuer des pratiques discriminatoires dans nos écosystèmes d'IA, il est urgent de mettre en place des barrières et de renforcer la responsabilité.

> « L'IA éthique », si elle n'est pas accompagnée d'une perspective critique stimulée par les besoins de responsabilité, peut légitimer et consolider des utilisations problématiques et même dangereuses de l'IA. En effet, si les normes éthiques sont définies, interprétées et certifiées par les mêmes acteurs en l'absence de normes juridiques ou de mécanismes de contrôle et de recours, il n'y a pratiquement aucune mesure de contrôle en place.
>
> – Vidushi Marda

L'expérience de Joanna Shields dans le secteur corrobore l'idée d'élargir les discussions sur l'éthique au-delà des murs de chaque entreprise. Elle note que les acteurs principaux du secteur ont historiquement choisi d'ignorer une réforme de l'équité.

> Au cours de la dernière décennie, je suis d'avis que le secteur privé n'a pas démontré la capacité à s'autoréglementer dans le cas des technologies émergentes. Nous avons été témoins de plusieurs points d'inflexion et le secteur des technologies dans son ensemble doit redoubler d'efforts et en faire plus pour régler de manière proactive les problèmes qui découlent des technologies qu'il crée.
>
> – Joanna Shields

Maintenant que nous avons soulevé trois lacunes sérieuses dans le tissu de l'IA éthique, nous discuterons des divers fils qui le retiennent. Notre discussion nous mènera vers les diverses autres parties prenantes, à l'extérieur des membres du secteur eux-mêmes, qui ont un intérêt pour les questions d'équité algorithmique. Nous présenterons les visions de nos experts et expertes pour une plus grande participation à des discussions élargies ainsi que des moyens pour sensibiliser et éduquer la société à prendre part à la construction d'un meilleur avenir pour l'IA. Enfin, nous explorerons brièvement ce à quoi ressemblerait ce processus participatif et ce qu'il pourrait permettre de réaliser.

QUE POUVONS-NOUS FAIRE

Les sections précédentes ont révélé que le fait de permettre aux organisations bénéficiant de la mise en œuvre rapide de l'IA d'être les seuls acteurs définissant l'équité est, ultimement, une pratique qui n'est pas durable. Les lacunes discutées s'entremêlent en un problème plus important que nous devons considérer : un manque de participation élargie à la gouvernance de l'IA. En laissant les entreprises définir le statu quo, nous avons créé un vide de pouvoir qui permet à de sérieux enjeux de passer inaperçus et de ne pas être traités. L'objectif de cette section est de fournir une orientation concrète sur la manière de traiter les enjeux soulevés précédemment dans ce chapitre. Ce faisant, nous explorons les diverses parties prenantes qui sont aussi impliquées dans le narratif de l'IA et démontrons comment elles peuvent jouer un rôle plus important dans la discussion sur l'équité dans le secteur de l'IA.

Avenue 1 : Intégrer les perspectives des diverses parties prenantes

Notre première orientation pour la mise en œuvre d'un changement porte sur la création d'un large cadre réglementaire. Bien que dans le discours courant, la réglementation soit souvent associée uniquement à des organes gouvernementaux, nos experts et expertes rejettent cette idée. Plusieurs soulignent la complexité de ces impulsions réglementaires et mettent en garde contre le fait qu'une seule partie prenante définisse la réglementation (que ce soit le gouvernement, un audit externe, etc.). En fait, dans nos discussions, nous constatons que la première orientation pour la promotion de l'équité en IA est une approche collaborative adoptée par tous les acteurs impliqués dans la création et la consommation de modèles d'IA. Comme Joanna Shields le soutient, « l'IA n'est pas un superpouvoir qui, un jour, démocratisera les avantages pour tous et toutes et nous devons travailler avec les gouvernements et le secteur technologique pour veiller à ce que l'IA que nous construisons profite à tous et à toutes, et non pas aux 1 % les plus puissants ». En se réunissant, ces acteurs peuvent se donner les outils les uns les autres pour créer un cadre réglementaire informé et robuste.

Margaret Mitchell affirme que la réglementation gouvernementale seule ne peut pas être une véritable panacée. Nous devons plutôt opter pour une approche collaborative qui encourage tous les membres du secteur technologique de tous les niveaux de l'écosystème d'IA à participer à sa responsabilisation. En outre, Rumman Chowdhury mentionne l'importance de comprendre ce que signifie de procéder à un audit d'une technologie avant de plonger, tête première, pour la réglementer.

> [Les entreprises technologiques] ont démontré qu'elles sont incapables de faire le minimum requis dans ce domaine. [...] [Mais] je ne crois pas que la réglementation est une solution miracle. [Cependant], en proposant des objectifs de haut niveau qui sont déployés du haut vers le bas dans une réglementation (gouvernements), [nous pouvons] commencer à offrir des incitatifs aux entreprises pour encourager des comportements conformes à [l'éthique].
>
> – Margaret Mitchell

> Nous ne sommes même pas dans une position où nous pouvons nous entendre sur ce à quoi ressemble un audit et cela, selon moi, est très très inquiétant étant donné que le monde réglementaire semble très excité à l'idée d'établir une réglementation avant même d'avoir fait le travail de base.
>
> – Rumman Chowdhury

Il paraît évident qu'aucune partie prenante (que ce soit un gouvernement ou autre) ne devrait à elle seule être responsable de la définition de l'équité, de l'élaboration de règlements et de leur mise en application par tous les autres acteurs impliqués en IA. Nous présenterons maintenant des extraits de nos discussions avec nos experts et expertes qui explorent les rôles et collaborations entre les diverses parties prenantes, chacune ayant un rôle unique à jouer pour favoriser le changement au sein du secteur de l'IA.

Acteurs étatiques

Vidushi Marda plaide en faveur d'une approche axée sur la gouvernance pour relever les défis en matière d'équité auxquels fait face le secteur. Elle définit un cadre dans lequel les différentes parties prenantes peuvent créer un réseau d'intérêts. Dans le même ordre d'idées, Joanna Shields présente les complexités liées à l'intégration d'organes gouvernementaux dans un paysage fragmenté en un cadre réglementaire commun. Néanmoins, elle soutient qu'une entente internationale est l'un des éléments clés dans la création d'une culture d'IA éthique.

> Dans une situation idéale, nous disposons de ce que j'appelle la « chambre de la gouvernance de l'IA ». Le plancher représente les normes relatives aux droits humains reconnus à l'échelle mondiale en tant qu'exigences minimales que doit satisfaire tout système d'IA. Le plafond, quant à lui, représente les normes éthiques indiquant vers où tout système d'IA doit tendre et dépeint ce que

> tout système d'IA aspire à être. Les murs représentent les divers leviers réglementaires et politiques qui décident de l'étendue et des limites dans lesquelles tout système d'IA devrait fonctionner et ce sont ces normes et spécifications techniques qui facilitent la manière dont nous atteignons toute autre partie de la chambre.
>
> – Vidushi Marda

> Le défi important en matière d'IA éthique découle de l'approche fragmentée adoptée à l'échelle mondiale en matière d'IA et du rythme inégal auquel les lois émergent. Les grandes puissances démontrent déjà de la divergence quant à leurs approches en matière de réglementation et une course à l'IA mondiale est en cours. Assurer le développement d'une IA bénéfique, digne de confiance et robuste nécessite de la collaboration entre des nations démocratiques partageant les mêmes idées et un ensemble de normes internationales qui obligent chaque gouvernement à rendre compte. Exercice d'équilibriste, ces normes doivent tenir compte de la nécessité d'un développement de l'IA conforme aux droits humains et aux valeurs fondamentales, sans pour autant freiner l'innovation. [...] Les nations à l'échelle planétaire doivent s'unir et bâtir un cadre mondial, puis établir des normes internationales sur l'utilisation de cette technologie. Il doit y avoir un organe intergouvernemental qui veille à ce que les entreprises et les gouvernements qui font une mauvaise utilisation de l'IA et qui la mettent en œuvre pour réprimer la vie privée, la dignité, la liberté et les droits humains en subissent les conséquences.
>
> – Joanna Shields

Arisa Ema souligne aussi l'importance de l'uniformité des normes éthiques. Elle aussi commente la profonde importance de créer une approche cohésive à l'équité qui transcende les localités géographiques.

> Pour mettre en œuvre une approche éthique à l'IA au sein d'une organisation, il est essentiel d'établir une structure de gouvernance appropriée. Cependant, en raison de différences de politiques, de valeurs et de structures industrielles entre chaque pays et région, l'état de la gouvernance n'est pas déterminé de manière uniforme. Bien que le respect des diverses valeurs soit important, des cadres de gouvernance trop fragmentés non seulement entravent l'innovation, mais peuvent aussi conduire à un arbitrage réglementaire.
>
> – Arisa Ema

Ulrich Aïvodji soulève un point crucial au sujet des disparités de pouvoir dans la géopolitique, ce qui a une incidence sur les discussions concernant la mise en œuvre de systèmes d'IA. En s'appuyant sur sa connaissance du développement technologique du continent africain, il souligne l'importance de donner aux pays qui ne sont pas encore entrés dans la « course à l'IA » (consultez Savage, 2020) une voix forte pour favoriser un paysage d'IA plus diversifié et plus approprié au contexte dans le but d'atténuer la dépendance croissante sur des technologies étrangères. Ceci permettrait l'émergence d'une pluralité de perspectives à l'échelle internationale et encouragerait un développement de l'IA indépendant dans les pays du Sud.

> Si un gouvernement n'investit pas dans des initiatives locales pour favoriser le développement de systèmes d'IA adaptés aux besoins locaux, des technologies développées en tenant compte des perspectives et des intérêts occidentaux seront importées et accroîtront encore davantage la dépendance du pays sur les technologies occidentales.
>
> – Ulrich Aïvodji

Rumman Chowdhury note aussi que l'harmonisation des perspectives ne devrait pas mener à l'adoption hégémonique d'une vision particulière de l'équité en IA. Dans sa discussion sur les mesures réglementaires, elle soulève une des lacunes de l'uniformisation des règlements à l'échelle mondiale.

> Je comprends qu'un environnement réglementaire sans friction facilite les choses du point de vue de la logistique de la mise en application. Or, je pense que nous courrons le risque d'ignorer la diversité de choix. Alors, lorsque nous parlons de l'état idéal, j'ignore de quelle manière créer un ensemble de lois pour gouverner tout ce qui concerne l'éthique et l'utilisation algorithmiques. Cela permettrait également l'équité dans la mesure où une personne comprend l'équité et nous donnerait aussi le choix et l'agence. Et cela n'a jamais existé dans l'histoire du monde, essentiellement un accord gouvernemental unifié sur quoi que ce soit.
>
> – Rumman Chowdhury

Auditeurs et auditrices externes

Un autre joueur important dans l'établissement d'un cadre réglementaire est les auditeurs et auditrices externes. Joanna Shields souligne l'importance de mener des audits indépendants des grandes entreprises technologiques. Elle explique que laisser les experts et expertes sans recours, quelle qu'en soit la forme, contre les entreprises peut créer une situation dangereuse dans laquelle leurs signaux d'alarme sont tout simplement ignorés.

> Dans le cas récent du congédiement de Dre Timnit Gebru de Google, l'éthique de mener des recherches avec de grandes entreprises de technologies a été remise en question et il en découle un plaidoyer en faveur d'une recherche indépendante, financée par des fonds publics, sur l'IA et ses préjudices potentiels ainsi que d'une législation robuste comme celle mise récemment de l'avant par l'UE afin de veiller à ce que la technologie soit développée de manière responsable.
>
> – Joanna Shields

Experts et expertes du domaine

Les experts et expertes du domaine sont un groupe de parties prenantes qui est étroitement lié aux autorités d'audit. Ils et elles ont un rôle crucial à jouer pour cadrer les modèles d'IA et façonner leur application à des domaines précis. Comme nous l'avons mentionné précédemment dans ce chapitre, il est communément admis que le caractère contextuel de l'application de l'IA nécessite une collaboration plus étroite avec les professionnels et professionnelles travaillant sur le terrain. Margaret Mitchell corrobore l'importance des experts et expertes du domaine dans la discussion sur la mise en œuvre de l'IA.

> Pour toutes ces technologies différentes, ceux et celles disposant d'une expertise pertinente (médecins, climatologues, etc.) ainsi que ceux et celles qui sont touchés (les personnes ayant des capacités différentes, les personnes vivant dans des régions plus isolées, les personnes susceptibles d'être déplacées par la technologie) devraient participer à façonner ce que la technologie accomplit, à établir la façon dont elle est utilisée et à déterminer si elle devrait même exister.
>
> – Margaret Mitchell

Margaret Mitchell suggère également la possibilité d'élargir la participation multipartite en proposant le renforcement des relations entre les experts et expertes en éthique et les autorités de réglementation. Elle crée aussi l'espace pour une autre partie prenante à cette collaboration réglementaire. Margaret Mitchell présente le potentiel d'avoir les auditeurs et auditrices externes jouer un rôle de renforcement des directives gouvernementales.

> Il y a un rôle à jouer ici, tant pour les entreprises et les organisations que pour la réglementation [...] Je pense que les gouvernements ont un rôle à jouer dans la définition des objectifs de haut niveau quant à ce qu'ils souhaitent retirer des systèmes : transparence, robustesse, peu importe, ce genre [d'objectifs] de haut niveau. Ensuite, les systèmes technologiques ou les personnes qui les développent peuvent fournir des éléments de preuve de cela en s'appuyant sur ce qu'ils comprennent au sujet de leurs systèmes. Il s'agit d'une [approche] haut vers le bas et bas vers le haut. La partie chargée de la réglementation dit « Nous souhaitons voir ceci et cela » et l'entreprise présente divers indicateurs qui, selon elle, sont appropriés. [...] Il s'agit donc de la rencontre d'une part d'autoréglementation et d'une autre part de réglementation externe.
>
> Actuellement, il n'y a pas d'experts et d'expertes au sein d'une entreprise. Vous trouvez des personnes avec du pouvoir au sein de l'entreprise. Donc, si vous disposez de chercheurs et chercheuses de pointe travaillant sur l'équité, ce sont eux et elles qui sont les bons experts et les bonnes expertes pour discuter d'équité avec les autorités de réglementation. Le fait de faire correspondre l'expertise aux besoins des représentants des gouvernements est essentiel. [...] Ensuite, cette idée d'audit indépendant où des chercheurs et chercheuses disent à l'auditeur ou à l'auditrice « voici les différents problèmes ici, [voici] ce que nous pensons voir ». Et l'auditeur ou l'auditrice peut s'en charger. Et cela permet également un certain degré de confidentialité.
>
> – Margaret Mitchell

Société civile

Finalement, le plus grand groupe de parties prenantes est la société civile. Cette catégorie nous comprend tous et toutes, soit en tant qu'utilisateurs et utilisatrices de la technologie ou en tant que personne qui pourrait dans diverses situations, sciemment ou non, être en contact avec les applications d'IA mises en œuvre par des tiers. Francisco Marmolejo-Cossío souligne l'importance d'intégrer les utilisateurs et utilisatrices dans le processus réglementaire, mais aussi d'être attentifs et attentives à la manière dont les préférences des utilisateurs et utilisatrices quant à leurs interactions avec la technologie sont fondées sur le contexte et dépendent de la sensibilisation et de la responsabilisation.

> Compte tenu de ces différentes préférences au sein de la population, je pense qu'attribuer une part de responsabilité aux utilisateurs et utilisatrices afin d'essayer d'opérer ce changement au sein du secteur sera fondamentalement différent. [...] Peut-être qu'il s'agit de quelque chose dans lequel des organisations externes peuvent intervenir et proposer quelque chose alors que tout le monde est sur un pied d'égalité. Donc, au lieu d'imposer complètement un ensemble d'indicateurs ou une conversation externe, par exemple, il pourrait être intéressant de fournir assez d'incitatifs à ces segments de la population qui ne participent pas nécessairement à la conversation sur la protection de la vie privée et sur l'équité pour qu'ils y participent. Cela pourrait donner un certain élan à cette approche basée sur le client et la cliente.
>
> – Francisco Marmolejo-Cossío

Il souligne le rôle que les groupes de la société civile peuvent jouer pour donner aux utilisateurs et utilisatrices les outils dont ils et elles ont besoin pour aplanir les disparités lorsqu'il est question de la sensibilisation à l'éthique de l'IA. De tels rôles peuvent aussi être élargis à tous les segments de la population soumis à la mise en œuvre de l'IA dans différentes sphères, avec d'autres acteurs du secteur privé ou public suivant l'exemple d'une approche de responsabilisation des utilisateurs et utilisatrices.

Ultimement, il est important de noter que, tout au long de nos conversations, les experts et expertes ont souligné que la réglementation n'est pas une question d'externalisation de la tâche consistant à attendre des organismes gouvernementaux qu'ils façonnent le paysage de l'IA, mais plutôt une question d'élaboration d'un ensemble de normes et de standards communs à adopter par l'ensemble du secteur technologique.

Avenue 2 : Sensibiliser par l'éducation

Une notion sous-jacente pour la mise en œuvre d'une participation élargie est la nécessité de conditions égales pour tous et toutes afin de permettre à toutes les parties de s'engager dans des discussions signifiantes et constructives. Plusieurs des personnes interviewées ont mentionné l'importance de la sensibilisation, de l'éducation et de la promotion de l'autonomie des personnes à faire des choix éclairés en ce qui a trait à leurs interactions avec les technologies d'IA. Compte tenu de la nature souvent isolée et opaque du secteur technologique, il est facile d'exclure les penseurs et penseuses non techniques des discussions concernant la création et la mise en œuvre de systèmes d'IA. Dans les faits, ce n'est qu'au cours des dernières années que les discussions sur l'IA éthique ont été intégrées au discours courant. Sans auditoire pouvant participer aux discussions relatives à l'IA, aucune réforme n'est possible. Élever les discussions portant sur l'équité nécessite une perspective critique de la part de la société qui ne peut être favorisée qu'en fournissant un accès à grande échelle à de l'éducation en matière d'équité. Cette éducation est nécessaire à tous les niveaux de notre société, tant technique que non technique. Comme le mentionne Arisa Ema, comprendre le rôle que jouent les préjugés dans le façonnement du développement technologique est nécessaire tant pour ceux et celles qui consomment la technologie que pour ceux et celles qui la créent.

> La technologie ne se développe pas d'elle-même. Le développement technologique doit avoir un objectif et il est influencé par les besoins de la société et les visions que les personnes ont d'elle. [...] Les craintes concernant les préjugés dans les algorithmes et les données d'IA sont maintenant partagées à l'échelle mondiale. L'une des raisons de ces préjugés est que notre société est elle-même partiale au départ. Il est difficile pour nous d'être conscients et conscientes des préjugés au sein de notre société.
>
> – Arisa Ema

Sans une connaissance des écueils dans le développement technologique et des préjugés qui les façonnent, il n'existe aucune possibilité de réévaluer nos visions communes et encore moins de formuler un cadre réglementaire approprié. Malheureusement, il est difficile de combler le fossé entre ceux et celles qui font partie du domaine technologique et ceux et celles qui n'en font pas partie. Comme le souligne Francisco Marmolejo-Cossío, il est grandement nécessaire de créer un dialogue entre ceux et celles qui sont touchés par l'IA et ceux et celles qui la créent.

> [Nous sommes] conscients et conscientes du manque d'équité et de la nécessité de proposer [plus] d'ateliers d'information sur l'IA ou de simplement diffuser de l'information. De cette manière, nous aurions un auditoire ou un impact auprès de certaines de ces communautés qui sont spécifiquement touchées par les questions d'équité... [Ceux et celles qui] pourraient ne pas être nécessairement au courant du fait qu'il existe des disparités dans les résultats ou les possibilités, compte tenu des mécanismes qui sont en place en arrière-plan. Il s'agit de prendre [une] approche de sensibilisation... [de prendre le temps] de considérer les segments de la population qui souffrent des résultats inéquitables et d'avoir des discussions avec eux.
>
> – Francisco Marmolejo-Cossío

Francisco Marmolejo-Cossío explique ensuite l'importance de sensibiliser les populations aux résultats inéquitables des modèles d'IA avec lesquels ils interagissent et souligne aussi l'importance d'éduquer les auditoires non techniques au processus d'automatisation et à la façon dont ils peuvent avoir une incidence sur leurs vies.

> [Nous pouvons] offrir d'autres formes d'éducation de ceci dans la sphère de l'éducation. À mesure que nous avançons et que nous réfléchissons à la manière de changer l'école primaire et l'école secondaire, cela pourrait certainement être quelque chose de fondamental dans la sphère de

> l'éducation. Réfléchir à l'automatisation, à ses impacts et à l'éthique qui se cache derrière tout cela, voilà quelque chose qui devrait être considéré au fur et à mesure qu'elle est plus présente [...].
>
> Un bon scénario pour l'avenir [serait] d'avoir une seule plateforme dans le système national d'éducation publique pour [avoir une discussion avec les décideurs et décideuses pédagogiques, mais aussi avec les personnes techniques comme vous et d'autres qui travaillent dans cet espace], parce que cela ne fera qu'être de plus en plus présent au cours des prochaines années.
>
> – Francisco Marmolejo-Cossío

Nous lui avons demandé de quelle manière un tel programme d'éducation devrait être défini et il a souligné l'importance d'une approche interdisciplinaire pour expliquer les techniques algorithmiques.

> [Créer un tel curriculum] est un exercice interdisciplinaire. Il ne s'agit pas simplement de l'aspect technique d'un point de vue des STIM, des techniques algorithmiques et des techniques d'optimisation que nous utilisons, il s'agit aussi du contexte sociétal qui est intégré dans les fonctions particulières ou des intrants dont nous disposons. Ultimement, il y a un processus décisionnel qui est facilité en partie par les techniques qui sont disponibles dans le système et ces dernières peuvent être incluses dans la discussion. Par exemple, réfléchir aux choix qui ont été faits pour créer l'ensemble de données, à ce que cela apporte et aux avantages et aux disparités qui sont générés. Et, ensuite, je pense qu'il y a beaucoup de choses que nous pouvons faire [avec] la [partie] technique de cette perspective pédagogique.
>
> – Francisco Marmolejo-Cossío

L'éducation est un élément crucial pour favoriser une grande participation. Francisco Marmolejo-Cossío fournit un compte rendu informatif de la manière dont l'éducation peut créer une conversation avec une grande participation de la communauté et permettre aux personnes de réfléchir à certains enjeux de transparence et de méfiance plus larges au sein de la sphère de l'IA.

> Ultimement, la confiance est toujours un enjeu avec ces choses. Faites-vous confiance aux personnes, au comité, au pouvoir que vous mettez entre les mains des personnes qui pourraient créer quelque chose qui, si mis en œuvre, touchera des millions de personnes ? Si nous nous concentrons uniquement sur les écoles primaires et secondaires, par exemple, et entamons une discussion de politique à l'échelle nationale, la politique pourrait avoir une grande incidence sur les perceptions négatives ou sur le manque potentiel de perception.
>
> – Francisco Marmolejo-Cossío

Ulrich Aïvodji commente aussi l'importance de la sensibilisation, portant son attention sur un autre but important d'une société informée en matière d'IA, soit d'éduquer la société au revers négatif des discussions sur l'IA éthique.

> Pour minimiser les risques de blanchiment éthique, la moindre des choses que nous pouvons faire est de sensibiliser les gens aux différentes façons dont les recommandations éthiques peuvent être facilement contournées par des entités bien motivées. [...] L'Afrique et la diaspora africaine comptent déjà plusieurs chercheurs et chercheuses travaillant activement à la sensibilisation aux préjudices que peuvent causer les systèmes décisionnels automatisés. Il est important de porter une attention à leurs travaux et de ne pas uniquement écouter les évangélistes technologiques aux points de vue toujours optimistes quant à l'IA.
>
> – Ulrich Aïvodji

Ultimement, sans éducation à l'importance et aux dangers de pratiques d'IA contraires à l'éthique et non réglementées, personne, mis à part les dirigeants et dirigeantes du secteur de l'IA, ne peut dénoncer les pratiques préjudiciables. Par ailleurs, sans éducation sur ce sujet, il est facile de se laisser prendre à des démonstrations superficielles de pratiques équitables destinées à dissimuler des problèmes structurels sous-jacents plus sinistres.

Avenue 3 : Élever la discussion sur les données et activer les réformes

Le débat grandissant dans l'espace public concernant la collecte de données est un très bon exemple de la mobilisation des parties prenantes en raison des conséquences de l'usage de l'IA et de la lutte en faveur d'une autonomie quant à la façon dont est mise en œuvre l'IA dans leurs vies. Après les scandales, comme Cambridge Analytica, les utilisateurs et utilisatrices non techniques sont devenus plus conscients et conscientes des façons problématiques utilisées pour recueillir leurs données (Confessore, 2018). Cette reconnaissance de l'impact de l'IA a catalysé un mouvement d'action beaucoup plus large, permettant à certaines personnes, qui autrement ne l'auraient pas fait, d'exprimer leurs opinions dans l'espace de l'IA (Garret, 2018). Lorsque le public est informé, chaque membre de notre société peut participer dans la création des éléments essentiels qui constituent le fondement de tout système d'IA, c'est-à-dire les données. Selon Rumman Chowdhury, être propriétaire de ses données signifie être capable de prendre des décisions éclairées quant aux données qui sont utilisées et à la manière dont elles sont utilisées.

> Il y a un certain contrat social lorsqu'il est question de l'IA. [...] « Je comprends que je vous fournis certaines données et informations et, qu'en retour, vous les utilisez pour faire une prédiction quelconque, qu'il s'agisse d'améliorer vos modèles afin d'améliorer votre produit en ce qui me concerne. » [...] La façon dont les organismes utilisateurs peuvent influencer l'équité algorithmique [est en] permettant aux deux groupes [utilisateurs et utilisatrices et entreprises] de comprendre et de reconnaître la valeur du contrat social comme il est écrit et de ne pas [permettre aux entreprises] de l'exploiter. Parce que, honnêtement, une grande partie de la mauvaise utilisation des algorithmes provient d'externalités négatives, celles qui vont au-delà de ce que l'utilisateur moyen ou l'utilisatrice moyenne accepterait probablement comme utilisation de ses données.
>
> Je pense que le meilleur scénario est un monde dans lequel les personnes sont propriétaires de leur expérience et ont le droit d'accepter ou de refuser de faire les choses. Elles ont le droit de partager ou non leurs données. Elles ont le droit de profiter de quelque chose lorsque cela leur convient, mais d'être laissées à elles-mêmes lorsque c'est ce qu'elles souhaitent. Je pense que parfois les politiques vont un peu loin, alors soit vous participez pleinement, soit vous vous retirez complètement. Je pense aussi qu'un monde idéal en est un dans lequel nous créons un spectre d'engagement permettant à chaque personne de choisir son degré d'engagement dans la société algorithmique.
>
> – Rumman Chowdhury

Francisco Marmolejo-Cossío partage une vision commune quant à l'autonomie sur nos données. Il mentionne que le développement durable est un aspect central de cette approche. Responsabiliser les citoyens et citoyennes a le potentiel de stimuler des changements dans l'écosystème de l'IA et de contribuer à un meilleur environnement pour les utilisateurs et utilisatrices ainsi que pour les entreprises. Traiter les utilisateurs et utilisatrices avec respect et faire progresser l'information, l'éducation et la transparence semblent essentiel pour un écosystème fondé sur les données.

> Dans le meilleur des cas, il s'agirait d'une prise de conscience dans ce contexte et d'une caractéristique importante que les consommateurs et consommatrices finissent par prendre en compte dans leurs décisions. Et je ne sais pas comment faire pression pour obtenir quelque chose comme ça par le biais de politiques, de l'éducation, de la sensibilisation. Mais je pense que ça irait loin et que ça

serait potentiellement plus durable. Et, dans un certain sens, tout ce type de réponses. Donc, une méthode transparente grâce à laquelle les utilisateurs et utilisatrices peuvent être conscients et conscientes des écueils potentiels des systèmes.

– Francisco Marmolejo-Cossío

CONCLUSION

En rédigeant ce chapitre, nous avons spécifiquement choisi trois enjeux dominants concernant le secteur de l'IA qui, selon nous et les experts et expertes avec qui nous avons discuté, mèneront, s'ils ne sont pas résolus, à une perpétration profonde et sérieuse de pratiques discriminatoires dans toutes les sphères de notre société. Nous avons fourni non seulement un compte rendu descriptif de ces enjeux, mais avons aussi réuni les principales suggestions de stratégies pour favoriser un changement dans une direction opposée et positive proposées par les personnes interviewées.

D'abord, nous avons considéré les défis associés à l'établissement de définitions de l'équité. Des tentatives frustrées de saisir des valeurs complexes dans une notation mathématique ne sont qu'un aspect de cette question. Le concept d'équité préexiste l'émergence de l'IA et est interprété de différentes manières selon les domaines, les cultures et d'autres facteurs contextuels. Les discussions avec les experts et expertes ont révélé à quel point le secteur est loin d'utiliser des méthodes d'équité dans son évaluation des modèles et que l'absence d'une définition appropriée de l'équité peut avoir des effets dévastateurs dans la mise en œuvre de modèles d'IA.

Ensuite, nous avons abordé les problèmes culturels au sein du secteur, soulignant le manque constant de diversité dans les pratiques d'embauche et dans l'organisation structurale des rôles et des fonctions tout au long du développement d'un produit. Qu'il s'agisse de diversité des genres, de multidisciplinarité ou autre, régler cette question nécessite un changement culturel à un niveau organisationnel supérieur. Les personnes interviewées ont présenté un compte rendu informatif sur la façon dont la diversité peut avoir un impact direct sur l'équité des modèles d'IA et se sont penchées sur sa cause. Il est important de noter à quel point la diversité peut être efficace pour libérer de grandes possibilités pour le secteur de l'IA d'une manière qui maximise les retombées sociétales positives.

Alors que des discussions de réglementation s'étendent au domaine de l'équité en IA, nous avons aussi discuté avec les experts et expertes des pratiques du secteur pour assurer une IA éthique et des perspectives de réglementations élargies. Dans presque tous nos entretiens, nous avons relevé des préoccupations quant à l'incapacité du secteur à s'autoréglementer. Les experts et expertes s'entendent aussi sur le fait qu'il est nécessaire de poser un regard critique sur les possibilités et les limites de la réglementation. Bien que la réglementation soit une avenue nécessaire pour en arriver à un écosystème d'IA plus équitable, elle ne doit pas être traitée simplement comme une solution miracle.

Enfin, en abordant les moyens pour aller de l'avant, nous avons souligné que, sans une participation élargie dans la création d'un cadre réglementaire, les enjeux actuels de l'IA continueront à s'aggraver, sans entrave. En nous appuyant sur ce sujet, nous avons présenté les groupes de parties prenantes qui devraient être mobilisés afin de collaborer à l'avancement de l'équité en IA, soit les gouvernements, les auditeurs et auditrices, les experts et expertes et la société civile.

En abordant la façon d'opérer les changements structuraux si nécessaires dans le domaine de l'IA, nous avons relaté les comptes rendus des experts et expertes sur la façon dont la sensibilisation et l'éducation peuvent aplanir les disparités pour les parties prenantes non techniques afin qu'elles participent à un processus décisionnel plus démocratique. En dernier lieu, nous avons discuté de la propriété des données et de la collecte des données en tant qu'exemples de sujets importants pour élever les discussions parmi les parties prenantes et activer une réforme des pratiques d'IA courantes. En puisant dans les commentaires des personnes interviewées, nous soutenons que l'évaluation des pratiques actuelles liées aux données peut être la première étape vers un changement mené par plusieurs parties prenantes en faveur d'une approche plus éthique à la conception et à la mise en œuvre de l'IA.

Ultimement, cerner les discriminations et lutter contre elles seront toujours des tâches incroyablement difficiles, tant à l'intérieur qu'à l'extérieur de l'espace de l'IA. C'est pourquoi la responsabilité d'une IA éthique devrait être partagée entre ceux et celles qui supervisent le développement de l'IA et ceux et celles qui l'utilisent. Comme l'a dit Arisa Ema, « quel que soit le degré de conscience en matière d'équité intégré dans la vision du développement de l'IA, s'il n'est pas accompagné de mesures concrètes, il ne vaut rien ». Pour renchérir, Joana Shields a exprimé notre sentiment d'urgence partagé d'aborder cette question. « Le temps est maintenant venu pour mettre en œuvre des cadres qui garantissent un développement technologique conforme à nos droits fondamentaux et qui empêchent les conséquences imprévues de l'IA de nuire aux vies et au tissu de notre société. »

RÉFÉRENCES

Angwin, J., Larson, J., Mattu, S. et Kirchner, L. 2016. Machine bias. *ProPublica*. 23 mai. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Cofessore, N. 2018. Cambridge Analytica and Facebook : The Scandal and the Fallout So Far. *New York Times*. 4 avril. https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html

Dastin, J. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. 10 octobre. ethttps://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Dietrich, W., Mendoza, C. et Brennan, T. 2018. *COMPAS Risk Scales : Demonstrating Accuracy Equity and Predictive Parity*. *Performance of the COMPAS Risk Scales in Broward County*. Northpointe Inc. Research Department. http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf

Garrett, G. 2018. The politics of data privacy in a post-Cambridge Analytica world. *Wharton Magazine*. 8 mai. https://magazine.wharton.upenn.edu/digital/the-politics-of-data-privacy-in-a-post-cambridge-analytica-world/

Grind, K., Schechner, S., McMillan, R. et West, J. 2019. How Google interferes with its search algorithms and changes your results. *Wall Street Journal.* 15 novembre. https://www.wsj.com/articles/how-google-interferes-with-its-search-algorithms-and-changes-your-results-11573823753

Larson, J., Mattu, S., Kirchener, L. et Angwin, J. 2016. How we analyzed the COMPAS recidivism algorithm. *ProPublica*. 23 mai. .https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Savage, N. 2020. The race to the top among world's leaders in artificial intelligence. *Nature*. 9 décembre. https://www.nature.com/articles/d41586-020-03409-8

Smith, B. et Browne, C. A. 2019. *Tools and Weapons : The Promise and the Peril of the Digital Age*. New York : Penguin Press.

Spielkamp, M. 2017. Inspecting algorithms for bias. *MIT Technology Review*. 12 juin. https://www.technologyreview.com/2017/06/12/105804/inspecting-algorithms-for-bias/

Skeem, J. et Lowenkamp, C. 2016. Risk, race, & recidivism : Predictive bias and disparate impact. *SSRN*. http://dx.doi.org/10.2139/ssrn.2687339

Telford, T. 2019. *Apple Card algorithm sparks gender bias allegations against Goldman Sachs. Washington Post*. 11 novembre. ethttps://www.washingtonpost.com/business/2019/11/11/apple-card-algorithm-sparks-gender-bias-allegations-against-goldman-sachs/

Verma, S. et Rubin, J. 2018. Fairness definitions explained. *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pp. 1-7. https://fairware.cs.umass.edu/papers/Verma.pdf

Yong, E. 2018. A popular algorithm is no better at predicting crimes than random people. *The Atlantic*. 17 janvier. https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646/

L'ATTENTION PARTIALE DANS LE DÉVELOPPEMENT DE L'IA : MENACES ET MESURES CORRECTIVES

ADJI BOUSSO DIENG

Informaticienne et statisticienne sénégalaise, ayant obtenu un diplôme de Ph. D. de l'Université Colombia. Professeure adjointe en informatique à l'Université Princeton, scientifique de recherche chez Google AI et fondatrice et présidente de l'organisme sans but lucratif The Africa I Know. Les membres de son groupe de recherche travaillent sur la conception de méthodes d'IA pour des applications scientifiques et de soins de santé. Ses travaux sont financés par la NSF et le Schmidt DataX Project.

ODD 5 - Égalité entre les sexes
ODD 9 - Industrie, innovation et infrastructure
ODD 10 - Inégalités réduites
ODD 11 - Villes et communautés durables
ODD 16 - Paix, justice et institutions efficaces
ODD 17 - Partenariats pour la réalisation des objectifs

L'ATTENTION PARTIALE DANS LE DÉVELOPPEMENT DE L'IA : MENACES ET MESURES CORRECTIVES

RÉSUMÉ

Tant au sein de l'industrie que des universités, les développements en intelligence artificielle (IA) qui dominent actuellement le narratif concernant ce domaine sont principalement centrés sur l'objectif de bâtir des systèmes d'IA qui peuvent faire certaines tâches plus efficacement que les humains. Une majorité de chercheurs, de chercheuses, d'ingénieurs et d'ingénieures, soutenus par les médias et les sources de financement, investissent toutes leurs énergies et leur attention sur la poursuite de l'intelligence artificielle générale (IAG) et l'étude des préjudices qui découlent des avancées faites dans cette poursuite. Cette « attention partiale » dans le domaine de l'IA entraîne de réelles menaces sociétales et socio-économiques et nous empêche de tirer parti de l'IA pour résoudre plusieurs problèmes pressants auxquels fait face l'humanité, des problèmes dont les solutions ne nécessitent pas le développement d'agents disposant d'intelligence superhumaine. Nous soutenons qu'il est possible pour l'IA d'être une technologie qui propulse l'humanité dans une direction positive, mais à condition de détourner l'attention de l'objectif de l'IAG et d'embrasser la philosophie et les pratiques d'une plus petite communauté au sein du domaine de l'IA, une communauté qui donne aux humains la faculté d'agir et qui s'intéresse à la prise en compte des connaissances et desiderata humains, de l'incertitude et de la contrôlabilité dans le développement des systèmes d'IA.

LE DOMAINE DE L'IA SOUFFRE D'UNE ATTENTION PARTIALE

Brossons un portrait global du domaine de l'intelligence artificielle (IA) comme il est aujourd'hui. La majorité des avancées dans le domaine qui sont relatées dans les médias sont produites par deux communautés différentes poursuivant des objectifs différents. Une communauté se concentre à atteindre l'intelligence artificielle générale (IAG) dont le but est de concevoir des systèmes d'IA aussi intelligents que les humains, sinon plus intelligents encore. Les principaux acteurs de ce programme

d'IAG se trouvent généralement dans l'industrie, mais aussi dans les universités. La deuxième communauté se concentre à souligner les préjudices que causent les avancées faites par la première communauté et étudie les implications de la réalisation de l'IAG. Ultimement, le travail de ces deux communautés est axé sur l'IAG. La première conçoit des outils pour atteindre l'IAG, tandis que l'autre agit en tant que garde-fou contre les préjudices et les répercussions futures de l'IAG. Nous appelons cette approche étroite dans le domaine, l'attention partiale.

Pour comprendre le problème qu'engendre l'attention partiale, nous devons établir les différentes parties prenantes en jeu ainsi que leurs motivations. Les acteurs dominants de la vision IAG de l'IA se trouvent principalement dans l'industrie, avec en tête de lice les entreprises comme Meta (ex-Facebook), DeepMind, Google AI et OpenAI. L'IA offre une occasion de croissance unique à ces entreprises en leur permettant d'attirer le peu de talents en IA dans le domaine, d'améliorer leurs produits existants ou d'entreprendre de nouvelles entreprises. Par exemple, Microsoft a investi, en 2019, un milliard de dollars américains dans un partenariat sur plusieurs années avec OpenAI pour que cette dernière « réalise la promesse de l'intelligence artificielle générale » (Microsoft, 2019). Depuis, OpenAI a conçu GPT-3, un système d'IA polyvalent qui peut traiter du texte de diverses sources dans le but de répondre à des questions générées par des humains. GPT-3 alimente plusieurs produits Microsoft, dont Azure, son service infonuagique lucratif, et GitHub, une plateforme de développement de logiciels d'envergure acquise par Microsoft en 2018 (Langston, 2021a, 2021b ; Nadella, 2018 ; Newman, 2021). En 2017, les chercheurs et chercheuses de Google AI ont conçu « Transformer », une architecture de réseau de neurones artificiels qui alimente plusieurs technologies d'IA, dont GPT-3 (Uszkoreit, 2017). Les « Transformers » sont maintenant au cœur du moteur de recherches et de l'outil de traduction de Google (Nayak, 2019 ; Raghavan, 2020 ; Caswell et Liang, 2020). Meta utilise l'IA dans ses différentes plateformes de réseaux sociaux dans la recommandation de contenu, l'étiquetage automatique de photos, la traduction multilingue, la modération de contenu, etc. Meta a récemment annoncé un nouveau projet qui tirera parti des vidéos des utilisateurs et utilisatrices de ses plateformes pour entraîner un système d'IA qui peut aider à la conception de plusieurs nouveaux produits (Zweig *et al*., 2021). DeepMind, quant à elle, s'est surtout concentrée sur la création d'un système d'IA qui peut battre des champions humains dans des jeux, comme Go ou les échecs (Gibbs, 2017 ; Hutson, 2017) et l'une des missions fondamentales de l'entreprise est de « résoudre l'intelligence » (DeepMind, n.d.). L'entreprise consacre depuis plus d'efforts à l'avancement de la science par le biais de l'IA grâce au développement d'AlphFord (équipe AlphaFold, 2020 ; Jumper *et al*., 2021), un système d'IA qui a réalisé des progrès importants dans la résolution de la question du repliement des protéines en biologie sur lequel se penchent les scientifiques depuis des décennies. S'appuyant sur le succès d'AlphaFord, le chef de la direction de DeepMind a récemment lancé une société dérivée à but lucratif pour tirer parti de la technologie pour la découverte de médicaments (Khan, 2021).

Ces incitatifs commerciaux stimulent le développement de l'IA, définissant les problèmes sur lesquels les recherches devraient se concentrer et les avancées qui devraient être valorisées. Cela se voit dans la dominance des sociétés technologiques quant au nombre d'articles publiés dans les principaux journaux portant sur l'IA, le Neural Information Processing Systems (NeurIPS) et l'International Conference on Machine Learning (ICML) (Rakicevic, 2021 ; Nguyen, 2021).

Bien que les entreprises technologiques d'envergure aient un rôle à jouer dans le problème de l'attention partiale en IA, elles partagent cette responsabilité avec d'autres acteurs, comme les médias. Tant les conséquences positives que négatives des avancées réalisées dans la poursuite de l'IAG reçoivent beaucoup d'attention de la part des médias, suscitant à la fois de l'intérêt et des craintes à l'égard de la technologie et menant également à une confusion quant à ses capacités et à ses objectifs (Jordan, 2018). En raison de la couverture médiatique entourant le fait que l'IA batte des champions du monde humains aux échecs ou au jeu de GO (Gibbs, 2017 ; Hutson, 2017), par exemple, et du sensationnalisme utilisé pour présenter les capacités des technologies, certaines personnes se demandent si l'IA les remplacera dans leur milieu de travail. D'autres s'inquiètent d'un avenir dystopique dans lequel les robots

prendront le contrôle du monde et tueront les êtres vivants sur la planète. Pourtant, outre la couverture médiatique, une autre manifestation de la crainte de l'impact de l'IA sur la société est l'émergence d'un domaine en pleine expansion, l'éthique de l'IA. Les discussions de plus en plus nombreuses sur l'établissement d'un cadre juridique pour l'IA par les gouvernements (Commission européenne, n.d. ; Candelon *et al*., 2021), le financement par des organismes à but non lucratif, comme Open Philanthropy (Beckstead et Muehlhauser, n.d.), la recherche sur l'atténuation des préjudices potentiels causés par l'IA et la prolifération des centres et des instituts universitaires en éthique de l'IA soulignent l'expansion récente du domaine. Ces efforts sont nécessaires pour atténuer les préjudices potentiels causés par les avancées vers l'IAG. Cependant, ils alimentent l'attention partiale en centrant le narratif sur l'IAG.

Comme mentionné au début de ce chapitre, l'attention partiale en IA est un problème. Cependant, elle suscite aussi beaucoup d'enthousiasme à l'égard des possibilités qu'offre l'IA. Des pays comme l'Égypte, le Brésil, le Canada et le Royaume-Uni travaillent sur leurs stratégies nationales en matière d'IA, signalant une conviction profonde quant au potentiel qu'a l'IA à stimuler une forte croissance dans les secteurs public et privé (Investir au Canada, n.d. ; GOV.UK, n.d.). Les agences de financement gouvernementales, comme la National Science Foundation aux États-Unis, investissent d'importantes sommes d'argent dans les universitaires qui poursuivent des recherches visant à faire progresser l'IA dans les différents secteurs (National Science Foundation, n.d.). Les étudiants et étudiantes, tant au premier cycle qu'aux cycles supérieurs, recherchent de plus en plus des programmes d'études liés à l'IA et choisissent de plus en plus l'IA comme champ d'études principal (Artificial Intelligence Index Report, 2021). Il y a une forte demande pour les emplois liés à l'IA et les salaires offerts sont relativement élevés (Chung, 2017). Les employeurs recherchent de plus en plus une maîtrise des sujets liés à l'IA (Columbus, 2019). Finalement, les entreprises en démarrage en IA prolifèrent et obtiennent beaucoup de financement des investisseurs et investisseuses en capital de risque et autres investisseurs et investisseuses (Weiss, 2021 ; Wilhelm et Heim, 2021).

POURQUOI DEVRAIT-ON SE PRÉOCCUPER DE L'ATTENTION PARTIALE DANS LE DOMAINE DE L'IA

La poursuite de l'IAG a favorisé une culture de recherche dans laquelle les innovations méthodiques et empiriques ainsi que l'élaboration des théories les sous-tendant sont centrées sur la résolution de tâches humaines. Par exemple, rédiger un texte cohérent à partir de rien, résumer des documents, entretenir des conversations, reconnaître des visages, répondre à des questions, décrire des images, jouer à des jeux, etc. La poursuite de l'IAG a mené à plusieurs avancées en IA, principalement en vision par ordinateur, en traitement automatique des langues et en systèmes de recommandation ainsi qu'à l'intersection de ces domaines.

En effet, on constate d'importantes améliorations dans la capacité des technologies langagières à produire des textes et des discours à des fins de conversation, de traduction, de résumé et autres qu'on ne peut distinguer de ceux rédigés par des humains. Le système GPT-3, un logiciel d'IA conçu par OpenIA qui peut, entre autres, écrire des textes cohérents de rien, est un exemple (Pilipiszyn, 2021). Les systèmes de recommandation alimentés par l'IA, qui sont utilisés dans toutes les plateformes de médias sociaux et autres services Internet et qui dictent le type de contenu que nous consommons en ligne, ainsi que les assistants vocaux, qui, quant à eux, répondent à nos questions, qu'elles concernent la météo, la culture ou les connaissances générales, et avec lesquels nous interagissons dans nos foyers et par l'entremise de nos appareils mobiles, sont d'autres exemples.

Néanmoins, ces avancées amènent aussi des désavantages : elles ont souvent une incidence négative sur les communautés marginalisées. Comme mentionné, cette réalité a entraîné l'émergence

de l'IA éthique, un domaine en forte croissance qui a attiré l'attention de chercheurs, de chercheuses, d'universitaires, d'activistes, de gouvernements, d'organisations à but non lucratif et de la société civile et dont l'objectif est de contrecarrer les menaces que posent les avancées en IA et de créer des cadres réglementaires juridiques pour le développement des technologies (Beckstead et Muehlhauser, n.d. ; Commission européenne, n.d. ; Candelon *et al.*, 2021).

Malgré les avantages qu'elles peuvent procurer, presque toutes les récentes avancées en IA, que ce soit en vision par ordinateur ou en traitement automatique des langues, posent des risques éthiques et drainent des ressources, tant humaines qu'économiques. Ces ressources pourraient être utilisées au développement d'une IA qui pourrait contribuer à résoudre les problèmes les plus pressants auxquels fait face l'humanité, comme les crises liées au climat et aux soins de santé. En examinant de près la façon dont la nouvelle IA est développée au sein de la communauté de recherche, nous pouvons mieux comprendre l'attention partiale, ou l'accent mal placé, dans le domaine de l'IA et ses nombreuses conséquences sociétales et socio-économiques.

Les nouvelles découvertes dont nous entendons parler dans les médias, comme la capacité de l'IA à produire des images de personnes et d'objets qui sont impossibles à distinguer d'images réelles ou à rédiger des histoires complètes lorsque demandé (Karras *et al.*, 2019 ; GPT-3, 2020), sont possibles grâce à des méthodes qui suivent le même parcours que nous appelons la *modélisation de tâches*. Cette méthode comprend quatre étapes. La première étape, la spécification de la tâche, consiste à décider quelle tâche nous souhaitons que le système d'IA apprenne à réaliser. Durant la deuxième étape, la collecte de données, de grandes quantités de données pertinentes à la tâche sont recueillies en recourant au moissonnage du Web, par exemple. Le développement d'un système est la troisième étape et comprend l'utilisation d'outils de modélisation, comme les réseaux de neurones artificiels, pour traiter et représenter les données et la conception d'un algorithme qui adapte ces outils à la tâche concernée à l'aide de données. Les modèles utilisés durant cette étape sont souvent très complexes. Ils possèdent plusieurs degrés de liberté (appelés des *paramètres*) et nécessitent d'importantes ressources informatiques. Enfin, la quatrième et dernière étape de la modélisation de la tâche sert à évaluer le système produit à l'étape précédente en ce qui a trait à la réalisation de la tâche précisée à la première étape. Cette évaluation consiste souvent à payer des humains pour qu'ils accèdent au système de manière à réaliser la tâche concernée, par exemple, en utilisant Amazon Mechanical Turk ou en soumettant le système à une évaluation des performances par rapport à un tableau de classement des mesures de référence. Chaque étape du parcours de modélisation des tâches présente des menaces potentielles abordées dans les sections suivantes.

Problèmes liés à la spécification de la tâche

Le paradigme dominant actuel, qui consiste à construire des systèmes d'IA capables de réaliser certaines tâches, nous empêche de réfléchir aux considérations éthiques liées à la spécification d'une tâche. Ce ne sont pas toutes les tâches qui devraient être réalisées. Certaines sont évidemment nuisibles et peuvent marginaliser certaines communautés et avoir une incidence négative sur l'humanité en général. Il existe plusieurs exemples de publications en IA, souvent dans des journaux prestigieux comme *Nature* ou dans le cadre de la Conference on Neural Information Processing Systems (NeurIPS), dont la proposition devrait susciter l'inquiétude. Par exemple, un article publié dans *Nature Communications* en 2020 proposait une méthode pour suivre les changements historiques en matière de fiabilité en utilisant des indices faciaux (Safra *et al.*, 2020). Un autre, publié dans le cadre de NeurIPS en 2019, proposait de reconstruire le visage d'une personne uniquement en fonction d'un enregistrement de sa voix (Wen *et al.*, 2019). Il existe de nombreux exemples similaires d'articles présentés dans des journaux ou des conférences sur l'IA de renom visant à prédire l'identité ou le caractère d'une personne sur la base de caractéristiques biologiques, telles que la voix ou les traits du visage. Ces types de méthodes alimentent des systèmes qui sont utilisés dans le monde réel, notamment des applications Internet fondées sur l'image ou des outils de police, dont l'impact négatif sur les communautés

marginalisées est largement documenté (Ryan-Mosley, 2021 ; General et Sarlin, 2021 ; Galston, 2020 ; Dunn *et al*., 2020). Le 9 janvier 2020, Robert Williams, un homme de race noire de 42 ans vivant à Farmington Hills au Michigan, a été arrêté et détenu pendant 30 heures après qu'il a été identifié à tort par reconnaissance faciale comme le suspect d'un crime qu'il n'avait pas commis (Ryan-Mosley, 2021). Il ne s'agit pas d'un cas isolé. Avant celle de M. Williams, des arrestations erronées avaient été rapportées. En 2019, par exemple, Nijeer Parks, un homme de race noire de 31 ans vivant au New Jersey a été arrêté après qu'il a été identifié à tort par reconnaissance faciale. Il a passé 11 jours en prison avant d'être finalement libéré (General et Sarlin, 2021). Il existe encore plusieurs autres cas d'arrestations injustifiées dues à l'utilisation de la reconnaissance faciale par les forces policières. Ces incidents ont suscité un tollé et conduit au lancement de plusieurs campagnes d'activistes et d'organisations de défense des droits civiques demandant l'interdiction de l'utilisation de la reconnaissance faciale par la police (Allyn, 2020 ; Snow, 2018) ainsi que des poursuites de la part de victimes (Harwell, 2021).

D'autres exemples de tâches qui ne devraient pas être réalisées par des systèmes d'IA sont celles dont les préjudices potentiels l'emportent sur les avantages qu'elles pourraient procurer à un secteur en particulier. Les hypertrucages, soient de fausses images ou vidéos de personnes qu'on ne peut distinguer de la réalité, ont beaucoup à offrir à l'industrie de la production vidéo et du cinéma en général en facilitant le montage de certains aspects d'une vidéo, tels que la voix, le ton ou l'accent de celui ou de celle qui parle. Or, les hypertrucages représentent partout une menace réelle à la démocratie puisqu'ils peuvent amplifier la désinformation en ligne. Ils permettent également de commettre des abus en ligne et ont été au cœur du débat sur le trafic sexuel en ligne et la violence à caractère sexuel (Galston, 2020 ; Dunn *et al*., 2020).

Enfin, une des limites d'une approche à l'IA axée sur les tâches est qu'elle nous empêche de tirer parti de l'IA pour résoudre des problèmes importants qui ne peuvent être réduits à la réalisation d'une tâche. Les problèmes qui requièrent une compréhension de certains mécanismes, comme ceux que l'on retrouve en sciences ou dans les soins de santé, et les problèmes qui nécessitent la compréhension de cause à effet ne se prêtent pas toujours bien à un cadre fondé sur la tâche. Par conséquent, une approche à l'IA axée sur les tâches constitue une occasion ratée de tirer parti de l'IA dans des domaines cruciaux comme les soins de santé.

LE POIDS DE LA DÉPENDANCE À UNE GRANDE QUANTITÉ DE DONNÉES

Bâtir des systèmes d'IA qui peuvent réaliser certaines tâches nécessite souvent la collecte de plusieurs téraoctets de données. Cette dépendance à l'égard des grands ensembles de données présente plusieurs inconvénients. Nous devons d'abord considérer les nombreux problèmes qui découlent du processus de collecte de données lui-même.

Les chercheurs, chercheuses, ingénieurs et ingénieures en IA ont recours au moissonnage du Web pour recueillir des données de différentes sources afin d'entraîner les systèmes d'IA. Parfois, ceci se fait sans égard à la vie privée et sans le consentement des utilisateurs et utilisatrices. L'entreprise de reconnaissance faciale, Clearview IA, fondée en 2017, moissonne apparemment des milliards de photos d'utilisateurs et d'utilisatrices des réseaux de médias sociaux et d'Internet sans leur consentement. Les photos sont utilisées pour entraîner son système de reconnaissance faciale, l'entreprise comptant parmi ses clients des organismes d'application de la loi (Hill, 2020). Il existe plusieurs autres bases de données semblables contenant des photos d'internautes qui ne sont pas conscients et conscientes de la violation de leur vie privée (McQuaid, 2021). Plus récemment, GitHub, la filiale de Microsoft, et OpenIA se sont associées pour développer et lancer GitHub Copilot, un outil générateur de codes fondé sur l'IA qui aide les codeurs et codeuses à coder plus facilement. Le système a été louangé quant à sa précision, mais a aussi fait sourciller en raison de problèmes de droit d'auteur (Taft, 2021). En effet, les données

d'entraînement de GitHub Copilot sont fondées sur des codes publics publiés sur GitHub, certains d'entre eux ayant des licences qui ne permettent pas de travaux dérivés (Taft, 2021).

Un autre problème du moissonnage du Web à la recherche de données est le manque d'attention portée à la qualité des données ainsi recueillies. Souvent, ces données contiennent des préjugés sous forme de stéréotypes négatifs envers certaines communautés, des propos nuisibles aux femmes et aux personnes de race noire, notamment. Par exemple, GPT-3, le système d'IA de pointe de langage naturel, a initialement été entraîné à l'aide de cinq sources de données différentes, totalisant près de 500 milliards de mots. Or, les sources de données contenaient un langage toxique qui a été amplifié par le modèle. GPT-3 produit, par exemple, des propos nuisibles aux musulmans et musulmanes et encode des préjugés fondés sur le genre et la race (Samuel, 2021 ; Lucy et Bamman, 2021).

Par ailleurs, une collecte aveugle de données engendre le problème de la qualité de ces données. Plus particulièrement, la qualité des données est compromise puisque les données ne sont souvent pas représentatives de tous et toutes et ne contiennent souvent pas d'informations au sujet de certaines communautés. Les ensembles de données utilisés, par exemple, pour entraîner les systèmes d'IA ont tendance à être partiaux aux hommes blancs et à la société occidentale. Des cultures entières ne sont pas actuellement représentées sur Internet, dont, notamment, le continent africain. Il s'agit d'un problème présent dans plusieurs sources de données.

Dans le domaine du traitement automatique des langues (TAL), plus particulièrement, le manque de représentativité dans les données est très sérieux. Bien que l'IA ait fait progresser les technologies langagières, ces avancées sont principalement en anglais et en d'autres langues qui sont bien représentées sur Internet, comme le mandarin, l'allemand et le français. Les langues du continent africain ne sont pas représentées dans les données utilisées pour entraîner les systèmes de langage fondés sur l'IA, privant ainsi un continent entier des avantages découlant des avancées en IA pour la compréhension des langues.

Enfin, bâtir des systèmes qui nécessitent de grandes quantités de données soulève aussi la question de l'énorme coût lié à leur collecte. D'abord, il y a le coût financier associé au stockage : des ressources informatiques sont nécessaires pour stocker des données. Ensuite, les données doivent souvent être étiquetées et des chercheurs et chercheuses doivent avoir recours à des outils comme Amazon Mechanical Turk pour l'étiquetage de données. Au-delà des coûts financiers, les coûts de main-d'œuvre humaine associés à l'étiquetage des données représentent une limite pour les domaines qui nécessitent une expertise, comme la science et les soins de santé. Étiqueter des images ou du texte est une tâche facile qui peut être externalisée à un grand bassin de professionnels et professionnelles. Cependant, l'étiquetage de molécules ou de radiographies, par exemple, nécessite une expertise du domaine que peu de personnes possèdent. Centrer le développement de l'IA sur ces pratiques de collecte de données peut, par conséquent, être limitant et discriminatoire.

LES PRÉJUDICES ET LES COÛTS DES GRANDS MODÈLES

Bâtir des systèmes d'IA qui peuvent réaliser certaines tâches requiert souvent des modèles très complexes. Un paradigme qui permet le développement de tels modèles complexes est *l'apprentissage profond* qui tire parti des réseaux de neurones artificiels, c'est-à-dire des couches de calculs qui imitent les neurones du cerveau, pour préciser des modèles flexibles qui peuvent être entraînés à l'aide de données pour réaliser des tâches. Au fur et à mesure que la tâche devient de plus en plus compliquée, une plus grande complexité doit être ajoutée au modèle. Ceci mène à de très grands modèles dont les décisions et les comportements ne peuvent être compris et, plus important encore, ne peuvent être contrôlés. Il existe plusieurs exemples d'échecs de ces grands modèles et des impacts négatifs qu'ils ont sur la société (Bender *et al.*, 2021).

Les systèmes de vision par ordinateur, par exemple, alimentant des technologies comme les voitures autonomes, encodent et amplifient des préjugés humains nuisibles, particulièrement en ce qui concerne les races. Effectivement, les systèmes de vision par ordinateur sont souvent incapables de reconnaître correctement les personnes de race noire. Facebook a récemment publié une invitation vidéo demandant à ses utilisateurs et utilisatrices, qui venaient tout juste de regarder une vidéo d'un homme de race noire publiée par le tabloïde britannique The Daly Mail, s'ils et si elles souhaitaient regarder d'autres vidéos de « primates » alors qu'aucun primate ne figurait dans la vidéo (Mac, 2021). Un autre exemple de ceci est survenu en 2015, lorsque l'application Google Photos a mal étiqueté plusieurs photos de deux Afro-Américains en indiquant qu'ils étaient des gorilles (Zhang, 2015). Il existe de nombreux autres exemples de ce type où des systèmes de vision par ordinateur échouent lamentablement lorsqu'il s'agit de personnes de race noire. Une voiture autonome utilisant un système de reconnaissance qui ne parvient pas à identifier certaines personnes peut provoquer des accidents mortels pour ces personnes qu'il ne parvient pas à identifier. Il s'agit souvent de personnes appartenant à des communautés marginalisées qui ne sont pas représentées dans les données utilisées pour entraîner le système de reconnaissance.

La vision par ordinateur n'est pas le seul domaine où de grands modèles ont échoué. Les technologies de langage naturel fondées sur l'IA ont aussi montré certaines limites et continuent d'être une source de préoccupation. Le 23 mars 2016, Microsoft a lancé un agent conversationnel, soit un logiciel qui entretient une conversation en ligne avec une personne par message texte ou à l'oral, nommé Tay, dont le compte Twitter a commencé à publier des tweets incendiaires et préjudiciables dès son lancement, contraignant Microsoft à cesser son utilisation seulement 16 heures après son lancement au public (Hunt, 2016). Le système de traduction alimenté par l'IA de Facebook est un autre exemple. Ce dernier a traduit à tort le message d'un utilisateur palestinien comme un appel à la violence, alors que le message original était un simple « bonjour ». Cette erreur de traduction a conduit à l'arrestation injustifiée de l'utilisateur par les forces israéliennes, qui ont été alertées après la publication du message (Ong, 2017).

Or, des systèmes d'IA ont également produit d'autres résultats inattendus lors de leur utilisation dans des assistants vocaux. J'ai une expérience personnelle de ce problème avec Siri, l'assistant vocal d'Apple, qui souvent ne reconnaît pas mon accent wolof. Ceci illustre bien que des progrès en IA ne profitent qu'à certaines personnes et pas à tous et à toutes. Enfin, il a été documenté que GPT-3 a un préjugé négatif envers les personnes musulmanes. Lorsqu'on lui demande de décrire les personnes musulmanes, GPT-3 débite des propos offensants qui amplifient les stéréotypes négatifs associant l'Islam à la violence (Samuel, 2021). GPT-3 a aussi des préjugés raciaux et fondés sur le genre. Un article publié en 2021 dans le cadre de la NLP Conference de l'Association of Computing Linguistics (ACL), un événement de premier ordre, rapportait que les histoires générées à l'aide de GPT-3 encodent et amplifient les préjugés sociaux relatifs à la race et au genre (Lucy et Bamman, 2021). Ces préjugés dommageables de GPT-3 devraient nous préoccuper, particulièrement en raison des nombreuses applications de cette technologie. Un billet de blogue d'OpenAI rapport que GPT-3 alimente plus de 300 applications dans plusieurs secteurs d'activité, dont l'éducation, la recherche, la conversation, l'achèvement de textes et bien d'autres encore (Pilipiszyn, 2021).

Une manière plus subtile dont les modèles fondés sur les grands réseaux de neurones ont un impact négatif sur les communautés marginalisées est qu'ils mémorisent souvent les rares échantillons d'entraînement dans leurs données d'entraînement (Feldman et Zhang, 2020). Ceci peut être exploité pour faire de l'ingénierie inverse de ces systèmes afin de déterminer les échantillons d'entraînement utilisés. Le problème est que ces rares échantillons d'entraînement correspondent souvent à des personnes issues de communautés marginalisées qui sont sous-représentées dans les données. Par conséquent, les grands modèles exposent davantage les personnes issues de ces communautés à des violations de leur vie privée, à la surveillance et à d'autres préjudices potentiels liés à leur identification.

Ces grands modèles entretiennent non seulement des préjugés contre les communautés marginalisées, leur empreinte carbone est aussi considérable (Strubell *et al.*, 2019). En 2019, une étude a démontré que les émissions de CO2 découlant de l'entraînement d'un modèle d'IA fondé sur un réseau de neurones de pointe pour le traitement automatique de langues (TAL) sont, en moyenne, 5 fois plus élevées que celles que génère une voiture à essence au cours de sa durée de vie et 56 fois plus élevées que celles que produit un humain au cours d'une année. Les systèmes TAL ne sont pas les seuls auxquels est associée une immense empreinte. Il s'agit d'un problème qui s'étend aux systèmes d'IA de pointe qui dépendent de plus en plus d'une architecture de réseau de neurones précise, nommée *Transformer* (Vaswani *et al.*, 2017 ; Patterson *et al.*, 2021), et les chercheurs et chercheuses se penchent sur des solutions de rechange plus écoénergétiques (Patterson *et al.*, 2021). Ces modèles ont non seulement un impact environnemental important, ils sont également coûteux à entraîner. Par conséquent, il place l'avenir de l'IA en tant que technologie entre les mains de ceux et celles qui disposent des moyens financiers pour la développer. Dans la même étude de 2019 portant sur l'empreinte carbone des modèles TAL fondés sur l'IA, Strubell *et al.* fournissaient aussi une estimation du coût financier pour les entraîner. Par exemple, le coût de l'entraînement d'un modèle TAL à usages multiples ayant reçu un prix du meilleur article dans le cadre de la prestigieuse conférence sur l'IA (EMNLP) en 2018 se situait entre 103 000 et 325 000 $ US. Étant donné que les modèles d'IA sont de plus en plus grands, les coûts ne cessent d'augmenter. En octobre 2021, Microsoft et NVIDIA ont introduit Megatron-Turing NLG 530B, le modèle de langage le plus puissant du monde (Alvi et Kharya, 2021). Le modèle fondé sur les « transformers » compte plus de 530 milliards de paramètres et le coût de son entraînement s'élève à 100 millions de dollars (Alvi et Kharya, 2021 ; Simon, 2021). Cette augmentation du coût de l'entraînement des modèles d'IA accentue l'écart de pouvoir au sein de la communauté de recherche en IA et entre les personnes qui peuvent tirer parti de la technologie. Le développement de systèmes d'IA qui requiert des sommes d'argent importantes exclut des communautés entières, notamment le continent africain, de l'écosystème de l'IA.

REFAÇONNER L'IA POUR EN FAIRE UNE TECHNOLOGIE QUI PROFITE À TOUS ET À TOUTES

Nous sommes actuellement témoins de l'émergence d'une révolution technologique stimulée par l'IA. Tout comme les révolutions qui l'ont précédée, elle transformera entièrement la façon dont nous faisons des affaires et dont nous interagissons avec le monde ainsi que nos vies au quotidien. Cependant, pour faire de l'IA une technologie dont tous et toutes profitent, nous devons nous éloigner de la poursuite de l'IAG et de la culture axée sur les tâches qui lui est associée. Quelles devraient être les caractéristiques définissant un domaine de l'IA qui profite à tous et à toutes ? Nous soutenons que l'IA doit être inclusive, sécuritaire et habilitante.

L'IA doit être plus inclusive en ce qui a trait à qui elle sert et à qui participe à son développement. Les coûts financiers très élevés associés au développement complet d'un modèle d'IA et à sa mise en œuvre, des coûts liés à la collecte de données aux coûts associés à l'entraînement et à l'évaluation, placent les avancées et les retombées de l'IA entre les mains d'une minorité qui dispose de ressources abondantes. Les innovations réalisées en IA sont stimulées par les intérêts de cette élite et centrées sur ceux et celles qui en font partie. Pour que l'IA soit plus inclusive, nous devons faire de l'accès à l'IA une priorité. Il existe des moyens concrets permettant d'améliorer l'accès à l'IA, comme la mise en place et le financement de ressources informatiques accessibles et gratuites pour tous et toutes.

En outre, il est possible d'améliorer l'accès à l'IA en finançant la création et l'entretien de bases de données pour différents domaines, comprenant de grands ensembles de données soigneusement choisis, représentatifs et respectueux des normes de confidentialité. Un tel financement inclurait la mise

à disposition gratuite des ensembles de données à tous et à toutes, en particulier dans les régions où les ressources sont limitées. L'importance des données de haute qualité est de plus en plus reconnue et les appels à une IA centrée sur les données gagnent du terrain. Mais nous devons aller au-delà de l'amélioration de la qualité des données pour les intérêts de quelques puissants et décentraliser plutôt la collecte et la conservation des données en encourageant et en soutenant les efforts des communautés locales, comme celles d'Afrique.

Une autre stratégie d'accès à l'IA consiste à centrer les efforts d'innovation sur des techniques qui sont plus efficaces sur le plan des ressources. Les chercheurs, chercheuses, ingénieurs et ingénieures devraient s'efforcer de concevoir des méthodologies qui sont de moins grandes consommatrices de données et de ressources informatiques. Les agences de financement devraient récompenser les travaux de recherche soutenant cette voie. Rendre l'IA plus inclusive en priorisant l'accès à l'IA fera en sorte qu'elle pourra s'attaquer à un ensemble plus diversifié de problèmes et faire avancer le domaine.

L'IA doit être sécuritaire pour profiter à tous et à toutes. La culture actuelle, axée sur l'objectif d'apprendre à réaliser des tâches humaines, a mené à des avancées qui profitent à quelques-uns et quelques-unes, mais qui nuisent à plusieurs au sein des communautés marginalisées. Une IA sécuritaire disposerait de cadres renforçant la transparence et la performance du développement de modèles, intégrant des façons, au sein du processus d'évaluation, de vérifier les attributs éthiques du modèle, dont l'équité et la protection de la vie privée, et de cadres permettant d'expliquer la cause de tout préjudice et d'intervenir facilement. Ce dernier point est possible si nous imposons la contrôlabilité dans les systèmes d'IA.

L'IA devrait donner du pouvoir à l'humanité. Ceci ne sera possible que si nous nous détournons de modèles fondés sur des tâches et collaborons avec les experts et expertes du domaine pour développer une IA qui résout les problèmes les plus pressants auxquels fait face l'humanité, comme la crise climatique et les soins de santé, qui améliore le sort des personnes marginalisées, qui mène à des découvertes scientifiques novatrices et qui améliore la société. Développer une IA qui nous permettra d'intégrer de manière transparente les informations provenant des données et des humains dans des solutions efficaces aux problèmes de l'humanité est une occasion que devrait saisir le secteur.

Une chose qui pourrait permettre à l'IA de profiter à tous et à toutes est de favoriser une adoption élargie des pratiques d'une plus petite communauté dans le domaine de l'IA dont les efforts sont moins relayés par les médias grand public, soit la communauté de l'apprentissage automatique axé sur des modèles probabilistes. Cette communauté se compose de chercheurs et de chercheuses, la majorité provenant d'universités, qui s'intéressent à l'intégration de l'incertitude et de la connaissance du domaine dans les systèmes de prise de décisions. Les méthodologies mises au point par cette communauté permettent d'apprendre même de très petites quantités de données. Leur approche à l'IA est donc inclusive des communautés ayant peu de ressources. C'est pour cette raison que nous retrouvons des techniques conçues par cette communauté, comme les processus gaussiens et les réseaux de neurones bayésiens, dans des domaines cruciaux comme les soins de santé et dans les sciences. Cette communauté traite les données comme un citoyen de première classe et non simplement comme un simple outil pour apprendre une tâche. Leurs travaux nécessitent la collaboration avec des experts et expertes du domaine afin de guider la création de méthodologies ciblées vers la résolution de problèmes. Donner à cette communauté les outils dont elle a besoin et adopter son approche à l'IA centrée sur les données et sur les humains nous rapprochera d'un domaine qui profite à tous et à toutes.

Refaçonner l'IA est possible et nécessitera des efforts de toutes les parties prenantes, des chercheurs, chercheuses, ingénieurs et ingénieures aux gouvernements, en passant par les médias, les agences de financement et le secteur privé.

RÉFÉRENCES

Allyn, B. 2020. Amazon halts police use of its facial recognition technology. *NPR*, 10 juin. https://www.npr.org/2020/06/10/874418013/amazon-halts-police-use-of-its-facial-recognition-technology

AlphaFold team, 2020. AlphaFold : a solution to a 50-year-old grand challenge in biology. DeepMind blog, 30 novembre. https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

Alvi, A. et Kharya, P. 2021. Using DeepSpeed and Megatron to train Megatron-Turing NLG 530B, the world's largest and most powerful generative language model. Microsoft Research Blog, 11 octobre. https://www.microsoft.com/en-us/research/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/

Artificial Intelligence Index Report. 2021. Chapter 4 : AI Education. https://aiindex.stanford.edu/wp-content/uploads/2021/03/2021-AI-Index-Report-_Chapter-4.pdf

Beckstead, N. et Muehlhauser, L. n.d. Potential risks from advanced artificial intelligence. Open Philanthropy. https://www.openphilanthropy.org/focus/global-catastrophic-risks/potential-risks-advanced-artificial-intelligence

Bender, E. M., Gebru, T., McMillan-Major, A., et Shmitchell, S. 2021. On the dangers of stochastic parrots : Can language models be too big ? *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Mars 2021, pp. 610–623. https://dl.acm.org/doi/10.1145/3442188.3445922

Candelon, F., di Carlo, R. C., De Bondt, M., et Evgeniou, T. 2021. AI regulation is coming. *Harvard Business Review*, edition de septembre-octobre. https://hbr.org/2021/09/ai-regulation-is-coming

Caswell, I. et Liang, B. 2020. Recent Advances in Google Translate. Google AI Blog, 8 juin. https://ai.googleblog.com/2020/06/recent-advances-in-google-translate.html

Chung, C. 2017. Tech giants are paying huge salaries for scare A.I. talent. *New York Times*, 22 octobre. https://www.nytimes.com/2017/10/22/technology/artificial-intelligence-experts-salaries.html

Columbus, L. 2019. AI skills among the most in-demand for 2020. *Forbes*, 27 novembre. https://www.forbes.com/sites/louiscolumbus/2019/11/27/ai-skills-among-the-most-in-demand-for-2020/?sh=69de02b36b44

DeepMind, n.d. Home page. https://deepmind.com/

Dunn, S., Carswell, N., Doagoo, B. C., Shoker, S. et Tatsis, S. 2020. Deepfakes and digital harms : Emerging technologies and gender-based violence. Online presentation and discussion (video), 25 novembre. Centre for International Governance Innovation. https://www.cigionline.org/events/deepfakes-and-digital-harms-emerging-technologies-and-gender-based-violence/

European Commission, n.d. Regulatory framework proposal on artificial intelligence. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Feldman, V. et Zhang, C. 2020. What neural networks memorize and why : Discovering the long tail via influence estimation. *arXiv :2008.03703 [cs.LG]*. https://arxiv.org/abs/2008.03703

Galston, W. A. 2020. Is seeing still believing ? The deepfake challenge to truth in politics. The Brookings Institution, January 8. https://www.brookings.edu/research/is-seeing-still-believing-the-deepfake-challenge-to-truth-in-politics/

General, J. et Sarlin, J. 2021. A false facial recognition match sent this innocent Black man to jail. *CNN Business*, 29 avril. https://www.cnn.com/2021/04/29/tech/nijeer-parks-facial-recognition-police-arrest/index.html

Gibbs, S. 2017. AlphaZero AI beats champion chess program after teaching itself in four hours. *The Guardian*, 7 décembre. https://www.theguardian.com/technology/2017/dec/07/alphazero-google-deepmind-ai-beats-champion-program-teaching-itself-to-play-four-hours

GOV.UK. n.d. Guidance : National AI Strategy. https://www.gov.uk/government/publications/national-ai-strategy

GPT-3. 2020. A robot wrote this entire article. Are you scared yet, human ? *The Guardian*, 8 septembre. https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3

Harwell, D. 2021. Amazon extends ban on police use of its facial recognition technology indefinitely. *Washington Post*, 18 mai. https://www.washingtonpost.com/technology/2021/05/18/amazon-facial-recognition-ban/

Hill, K. 2020. The secretive company that might end privacy as we know it. *New York Times*, 18 janvier. https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html

Hunt, E. 2016. Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter. *The Guardian*, 24 mars. https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter?CMP=

Hutson, M. 2017. This computer program can beat humans at Go—with no human instruction. *Science*, 18 octobre. https://www.science.org/content/article/computer-program-can-beat-humans-go-no-human-instruction

Invest in Canada. n.d. Pan-Canadian AI Strategy. https://www.investcanada.ca/programs-incentives/pan-canadian-ai-strategy

Jordan, M. 2018. Artificial intelligence—The revolution hasn't happened yet. Medium.com (personal blog), 19 avril. https://medium.com/@mijordan3/artificial-intelligence-the-revolution-hasnt-happened-yet-5e1d5812e1e7

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. et Hassabis, D. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* Vol. 596, pp. 583–589. https://www.nature.com/articles/s41586-021-03819-2

Kahn, J. 2021. DeepMind spins out new Alphabet company focused on drug discovery. *Fortune*, 4 novembre. https://fortune.com/2021/11/04/deepmind-spins-out-alphabet-company-isomorphic-drug-discovery-company/

Karras et al. et Nvidia. 2019. Imagined by a GAN (generative adversarial network) StyleGAN2. https://thispersondoesnotexist.com/

Langston, J. 2021a. From conversation to code : Microsoft introduces its first product features powered by GPT-3. Microsoft/The AI Blog, 25 mai. https://blogs.microsoft.com/ai/from-conversation-to-code-microsoft-introduces-its-first-product-features-powered-by-gpt-3/

Langston, J. 2021b. New Azure OpenAI Service combines access to powerful GPT-3 language models with Azure's enterprise capabilities. Microsoft/The AI Blog, 2 novembre. https://blogs.microsoft.com/ai/new-azure-openai-service/

Lucy, L. et Bamman, D. 2021. Gender and representation bias in GPT-3 generated stories. *Proceedings of the 3rd Workshop on Narrative Understanding, Association for Computational Linguistics*, 11 juin. pp. 48–55. https://aclanthology.org/2021.nuse-1.5.pdf

Mac, R. 2021. Facebook apologizes after A.I. puts "primates" label on video of Black men. *New York Times*, 3 septembre. https://www.nytimes.com/2021/09/03/technology/facebook-ai-race-primates.html

McQuaid, J. 2021. Limits to growth : Can AI's voracious appetite for data be tamed ? *Undark*, 18 octobre. https://undark.org/2021/10/18/computer-scientists-try-to-sidestep-ai-data-dilemma/

Microsoft. 2019. OpenAI forms exclusive computing partnership with Microsoft to build new Azure AI supercomputing technologies. Microsoft News Center, 22 juillet. https://news.microsoft.com/2019/07/22/openai-forms-exclusive-computing-partnership-with-microsoft-to-build-new-azure-ai-supercomputing-technologies/

Nadella, S. 2018. Microsoft + GitHub = empowering developers. Microsoft/Official Microsoft Blog, 4 juin. https://blogs.microsoft.com/blog/2018/06/04/microsoft-github-empowering-developers/

National Science Foundation (NSF). n.d. Artificial intelligence at NSF. https://www.nsf.gov/cise/ai.jsp

Nayak, P. 2019. Understanding searches better than ever before. Google/The Keyword blog, 25 octobre. https://blog.google/products/search/search-language-understanding-bert/

Newman, J. 2021. GitHub's new tool uses AI to craft code. Some developers are furious. *Fast Company*, 9 juillet. https://www.fastcompany.com/90653878/github-copilot-microsoft-openai-coding-tool-backlash

Nguyen, T. 2021. An overview of ICML 2021's publications. VinAI Research/Achievements blog, 18 juillet. https://www.vinai.io/an-overview-of-icml-2021s-publications

Ong, T. 2017. Facebook apologizes after wrong translation sees Palestinian man arrested for posting "good morning." *The Verge*, 24 octobre. https://www.theverge.com/us-world/2017/10/24/16533496/facebook-apology-wrong-translation-palestinian-arrested-post-good-morning

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M. et Dean, J. 2021. Carbon emissions and large neural network training. *arXiv :2104.10350 [cs.LG]*. https://arxiv.org/abs/2104.10350

Pilipiszyn, A. 2021. GPT-3 Powers the Next Generation of Apps. OpenAI blog, 25 mars. https://openai.com/blog/gpt-3-apps/

Raghavan, P. 2020. How AI is powering a more helpful Google. Google/The Keyword blog, 15 octobre. https://blog.google/products/search/search-on/

Rakicevic, N. 2021. NeurIPS Conference : Historical data analysis. Medium.com/Towards Data Science, 27 février. https://towardsdatascience.com/neurips-conference-historical-data-analysis-e45f7641d232

Ryan-Mosley, T. 2021. The new lawsuit that shows facial recognition is officially a civil rights issue. *MIT Technology Review*, 14 avril. https://www.technologyreview.com/2021/04/14/1022676/robert-williams-facial-recognition-lawsuit-aclu-detroit-police/

Safra, L., Chevallier, C., Grèzes, J. et Baumard, N. 2020. Tracking historical changes in trustworthiness using machine learning analyses of facial cues in paintings. *Nature Communications* Vol. 11, Art. No. 4728. https://www.nature.com/articles/s41467-020-18566-7

Samuel, S. 2021. AI's Islamophobia problem. *Vox*, 18 septembre. https://www.vox.com/future-perfect/22672414/ai-artificial-intelligence-gpt-3-bias-muslim

Simon, J. 2021. Large language models : A new Moore's Law ? Hugging Face blog, 26 octobre. https://huggingface.co/blog/large-language-models

Snow, J. 2018. Amazon's face recognition falsely matched 28 Members of Congress with mugshots. ACLU blog, 26 juillet. https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28

Strubell, E., Ganesh, A. et McCallum, A. 2019. Energy and Policy Considerations for Deep Learning in NLP. In the 57th Annual Meeting of the Association for Computational Linguistics (ACL). Florence, Italy. Juillet 2019. *arXiv:1906.02243 [cs.CL]* https://arxiv.org/abs/1906.02243

Taft, D. K. 2021. GitHub Copilot : A powerful, controversial autocomplete for developers. *The New Stack*, 1er juillet. https://thenewstack.io/github-copilot-a-powerful-controversial-autocomplete-for-developers/

Uszkoreit, J. 2017. Transformer : A novel neural network architecture for language understanding. Google AI Blog, 31 août. https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. et Polosukhin, I. 2017. Attention is all you need. *arXiv:1706.03762 [cs.CL]*. https://arxiv.org/abs/1706.03762

Weiss, T. R. 2021. AI startups continue to rack up millions in VC funding in August 2021. *EnterpriseAI*, 26 août. https://www.enterpriseai.news/2021/08/26/ai-startups-continue-to-rack-up-millions-in-vc-funding-in-august-2021/

Wen, Y., Raj, B., et Singh, R. 2019. Face Reconstruction from Voice using Generative Adversarial Networks. *NeurIPS Proceedings*. https://proceedings.neurips.cc/paper/2019/hash/eb9fc349601c69352c859c1faa287874-Abstract.html

Wilhelm, A. et Heim, A. 2021. Huge deals are pushing more AI startups into IPO territory. *TechCrunch*, 6 novembre. https://techcrunch.com/2021/11/16/huge-deals-are-pushing-more-ai-startups-into-ipo-territory/

Zhang, M. 2015. Google Photos tags two African-Americans as gorillas through facial recognition software. *Forbes*, 1er juillet. https://www.forbes.com/sites/mzhang/2015/07/01/google-photos-tags-two-african-americans-as-gorillas-through-facial-recognition-software/?sh=90e05f713d8b

Zweig, G., Auli, M. et Fagan, F. 2021. Learning from videos to understand the world. Meta AI, 12 mars. https://ai.facebook.com/blog/learning-from-videos-to-understand-the-world

L'IA TEND À CENTRALISER LA PRISE DE DÉCISIONS ET LE POUVOIR, ET C'EST UN PROBLÈME

ERIK BRYNJOLFSSON

Laboratoire d'économie numérique de l'Université de Stanford et National Bureau of Economic Research (États-Unis).

ANDREW NG

Foundateur et PDG des sociétés DeepLearning.AI et Landing AI. Projesseur Adjoint au Département de Sciences Informatiques de l'Université de Stanford.

ODD 9 - Industrie, innovation et infrastructure

ODD 10 - Inégalités réduites

ODD 11 - Villes et communautés durables

ODD 16 - Paix, justice et institutions efficaces

ODD 17 - Partenariats pour la réalisation des objectifs

L'IA TEND À CENTRALISER LA PRISE DE DÉCISIONS ET LE POUVOIR, ET C'EST UN PROBLÈME

RÉSUMÉ[11]

Au cours des dernières années, les systèmes d'intelligence artificielle (IA) ont considérablement gagné en ampleur et en puissance. Ils ont maintenant le potentiel de centraliser la prise de décisions de manière substantielle. Centraliser la prise de décisions peut certes s'avérer efficace, mais peut également mener à une concentration de la richesse et du pouvoir. Bien que la technologie évolue actuellement de telle sorte qu'elle fait place à une augmentation du degré de centralisation de la prise de décisions et du pouvoir, cette issue ne nous semble pas inéluctable. Il serait possible de concevoir des technologies décentralisées, interopérables ou fédérées qui assureraient le maintien d'une prise de décisions et d'un exercice du pouvoir décentralisés. Fondamentalement, nous pouvons renforcer la démocratie et les institutions politiques pour qu'elles supervisent la prise de décisions effectuées par des machines.

Dans ce chapitre, nous discutons de certaines tendances actuelles touchant l'IA et d'autres technologies, tendances qui pourraient faire pencher la balance vers la centralisation ou la décentralisation, comparons la prise de décisions qu'effectuent des systèmes automatisés à celle des humains, examinons certaines données empiriques sur la concentration observée jusqu'à maintenant et présentons des solutions technologiques, économiques et politiques. Nous ne cherchons pas à prédire l'avenir, mais à formuler des mises en garde quant au niveau sans précédent de concentration du pouvoir décisionnel que pourrait entraîner l'IA si nous n'agissons pas de manière responsable.

11. Remerciements : Nous tenons à remercier Lynn He pour son aide considérable dans l'édition et la révision de ce chapitre.

INTRODUCTION

Au cours des dernières années, les systèmes d'intelligence artificielle (IA) ont considérablement gagné en ampleur et en puissance. Le modèle linguistique GPT-3 d'OpenAI compte plus de 175 milliards de paramètres, et Google a pour sa part entraîné des modèles comptant plus de 1 billion de paramètres (Talagala, 2021). De tels modèles atteignent le niveau d'une intelligence humaine, voire surhumaine, dans de plus en plus de domaines, dont certains bien pointus. Par exemple, AlphaZero de Deepmind a joué 4,9 millions de parties d'échecs contre lui-même et a atteint en une journée un classement supérieur à celui de tout être humain (Silver *et al.*, 2017). À partir de données historiques ou simulées, les modèles d'apprentissage automatique sont entraînés à reconnaître des formes et à prévoir des résultats. Plus un modèle comporte un nombre élevé de paramètres et plus la capacité à traiter les données est grande, mieux le modèle parviendra à généraliser. AlphaZero est meilleur que les humains non seulement aux échecs, mais aussi à d'autres jeux, comme le go et le shogi. En outre, il peut battre des systèmes d'IA spécialisés comme Stockfish, mis au point pendant de nombreuses années dans le but précis de jouer aux échecs. De la même manière, le modèle GPT-3 produit des textes qu'il est souvent impossible de distinguer de ceux qu'auraient écrits des humains, dépassant ainsi la capacité d'analyse purement linguistique de la plupart des modèles qui l'ont précédé.

Ces grands et puissants modèles entraîneront potentiellement une hausse considérable de la productivité et du niveau de vie. Ils s'appliquent à des domaines aussi divers que le développement logiciel (Belton, 2021), le diagnostic médical (Ronneberger *et al.*, 2015) et la prévision des catastrophes naturelles (Devries *et al.*, 2018). Ils peuvent aussi conduire à une nette centralisation de la prise de décisions. En raison des limites qu'a tout cerveau humain en ce qui a trait à la quantité d'information assimilable et au nombre de décisions traitables en une journée, la prise de décisions a historiquement été décentralisée vers les marchés et d'autres structures organisationnelles hiérarchiques. Sauf que la capacité des machines à analyser de plus en plus d'information ainsi qu'à prendre des milliers de décisions chaque seconde ne cesse d'augmenter, ce qui laisse en principe présager une centralisation de la prise de décisions par des machines.

Centraliser la prise de décisions peut certes s'avérer efficace, notamment parce que le processus tient alors compte de liens entre différentes instances, mais cette mesure peut également mener à une concentration de la richesse et du pouvoir. Il ne s'agit pas d'une solution souhaitable pour ceux et celles qui perdent alors de leur pouvoir décisionnel, en tout ou en partie.

Bien que la technologie évolue actuellement de telle sorte qu'elle fait place à une augmentation du degré de centralisation de la prise de décisions et du pouvoir, cette issue ne nous semble pas inéluctable. Il serait possible de concevoir des technologies décentralisées, interopérables ou fédérées qui assureraient le maintien d'une prise de décisions et d'un exercice du pouvoir décentralisés. En effet, un certain nombre de penseurs et penseuses soutiennent que des innovations telles qu'Internet, les chaînes de blocs et des technologies connexes favorisent au contraire une décentralisation (Malone, 2003 ; Pueyo, 2021 ; Srinivasan, 2019 ; Lera *et al.*, 2020). Il est également possible de soutenir la décentralisation du pouvoir économique en veillant à une distribution générale du capital humain, physique et financier, par exemple en utilisant des outils comme le revenu de base ou encore des allocations favorisant l'acquisition de compétences ou permettant de faire des contributions politiques. Fondamentalement, nous pouvons aussi renforcer la démocratie et les institutions politiques pour qu'elles supervisent la prise de décisions effectuées par des machines.

Dans ce chapitre, nous discutons de certaines tendances actuelles touchant l'IA et d'autres technologies, tendances qui pourraient faire pencher la balance vers la centralisation ou la décentralisation, comparons la prise de décisions qu'effectuent des systèmes automatisés à celle des humains, examinons certaines données empiriques sur la concentration observée jusqu'à maintenant et présentons des solutions

technologiques, économiques et politiques. Nous ne cherchons pas à prédire l'avenir, mais à formuler des mises en garde quant au niveau sans précédent de concentration du pouvoir décisionnel que pourrait entraîner l'IA si nous n'agissons pas de manière responsable.

LES SYSTÈMES D'IA PRENNENT DE L'AMPLEUR

Les systèmes d'IA sous-tendent de nombreuses activités essentielles de l'économie moderne. De la simple et familière droite de régression $y = ax + b$, qui ne comporte que deux paramètres, aux réseaux neuronaux de pointe, qui en comprennent des milliards, les modèles d'apprentissage automatique et les ensembles de données qui les alimentent atteignent une ampleur qui dépasse souvent l'entendement.

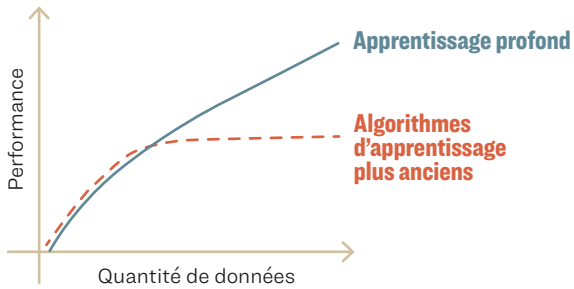
Cette croissance explosive s'explique de manière toute logique : des modèles plus importants réussissent à traiter une quantité supérieure de données. Si l'on considère l'apprentissage automatique comme le fait d'entraîner des modèles pour qu'ils établissent des prédictions en fonction de données historiques, alors il faut conclure que plus le modèle reçoit de données et plus celles-ci sont diversifiées, meilleure sera la performance de l'algorithme. Ce raisonnement a entraîné une course à la conception de modèles de taille, par les universités et l'industrie, et par le fait même à un accroissement des données. L'exploitation de puissants ordinateurs a levé de nombreux obstacles à l'innovation relative aux systèmes et à leur taille. En 2009, Andrew Ng et son équipe de l'Université de Stanford ont reconnu le potentiel des unités de traitement graphique (UTG) – des processeurs créés pour les jeux vidéo – dans l'adaptation de calculs que font des réseaux neuronaux profonds, ce qui a fait passer la durée d'entraînement de plusieurs semaines à quelques jours (Ng *et al.*, 2009). Cette avancée a contribué à l'avènement d'une nouvelle ère de l'apprentissage automatique, soit celle de l'apprentissage profond, qui a permis d'innover dans l'architecture des modèles et la taille des ensembles de données, et a abouti à des résultats remarquables[12].

L'utilisation d'UTG est de nos jours banale, voire élémentaire, et l'apprentissage profond intervient dans plusieurs de nos activités courantes par de nombreuses applications, notamment la recommandation de contenus dans des plateformes de réseaux sociaux, l'organisation de services de covoiturage et l'achat en ligne (Brynjolfsson et McAfee, 2017a). La nette dominance de cette technique découlant de la science des données s'explique en partie par le fait que l'apprentissage profond, plus que toute technologie antérieure, évolue en fonction des données (voir la figure 1).

12. L'apprentissage profond est par exemple parvenu à une bien meilleure classification d'images que des approches antérieures. Voir Krizhevsky *et al.* (2012).

| FIGURE 1 |

Pourquoi l'apprentissage profond ? La performance des algorithmes s'améliore sans cesse au fil de l'accroissement des données, alors que celle des anciens algorithmes d'apprentissage atteint un plateau. *Source :* Ng (2015).

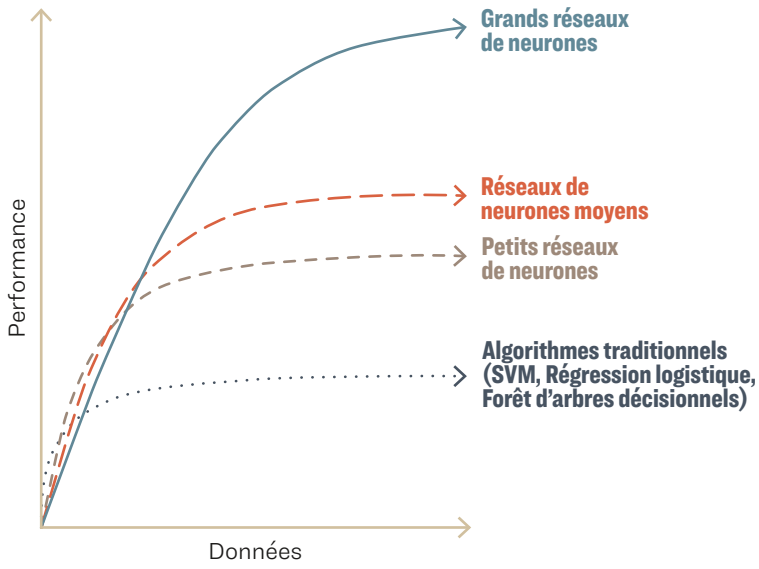


Comment les techniques de science des données s'adaptent-elles à la quantité de données ?

Cette progression suit la logique énoncée précédemment selon laquelle un modèle présentera de meilleurs résultats s'il est entraîné au moyen d'une grande variété de données. Plus un modèle est imposant, plus il emploie de paramètres pour inférer un apprentissage, ce qu'il fait à partir d'ensembles de données qui ne cessent de croître. Cette logique, largement reprise en IA, a stimulé la concurrence en ce qui a trait tant à la puissance de calcul des modèles qu'à l'ampleur de ceux-ci. De la même manière que le passage des processeurs centraux (1 million de connexions) aux processeurs graphiques ou UTG (10 millions de connexions) a comblé un écart entre les algorithmes de l'apprentissage automatique courant et les techniques d'apprentissage profond, l'exploitation de l'informatique en nuage (nombreux serveurs, 1 milliard de connexions) et du calcul haute performance (plusieurs UTG, 100 milliards de connexions) entraînera une expansion de la structure des modèles (voir figure 2).

| FIGURE 2 |

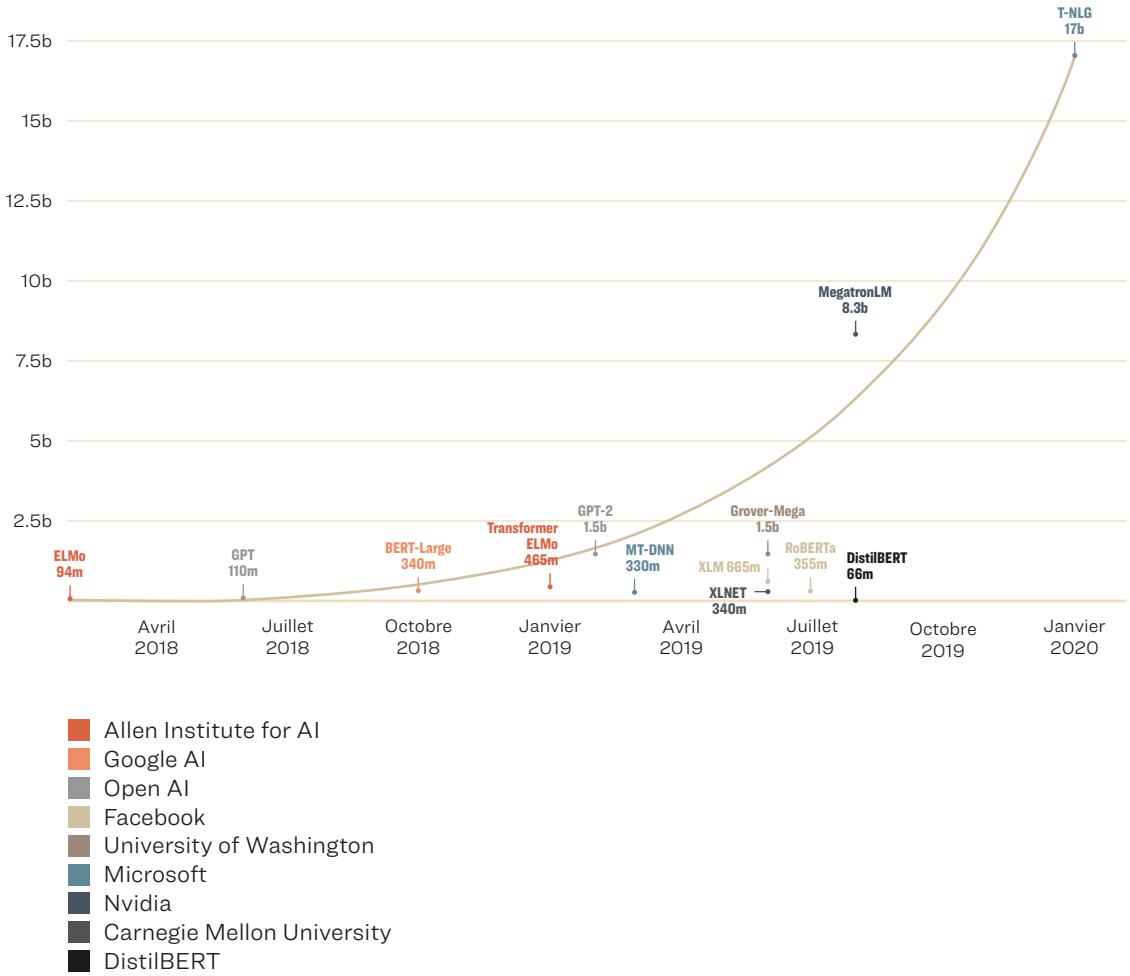
La performance d'un modèle plafonne si les ensembles de données ne contiennent pas assez de paramètres à partir desquels inférer un apprentissage. *Source :* Ng (2016).



L'augmentation rapide de la taille des modèles linguistiques au cours des dernières années reflète d'ailleurs déjà l'accroissement de la puissance de calcul. Elle traduit les aspirations des technologues, qui cherchent à entraîner les systèmes d'apprentissage automatique en leur fournissant toujours plus de données. La figure 3 explicite que l'augmentation de la taille des modèles de traitement automatique des langues naturelles s'est accélérée au cours des dernières années. La figure ne montre pas le modèle GPT-3 d'OpenAI (175 milliards de paramètres), dont la taille est dix fois supérieure à celle du T-NLG de Microsoft.

| FIGURE 3 |

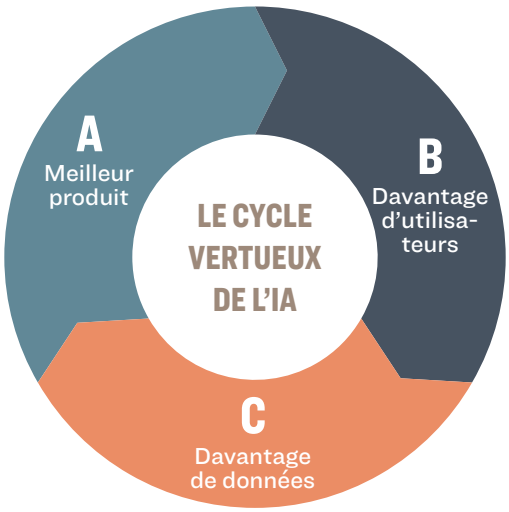
Tailles des modèles linguistiques. *Source :* Ng (2020).



Puisque l'accumulation de données entraîne des coûts quasi négligeables, la tendance à accroître la taille des modèles s'accentue. Les modèles d'apprentissage profond parviennent constamment à intégrer de nouvelles données d'entraînement sans que les calculs engendrent de coûts importants. Cet avantage s'inscrit dans le « cycle vertueux » de l'IA : les applications offertes à la clientèle engendrent organiquement un flux constant de données, l'intégration de celles-ci permet à faible coût d'améliorer la capacité de prévision du système, la version améliorée du modèle attire de nouveaux consommateurs et consommatrices et ces derniers contribuent à l'accroissement des données d'utilisation (voir figure 4). Si d'éventuelles entreprises concurrentes ne profitent pas d'un tel accès à des données de qualité, il leur sera difficile de s'introduire dans la boucle de rétroaction positive liée à l'accumulation de données.

| FIGURE 4 |

Cycle vertueux de l'IA. *Source :* Ng (2020).



Une telle dynamique ressemble à celle de la recherche sur le Web. Les principaux moteurs de recherche tels que Google, Bing et Baidu disposent de vastes quantités de données leur indiquant quels hyperliens reçoivent le plus grand nombre de clics en fonction des requêtes effectuées. Ces données aident les entreprises à perfectionner leur moteur de recherche (A), ce qui fait en sorte qu'on les utilise davantage (B) et qu'on fournit ainsi aux entreprises davantage de données d'usagers les plus pertinentes et profitables disponible dans le marché (C).

Tant que les avantages liés à la performance des modèles se maintiennent, ce cycle produit de manière inhérente une dynamique de marché naturelle dans laquelle l'entreprise qui s'est creusé une avance emporterait toute la mise, et même les petits avantages s'avèrent parfois décisifs (Frank et Cook, 2013). Un tel marché favorise des produits vedettes puisque la clientèle a peu intérêt, sinon aucun, à choisir un produit qui serait à peine moins bon que le meilleur de sa catégorie. Pour reprendre l'exemple de la recherche en ligne, aucune mesure n'incite les internautes à choisir un moteur de recherche qui serait légèrement moins performant que Google. Même les récompenses virtuelles comme celles qu'offre Bing pour encourager la recherche dans son moteur, et donc la production de données, ne parviennent pas à accélérer l'accumulation de données découlant naturellement du cycle vertueux.

Pour maximiser les effets positifs de la boucle de rétroaction, les entreprises doivent également employer une stratégie appropriée à l'égard des données. La centralisation de celles-ci s'avère notamment cruciale pour tirer profit de l'effet multiplicateur de l'apprentissage profond. Si la gestion des bases de données relève de responsables ou de services différents travaillant en vase clos, les ingénieurs et ingénieures pourront difficilement récupérer des données qu'il leur sera possible de recouper de manière fructueuse. Établir un entrepôt de données commun favorise la création d'un seul grand système d'IA qui exploite un maximum d'inférences et dont la performance dépasse, pour des raisons d'échelle, celle de nombreux systèmes de taille inférieure (voir figure 2).

Bien que les avantages liés à la performance des modèles ne diminuent pas en fonction de la taille de ceux-ci, les innovations touchant la puissance des calculs, les données et la structure des modèles ont tendu vers une centralisation. Les UTG ont permis d'améliorer les ordinateurs avant d'être elles-mêmes supplantées par une forme de centralisation : l'informatique en nuage (plusieurs serveurs) et le calcul haute performance (plusieurs UTG). Le cycle vertueux de l'IA favorise l'extraction et l'agrégation constantes de données alors que la taille des modèles s'amplifie de façon exponentielle de manière à intégrer des ensembles de données en expansion et à en généraliser l'application. La possibilité de déployer des systèmes d'IA et d'en tirer des bénéfices a suivi la même tendance que l'innovation. Elle risque de profiter davantage à quelques parties prenantes ainsi qu'à des travailleuses et travailleurs hautement éduqués. Cette inclination vers une centralisation de l'IA de même que les conséquences qu'elle entraînerait méritent par conséquent d'être examinées.

LA PRISE DE DÉCISIONS HUMAINE EST DÉCENTRALISÉE

Le potentiel de centralisation de la prise de décisions automatisée contraste avec la relative décentralisation de la prise de décisions humaine dans les marchés et les organisations. Cette situation reflète l'idée centrale selon laquelle les humains ont une capacité déterminée à traiter de l'information. Comme l'a fait remarquer Simon (1955, p. 99), tout modèle de prise de décisions qui se veut réaliste doit prendre en considération une telle limite :

> Broadly stated, the task is to replace the global rationality of economic man with the kind of rational behavior that is compatible with the access to information and the computational capacities that are actually possessed by organisms, including man, in the kinds of environments in which such organisms exist.

Il a par exemple été estimé que la mémoire humaine a une capacité de 2,5 pétaoctets (Reber, 2010) et que le cerveau présente une capacité à traiter inconsciemment de l'information s'élevant à 11 millions de bits par seconde (Wilson, 2004). Cet organe présente cependant certains « goulots d'étranglement » (Marois et Ivanoff, 2005). Ainsi, le traitement conscient ou « intelligent » de l'information – par exemple, le processus de lecture – ne s'effectuerait qu'à 50 ou 60 bits par seconde (Markowsky, 2021 ; Emerging Technology from the arXiv, 2020).

Peu importent les chiffres exacts, la capacité du cerveau est bien sûr finie et si limitée qu'aucun être humain, aussi intelligent soit-il, ne peut à lui seul prendre toutes les décisions stratégiques et financières relatives à une petite organisation, encore moins à une industrie ou à toute une économie.

Ce fait se répercute dans la manière de structurer les organisations. Il y a répartition des sphères de prise de décisions au sein des entreprises. Certaines personnes interagissent avec la clientèle, tandis que d'autres se chargent de la conception des produits. Leurs équipes respectives assument un nombre limité de responsabilités, et chacune se concentre même sur une certaine clientèle ou certains produits. D'autres personnes encore s'occupent de la gestion des stocks, de la fabrication ou des chaînes d'approvisionnement, ou elles prennent des décisions concernant les finances, le marketing, le recrutement ou l'orientation stratégique générale de l'entreprise. Certaines décisions sont très circonscrites (p. ex., faut-il vider telle corbeille à papier ?), alors que d'autres se répercutent plus largement sur l'entreprise (p. ex., doit-on percer un nouveau marché ?).

Comme l'ont fait remarquer Smith (2008 [1776]), Hayek (1986 [1945]) et plusieurs autres, les marchés et le système des prix présentent notamment l'avantage de permettre une décentralisation et une répartition de la prise de décisions au-delà des limites de l'entreprise. Un fabricant de crayons n'a pas à prendre une myriade de décisions pour produire le bois, le graphite, l'étain et la gomme à effacer dont ses produits sont faits, pas plus qu'à procéder à tout moment à une analyse coûts-avantages quant aux

sources d'approvisionnement de chacun de ces matériaux. Le ou la propriétaire de l'usine n'a qu'à savoir que le cours du marché de ces matières premières est une donnée utile pour en négocier l'achat. Dans le même ordre d'idées, un consommateur ou une consommatrice se contentera souvent d'estimer les avantages que lui procure un crayon, puis d'évaluer son prix sur le marché, lequel reflète d'innombrables compromis effectués par d'autres preneurs et preneuses de décisions en ce qui a trait à la conception, à la production, à la vente et à la livraison de ce bien. Le premier des théorèmes sur lequel se fonde l'économie du bien-être énonce que s'il n'y a pas d'externalités, d'information parfaite et de pouvoir de marché, il en résultera un optimum de Pareto, soit un équilibre général (Hammond, 1997)[13].

En effet, des millions de décideurs et décideuses peuvent focaliser leur attention – leur « rationalité limitée », pour reprendre l'expression de Simon (1955) – sur seulement un aspect pointu d'une vaste question à résoudre pour faire rouler l'économie, et tout ignorer du reste de la question (ou justement présumer que ces autres facteurs se résument à des prix).

Un débat animé a eu cours dans les années 1930 et 1940 afin de déterminer s'il était possible, en principe, de centraliser complètement de telles décisions. Dans ce débat sur le « calcul socialiste », un camp mené par Lange (1936), Lerner (1938) et d'autres faisait valoir que toute l'information requise pouvait être transmise à une personne ou une équipe chargée de prendre les décisions, qui calculerait les coûts et les avantages y étant associés, déterminerait les ressources optimales à allouer, et transmettrait aux autres acteurs de l'économie des instructions sur ce qu'il faut produire, transporter et consommer. L'autre camp, mené par von Mises (1951), Hayek (1986 [1945]) et d'autres, soutenait en substance qu'il y avait tout simplement trop d'information, ce qui rendait le calcul inexécutable.

Dans son fameux article « L'utilisation de l'information dans la société », Hayek (1986 [1945], p. 122) fait remarquer que chaque personne détient des bribes d'une seule information pouvant être utile dans certaines circonstances : « Connaître et utiliser une machine en partie inemployée, ou savoir comment mieux utiliser le talent de quelqu'un, avoir connaissance d'un stock sur lequel on pourra tirer durant une interruption d'approvisionnement [...] »

La dispersion de l'information défie les principes de l'agrégation statistique. Savoir qu'à un endroit précis, un camion en particulier est vide vaut quelque chose. Savoir que de nombreux camions rentrent souvent vides au terme de leur voyage quelque part en Amérique ou qu'une formule générale optimise l'itinéraire des livraisons ne se substitue pas à l'information précisant quel camion est vide. Hayek a donc souligné que le système économique joue un rôle important consistant à accorder le pouvoir décisionnel et les mesures qui encouragent l'exercice de ce pouvoir aux personnes disposant de l'information pertinente pour le faire. Et puisque cette information est dispersée, le pouvoir décisionnel devrait également l'être. Jensen et Meckling (1992) ont appliqué cette idée au pouvoir décisionnel qui s'exerce au sein d'une grande entreprise, où la difficulté à attribuer des droits de propriété inaliénables complique le problème.

Il y a malgré tout eu une tentative de centralisation des décisions économiques dès le début des années 1970 dans le cadre du projet Cybersyn, qui visait à gérer l'économie chilienne (Morozov, 2014). Un ordinateur central se trouvant dans la capitale, Santiago, était relié à des usines de tout le pays par un réseau télex national. Des renseignements relatifs, par exemple, à l'approvisionnement en matières premières ou à la productivité de la main-d'œuvre alimentaient un logiciel de simulation économique, qui renvoyait quant à lui des directives aux usines ou à d'autres organisations.

13. Dans la pratique, tous ces éléments ne sont pas réunis, ce qui limite dans une certaine mesure l'importance qu'ils ont concrètement.

Le projet Cybersyn a été abandonné en 1973[14]. Après l'effondrement de l'Union soviétique en 1989, il a semblé que le débat sur le « calcul socialiste » se réglait de manière décisive à l'avantage du camp qui défendait une décentralisation de la prise de décisions. Même si elle fonctionnait relativement bien dans certains secteurs comme l'industrie lourde, l'approche préconisant une planification centralisée semblait dépassée dans des secteurs de l'économie faisant preuve d'une grande innovation et évoluant rapidement, notamment celui du matériel informatique, des logiciels et des réseaux numériques, que les États-Unis en sont venus à dominer étant donné leur approche décentralisée.

Cette même industrie de la technologie numérique qui a évolué grâce à cette approche admet toutefois aujourd'hui un degré de centralisation de la prise de décisions beaucoup plus important que par le passé. Comme l'ont mentionné Brynjolfsson et Mendelson (1993), alors qu'une façon de rapprocher information et pouvoir décisionnel consiste à ramener ce pouvoir là où se trouve l'information, une autre option serait plutôt de déplacer l'information. D'importants systèmes, comme ceux permettant la planification des ressources en entreprise, ou d'importants réseaux, comme Internet, rendent ce déplacement possible pour divers types d'information. Des technologies, des capteurs – par exemple les étiquettes d'identification par radiofréquence (en anglais, RFID) – et l'Internet des objets facilitent l'enregistrement et la numérisation d'une telle information dispersée à laquelle Hayek faisait référence dans son fameux article.

Présentons un exemple concret. Durant la majeure partie du 20e siècle, le ou la propriétaire d'une petite épicerie familiale savait mieux que quiconque quels parfums de gomme à mâcher ou de glace étaient les plus populaires dans le voisinage, et quels commerces locaux livraient ces produits de manière fiable et abordable, ce qui constitue pour Hayek un élément d'information. Dans les années 1990, l'entreprise Walmart a conçu et mis en œuvre des systèmes sophistiqués effectuant un suivi de sa chaîne de vente dans tout le pays, de la demande au point de vente jusqu'à l'expédition par un détaillant. Aujourd'hui, la base de données qui se trouve dans ses installations de Bentonville, en Arkansas, « connaît » probablement les achats de glace à la menthe effectués dans chaque quartier plus en détail que n'importe quel épicier local ne le pouvait hier, sans compter que l'analyse prédictive parvient à déterminer quelle sera la demande demain en fonction des tendances saisonnières, des campagnes de marketing interactif, de données de téléphonie cellulaire, des prévisions météorologiques et d'une quantité impressionnante d'autres variables (Brynjolfsson *et al.*, 2021). C'est sans surprise que de petites épiceries ou de modestes détaillants en tous genres perdent des parts de marché au profit des grandes chaînes (Decker *et al.*, 2020).

Une poignée d'importants détaillants en ligne poussent encore plus loin l'utilisation d'information très précise. Il serait surprenant qu'un ou une libraire faisant du commerce traditionnel recommande des livres avec autant de finesse et de perspicacité que le font les outils d'Amazon, qui non seulement tirent profit d'un historique détaillé des achats et des préférences de chacun et chacune, mais recoupent également l'information touchant plusieurs utilisateurs et utilisatrices, grâce à des systèmes d'apprentissage automatique à la fine pointe entraînés par des téraoctets de données. De façon analogue, on conçoit des systèmes qui prédisent des pannes de moteur avant qu'elles ne surviennent (GE Research, n.d.), qui maintiennent le niveau des stocks dans les entrepôts (Chang, 2020), ou qui surveillent la circulation routière (Lau, 2020) ou une foule d'autres éléments qui nécessitaient auparavant de recueillir des données, de posséder l'expertise et de prendre des décisions sur le terrain.

14. Malgré la fin réelle des activités, le projet s'est poursuivi dans la science-fiction. Dans le roman *Synco*, par exemple, publié en 2008, Jorge Baradit l'a décrit comme « l'établissement du premier état cybernétique, un exemple universel, la véritable troisième voie, un miracle ».

Il n'y a pas que l'information qui est devenue un objet de centralisation. Comme nous l'avons mentionné dans la section précédente de ce chapitre, des décisions peuvent essentiellement être prises par des machines. Alors que la puissance de traitement du cerveau humain n'a pas changé au fil de milliers d'années, celle des ordinateurs double environ tous les deux ans depuis que Gordon Moore a énoncé la loi qui porte son nom, en 1965.

Grossman, Hart et Moore ont fourni un cadre afin de schématiser le pouvoir décisionnel, la propriété, le pouvoir de négociation et les frontières de l'entreprise (Grossman et Hart, 1986 ; Hart et Moore, 1990). Ils ont notamment montré que, dans toute décision de nature économique, la meilleure option consiste à attribuer la propriété des actifs ainsi que les droits résiduaires de contrôle qui s'y rapportent aux principales autorités décisionnelles. Il s'agit d'une manière de procurer à celles-ci un pouvoir de négociation pour qu'elles réclament une part de la valeur découlant de leurs décisions, ce qui contribue à les motiver à créer encore plus de valeur.

Brynjolfsson (1994) a étendu la portée de ce cadre pour montrer que, lorsque l'information passe de cerveaux humains à des actifs non humains (par exemple, une base de données ou un système d'IA), il devient possible et même souvent profitable, d'un point de vue économique, de centraliser non seulement l'information, mais aussi la propriété des actifs qui y sont liés. Le cadre de Grossman, Hart et Moore associe la dispersion de la propriété des actifs à des transactions qui auraient lieu entre différentes entreprises du marché. À l'opposé, le regroupement de la propriété des actifs serait une caractéristique propre aux transactions ayant lieu au sein d'une même entreprise. Ainsi, en centralisant l'information et la prise de décisions, les systèmes d'IA font parfois en sorte qu'il est préférable de centraliser aussi la propriété. Selon ce cadre, il faut pour ce faire ramener les transactions à l'intérieur des frontières de l'entreprise et limiter les transactions sur le marché.

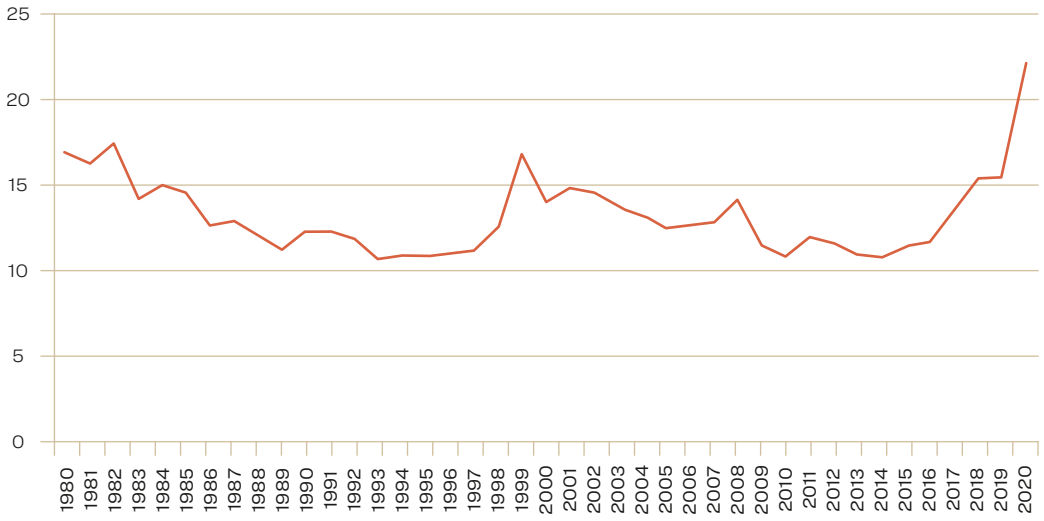
LA CENTRALISATION S'IMPLANTE DÉJÀ DANS DE NOMBREUX DOMAINES

Nous avons précédemment fait valoir que l'IA peut mener à une centralisation de la prise de décisions, ce qui survient en fonction du transfert de deux éléments des cerveaux humains vers les machines : d'une part, le savoir ou l'information et d'autre part, la puissance de traitement.

Divers exemples illustrent ce phénomène, qui s'observe au-delà de l'essor des chaînes de magasins et des détaillants en ligne venus remplacer les commerces locaux. Pour la première fois de l'histoire des États-Unis, cinq entreprises présentent une capitalisation boursière dépassant le billion de dollars (Apple, Amazon, Alphabet, Microsoft et Facebook). Comme le montre la figure 5, les cinq entreprises les plus importantes dont tient compte l'indice S&P 500 représentent désormais plus de 22 % de sa valeur, un record des temps modernes (Scheid, 2020).

| FIGURE 5 |

Part moyenne de l'indice S&P 500 attribuable à la capitalisation boursière des cinq entreprises les plus importantes, par année. *Source :* Scheid, 2020 (données recueillies le 24 juillet).



La concentration ne prend pas de l'ampleur que dans l'industrie du numérique, mais également dans l'ensemble de l'économie états-unienne. À titre d'exemple, quatre grossistes ont la mainmise sur 80 % du marché de la viande de bœuf (MacDonald *et al.*, 1999) et de nombreux foyers américains doivent se tourner vers le seul fournisseur qui leur offre un accès Internet à haut débit (Rogers, 2017). Des études étayent ces chiffres et montrent que les grandes entreprises ont vu leurs profits croître au cours des 40 dernières années, ce qui a réduit les chances des petites entreprises d'imposer une réelle concurrence (Boushey et Knudsen, 2021). En 2020, la concentration du marché aux mains des 3 000 entreprises ayant la plus forte capitalisation boursière a atteint un sommet inégalé depuis 1986, selon la Bank of America (The Economist, 2021).

Cela dit, la centralisation accrue qui s'observe aux États-Unis ces dernières années n'a généralement rien à voir avec l'IA, mais découle plutôt de l'influence d'autres technologies et de choix politiques tels que des changements aux lois antitrust et une certaine déréglementation (Stuck et Ezrachi, 2017). En ce qui a trait particulièrement aux technologies, des économies d'échelle relatives tant à l'offre qu'à la demande avantagent les grandes entreprises.

En ce qui concerne l'offre, la production de matériel informatique et surtout de logiciels tend à comporter des coûts fixes élevés et des coûts marginaux faibles, voire nuls. Produire le premier exemplaire d'un tout nouveau type de microprocesseur peut représenter des millions ou des milliards de dollars[15] en recherche et développement et en d'autres coûts de production initiaux. Les exemplaires suivants seront toutefois fabriqués à très faible coût, voire sans rien payer de plus. Cela signifie que les grandes entreprises qui vendent une importante quantité de leurs produits bénéficient de coûts moyens relativement faibles, ce qui leur permet d'imposer un prix avantageux tout en engrangeant de forts profits. Une telle structure de coûts est courante dans l'industrie des technologies numériques.

Puisque « les logiciels mangent le monde », pour reprendre les mots de Marc Andreessen, l'enjeu s'applique aussi à d'autres industries qui prennent de plus en plus le virage numérique (Brynjolfsson *et al.*, 2008). Ainsi, dans la fabrication d'automobiles, le coût des composantes électroniques est passé d'un ratio du coût de production de 18 % à 40 % au cours des 20 dernières années (Deloitte, 2019). Des détaillants tels que les pharmacies CVS aux États-Unis ont numérisé leurs processus d'affaires, ce qui leur a permis de croître très rapidement en nombre et en taille (McAfee et Brynjolfsson, 2008).

En ce qui concerne la demande, les entreprises sont de plus en plus nombreuses à tirer profit de leur envergure. Prenons Facebook, Apple, Uber ou Airbnb. Grâce aux avantages de la mise en réseau, la valeur de leurs produits et services augmente à mesure que croît leur clientèle. Il en découle des retombées directes, par exemple quand des proches s'inscrivent tous à un même réseau social et y partagent des publications, de même que des retombées indirectes (et bilatérales), par exemple quand la clientèle d'Uber bénéficie d'une augmentation du nombre de chauffeurs et de chauffeuses ou encore que des annonceurs profitent d'une hausse de la clientèle. On évalue souvent les effets d'échelle de manière approximative en suivant la loi de Metcalfe, selon laquelle l'utilité d'un réseau est proportionnelle au carré du nombre d'utilisateurs et utilisatrices, alors que les coûts ne progressent que de manière linéaire. Bien souvent, un modèle d'attachement préférentiel (faisant en sorte qu'une nouvelle clientèle est plutôt susceptible de se joindre à des réseaux ayant déjà un grand nombre d'utilisateurs et utilisatrices) mènera à une dynamique dans laquelle l'entreprise qui s'est creusé une avance emportera toute la mise, un phénomène que décrit bien la courbe de la loi de puissance. Bien que l'IA ne soit pas au cœur de la mise en réseau, elle en amplifie souvent les effets, par exemple quand elle facilite l'accès à des services de transport personnalisés.

Les économies d'échelle relatives à l'offre tout comme les effets de la mise en réseau (qui deviennent en fait des économies d'échelle relatives à la demande) s'observent couramment dans les industries ayant recours aux technologies numériques. Ils favorisent l'établissement d'une seule grande installation de production ou d'un seul large réseau, et n'entraînent pas nécessairement une centralisation de la prise de décisions. Il arrive que plusieurs entreprises partagent les mêmes infrastructures, par exemple des autoroutes, ou qu'elles rendent des réseaux distincts interopérables, ce qui leur permet de bénéficier des avantages de la mise en réseau même si chaque composante est la propriété de différentes entités[16]. Donc, d'un point de vue technique et économique, des approches réglementaires visant à éviter la centralisation du pouvoir décisionnel fonctionnent, par exemple le fait d'assurer la transférabilité des numéros de téléphone d'un fournisseur de services à un autre (Federal Communications Commission, 1996).

L'IA ne suit pas, quant à elle, cette tendance. Elle ne centralise pas que des activités, mais bien la prise de décisions elle-même. Alors qu'une répartition du pouvoir décisionnel reste possible même quand on centralise des activités, elle s'avère très difficile quand il y a centralisation de la prise de décisions. Si un important système d'IA se fondant sur de grandes quantités de données parvient systématiquement

15. La nouvelle usine de fabrication de puces d'Intel, par exemple, coûtera de 60 à 120 milliards de dollars (Shilov, 2021).

16. La manière dont de nombreux gestionnaires de réseaux indépendants coordonnent le trafic de données sur Internet en constitue un exemple.

à prendre de meilleures décisions que n'importe quelle entité qui l'administre, quelles mesures fiables de contrôle et d'équilibre pouvons-nous mettre en place ? Comment nous assurer que le système chargé de prendre des décisions partage les valeurs et les objectifs de la population en général[17] ?

UNE FORTE CENTRALISATION DE LA PRISE DE DÉCISIONS NUIT À LA SOCIÉTÉ

Centraliser la prise de décisions comporte de nombreux avantages, notamment celui de mieux saisir l'interdépendance et les interactions qui s'observent entre des parties d'un ensemble et ainsi de procéder à une optimisation globale et non locale.

La centralisation comporte toutefois également des risques. Centraliser la prise de décisions, c'est aussi centraliser le pouvoir. Il ne s'agit pas d'une mauvaise chose, dans la mesure où l'autorité qui bénéficie du pouvoir est bienveillante et où elle cherche à créer de la valeur pour toutes les parties en cause. Comme l'a fait remarquer Lord Acton (1887) : « Le pouvoir tend à corrompre, le pouvoir absolu corrompt absolument. » Il serait donc risqué de se fier aux bonnes intentions d'un noyau de décideurs et décideuses. Même si ces derniers cherchent au départ à protéger la liberté et l'égalité des droits, leurs intentions pourraient s'envoler faute de mesures visant à encadrer ce pouvoir.

Une concentration de la prise de décisions en matière économique, en particulier, ébranle parfois la démocratie. Comme le juge Louis Brandeis l'a formulé (Dilliard, 1941, p. 42-45) : « Nous pouvons avoir la démocratie dans ce pays ou nous pouvons avoir une grande richesse concentrée dans les mains de quelques-uns, mais nous ne pouvons pas avoir les deux. » Souvent, les outils de l'IA servent d'ailleurs précisément à orienter les flux d'information de manières qui nuisent à la démocratie (Acemoglu, 2021).

D'un point de vue historique, la concentration du pouvoir à l'échelle d'un pays ou d'un système a aussi nui aux droits fondamentaux. Il ne s'agit pas d'une règle universelle, mais les personnes qui n'ont aucun rôle à jouer dans la prise de décisions ne reçoivent généralement pas un traitement adéquat (Acemoglu et Robinson, 2012). Des politologues ont donc fait valoir que, pour mieux défendre les droits des gens ordinaires, mieux vaut un système moins efficace allant de pair avec des mesures de contrôle et d'équilibrage qu'un système très raffiné (voir par exemple Reich *et al.*, 2021). Cette inefficacité s'avère certes frustrante, mais il ne semble pas y avoir de meilleure solution. Pour reprendre les paroles de Winston Churchill : « La démocratie est le pire système de gouvernement, à l'exception de tous les autres qui ont pu être expérimentés dans l'histoire. »

RECOMMANDATIONS

Comme il en est fait mention dans l'introduction, la centralisation de la prise de décisions n'est pas inéluctable. De puissantes forces s'y opposent, et l'issue de cet enjeu dépendra largement des choix que nous faisons aujourd'hui.

Trois séries d'approches permettent notamment de limiter la centralisation de la prise de décisions ou d'atténuer les risques qu'elle pose : 1) des approches technologiques, 2) des approches économiques et 3) des approches politiques.

17. Russel (2009), Bostrom (2016) et Yudkowsky (2016) font partie de ceux et celles qui ont mis l'accent sur ce questionnement.

Approches technologiques

Nous pouvons concevoir une technologie qui contribue à la décentralisation de la prise de décisions en faisant en sorte qu'il soit plus facile qu'avant pour une variété de personnes d'innover, d'améliorer des biens et services, ou encore d'en créer de nouveaux. Nous pouvons aussi concevoir des plateformes qui favorisent l'émergence d'idées et l'esprit d'entreprise, et qui contribuent à donner le pouvoir décisionnel et des rétributions financières à des personnes qui innovent (Phelps, 2015).

L'interopérabilité et des normes cohérentes peuvent maintenir les avantages de la mise en réseau, tout en décentralisant les droits de propriété. Souvent, des normes très claires et strictes stimulent paradoxalement la créativité, en forçant la conception de composantes qui se conforment à leurs exigences. Le protocole TCP/IP, soit la norme au cœur d'Internet, constitue un exemple convaincant de cet effet de stimulation. Son adoption a encouragé l'innovation. Des millions d'entrepreneurs et entrepreneuses, de chercheurs et chercheuses, et de développeurs et développeuses de logiciels ont créé des applications technologiques variées – courrier électronique, World Wide Web, voix sur IP, Internet des objets, etc. – qui ont rapidement pris de l'ampleur et de l'importance. Il suffisait qu'elles respectent quelques normes de base.

À l'instar de l'interopérabilité, qui préserve les avantages de la mise en réseau tout en décentralisant la propriété et le pouvoir, la réplication et le partage des données permettent à plusieurs systèmes d'apprentissage automatique de tirer profit de vastes ensembles de données. En fait, l'un des avantages majeurs des données numériques est que leur réplication s'effectue à très faible coût. Ainsi, nul besoin de trop dépenser pour produire plusieurs exemplaires de fichiers contenant des images, du texte ou d'autres types de données dont des équipes et des organisations se serviront pour entraîner leur propre modèle.

Parallèlement, une large répartition des outils d'IA (tels que de grands modèles préentraînés) et des mégadonnées (*big data*) serait profitable, de petites entités se montreraient alors plus compétitives devant les entreprises dominantes, ce qui viendrait contrecarrer l'effet d'attachement préférentiel. Lera *et al.* (2020) proposent par exemple de concevoir un système fédéré dans lequel il y aurait partage, au sein d'un réseau d'entités participantes, de la description des données dans son ensemble (et non des données sous-jacentes). Chacune des entités combinerait de son côté ses propres données à la description commune.

D'autres innovations technologiques ont été conçues dans le but précis de décentraliser le pouvoir. La chaîne de blocs, notamment, évite le recours à un contrôleur d'accès qui approuve les transactions – ce qui est nécessaire pour les bases de données traditionnelles ou les grands livres – et décentralise plutôt ce processus. Elle s'est accompagnée de fascinantes promesses, certains ayant même estimé qu'elle mènerait à la réémergence d'une infrastructure financière décentralisée (Pentland, 2015), voire à la disparition de l'État-nation (Pueyo, 2021). Concrètement, la chaîne de blocs a souvent eu un effet plus centralisateur encore que l'infrastructure conventionnelle. À un certain moment, seulement quatre mineurs et mineuses de bitcoins se trouvant en Chine accaparaient plus de 50 % de l'activité de minage (Sharma, 2019).

Nous pouvons par ailleurs également concevoir des systèmes d'IA visant à « augmenter » la prise de décisions humaine plutôt qu'à la remplacer et ainsi peut-être soutenir la décentralisation du pouvoir[18]. En effet, même si le nombre de décisions que peuvent prendre des machines ne cesse de croître, les humains exécutent mieux la plupart des tâches, et nombre de celles-ci s'accomplissent encore mieux si l'on combine l'humain et la machine. La prise de plusieurs types de décisions restera donc décentralisée. Une option susceptible de s'appliquer un certain temps serait que des technologues s'affairent à concevoir

18. Pour en apprendre davantage au sujet de cette approche, voir par exemple Brynjolfsson et McAfee (2011) ou Brynjolfsson (2022).

des systèmes qui soutiennent les humains plutôt que de les remplacer. Ces derniers restent souvent meilleurs que les systèmes d'apprentissage automatique pour s'acquitter de tâches non structurées et définir des problèmes, ce qui est particulièrement important en matière d'innovation et d'esprit d'entreprise (Bryanjolfsson *et al.*, 2017).

Ces approches technologiques s'avèrent certes utiles, mais il est probable que des entités ayant intérêt à empêcher leur mise en œuvre s'y opposent. La transférabilité des numéros de téléphone, par exemple, est une mesure nettement avantageuse pour la clientèle, mais les fournisseurs de services de téléphonie mobile ont refusé de l'appliquer et l'ont contestée pendant de nombreuses années (Douglass, 2002). Ce n'est qu'après une décision qu'a prise le Congrès des États-Unis en 2003 que les fournisseurs se sont conformés à l'exigence de permettre la transférabilité (Federal Communications Commission, 2009). Mettre en œuvre les formes d'interopérabilité et de partage des données et de l'information nécessaires à une répartition des systèmes d'apprentissage automatique risque d'être beaucoup plus complexe que d'imposer la transférabilité des numéros de téléphone. Cette complexité peut en retarder la mise en œuvre et mener à des conséquences insoupçonnées. Certains diront que le Règlement général sur la protection des données, qui vise à protéger les renseignements personnels des consommateurs et consommatrices, a peut-être renforcé le pouvoir de grandes plateformes numériques (Nouwens *et al.*, 2020).

En fin de compte, les approches technologiques présentent une principale faiblesse. Bien qu'elles contribuent à atténuer certaines forces centralisatrices tels les effets de la mise en réseau (par l'interopérabilité), elles parviennent difficilement à contrer la centralisation de la prise de décisions en elle-même[19].

Approches économiques

Une autre série d'approches visant à lutter contre la centralisation du pouvoir relève de l'économie. En investissant dans l'éducation générale, une société améliore dans l'ensemble l'efficacité de la prise de décisions. Plus il y a de personnes qui possèdent le capital humain – soit les compétences et le savoir – requis pour prendre des décisions éclairées, plus le pouvoir de décider risque d'être largement réparti. De la même manière, une vaste répartition de la propriété relative aux capitaux matériels et financiers stimulerait l'esprit d'entreprise et étendrait ainsi la prise de décisions. Si elles sont bien menées, de telles politiques s'avèrent doublement gagnantes : elles distribuent le pouvoir de négociation, tout en stimulant l'innovation, la productivité et la croissance. Au cours de l'histoire, l'équilibre de nombreux marchés a reposé sur la présence de deux ou trois entreprises dominantes plutôt que sur un monopole, malgré les importantes économies d'échelles qui auraient profité à celui-ci. Et Schumpeter (1942) l'a fait remarquer : même les monopoles se font parfois renverser par le processus de destruction créatrice qui survient quand de nouvelles plateformes et de nouveaux paradigmes apparaissent. Un ferme soutien aux nouveaux joueurs désirant percer le marché et à la concurrence contribuerait à limiter le pouvoir d'éventuels monopoles.

Une approche complémentaire consisterait à encadrer les organisations qui centralisent le pouvoir, de la même manière dont les organismes de réglementation ont longtemps cherché à tenir en bride les entreprises monopolistiques de domaines tels que la production d'électricité ou les services téléphoniques. Cela pourrait passer par une réglementation visant directement la prise de décisions et les bénéfices – une mesure souvent appliquée aux services publics fournissant de l'électricité – ou par une approche plus souple comme l'imposition d'une taxe sur les revenus de la publicité numérique, dont le taux irait grandissant dans certains domaines en fonction des revenus des organisations, selon

19. Jusqu'à maintenant du moins, les tentatives d'utilisation de l'« apprentissage fédéré » n'ont pas réussi à répartir l'apprentissage autant que l'on s'y attendait.

ce qu'a proposé Romer (2019). Étant donné les profits importants que représente une faible concurrence ou une situation de monopole, les entreprises sont souvent amenées à limiter l'interopérabilité ou le partage des données. Voilà qui justifie que les organismes de réglementation encouragent ou imposent un certain partage des données, tout comme les autorités antitrust interviennent pour maintenir ou stimuler la concurrence.

Subventionner la décentralisation de la prise de décisions ou pénaliser la prise de décisions centralisée sont deux approches qui présentent une même faiblesse : elles risquent de saper les économies d'échelle. Si la prise de décisions centralisée devient réellement la meilleure manière de répartir des ressources, alors les sociétés qui cherchent à la freiner pourraient accuser un retard sur le plan économique.

Dans ce cas, il paraît raisonnable de répartir le pouvoir économique au moyen d'un revenu de base universel (voir par exemple Lowrey, 2018). Si les dollars s'apparentent à des « votes » qui orientent les décideurs et décideuses économiques vers certains domaines pour déterminer le type de produits et services en demande ou le genre d'innovations qui semblent rentables, alors le revenu de base distribue largement les « votes » et donc le pouvoir décisionnel qu'ils représentent. De cette façon, même si un marché débridé accordait toujours moins de pouvoir aux personnes qui ne possèdent pas de grands systèmes d'apprentissage automatique ou qui n'ont aucune influence sur ceux-ci, le revenu de base universel rétablirait une partie de ce pouvoir, du moins dans la sphère économique.

L'attribution d'un revenu de base viserait à répartir plus largement le capital humain si une partie de l'allocation était destinée à l'acquisition de connaissances ou même conditionnelle à celle-ci. Il serait alors question d'un revenu de base conditionnel, et non universel. Des administrations pourraient aussi choisir d'employer le revenu de base pour redistribuer le pouvoir politique en y rattachant des fonds à verser en contributions politiques. L'attribution généralisée d'allocations servant à contribuer aux campagnes politiques diluerait le pouvoir de grandes entreprises et de riches individus, et elle redonnerait plutôt de la visibilité à certains messages ou à des figures politiques.

Les promesses dont sont porteuses toutes ces approches s'effritent toutefois si les détenteurs et détentrices du pouvoir économique n'ont pas besoin de la contribution décisionnelle ou financière des autres. Ceux-là pourraient même décider de réécrire les règles pour renforcer ce pouvoir qu'ils détiennent. Il existe un réel risque de voir la concentration du pouvoir économique mener à une concentration du pouvoir politique. Le problème de la capture réglementaire est récurrent quand des organismes de réglementation tombent sous l'emprise des entités qu'ils sont censés réguler. *Quis custodiet ipsos custodes ?*[20]

Approches politiques

En fin de compte, maintenir et renforcer la démocratie constitue peut-être le meilleur contrepoids à la centralisation du pouvoir dans d'autres sphères. Alors que l'apprentissage automatique entraîne une concentration de la prise de décisions dans de nombreux domaines, c'est bien une décision politique – et non technologique ou économique – de remettre ultimement le pouvoir au peuple grâce à des principes et à des institutions démocratiques. Les forces du marché et le capitalisme tendent, pour des questions d'efficacité, à centraliser la prise de décisions au moyen d'importants systèmes d'apprentissage automatique, ce qui creuse davantage les disparités en ce qui a trait à la richesse et au pouvoir. Ajoutons, pour reprendre des propos de Khosla (2017), que la démocratie admet néanmoins le capitalisme et qu'elle devrait disposer d'outils pour pallier les disparités.

20. Mais qui gardera ces gardiens ?

Un principe fondamental de la démocratie est « une personne, un vote », et non « un dollar, un vote ». Tandis que le marché récompense les gens et leur accorde du pouvoir décisionnel en fonction d'une certaine estimation de leur pouvoir de négociation économique, la démocratie traite quant à elle les gens comme des fins plutôt que des moyens. La voix de chaque personne a le même poids, peu importe que celle-ci détienne de précieux renseignements, de bons moyens financiers ou des actifs importants, ou qu'elle n'apporte aucune contribution.

La simple mise sur pied d'institutions au sein d'une démocratie ou d'une république ne suffit toutefois pas à maintenir une gouvernance répartie[21]. Il existe une foule de manières de brimer le droit de vote ou d'empêcher l'exercice de ce droit, ou de détourner le pouvoir de certaines personnes ou de certains groupes au profit d'autres. L'apprentissage automatique peut même soutenir des initiatives antidémocratiques, par exemple servir à analyser minutieusement des données afin de prévoir les intentions de vote, puis de procéder à un découpage électoral partisan, d'adapter des stratégies de marketing, ou de présenter des candidatures plaisant davantage à ceux et celles qui ont la mainmise sur les données et la technologie. La démocratie dépend également d'une foule d'institutions et de normes, par exemple d'une presse libre, de la liberté d'association et du droit à la dissidence. L'emploi des systèmes d'IA de plus en plus puissants et le recours abusif à de tels systèmes menacent également ces principes.

De plus, comme si le projet d'établir une véritable démocratie n'était pas assez complexe, la puissance grandissante de l'apprentissage automatique est un enjeu qui dépasse les frontières nationales et s'étend à toute la planète. Les réseaux de transmission des données et les réseaux sociaux actuels se déploient au-delà des frontières, ce qui permet à des entités – qu'il s'agisse d'États ou d'autres parties prenantes – d'étendre leur influence à l'échelle internationale. Une concentration de la prise de décisions peut nuire aux organisations au-delà des frontières politiques et géographiques, alors un système efficace devrait comprendre un cadre global permettant d'établir des règles équitables.

Si les systèmes d'apprentissage automatique accentuent la centralisation de la prise de décisions au sein des marchés et des entreprises, il importera de consolider la démocratie dans la sphère politique. La démocratie représente l'ultime contrepoids à un pouvoir centralisé et une mesure de protection de la liberté individuelle.

CONCLUSION

D'aussi loin que les humains se sont associés, la centralisation et la décentralisation de la prise de décisions ont fait l'objet de tensions. Si la prise de décisions centralisée tient compte de l'interdépendance des parties d'un ensemble et améliore ainsi l'efficience, elle comporte néanmoins historiquement deux inconvénients majeurs : 1) la puissance de calcul connaît des limites, ce qui empêche toute entité, qu'il s'agisse d'un humain ou d'une machine, de prendre des décisions au-delà d'un sous-ensemble fini de dilemmes et 2) aucune entité ne possède toute l'information ou l'expertise pertinentes à la prise de décisions.

Les ordinateurs, l'apprentissage automatique et les systèmes de collecte de données se faisant de plus en plus puissants, ces deux éléments semblent moins contraignants. Alors que la prise de décisions humaine est intrinsèquement décentralisée en raison des limites qu'a l'esprit humain en matière de calcul ainsi que de collecte et de partage de l'information, ces capacités s'améliorent

21. Selon une version non authentifiée de la convention constitutionnelle des États-Unis, une dame aurait demandé à Benjamin Franklin : « Eh bien, docteur, qu'avons-nous : une république ou une monarchie ? », ce à quoi il lui aurait répondu : « Une république, si vous pouvez la garder. »

sans cesse dans les systèmes d'IA modernes. Dans un éventail de domaines de plus en plus large, de tels systèmes tiennent compte d'une quantité accrue d'information et prennent de meilleures décisions que des individus, certes, mais aussi que des groupes de personnes travaillant ensemble.

L'évolution de ces systèmes pourrait entraîner des répercussions considérables sur l'économie, notre gouvernance et même l'équilibre mondial. Des années 1950 à 1980, une lutte constante a opposé, d'une part, les systèmes économique et politique de l'Ouest, qui s'appuyaient largement sur une répartition de la gouvernance (démocratie) et de la propriété des moyens de production (capitalisme), et, d'autre part, le système soviétique, qui centralisait fortement ces deux types de pouvoir décisionnel. En 1989, il est devenu manifeste que la première combinaison était gagnante. Bien que l'explication qui prévaut mette souvent l'accent sur les vertus de la liberté et de la démocratie, le facteur décisif de cette victoire est vraisemblablement la capacité d'innovation supérieure dont jouit la libre entreprise, particulièrement quand il est question d'innovation technologique de pointe et de création globale de richesse.

Sachant que, même s'il vise au départ des objectifs nobles, l'exercice centralisé du pouvoir risque de dériver vers un totalitarisme, la plupart d'entre nous se réjouissent de constater qu'une approche décentralisée l'a emporté. Le résultat serait-il le même si la lutte reprenait en 2030 ou en 2040 ? La capacité qu'ont les technologies à prendre des décisions n'est plus ce qu'elle était il y a 40 ans. Elle évolue à une vitesse vertigineuse, et les systèmes de prise de décisions automatique gagnent en ampleur et en puissance, d'une manière inégalée.

Tout dénouement ne nous semble pas inéluctable, mais la complaisance n'est pas de mise. Les mesures favorisant la prise de décisions centralisée prendront souvent le dessus sur les avantages que la société tirerait d'une décentralisation. Ainsi, nous ne pouvons pas nécessairement compter sur un marché débridé pour empêcher que s'accentue la centralisation de la prise de décisions, du pouvoir et de la richesse. À mesure que la technologie évolue, il nous incombe de soupeser les avantages et les risques que présente la prise de décisions par des machines de plus en plus puissantes et de veiller consciencieusement à favoriser la liberté et l'épanouissement des êtres humains.

RÉFÉRENCES

Acemoglu, D. 2021. *Redesigning AI*. Cambridge, MIT Press.

Acemoglu, D. et Robinson, J. A. 2012. *Why Nations Fail : The Origins of Power, Prosperity, and Poverty*. New York, Currency.

Acton, J. 1887. *Acton-Creighton Correspondance.* Éditeur inconnu.

Baradi, J. 2009. Synco : El juego del revés. *El Mercurio Revista de Libros* (article en espagnol).

Begenau, J. 2018. Big data in finance and the growth of large firms. *Journal of Monetary Economics*, vol. 97, pp. 71-87.

Belton, P. 2021. *Why coders love the AI that could put them out of a job.* BBC News. 7 spetembre. https://www.bbc.com/news/business-57914432

Bostrom, N. 2016. *Superintelligence : Paths, Dangers, Strategies.* Oxford : University Press.

Boushey, H. et Knudsen, H. 2021. *The importance of competition for the american economy*. The White House. Bloque. 9 juillet. https://www.whitehouse.gov/cea/blog/2021/07/09/the-importance-of-competition-for-the-american-economy/

Brynjolfsson, E. 1994. Information assets, technology and organization. *Management Science*, vol. 40, nº 12, pp. 1645-1662.

Brynjolfsson, E. 2022. The Turing trap : The promise and peril of creating human level intelligence. *Daedalus*. Numéro du printemps. https://www.amacad.org/publication/turing-trap-promise-peril-human-artificial-intelligence

Brynjolfsson, E., Jin, W. et McElheran, K. S. 2021. *The power of prediction : Predictive analytics, workplace complements, and business performance.* SSRN. http://dx.doi.org/10.2139/ssrn.3849716

Brynjolfsson, E. et McAfee, A. 2011. *Race against the machine : How the digital revolution is accelerating innovation, driving productivity, and irreversibly transforming employment and the economy*. Digital Frontier Press.

––––. 2017a. Artificial intelligence, for real. *Harvard Business Review*, pp. 1-31. https://store.hbr.org/product/artificial-intelligence-for-real/BG1704

––––. 2017b. What's driving the machine learning explosion ? *Harvard Business Review,* The Big Idea Series. 18 juillet. https://hbr.org/2017/07/whats-driving-the-machine-learning-explosion

Brynjolfsson, E., McAfee, A., Sorell, M. et Zhu, F. 2008. Scale without mass : Business process replication and industry dynamics. *Harvard Business School Technology & Operations Mgt. Unit Research Paper,* nº 07-016.

Brynjolfsson, E. et Mendelson, H. 1993. Information systems and the organization of modern enterprise. *Journal of Organizational Computing*, vol. 3, nº 3, pp. 245-255.

Brynjolfsson, E. et Mitchell, T. 2017. What can machine learning do ? Workforce implications. *Science*, vol. 358, nº 6370, pp. 1530-1534.

Chang, C. 2020. *3 ways AI can help solve inventory management challenges*. IBM Supply Chain Blog. 4 mars. https://www.ibm.com/blogs/supply-chain/3-ways-ai-solves-inventory-management-challenges/

Decker, R. A., Haltiwanger, J., Jarmin, R. S. et Miranda, J. 2020. Changing business dynamism and productivity : Shocks versus responsiveness. *American Economic Review*, vol. 110, nº 12, pp. 3952-3990.

Deloitte, 2019. *Semiconductors – The Next Wave : Opportunities and Winning Strategies for Semiconductor Companies*. https://www2.deloitte.com/content/dam/Deloitte/tw/Documents/technology-media-telecommunications/tw-semiconductor-report-EN.pdf

Devries, P. M., Viégas, F., Wattenberg, M. et Meade, B. J. 2018. Deep learning of aftershock patterns following large earthquakes. *Nature,* vol. 560, nº 7720, pp. 632-634.

Dilliard, I. 1941. Mr. Justice Brandeis : Great American. *The Modern View Press*. 128 p.

DiSalvo, D. 2013. Your brain sees even when you don't. *Forbes*. 22 juin. https://www.forbes.com/sites/daviddisalvo/2013/06/22/your-brain-sees-even-when-you-dont/?sh=6c44bbe6116a

Douglass, E. 2002. Carriers aim to kill number portability. *Los Angeles Times*. 16 janvier. https://www.latimes.com/archives/la-xpm-2002-jan-16-fi-cell16-story.html

Economist (The). 2021. Is America Inc getting less dynamic, less global and more monopolistic ? 18 septembre. https://www.economist.com/business/is-america-inc-getting-less-dynamic-less-global-and-more-monopolistic/21804757

Emerging Technology from the arXiv. 2009. New measure of human brain processing speed. *MIT Technology Review*. 25 août. https://www.technologyreview.com/2009/08/25/210267/new-measure-of-human-brain-processing-speed/

Federal Communications Commission (USA). 1996. Telephone number portability. https://www.fcc.gov/document/telephone-number-portability-19

Federal Communications Commission (USA). 2009. Wireless local number portability. https://docs.fcc.gov/public/attachments/FCC-96-286A1.pdf

Frank, R. et Cook, P. 2013. Winner-take-all markets. *Studies in Microeconomics*, vol. 1, nº 2, pp. 131-154.

GE Research. n.d. Predictive maintenance. https://www.ge.com/research/project/predictive-maintenance

Grossman, S. J. et Hart, O. D. 1986. The costs and benefits of ownership : A theory of lateral and vertical Integration. *Journal of Political Economy*, vol. 94, nº 4, pp. 691-719.

Hammond, P. 1997. The efficiency theorems and market failure. Preprint chapter. Kirman, A (ed.). *Elements of General Equilibrium Analysis*. Toronto : Wiley. 1998..

Hart, O. 1989. An economist's perspective on the theory of the firm. *Columbia Law Review*, vol. 89, nº 7, pp. 1757-1774.

Hart, O. et Moore J. 1990. Property rights and the nature of the firm. *Journal of Political Economy*, vol. 98, nº 6, pp. 1119-1158.

Hayek, F. A. 1986 [1945]. L'utilisation de l'information dans la société. *Revue française d'économie*, vol. 1, nº 2, pp. 117-140. https://doi.org/10.3406/rfeco.1986.1120 (consulté le 2 novembre 2021).

Jensen, M. et Meckling, W. 1992. Specific and general knowledge and organizational structure. L. Werin et H. Wijkander (ed.), *Contract Economics*. Oxford, Blackwell Publishers.

Khosla, V. 2017. AI : Scary for the right reasons. Blogue Khosla Ventures, 12 septembre. https://www.khoslaventures.com/ai-scary-for-the-right-reasons

Krizhevsky, A., Sutskever, I. et Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, vol. 60, nº 6, pp. 84-90.

Lange, O. 1936. On the economic theory of socialism. *The Review of Economic Studies*, vol. 4, nº 1.

Lau, J. 2020. Google Maps 101 : How AI helps predict traffic and determine routes. Blogues Google, 3 septembre. https://blog.google/products/maps/google-maps-101-how-ai-helps-predict-traffic-and-determine-routes/

Lera, S. C., Pentland, A. et Sornette, D. 2020. Prediction and prevention of disproportionally dominant agents in complex networks. *Proceedings of the National Academy of Sciences*, vol. 117, nº 44, pp. 27090-27095.

Lerner, A. 1938. Theory and practice in socialist economics. *Review of Economic Studies*, vol. 6, nº 1.

Lowrey, A. 2018. *Give People Money : How a Universal Basic Income Would End Poverty, Revolutionize Work, and Remake the World.* New York, Broadway books.

MacDonald, J. M., Ollinger, M. E., Nelson, K. E. et Handy, C. R. 1999. *Consolidation in U.S. Meatpacking.* Food and Rural Economics Division, Economic Research Service, U.S. Department of Agriculture, Agricultural Economic Report No. 785.

Malone, T. 2003. The decentralization imperative. *MIT Technology Review.* 24 octobre. https://www.technologyreview.com/2003/10/24/274985/the-decentralization-imperative/

Markowsky, G. 2021. Physiology. *Britannica.* https://www.britannica.com/science/information-theory/Physiology

Marois, R. et Ivanoff, J. 2005. Capacity limits of information processing in the brain. *Trends in Cognitive Sciences*, vol. 9, nº 9, pp. 415.

McAfee, A. et Brynjolfsson E. 2008. Investing in the IT that makes a competitive difference. *Harvard Business Review*. Juillet. https://hbr.org/2008/07/investing-in-the-it-that-makes-a-competitive-difference

Morozov, E. 2014. The planning machine : project cybersyn and the origins of the Big Data nation. *The New Yorker.* 13 octobre. https://www.newyorker.com/magazine/2014/10/13/planning-machine

Ng, A., Madhavan, A. et Raina, R. 2009. Large-scale deep unsupervised learning using graphics processors. Proceedings of the 26th International Conference on Machine Learning.

––––. 2015. What data scientists should know about deep learning. Speech presented at Extract Data Conference. 24 novembre. https://www.slideshare.net/ExtractConf/andrew-ng-chief-scientist-at-baidu

––––. 2016. Presentation given at Bay Area Deep Learning School.

––––. 2020a. *Landing AI Transformation Playbook 5*. https://landing.ai/wp-content/uploads/2020/05/LandingAI_Transformation_Playbook_11-19.pdf

––––. 2020b. State of AI. Presentation at AI Fund Update.

Nouwens, M., Liccardi, I., Veale, M., Karger, D. et Kagal, L. 2020. Dark patterns after the GDPR : Scraping consent pop-ups and demonstrating their influence. *CHI '20 : Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. pp. 1-13.

Pentland, A., Nathan, O. et Zyskind, G. 2015. Enigma : Decentralized computation platform with guaranteed privacy. *arXiv :1506.03471.*

Phelps, E. 2015. *Mass Flourishing : How Grassroots Innovation Created Jobs, Challenge and Change*. Princeton, New Jersey, Princeton University Press.

Pueyo, T. 2021. Internet and blockchain will kill nation-states. Blogue Uncharted Territories, 29 août. https://unchartedterritories.tomaspueyo.com/p/internet-blockchain-kill-nation-states

Reber, P. 2010. What is the memory capacity of the human brain ? *Scientific American.* 1er mai. https://www.scientificamerican.com/article/what-is-the-memory-capacity/

Reich, R., Sahami, M. et Weinstein, J. M. 2021. *System Error : Where Big Tech Went Wrong and How We Can Reboot*. New York, Harper Collins.

Rogers, K. 2017. More than 100 million Americans can only get internet service from companies that have violated net neutrality. *Vice.* 11 décembre. https://www.vice.com/en/article/bjdjd4/100-million-americans-only-have-one-isp-option-internet-broadband-net-neutrality

Romer, P. 2019. A tax that could fix big tech. *The New York Times.* 6 mai. https://www.nytimes.com/2019/05/06/opinion/tax-facebook-google.html

Ronneberger, O., Fischer, P. et Brox, T. 2015. U-Net : Convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science Medical Image Computing and Computer-Assisted Intervention – MICCAI* 2015, pp. 234-241.

Russell, S. 2019. *Human Compatible : Artificial Intelligence and the Problem of Control*. New York : Penguin Books.

Schumpeter, J. 1942. *Capitalism, Socialism and Democracy*. University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship. New York : Harper & Brothers.

Sharma, R. 2019. Bitcoin won't win worldwide adoption because China controls it : Ripple CEO. *Investopedia*. 25 juin. https://www.investopedia.com/news/bitcoin-wont-win-worldwide-adoption-because-china-controls-it-ripple-ceo/

Shilov, A. 2021. Intel : Upcoming US fab will be a small city, to cost $60 to $120 billion. *Tom's Hardware*. 6 août. https://www.tomshardware.com/news/intel-to-spend-up-to-120-billion-on-new-us-manufacturing-hub

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A. (...) 2017. Mastering the game of Go without human knowledge. *Nature,* vol. 550, nº 7676, pp. 354-359.

Simon, H. A. 1955. A Behavioral Model of Rational Choice. *Quarterly Journal of Economics,* vol. 69, nº 1, pp. 99-118.

Smith, A. 1776 (2008). *An Inquiry into the Nature and Causes of the Wealth of Nations.* Oxford : Oxford University Press.

Srinivasan, B. 2019. Balaji Srinivasan on the argument for decentralization – Part 1. The Pomp Podcast, episode 295.

Stucke, M. E. et Ezrachi, A. 2017. The rise, fall, and rebirth of the U.S. antitrust movement. *Harvard Business Review*. 15 décembre. https://hbr.org/2017/12/the-rise-fall-and-rebirth-of-the-u-s-antitrust-movement

Schied, B. 2020. Top 5 tech stocks' S&P dominance raises fear of bursting bubble. *S&P Global Market Intelligence.* https://www.spglobal.com/marketintelligence/en/news-insights/latest-news-headlines/top-5-tech-stocks-s-p-500-dominance-raises-fears-of-bursting-bubble-59591523

Talagala, N. 2021. Google built a trillion parameter AI model. 7 things you should know. *Forbes.* 8 juillet. https://www.forbes.com/sites/nishatalagala/2021/07/08/google-built-a-trillion-parameter-ai-model-7-things-you-should-know/?sh=4909de7b7974

Tambe, P., Hitt, L., Rock, D. et Brynjolfsson, E. 2020. Digital capital and superstar firms. National Bureau of Economic Research. NBER Working Paper Series. https://www.nber.org/papers/w28285

von Mises, L. 1951. *Socialism : An Economic and Sociological Analysis.* New Haven, Yale University Press. Ludwig von Mises Institute.

Werin, L. et Wijkander, H. (dir.). 1992. *Journal of Applied Corporate Finance*, vol. 8, nº 2.

Wilson, T. D. 2004. *Strangers to Ourselves : Discovering the Adaptive Unconscious*. Cambridge : Harvard University Press.

Yudkowsky, E. 2016. *The AI alignment problem : Why it's hard, and where to start*. Allocution présenté à l'Université Stanford, 5 mai. Vidéo. https://intelligence.org/stanford-talk/

LES DILEMMES DANS L'ANGLE MORT DU DÉVELOPPEMENT RESPONSABLE DE L'INTELLIGENCE ARTIFICIELLE EN TEMPS DE PANDÉMIE

NATHALIE VOARINO

Chercheuse postdoctorale, Hub Santé – Politique, Organisations et Droit (H-POD), Faculté de droit de l'Université de Montréal, Observatoire international sur les impacts sociétaux de l'IA et du numérique (OBVIA – Fonds de recherche du Québec).

CATHERINE RÉGIS

Professeure titulaire, Université de Montréal, Chaire de recherche du Canada sur la culture collaborative en droit et politiques de la santé, chercheuse à Mila (Institut québécois d'intelligence artificielle) et à l'Observatoire international sur les impacts sociétaux de l'IA et du numérique, co-directrice du Hub Santé – Politique, Organisations et Droit (H-POD).

- ODD 3 - Bonne santé et bien-être
- ODD 8 - Travail décent et croissance économique
- ODD 9 - Industrie, innovation et infrastructure
- ODD 10 - Inégalités réduites
- ODD 11 - Villes et communautés durables
- ODD 13 - Mesures relatives à la lutte contre les changements climatiques
- ODD 14 - Vie aquatique
- ODD 15 - Vie terrestre
- ODD 16 - Paix, justice et institutions efficaces
- ODD 17 - Partenariats pour la réalisation des objectifs

LES DILEMMES DANS L'ANGLE MORT DU DÉVELOPPEMENT RESPONSABLE DE L'INTELLIGENCE ARTIFICIELLE EN TEMPS DE PANDÉMIE

RÉSUMÉ

Se pencher sur les dilemmes de l'éthique de l'IA et sur la manière de concilier les tensions potentielles qui émergent de la mise en œuvre des principes permet d'identifier certains des angles morts de la gouvernance de l'IA. La pandémie actuelle constitue un cas paradigmatique de leur étude, plusieurs pays ayant choisi de s'appuyer sur ces technologies pour contribuer aux efforts de santé publique en cours. Guidées par les dix principes éthiques de la Déclaration de Montréal pour un développement responsable de l'IA, nous présentons dans ce texte deux exemples de dilemmes clés qui émergent de l'usage de l'IA pour contrer la COVID-19 soit : 1) le dilemme entre le principe de protection de la vie privée et de l'intimité et le principe de solidarité ; 2) le dilemme entre le principe d'équité et le principe de développement soutenable. Nous dégageons ensuite quelques pistes de solution, basées sur l'approche des capabilités, en vue d'y répondre. Résoudre ces dilemmes est essentiel afin d'exploiter le plein potentiel des SIA pour lutter contre les pandémies – qu'il s'agisse de celle-ci ou des prochaines, et ce, en vue d'assurer que l'IA bénéficie à la santé de toutes et tous.

INTRODUCTION

Le domaine de l'intelligence artificielle (IA) a connu dans la dernière décennie des avancées majeures provoquées, entre autres, par la sophistication des outils informatiques et le nombre croissant de données disponibles (Cardon, Cointet et Mazières, 2018). Qu'il s'agisse de la santé, des ressources humaines, de l'environnement ou de l'éducation : les bénéfices potentiels de ce nouveau printemps de l'IA n'épargnent aucun secteur de la société. L'implémentation systémique de systèmes d'IA (SIA) de plus en plus autonomes en vue d'automatiser des tâches répétitives jusqu'ici confiées à des humains soulève cependant de nombreux enjeux largement identifiés au cours des dernières années. Pour ne citer qu'eux, il s'agit d'un risque d'atteinte à la vie privée (Stahl et Wright, 2018) ; à la justice sociale

relativement aux biais que pourraient perpétuer les algorithmes (Kim, 2016 ; Risse, 2019) ; à la déshumanisation des activités (ex. Coeckelbergh, 2015) – considérant notamment la diminution de la supervision humaine ; ou encore d'érosion de la responsabilité des individus qui ont recours à l'IA (Noorman, 2016) en raison du manque de transparence des décisions algorithmiques (favorisée par la fameuse « boîte noire » des algorithmes) (Ananny et Crawford, 2018).

En réponse à ces risques et enjeux, de nombreuses initiatives à travers le monde ont défini des principes éthiques directeurs pour un développement responsable de l'IA : 84 documents étaient recensés en 2019 (Jobin, Ienca et Vayena, 2019), 167 ont été recensés en 2020 (AlgorithmWatch, 2020), incluant la Déclaration de Montréal pour un développement responsable de l'IA (2018). Que ces documents aient une portée nationale ou internationale, leurs principes éthiques ont pour vocation de guider la gouvernance de l'IA soit, d'orienter le développement de différents mécanismes comme l'élaboration de politiques publiques, de lois et règlements ou de normes techniques (AI HLEG, 2019).[22]

Si le travail des dernières années concernant l'identification des enjeux et la définition des principes éthiques a été considérable, un écart entre ces principes et leur mise en œuvre pratique reste difficile à surmonter, comme le soulignent plusieurs expertes et experts (Mittelstadt, 2019 ; Morley *et al.*, 2020 ; Hagendorff, 2020 ; Shneiderman, 2020 ; Siau et Wang, 2020 ; Langlois et Régis, 2021). Cet écart est notamment lié au fait que ces principes seraient trop abstraits ou trop vagues, ce qui rend difficile leur interprétation en vue de guider le développement des mécanismes susmentionnés (Mittelstadt, 2019 ; Morley *et al.*, 2020). L'écart est également alimenté par la difficulté de leur hiérarchisation lorsque ces principes s'opposent en dilemmes (Whittlestone *et al.*, 2019 ; Yeung, Howes et Pogrebna, 2020), soit lorsque, dans une situation donnée, il n'est pas possible de suivre un principe sans entraver le respect d'un autre.[23] Si ce niveau d'abstraction est bien le propre des principes éthiques (Massé, 2003),[24] l'apparition de dilemmes lorsqu'il est question de les mettre en œuvre est un enjeu majeur. Comment les principes pourraient-ils guider le développement de politiques ou de normes lorsqu'ils se retrouvent en contradiction ? Cet enjeu pourrait mener à un rejet ou un désintérêt envers ces principes, qui risqueraient de ne plus être considérés comme pertinents pour guider l'action, amenuisant ainsi considérablement leur contribution potentielle. Ce risque de désintérêt envers l'éthique – notamment en raison d'une absence de définition claire d'allocation de la responsabilité – est parfois qualifié d'esquive éthique (« ethics shirking ») soit, le fait de ne plus recourir à des pratiques éthiques car celles-ci sont jugées peu efficaces dans un contexte donné (Floridi, 2019).

Selon Whittlestone *et al.* (2019), l'identification de ces dilemmes est en effet l'une des prochaines étapes essentielles de l'éthique de l'IA en vue d'une gouvernance effective. Outre le fait qu'elle contribuerait à combler l'écart entre les principes et la pratique, cette identification permettrait, selon ces auteures et ces auteurs, de mettre en évidence les situations qui nécessitent la mise en œuvre de nouvelles solutions – quelle qu'en soit la nature – là où les seuls principes éthiques ne sauraient suffire pour guider l'action (Whittlestone *et al.*, 2019). Se pencher sur les dilemmes de l'éthique de l'IA et sur la manière de concilier les tensions potentielles qui émergent de la mise en œuvre des principes permettrait alors d'identifier certains des angles morts de la gouvernance de l'IA.

---

22. Selon ce rapport du Groupe d'experts de haut niveau sur l'intelligence artificielle de la Commission Européenne, il existe différentes manières de mettre en œuvre les principes éthiques de l'IA : les méthodes techniques (par exemple, les procédures incluses dans les architectures d'algorithmes) et non-techniques (par exemple, les mécanismes juridiques) – nous traitons ici essentiellement de la deuxième catégorie.

23. Un dilemme éthique apparait lorsque l'application d'un principe (ou la poursuite d'une valeur) empêche la mise en œuvre d'un autre principe (ou d'une autre valeur) – et qu'aucun des principes en conflit ne s'impose de lui-même car de « bons » arguments existent en faveur des deux alternatives (Durand, 2007).

24. En effet, « les principes ont pour vocation d'être suffisamment abstraits pour permettre leur pérennité et la flexibilité de leur interprétation en vue d'une large appropriation (voire d'une appropriation universelle) » (Voarino, 2020 p. 182).

Si ces dilemmes s'observent lors de l'opérationnalisation des principes dans des situations concrètes, la pandémie de COVID-19 constitue un cas paradigmatique de leur étude. En effet, parce qu'elles demandent d'agir dans l'urgence et souvent en situation d'incertitude, les crises limitent le temps et les données probantes disponibles pour évaluer les risques associés à de nouveaux usages (Tzachor *et al.*, 2020 ; Cave *et al.*, 2021). Également, l'ampleur du déploiement (local ou international) des solutions (technologiques ou autres) accroît l'impact des conséquences néfastes inattendues (Tzachor *et al.*, 2020 ; Cave *et al.*, 2021) tout comme celui des bénéfices attendus.

La pandémie peut donc agir comme un puissant révélateur de dilemmes éthiques, notamment ceux qui naissent de l'utilisation de l'IA. Plusieurs SIA ont en effet été identifiés et déployés pour contribuer aux efforts de santé publique en cours – qu'ils interviennent à l'échelle moléculaire (par exemple, avec l'optimisation du développement de vaccins), clinique (par exemple, en support au diagnostic) ou sociétale (par exemple, en modélisant l'épidémie) (Bullock *et al.*, 2020) – offrant ainsi plusieurs perspectives prometteuses dans la lutte contre la propagation du virus. Cependant, leur usage a soulevé de nombreuses préoccupations relatives, notamment, à la protection des données personnelles, au respect du consentement et de l'autonomie des citoyennes et citoyens ou à l'atteinte de différentes libertés individuelles et de différents droits humains fondamentaux (Gasser *et al.*, 2020 ; Naudé, 2020 ; Cave *et al.*, 2021 ; von Struensee, 2021). Pour guider les réponses à ces préoccupations, il était donc particulièrement à propos de se référer aux principes éthiques qui ont fait l'objet de tant d'attention avant le début de la pandémie. Cependant, plusieurs dilemmes importants apparaissent lorsqu'on se réfère à ces principes pour guider le développement responsable de l'IA dans le contexte de la lutte contre la propagation de la maladie à COVID-19.

Guidées par les dix principes éthiques de la Déclaration de Montréal (2018), nous présentons dans ce texte deux exemples de dilemmes clés qui émergent de l'usage de l'IA pour contrer l'actuelle pandémie. La Déclaration de Montréal, basée sur un processus de coconstruction ayant impliqué plus de 500 citoyens, a reçu une attention scientifique et internationale (Else, 2018 ; Fjeld *et al.*, 2020) et a été identifiée comme un outil important du développement responsable en IA (The Future Society, 2020). Cet exercice permet de mettre en évidence certains des angles morts de l'éthique de l'IA, issus notamment de la tendance de certains principes à en éclipser d'autres. Nous dégageons ensuite quelques pistes de solution en vue d'y répondre. Nous pensons que résoudre ces dilemmes est essentiel afin d'exploiter le plein potentiel des SIA pour lutter contre les pandémies – qu'il s'agisse de celle-ci ou des prochaines, et ce, en vue d'assurer que l'IA bénéficie à toutes et tous.

DEUX EXEMPLES DE DILEMMES CLÉS LIÉS À L'UTILISATION DE L'IA POUR CONTRER LA PANDÉMIE DE COVID-19

La solidarité dans l'ombre de la protection de la vie privée

Un premier dilemme mis en lumière lors de l'utilisation de l'IA pour contrer la pandémie de COVID-19 est celui qui oppose la protection de la vie privée à la solidarité, particulièrement débattue dans le contexte du partage de données en vue de mettre en place des mécanismes de surveillance de santé publique – par exemple, via des applications de traçage.[25]

25. Ces applications peuvent concerner le traçage de contact ou le traçage de position, lesquelles permettent d'identifier les utilisateurs qui représentent un risque de contagion (respectivement, en établissant un historique de contacts ou en pistant la localisation de personnes testées positives) (Mondin et de Marcellis-Warin, 2020). Ce type d'application n'est pas toujours soutenu par des systèmes d'IA et d'autres systèmes d'IA sont susceptibles de contribuer à la surveillance de la propagation du virus.

Le risque d'entrave à la vie privée est, indépendamment de l'actuelle pandémie, l'une des préoccupations majeures de l'avènement de l'IA en santé (Christen *et al.*, 2016 ; Azencott, 2018 ; Iyengar, Kundu et Pallis, 2018 ; Hager *et al.*, 2019) et un des enjeux les plus discuté de la littérature des domaines apparentés tels que celui des données massives (voir par exemple Mittelstadt et Floridi, 2016 ou Stahl et Wright, 2018). Faisant échos à un droit humain fondamental – présent, par exemple, dans l'article 12 de la *Déclaration universelle des droits de l'Homme* et le Règlement général sur la protection des données (RGPD) européen – le principe 3 de la Déclaration de Montréal de *Protection de l'intimité et de la vie privée* invite, au-delà de la simple garantie de protection de la confidentialité et d'anonymisation des données personnelles, à protéger « des espaces d'intimité dans lesquels les personnes ne sont pas soumises à une surveillance » ; et stipule que « toute personne doit pouvoir garder un contrôle étendu sur ses données personnelles, en particulier par rapport à leur collecte, usage et dissémination ».

Si ce principe a pu être mis à mal lors de l'actuelle pandémie, c'est que l'usage de l'IA et, ultimement, ses performances, sont hautement dépendants de l'accès aux données des individus (Bullock *et al.*, 2020).[26] Souvent, les données qui font l'objet de discussions et préoccupations relativement à la protection de la vie privée sont celles collectées en dehors du système de santé, comme celles provenant d'Internet, générées sur les médias sociaux ou issues des téléphones intelligents (Mittelstadt et Floridi, 2016 ; Ienca et Vayena, 2020 ; Scassa, Millar et Bronson, 2020 ; Kassab et Graciano Neto, 2021). Leur collecte fait en effet entrer la surveillance dans *des espaces d'intimité*, ce qui avait déjà été soulevé par plusieurs experts avant la pandémie alors que la portabilité des systèmes a introduit une collecte de données de santé horizontale et ubiquitaire, sortant des espaces traditionnels de soins, et potentiellement intrusive même quand les données sont anonymes ou de nature peu sensible (Mittelstadt et Floridi, 2016 ; IEEE, 2017 ; Villani, 2018). De plus, ce type de collecte *limite le contrôle* possible des citoyennes et citoyens sur leurs données, en particulier vis-à-vis de ce qui en est fait lors de leur réutilisation, qui devient alors quasiment infinie (Christen *et al.*, 2016 ; Rial-Sebbag, 2017). Sur ce point, l'entrave a été plus ou moins importante selon les pays, notamment en fonction de la nature obligatoire (ou non) du recours à ces SIA ou du recours à la collecte de ces données, mais également en fonction du niveau de transparence des autorités ou des outils numériques quant aux fins de leur utilisation (Mondin et de Marcellis-Warin, 2020). Cette surveillance risque alors également d'entraver le principe 2 de *Respect de l'autonomie* – lequel stipule notamment que ne soient pas « développés ni utilisés pour prescrire aux individus un mode de vie particulier, soit directement, soit indirectement en mettant en œuvre des mécanismes de surveillance, d'évaluation ou d'incitation contraignants » ; et que « les institutions publiques ne doivent pas utiliser les SIA pour promouvoir ni défavoriser une conception de la vie bonne ».

Selon Mello et Wang (2020), s'il n'est pas nouveau de recourir à ce genre de données pour la surveillance de maladies, de nombreux pays ont « fait passer l'épidémiologie numérique à un niveau supérieur en répondant à la COVID-19 »[27] (p. 951, traduction libre) avec la collection à grande échelle de données de millions d'utilisateurs (Ienca et Vayena, 2020). Ce phénomène a soulevé des préoccupations relatives à la vie privée dans plusieurs pays du monde comme, par exemple, au Canada (CEST, 2020) ; en Chine (Ienca et Vayena, 2020 ; Mello et Wang, 2020 ; Shachar, Gerke et Adashi, 2020) ; aux États-Unis (Shachar, Gerke et Adashi, 2020) ou encore au Zimbabwe (Mbunge *et al.*, 2021).

Un risque de glissement vers un excès de traçage et de surveillance a notamment été soulevé (voir par exemple : Scassa, Millar et Bronson, 2020 ; Mbunge *et al.*, 2021 ; Tran et Nguyen, 2021 ; CEST, 2020) alors que le principe 3 demande de limiter l'intrusion potentielle des SIA dans la vie des individus quand

26. Plusieurs ensembles de données utiles pour l'analyse de système d'IA en vue de contrer la propagation du COVID-19 ont été recensés par Bullock *et al.* (2020) dans leur revue de littérature, notamment des données relatives au nombre de cas ou à leur localisation.

27. 'Several countries have taken digital epidemiology to the next level in responding to COVID-19' (Mello et Wang, 2020, p.951).

ces systèmes sont susceptibles de « faire du tort » lors d'usages visant « à juger moralement des personnes ou de leur choix de vie » (Déclaration de Montréal, 2018). Cet aspect dudit principe semble peu compatible avec l'utilisation de SIA pour observer l'adhésion aux mesures de santé publique, comme ce fut le cas, selon Mello et Wang (2020), en Chine, en Pologne ou en Russie.

Cependant, de tels SIA pourraient contribuer à limiter la propagation du virus en identifiant l'apparition de futurs foyers de contamination (Vaishya *et al.*, 2020) ou en aidant à mieux comprendre la propagation virale (Alimadadi *et al.*, 2020) et ainsi rendre plus efficace la mise en place des mesures de santé publique. Ceci pourrait ainsi accélérer la fin de mesures liberticides comme le confinement (Shachar, Gerke et Adashi, 2020) ou la fin de l'accès restreint à l'éducation, aux activités économiques et culturelles. Ne pas avoir recours à ces SIA en raison de la protection de la vie privée risquerait alors de nuire au principe 4 de *Solidarité* de la Déclaration de Montréal, selon lequel le développement de l'IA doit « être compatible avec le maintien de liens de solidarité entre les personnes et les générations » et « améliorer la gestion des risques et créer les conditions d'une société de mutualisation des risques individuels et collectifs plus efficace ». Dans le cas de la présente pandémie, toutes mesures efficaces visant à limiter la propagation du virus contribuent en effet à la *solidarité entre les générations* (par exemple, envers les personnes âgées ayant été particulièrement touchées – Jackman, 2020 ; Lagacé, Garcia et Bélanger-Hardy, 2020) ou entre différents groupes (par exemple, envers les travailleuses et travailleurs essentiels qui n'ont pu être assignés au télétravail). La *mutualisation des risques* semble de son côté encourager le partage de données (qu'elles soient personnelles ou non) en vue des bénéfices collectifs susmentionnés (soit, d'améliorer la santé de toutes et tous), alors que Naudé (2020) a identifié les considérations relatives à la vie privée comme un des obstacles à l'efficacité des SIA utilisés pour contrer la pandémie.

Cette tension entre les principes de protection de la vie privée et de solidarité a été soulevé qu'il soit question de la gestion de la pandémie de COVID-19 en général (ex. en Colombie, voir de la Espriella, Llanos et Hernandez, 2021) ou concernant le recours aux applications de traçage en particulier (voir Kudina, 2021). Cependant, cette tension était déjà marquée lorsqu'il était question d'IA dans le domaine de la santé avant la pandémie actuelle. En effet, certains défendaient que la protection de la vie privée est dépassée à l'heure où le partage de données (personnelles) sur les réseaux sociaux est omniprésent (Spiekermann, Korunovska et Langheinrich, 2018) et d'autres soulevaient que de telles entorses à la vie privée étaient justifiées dans des situations de crises (O'Doherty *et al.*, 2016 ; Fiore et Goodman, 2016). Dans un contexte de santé publique, le partage de données est alors pour certains un devoir moral qui justifie une entrave à la vie privée considérant les bénéfices pour le bien commun (Fiore et Goodman, 2016 ; Hand, 2018 ; Mello et Wang, 2020). Terry et Coughlin (2021) proposent même une « recalibration » de la protection de la vie privée sur la base de considérations solidaires, lesquelles ont pu être observées dans le contexte de la pandémie de COVID-19.

Ainsi, le respect du principe 3 de *Protection de l'intimité et de la vie privée*, dont l'importance éthique n'est plus à démontrer, entraverait le respect du principe 4 de *Solidarité*, qui invite au partage des données (personnelles) du plus grand nombre en vue de potentialiser les bénéfices des SIA pour la santé de toutes et tous.

Le développement soutenable dans l'ombre de l'équité

De l'usage de l'IA dans un contexte de santé mondiale naît également une tension entre le devoir moral d'assurer un accès à toutes et tous aux technologies qui supportent des SIA (et aux bénéfices sanitaires de leur usage) selon le principe 6 d'*Équité* tout en limitant l'impact environnemental de ces SIA suivant le principe 10 de *Développement soutenable*.

Selon la Déclaration de Montréal, le principe 6 d'*Équité* requiert que « le développement et l'utilisation des SIA doivent contribuer à la réalisation d'une société juste et équitable ». Ceci implique notamment

que les SIA bénéficient « économiquement et socialement à tous en faisant en sorte qu'ils réduisent les inégalités et la précarité sociales » ; que « l'accès aux ressources, aux savoirs et aux outils numériques fondamentaux » soit « garanti pour tous » et de soutenir « le développement de communs algorithmiques et de données ouvertes pour les entraîner ».

Dans un contexte de santé mondiale, l'IA a été identifiée (dans la veine de la santé numérique) comme un outil particulièrement prometteur en vue d'atteindre la couverture universelle de santé (Global observatory for eHealth, 2015, 2016 ; OMS, 2018), faisant échos aux impératifs d'équité susmentionnés. Dès lors, il s'agirait d'équiper les populations et les groupes (exclus ou marginalisés) qui ont peu ou pas accès aux technologies et infrastructures en outils capables de supporter des SIA pour leur assurer un meilleur accès aux soins et aux services de santé. Selon l'OMS (2021a) cela suppose en effet d'assurer l'accès à des ressources techniques et humaines ainsi qu'aux infrastructures nécessaires telles que l'électrification, la connexion à internet, les réseaux et dispositifs sans fil ou mobiles. Cet objectif s'inscrit dans un chantier plus large de la scène internationale qui vise à dépasser la « fracture numérique » définie par l'Organisation mondiale de la Santé (OMS) comme « la répartition inégale de l'accès aux technologies de l'information et de la communication, de leur utilisation ou de leurs effets entre un certain nombre de groupes distincts » (OMS, 2021.a, p. 34, traduction libre).[28] En effet, tel que le recommandait le Groupe de haut niveau du Secrétaire général des Nations Unies sur la coopération numérique :

> **D'ici à 2030, chaque adulte devrait avoir un accès abordable aux réseaux numériques, ainsi qu'à des services financiers et de santé accessibles par voie numérique, afin de contribuer de manière substantielle à la réalisation des objectifs de développement durable (Nations Unies, 2019 dans OMS, 2021.a, p. 34, traduction libre).**[29]

Cette fracture numérique peut aussi bien s'observer entre différents pays du monde (Makri, 2019) qu'entre différents groupes au sein d'une même société. Si l'enjeu de fracture numérique existe depuis près d'un quart de siècle,[30] ses effets ont été exacerbés lors de la pandémie de COVID-19, alors que le recours au numérique s'est davantage généralisé en santé comme dans d'autres secteurs (Davis, 2020 ; Ramsetty et Adams, 2020). En effet, les téléconsultations ont par exemple été favorisées sur les consultations en présentiel pour limiter la propagation du virus. Dans ce contexte, de nombreuses personnes n'ayant pas accès aux technologies et infrastructures numériques (et, *a fortiori*, aux technologies et infrastructures susceptibles de supporter des SIA) ont été mises à l'écart des solutions de santé offertes, qu'il s'agisse de personnes âgées (Martins Van Jaarsveld, 2020), ruralisées (Lai et Widmar, 2021) ou aux revenus limités (News, 2020).

Répondre de cette fracture numérique supposerait ainsi d'outiller une partie non négligeable de la population mondiale (pour ne pas dire l'ensemble dans un idéal éthique) et s'accompagnerait inexorablement d'un nombre plus important de technologies et d'infrastructures essentielles au déploiement de SIA et à l'entrainement des algorithmes. Parce que ces derniers sont dépendants du nombre de données disponibles, ce « virage mondial vers les technologies numériques en santé » (« global shift toward new digital technologies in health ») (pour reprendre les termes de Davis, 2020) risque également de s'accompagner d'un accroissement des données générées, collectées, stockées

28. 'The uneven distribution of access to, use of or effect of information and communication technologies among any number of distinct groups' (OMS, 2021.a, p. 34).

29. 'By 2030, every adult should have affordable access to digital networks, as well as digitally enabled financial and health services, as a means to make a substantial contribution to achieving the Sustainable Development Goals' (the United Nations Secretary-General's High-level Panel on Digital Cooperation in OMS, 2021.a, p. 34)

30. Le terme « *digital divide* » (traduit ici par « fracture numérique ») a été utilisé pour la première fois en 1995 aux États-Unis (Dijk, 2020).

et analysées. Ce serait par exemple le cas avec la création de très larges ensembles de données spécifiques aux pandémies, tel que le projet de Centre de renseignements sur les pandémies et les épidémies de l'OMS (« WHO Hub for Pandemic et Epidemic Intelligence ») (OMS, 2021b). Cette initiative devrait prendre la forme d'une plateforme mondiale de collecte et d'analyse de données qui pourraient être utile à la prévention et à la gestion de futures pandémies, avec notamment pour objectif de dépasser les restrictions étatiques relatives à la confidentialité et à la protection de la vie privée en vue d'assurer un partage pertinent et efficaces des données pour le bien commun (OMS, 2021b).

Or, les technologies numériques ne sont pas neutres d'un point de vue environnemental. Outre le niveau important de déchets électroniques qui accompagne l'innovation numérique (Dwivedi et al., 2021), le fonctionnement des centres de données, ainsi que la production d'ordinateurs et de téléphones intelligents, consomment une quantité importante d'énergie et pourraient contribuer de manière non négligeable au réchauffement climatique (Gmach *et al.*, 2010 ; The Shift Project, 2020). C'est également le cas de l'entrainement de modèles d'IA qui s'accompagne de plus en plus d'émission de gaz à effet de serre (GES) (Ligozat et al., 2021). Pour d'autres, si la numérisation des activités est parfois considérée comme une solution pour réduire ces émissions de GES (Patsavellas et Salonitis, 2019 ; Ghobakhloo, 2020 ; IEA, 2021), il est reconnu que cette transition énergétique requiert beaucoup de minéraux critiques et terres rares (Commission Européenne, 2020 ; Hund *et al.*, 2020 ; IEA, 2021). Cette transition énergétique s'accompagne alors d'autres conséquences dommageables pour l'environnement. Les téléphones intelligents, comme d'autres dispositifs informatiques supportant des SIA, nécessitent ce type de minéraux (notamment le lithium, utile par exemple au développement de batteries), dont le forage connaît des conséquences désastreuses pour les écosystèmes (Crawford, 2021 ; IEA, 2021). Comme le présente le rapport de 2020 de l'Agence Internationale de l'Énergie, le forage de ce type de matériaux : 1) peut impacter la biodiversité et provoquer une perte d'habitats pour les espèces animales (notamment, pour celles en voie de disparition) ; 2) requiert de larges volumes d'eau (ce qui est peu soutenable dans un contexte de pénurie) ; 3) peut engendrer une contamination acide des eaux usées, et ; 4) génère des déchets dangereux qui pourraient augmenter avec la diminution de la qualité des minéraux (IEA, 2021).

Respecter le principe 3 d'*Équité* pourrait alors entraver le respect du principe 10 de *Développement soutenable* de la Déclaration de Montréal, qui requiert notamment que le développement et l'utilisation de SIA soient réalisés « de manière à assurer une soutenabilité écologique forte de la planète ». Ceci implique, entre autres, de « minimiser les émissions de gaz à effet de serre » ; « viser à générer un minimum de déchets électriques et électroniques » et de « minimiser les impacts sur les écosystèmes et la biodiversité ». Bien que plusieurs pistes de solution se dégagent pour limiter les conséquences environnementales des technologies numériques (The Shift Project, 2020 ; IEA, 2021), plusieurs experts questionnent si la transition numérique et la transition écologique sont compatibles[31] et dans quelles mesures ces solutions sont suffisantes et efficaces à court terme, considérant l'urgence d'agir en matière climatique (GIEC, 2021). Cette tension est notamment ressortie de la conférence des Nations Unies sur les changements climatiques (COP26) de 2021, lors de laquelle plusieurs expertes et experts ont questionnés dans quelle mesure les technologies numériques permettent de contribuer à la réponse aux changements climatiques ou bien font partie intégrante du problème (Dwivedi et al., 2021).

Ce dilemme est d'autant plus important dans un contexte de santé mondiale, alors que respecter le développement soutenable est directement lié à la santé des populations (Patz *et al.*, 2014 ; Solomon et LaRocque, 2019). La dégradation de l'environnement, de la biodiversité ainsi que le réchauffement climatique pourraient notamment favoriser l'apparition de nouvelles pandémies (Mackenzie et Jeggo, 2019 ; Solomon et LaRocque, 2019 ; Charlier *et al.*, 2020 ; Hébert, 2021). Ces préoccupations sont au cœur des principes de Manhattan, développés en 2004 lors d'un symposium regroupant des experts internationaux en vue de réfléchir, entre autres, à la prévention de l'apparition de maladies infectieuses

31. Voir par exemple les publications sur le sujet du projet « Chemins de transition » : https://cheminsdetransition.org/numerique/

telles que les zoonoses (Principes de Manhattan, 2004). Ces principes défendent une approche globale liant préoccupations environnementales et sanitaires, dirigée vers « un monde, une santé » (« One World, One Health ») (Principes de Manhattan, 2004). L'importance de ces enjeux a conduit plusieurs expertes et experts internationaux à écrire une « lettre ouverte à l'OMS » (voir Charlier *et al.*, 2020), faisant le point sur les conséquences sanitaires du réchauffement climatique (incluant le risque de pandémie) et incitant les organisations (internationales) à se pencher sur ce problème.

Équité et durabilité semblent alors difficilement conciliables, en particulier dans un contexte de santé mondiale. Si cette conciliation est en partie le but des Objectifs de développements durables, ces derniers perpétuent ce dilemme en indicateurs potentiellement contradictoires lorsqu'il est question du numérique (par exemple, comment concilier l'indicateur 5.b.1. « proportion de la population possédant un téléphone portable, par sexe » avec la cible 12.2. « d'ici à 2030, parvenir à une gestion durable et à une utilisation rationnelle des ressources naturelles ») (Nations Unies, 2021). Le respect du principe 10 de *Développement soutenable* se heurte alors à l'objectif (plus que louable) de dépasser la fracture numérique selon le principe 6 d'*Équité*.

DÉPASSER LES DILEMMES ÉTHIQUES POUR ÉCLAIRER LES ANGLES MORTS AU TRAVERS DU PRISME DE L'APPROCHE DES CAPABILITÉS

La résolution des dilemmes survenus lors de l'usage de SIA pour contrer la pandémie de COVID-19 est particulièrement pertinente pour renseigner la gouvernance éthique de la santé mondiale. Ce sujet est par ailleurs négligé par la communauté de recherche sur l'éthique de l'IA (Murphy *et al.*, 2021). Ces dilemmes s'inscrivent, entre autres, dans la mire d'un dilemme classique de ce champ d'action, à savoir : « comment trouver le juste équilibre entre les besoins de tous quand ceux-ci s'opposent à la protection des droits individuels » (Stapleton *et al.*, 2014, p. 4, traduction libre)[32] ou, en d'autres termes, comment concilier la santé de chacune et chacun avec celle de la collectivité. La frontière entre les deux n'est pas toujours étanche, la protection des droits individuels pouvant évidemment contribuer à l'atteinte d'objectifs collectifs. Cette précision faite, dans les exemples de dilemmes ici présentés, on retrouve en effet des principes dont les dimensions individuelles font échos à des droits fondamentaux (soit, le principe 3 de *Protection de l'intimité et de la vie privée* ou le principe 6 d'*Équité*) qui entre en conflit avec des principes guidés par des objectifs qui relèvent de considérations plus collectives (soit, le principe 4 de *Solidarité* et le principe 10 de *Développement soutenable*).

32. 'How to balance the needs of 'the many' against the rights of 'the individual' ' (Stapleton *et al.*, 2014, p. 4).

En santé publique, il s'agit d'un dilemme récurrent qui oppose une « éthique individualiste » (*individualistic ethics*) nourrie par des traditions d'autonomie et de droits individuels avec une éthique plus collective, basée sur le bien commun et la solidarité (Kenny, Sherwin et Baylis, 2010).[33] Cette tension entre les intérêts individuels et collectifs est à la base des enjeux éthiques qui surviennent des usages de l'IA en santé (Voarino, 2020). Elle a été accentuée dans le contexte de la pandémie de COVID-19 (Anet *et al.,* 2020) tel que le mentionne Biggeri (2020, p. 277, traduction libre) :

> Nous avons été prêts à renoncer à nos libertés (individuelles) de mouvement et d'association afin de préserver la santé et la longévité des plus vulnérables. Nous réalisons que la gouvernance et les systèmes de santé publique doivent prêter une attention bien plus grande au bien-être individuel et collectif.[34]

Résoudre les dilemmes susmentionnés demande en partie de se pencher sur le juste équilibre entre les dimensions individuelles et collectives des préoccupations relatives à l'utilisation de l'IA pour contrer la pandémie.[35] Pour réfléchir à l'atteinte de cet équilibre, nous pensons ici que l'approche des capabilités est une piste intéressante. L'approche des capabilités est issue des travaux d'Amartya Sen qui a remis en question les indicateurs économiques traditionnels pour évaluer le développement humain (Sen, 1983). Selon cette approche, l'évaluation du développement ne se mesure pas en termes de possession de ressources ou de revenus, mais plutôt vis-à-vis de ce que les individus sont réellement capables de faire et d'être, soit en termes de capabilités (Oosterlaken, 2015, résumant plusieurs études de Sen et Nussbaum). Cette approche est depuis très souvent utilisée en ce qui a trait au développement, notamment par des organisations internationales comme, par exemple, par le Programme de développement des Nations Unies de 2020 (UNDP, 2020). L'approche des capabilités est particulièrement pertinente en ce qui a trait à l'évaluation des technologies. C'est notamment ce que soutient le mouvement des technologies appropriées (pour *Appropriate Technology Movement* – ATM) (Oosterlaken, 2015). Sur la base de l'approche des capabilités, l'ATM est dirigé par la question fondamentale suivante en ce qui concerne l'évaluation des technologies : « Ces initiatives donnent-elles vraiment aux individus – dans toute leur diversité humaine – les moyens de mener les vies qu'ils jugent bonnes ? » (Oosterlaken, 2015, p. 41, traduction libre).[36] En d'autres termes, selon l'ATM, un développement technologique approprié devrait assurer l'expansion des capabilités humaines (Oosterlaken, 2015).

33. Ce dilemme éthique a par exemple été très largement discuté dans le contexte de l'obligation vaccinale, indépendamment de l'actuelle pandémie (voir par exemple : Krantz, Sachs and Nilstun, 2004 ; Dawson, 2015 ; Boas, Rosenthal and Davidovitch, 2016 ; Sim, 2017).

34. 'We have been willing to renounce (individual) freedom of movement and association to preserve the health and longevity of the most vulnerable. We realise that public health systems and governance need to pay far greater attention to collective and individual well-being' (Biggeri, 2020, p.277).

35. Ceci demande d'aller au-delà de la simple hiérarchisation des principes (soit, favoriser un principe sur un autre) comme cela est cependant possible dans certaines situations, comme le mentionne la Déclaration de Montréal (2018) : les principes ne sont pas hiérarchisés bien qu'il soit possible, selon les contextes, d'accorder plus de poids à un principe qu'à un autre tant que « l'interprétation qui en est faite est cohérente » (Déclaration de Montréal, 2018).

36. 'Do such initiatives truly empower people – in all their human diversity – to lead the lives they have reason to value ?' (Oosterlaken, 2015, p. 41).

Si cette approche est pertinente, c'est qu'elle permet, d'abord, de se pencher sur les enjeux relatifs à la gestion de la pandémie alors que celle-ci a introduit une perte importante de capabilités vis-à-vis de nombreux aspects de la vie (Anand *et al.*, 2020 ; Biggeri, 2020). En effet, selon Anand et al., (2020) les capabilités de base comme la santé, l'éducation, la nutrition et le lien social ont été compromises lors de la pandémie de COVID-19. Qu'il s'agisse de choix individuels ou de décisions gouvernementales : « plusieurs populations ont dû renoncer temporairement à certaines libertés pour protéger d'autres libertés qu'elles jugeaient bonnes » (Anand et al., 2020, p. 294, traduction libre).[37]

L'approche des capabilités permet ensuite d'embrasser et de dépasser l'opposition binaire les dimensions individuelles et collectives des dilemmes présentés. Si l'approche des capabilités a cependant également parfois été critiquée pour son emphase individualiste, plusieurs défendent qu'elle permet de considérer le bien-être social comme une production organisée de bien-être collectifs (Doucin, 2009) ou comme une responsabilité collective envers des libertés individuelles (Fusulier and Sirna, 2010). Une entrave à l'équité ou à la vie privée, dans un contexte de santé mondiale, pourrait également nuire au bien-être collectif des populations et ainsi opposer des dimensions collectives entre elles. Nous pensons que l'approche des capabilités permet de dépasser une approche distributive de résolution des tensions – laquelle vise à résoudre l'opposition d'idées par le choix d'une solution proportionnelle au rapport de force ou au mérite – et de réfléchir aux dilemmes selon une approche intégrative – laquelle vise à définir une norme commune d'arbitrage rassembleuse qui crée une valeur additionnelle pour les deux idées en tension au départ. L'approche des capabilités permet d'identifier cette norme commune comme étant celle de l'accroissement des capabilités humaines.

Selon l'approche des capabilités, il est nécessaire de considérer au moins deux dimensions pour assurer que des ressources (ici, les SIA) soient converties par les individus en « fonctionnements » (functionings) effectifs (soit, ce qui est effectivement réalisé ou atteint par les individus) : il faut, d'une part, que ces ressources introduisent des possibilités supplémentaires réelles et, d'autre part, que les individus soit libres d'y avoir recours et choisissent de le faire (Bonvin et Farvaque, 2007 ; Fusulier et Sirna, 2010).

Concernant la première dimension, soit les possibilités supplémentaires réelles introduites par les SIA, il est nécessaire de mentionner que plusieurs expertes et experts ont souligné que peu des SIA développés pour contrer la propagation de la maladie à COVID-19 n'ont été réellement efficaces (Naudé, 2020 ; Wynants *et al.*, 2020 ; Douglas Heaven, 2021). Leur potentiel s'est vu limité (selon le type de SIA en jeu) par différents facteurs comme : le manque de données, des données de mauvaise qualité (peu opportunes ou insuffisamment robustes), des modèles présentant un risque élevé de biais, une incapacité à être utilisé par des non-spécialistes ou dans des environnements limités en ressources, ou encore des limitations éthiques et juridiques (Chen et See, 2020 ; Naudé, 2020 ; Wynants *et al.*, 2020). La majorité des SIA utilisés pour contrer la pandémie n'étaient qu'à un jeune stade de développement, pas suffisamment matures pour une utilisation dans des contextes « réels » – en particulier, dans le contexte clinique – limitant leur portée et la possibilité de généraliser leurs usages (Gunasekeran et al., 2021 ; Hashiguchi et al., 2022 ; Bullock et al., 2020), amenant l'OMS à considérer que l'impact réel des SIA pour lutter contre la pandémie de COVID-19 a pour l'instant été « modeste » (OMS, 2021a).

Dans le contexte des dilemmes présentés, ceci demande alors de se pencher avant toutes choses sur les moyens de dépasser les limites à l'efficacité des SIA, sans quoi ces derniers ne permettraient pas ou peu d'atteindre une plus-value solidaire ou équitable effective – laissant ainsi les considérations relatives à l'entrave à la vie privée et au développement soutenable justifier une éventuelle restriction des usages. L'ATM présupposant que toutes les technologies ne représentent pas un progrès en elles-mêmes

37. 'Many populations have had to give up certain freedoms temporarily to protect other freedoms that they have reason to value' (Anand et al., 2020, p. 294).

(Oosterlaken, 2015), il est également important de tenir compte de l'existence d'une hyperbole autour du développement des SIA (Gibert, 2019), qui pourrait conduire à en surestimer les bénéfices, et d'éventuellement considérer d'autres alternatives si leur usage s'avère prématuré.

Concernant la deuxième dimension, l'évaluation de la transformation de capabilités réelles en fonctionnements effectifs demandent d'identifier les choix qu'ont effectivement faits les individus – ainsi que les valeurs et les préférences qui les motivent. Selon l'approche des capabilités (utilisée dans le contexte de l'ATM) : des ressources (des technologies) se transforment en capabilités ou libertés réelles par le biais de l'existence de « facteurs de conversion » (« conversion factors ») soit, des conditions préalables essentielles au développement des capabilités, que ces conditions soient environnementales, sociales ou culturelles (Bonvin et Farvaque, 2007 ; Oosterlaken, 2015). Ces capabilités ou libertés réelles se transforment en fonctionnements effectifs lorsque les individus choisissent d'y avoir recours (notamment selon leurs préférences, une fois que la possibilité existe) (Bonvin et Farvaque, 2007 ; Oosterlaken, 2015). Ceci demande, entre autres, d'identifier les attentes et les craintes citoyennes vis-à-vis de l'usage des différents SIA développés pour lutter contre la pandémie, mais également d'évaluer le recours réel à ces SIA une fois la possibilité d'usage introduite, et les raisons d'une faible adoption. Ceci est particulièrement pertinent dans le contexte de l'utilisation de l'IA en santé, alors que plusieurs facteurs affectant possiblement la confiance des professionnels de santé en ces dispositifs ont été identifiés (impactant directement leur appropriation et utilisation en milieu clinique) (Asan, Alparslan et Avishek, 2020). Également, des enquêtes européennes ont démontré que tous les citoyennes et citoyens n'étaient pas prêts à utiliser une application de traçage de contact en raison de préoccupations en matière de confidentialité et de sécurité et un certain scepticisme vis-à-vis de leur efficacité (Craglia et al., 2020). Plusieurs des pays ayant eu recours à ce genre d'application sur une base volontaire ont par ailleurs observés un faible taux d'adoption (par exemple, 16 % de la population de Singapour et 4 % de la population australienne en avril 2021 – Akinbi, Forshaw et Blinkhorn, 2021). L'exemple des applications de traçage, même si elles ne relèvent pas toutes de l'IA, est particulièrement à propos, car leur efficacité est très dépendante de la propension des citoyennes et citoyens à l'utiliser (il est estimé que 50 % à 70 % de la population doit y avoir recours pour qu'elle soit efficace) (Akinbi, Forshaw et Blinkhorn, 2021). Si le nombre de personnes qui installent l'application est un indicateur, il ne saurait être suffisant. Par exemple, seulement 14 notifications avaient été gérées par l'application française en août 2020 avec 1,9 millions de téléchargement (Akinbi, Forshaw et Blinkhorn, 2021).

Cette seconde dimension de l'approche des capabilités invite cependant à considérer quelques pistes vis-à-vis de la résolution des dilemmes, notamment concernant ce qui peut pousser les individus à choisir ou non d'avoir recours aux SIA. Puisque l'ATM (suivant l'approche des capabilités) accorde une importance particulière à la diversité des individus, elle demande de faire participer les populations concernées dans le développement de solutions technologiques (Oosterlaken, 2015), et ce, en vue d'embrasser la diversité humaine et donc la diversité des préférences. Comme le reconnait Doucin (2009) :

> **Développer les capabilités, ce n'est pas seulement faire de la formation [...] c'est partir d'un dialogue [...] avec les populations identifiées, en s'adressant à des groupes, mais en veillant à ce qu'ils n'écrasent pas les individus, pour ensuite construire avec elles les outils d'une politique (p. 447).**

Ceci est particulièrement à propos vis-à-vis de l'empiètement potentiel sur la vie privée au nom de considérations solidaires, alors que le type de données collectées, les fins pour lesquelles elles le sont et les acteurs qui y auront accès ont changés lors de l'usage de SIA pour contrer la pandémie : ceci suppose une forme de renégociation du contrat social vis-à-vis des données de santé. Le Comité consultatif national d'éthique français mettait déjà en avant, avant l'actuelle pandémie, différents points de rupture entre la gestion des données de santé traditionnelles et l'avènement des données massives en santé, notamment : un changement d'échelle, la pérennité de conservation, la diffusion rapide au-delà des équipes médicales et des frontières (CNNE, 2019). Cette rupture s'est accentuée avec la pandémie, notamment avec la collecte de données de géolocalisation des citoyennes et citoyens et de leurs déplacements à des fins de santé, alors que des données générées sur les médias sociaux ont pu être

utile à la décision publique (comme l'analyse de sentiments vis-à-vis de la vaccination – voir par exemple Wilson et Wiysonge, 2020). Les données de santé plus traditionnelles (ex. un diagnostic) n'étaient plus collectées uniquement pour soigner le patient concerné mais aussi pour d'autres fins (comme assigner au confinement).

L'ATM et l'approche des capabilités demande également d'accorder une attention particulière aux inégalités sociales qui pourrait influencer la conversion d'une ressource en fonctionnement effectif pour toutes et tous, au-delà de la simple création de ressources ou de moyens (Fusulier et Sirna, 2010 ; Oosterlaken, 2015). Ceci peut amener à contester la pertinence contextuelle d'équiper toutes personnes en technologies numériques au nom d'un principe d'équité, y compris celles qui n'auraient pas accès à des ressources de base nécessaires, par exemple, à la survie. L'accès au numérique est-il prioritaire ou pertinent dans tous les contextes ? Se pencher sur les inégalités nécessite également que l'empiètement sur l'intimité et la vie privée soit justifié seulement si une réelle plus-value solidaire en émane, et ce pour toutes celles et tous ceux concernés par cet empiètement.

Or, au-delà de la simple fracture numérique, des inégalités persistent relativement au partage des bénéfices de l'analyse des données : une importante asymétrie demeure, d'une part, entre les personnes qui collectent, stockent et exploitent les données massives et, d'autre part, celles qui les génèrent ou que la collecte de données cible. Ce phénomène est appelé « fracture des données massives » (« big data divide ») (Andrejevic, 2014 ; McCarthy, 2016). Il a également été souligné que la pandémie de COVID-19, en exacerbant les inégalités préexistantes, a eu des conséquences bien plus dommageables sur les populations précaires – en particulier en ce qui a trait aux conséquences sur les capabilités (Biggeri, 2020). Ce sont également les populations des pays du Sud qui sont les plus immédiatement et les plus fortement touchées par les conséquences du réchauffement climatique (Goodman, 2009) ou celles de la dégradation de l'environnement issue de l'extraction des minéraux critiques. Par exemple, les principaux fournisseurs en éléments nécessaires au développement de technologies numériques sont la Chine (41 %) et les pays d'Afrique (30 %) (Commission Européenne, 2020). L'Europe serait également grandement dépendante de l'Asie du Sud-Est vis-à-vis des composantes et de l'assemblage de ces technologies (Commission Européenne, 2020). Ceci peut amener à interroger la réelle plus-value équitable des SIA pour les populations exclues du numérique, si celles-ci sont celles qui souffrent le plus des conséquences environnementales de leur développement.

Enfin, Sen (2013) et Dubois (2006) invitent à repenser l'impact de l'augmentation des capabilités sur la durabilité. L'approche des capabilités permet en effet d'envisager les impératifs d'équité non seulement entre pays ou entre groupes qui auraient plus ou moins accès aux technologies numériques, mais également entre les générations humaines. Le développement soutenable se comprend alors en termes d'équité intergénérationnelle, visant à ce que les prochaines générations aient accès au moins aux mêmes capabilités que les générations actuelles (Dubois, 2006). Ce maintien des capabilités entre générations ne doit pas se limiter selon Sen (2013) à un maintien de « notre capacité à satisfaire nos besoins ressentis » (« our ability to fulfil our felt needs »), mais doit plutôt viser « à maintenir les libertés humaines » (« sustaining human freedoms »). Dans cette optique, le principe 3 d'*Équité* ne s'oppose alors plus au principe 10 de *Développement soutenable*, mais en fait partie intégrante, étendant la portée du principe « ne laisser personne pour compte » (« leave no one behind ») des Nations Unies aux individus des générations à venir (United Nations, n.d.).

CONCLUSION

Ainsi, la mise en œuvre de principes éthiques pour guider un développement responsable de l'IA en vue de contrer la pandémie de COVID-19 révèle l'existence de plusieurs dilemmes. Suivant les principes de la Déclaration de Montréal, deux dilemmes clés ont notamment été mis en évidence : le respect du principe 3 de *Protection de l'intimité et de la vie privée* pourrait entraver le respect du principe 4 de *Solidarité* qui invite au partage des données personnelles du plus grand nombre et ; le respect du principe 10 de *Développement soutenable* se heurte à l'objectif de dépasser la fracture numérique selon le principe 6 d'*Équité*. L'intérêt de résoudre ces dilemmes est double. Vis-à-vis de la gouvernance de l'IA, la résolution des dilemmes pourrait contribuer à prévenir un potentiel rejet ou désintérêt éthique en assurant plus de cohérence eu égard aux lignes directrices existantes. Vis-à-vis de la santé mondiale, la résolution des dilemmes est nécessaire en vue d'assurer un développement responsable de l'IA en santé et ainsi contribuer au mieux à la gestion des pandémies.

L'approche des capabilités de Sen semble prometteuse pour dépasser la binarité des dilemmes présentés. Selon l'approche des capabilités, il est nécessaire de considérer au moins deux dimensions pour assurer un développement « approprié » des SIA, soit qu'ils permettent aux individus, dans toute leur diversité, de mener les vies qu'ils jugent bonnes (pour reprendre les mots d'Oosterlaken, 2015). D'une part, il faut évaluer dans quelle mesure les SIA introduisent des possibilités supplémentaires réelles, ce qui demande de dépasser les limites actuelles concernant leur efficacité et éventuellement considérer d'autres alternatives – sans quoi il n'est pas possible d'assurer une réelle plus-value solidaire (principe 4) ou équitable (principe 6) des SIA. D'autre part, il faut se pencher sur les conditions qui permettent aux individus de choisir d'avoir recours ou non aux SIA. Ceci nécessite d'impliquer les citoyennes et citoyens, d'identifier leurs préférences, et de tenir compte du contexte dans lequel les SIA sont implémentés – notamment, des inégalités préexistantes entre différents groupes et entre les générations actuelles et futures – en vue de collectivement définir les attentes en termes de vie privée (principe 3) et de soutenabilité (principe 10).

Si nous reconnaissons qu'il ne s'agit ici que d'un premier niveau d'analyse, celui-ci nous invite cependant à formuler ou à réitérer l'importance de quelques pistes de solution, qu'elles visent à résoudre les dilemmes présentés ou plus généralement à éclairer les potentiels angles morts de l'éthique de l'IA :

Encourager, avant un déploiement à large échelle des SIA (que ce soit dans le cadre de la pandémie actuelle ou de futures pandémies), **le financement et la réalisation de recherches** sur : 1) les limites qu'ont rencontrés les SIA utilisés pour contrer la pandémie de COVID-19 et ; 2) l'impact sur la santé des populations des conséquences environnementales des technologies numériques, et ce aux différentes étapes de leur cycle de vie. Les connaissances acquises sur ces limites et sur cet impact devraient permettre d'augmenter les possibilités effectives qu'introduisent les SIA, visant ainsi à augmenter les capabilités des populations actuelles et futures.

Systématiser la mise en place de coconstruction, avec les citoyennes et citoyens concernés, des solutions numériques et des politiques publiques relatives au développement de SIA dans un contexte de santé mondiale. Plus que la simple consultation, la co-construction implique la participation active des populations et est essentielle à un développement technologique approprié. Ceci permettrait d'aligner le développement des SIA avec les valeurs et les préférences citoyennes, dimensions essentielles de l'approche des capabilités. Si une atteinte proportionnée aux droits et libertés individuels se justifie au nom du bien commun, il est essentiel de collectivement évaluer la forme que ce bien commun devrait prendre. Cette co-construction doit être réalisée en prêtant une attention particulière aux populations locales, marginalisées et groupes exclus au sein des sociétés concernées, et doit assurer un échange bidirectionnel entre les pays du Nord et les pays du Sud.

Faire le choix de la pertinence dès la conception (*by-design)* des SIA. Envisager d'autres options que le numérique lorsque celui-ci ne représente pas un moyen d'augmenter les capabilités réelles, en vue d'atteindre un équilibre soutenable. Ceci demande de remettre en question l'atteinte de l'équité en vue de dépasser la fracture numérique uniquement en augmentant l'accès aux technologies numériques pour les populations qui y ont peu ou pas accès, et de considérer par exemple des mesures visant à limiter la surconsommation en technologies numériques au sein, notamment, des pays du Nord.

Favoriser les approches globales pour répondre des enjeux relatifs à l'IA, en particulier dans un contexte de santé mondiale. Ceci implique de limiter les approches en silos, par projet, par programme ou par discipline, qui favorisent les angles morts (par exemple, qui traiteraient de la fracture numérique d'un côté et du développement durable de l'autre). Ceci nécessite également de ne pas se limiter à une conception locale des enjeux décrits, considérant la mondialisation des échanges et de la numérisation, la diffusivité de l'encadrement de l'IA[38] ou le caractère transfrontalier et transsectoriel des pandémies.

La mise en œuvre de ces différents mécanismes, même dans l'urgence associée à toutes situations de crise, est essentielle à moyen et long terme. Nous pensons qu'ils contribueraient à créer des solutions viables dans une perspective d'un monde, une santé (« one world, one health ») en vue que l'IA bénéficie à toutes et tous.

38. Ou « diffuseness problem », décrit par Danaher (2015) comme le « problème qui se pose lorsque les systèmes d'IA sont développés par des équipes de chercheurs qui sont organisationnellement, géographiquement, et [...] juridiquement distinctes » (traduction libre) ('the problem that arises when AI systems are developed using teams of researchers that are organisationally, geographically, and perhaps more importantly, jurisdictionally separate') permettant d'échapper à la réglementation d'un pays en profitant de cette diffusivité juridictionnelle.

RÉFÉRENCES

AIHLEG (High-Level Expert Group on Artificial Intelligence). 2019. *Ethics Guidelines for Trustworthy AI*. Brussels : European Commission. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.

Akinbi, A. Forshaw, M. et Blinkhorn, V. 2021. Contact Tracing Apps for the COVID-19 Pandemic : A Systematic Literature Review of Challenges and Future Directions for Neo-Liberal Societies. *Health Information Science and Systems* 9 (1) : 18. https://doi.org/10.1007/s13755-021-00147-7.

AlgorithmWatch, 2020, AI Ethics Guidelines Global Inventory, https://inventory.algorithmwatch.org/

Alimadadi, A. *et al.* 2020. Artificial intelligence and machine learning to fight COVID-19. *Physiological Genomics*, 52(4), pp. 200–202. doi :10.1152/physiolgenomics.00029.2020.

Anand, P., Ferrer, B. Gao, Q., Nogales, R. et Unterhalter, E. 2020. COVID-19 as a Capability Crisis : Using the Capability Framework to Understand Policy Challenges. *Journal of Human Development and Capabilities*, 21(3), pp. 293–299. doi :10.1080/19452829.2020.1789079.

Ananny, M. and Crawford, K. 2018. Seeing without knowing : Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), pp. 973–989. doi :10.1177/1461444816676645.

Andrejevic, M. 2014. Big Data, Big Questions| The Big Data Divide. *International Journal of Communication*, 8(0), p. 17.

Asan, O., Bayrak A. E., et Choudhury, A. 2020. Artificial Intelligence and Human Trust in Healthcare : Focus on Clinicians. *Journal of Medical Internet Research*, 22 (6) : e15154. https://doi.org/10.2196/15154.

Azencott C.-A. 2018. Machine learning and genomics : precision medicine versus patient privacy. *Philosophical Transactions of the Royal Society A : Mathematical, Physical and Engineering Sciences*, 376(2128), p. 20170350. doi :10.1098/rsta.2017.0350.

Biggeri, M. 2020. Introduction : Capabilities and Covid-19. *Journal of Human Development and Capabilities*, 21(3), pp. 277–279. doi :10.1080/19452829.2020.1790732.

Boas, H., Rosenthal, A. et Davidovitch, N. 2016. Between individualism and social solidarity in vaccination policy : the case of the 2013 OPV campaign in Israel. *Israel Journal of Health Policy Research*, 5(1), p. 64. doi :10.1186/s13584-016-0119-y.

Bonvin, J.-M. et Farvaque, N.2007. L'accès à l'emploi au prisme des capabilités, enjeux théoriques et méthodologiques. *Formation emploi. Revue française de sciences sociales*, (98), pp. 9–22. doi :10.4000/formationemploi.1550.

Bullock, J. *et al.* 2020. Mapping the landscape of Artificial Intelligence applications against COVID-19. *Journal of Artificial Intelligence Research*, 69, pp. 807–845. doi :10.1613/jair.1.12162.

Cardon, D., Cointet, J.-P. et Mazières, A. 2018. La revanche des neurones. *Reseaux*, n° 211(5), pp. 173–220.

Cave, S. *et al.* 2021. Using AI ethically to tackle covid-19. *BMJ (Clinical research ed.)*, 372, p. n364. doi :10.1136/bmj.n364.

CCNE. 2019. Données massives (big data) et santé : une nouvelle approche des enjeux éthiques. Avis 130. Comité Consultatif National d'éthique français. https://www.ouvrirlascience.fr/wp-content/uploads/2019/06/CCNE_Donnees-massives-et-sant%C3%A9_avis130_29mai2019.pdf.

Charlier, P. *et al.* 2020. Global warming and planetary health : An open letter to the WHO from scientific and indigenous people urging for paleo-microbiology studies. *Infection, Genetics and Evolution : Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, 82, p. 104284. doi :10.1016/j.meegid.2020.104284.

Chen, J. et See, K.C. 2020. Artificial Intelligence for COVID-19 : Rapid Review. *Journal of Medical Internet Research*, 22(10), p. e21476. doi :10.2196/21476.

Christen, M. *et al.* 2016. On the Compatibility of Big Data Driven Research and Informed Consent : The Example of the Human Brain Project. in Mittelstadt, B.D. and Floridi, L. (eds) *The Ethics of Biomedical Big Data*. Cham : Springer International Publishing (Law, Governance and Technology Series), pp. 199–218. doi :10.1007/978-3-319-33525-4_9.

Coeckelbergh, M. 2015. Artificial agents, good care, and modernity. *Theoretical Medicine and Bioethics*, 36(4), pp. 265–277. doi :10.1007/s11017-015-9331-y.

Craglia M., de Nigris S., Gomez-Gonzalez E., Gomez E., Martens B., Iglesias Portela M., Vespe M., Schade S., Micheli M., et Kotzev A. 2020. Artificial Intelligence and Digital Transformation : early lessons from the COVID-19 crisis. *JRC Science for policy report. Publications Office of the European Union*.

Crawford, K. 2021. *Atlas of AI*. Yale University Press.

Danaher, J. 2015. Philosophical Disquisitions : Is effective regulation of AI possible ? Eight potential regulatory problems. *Philosophical Disquisitions*, http://philosophicaldisquisitions.blogspot.com/2015/07/is-effective-regulation-of-ai-possible.html

Davis, S.L.M. 2020. The Trojan Horse. *Health and Human Rights*, 22(2), pp. 41–47.

Dawson, A.J. 2015. Ebola : what it tells us about medical ethics. *Journal of Medical Ethics*, 41(1), pp. 107–110. doi :10.1136/medethics-2014-102304.

Déclaration de Montréal. 2018. *Déclaration de Montéal pour un développement responsable de l'intelligence artificielle*. Université de Montréal. https://www.declarationmontreal-iaresponsable.com/la-declaration.

Dijk, J. van. 2020. The Digital Divide. *John Wiley & Sons*. ISBN : 978-1-5095-3446-3.

Doucin, M. 2009. Review of Repenser l'action collective : une approche par les capabilités, (Réseau Impact, coll. « Éthique économique »). *Revue Tiers Monde*, 50(198), pp. 444–448.

Douglas Heaven, W. 2021. Hundreds of AI tools have been built to catch covid. None of them helped. *MIT Technology Review*. https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/.

Dubois, J.-L. 2006. Approche par les capabilités et développement durable : La transmission intergénérationnelle des capabilités. *Amartya Sen : un économiste du développement*, pp. 201–213.

Dwivedi, Y. K., Hughes, L. Kar, A.K., Baabdullah, A.M., Grover, P. Abbas, R. Andreini, D. et al. 2022. Climate Change and COP26 : Are Digital Technologies and Information Management Part of the Problem or the Solution ? An Editorial Reflection and Call to Action. *International Journal of Information Management* 63 (avril) : 102456. https://doi.org/10.1016/j.ijinfomgt.2021.102456.

El-Sayed, A. et Kamel, M. 2020. Future threat from the past. *Environmental Science and Pollution Research International*, pp. 1–5. doi :10.1007/s11356-020-11234-9.

Else, H. 2018. Europe's AI researchers launch professional body over fears of falling behind. *Nature*. doi :10.1038/d41586-018-07730-1.

de la Espriella, F.R.M., Llanos, A.Z.B. et Hernandez, J.C. 2021. Privacy as a human right and solidarity as a constitutional value in the era of Covid-19. *Juridicas Cuc*, pp. 17–17.

Commission Européenne. 2020. *Critical raw materials for strategic technologies and sectors in the EU – A foresight study*. Luxembourg : Publications Office of the European Union. https://ec.europa.eu/docsroom/documents/42881.

Fiore, R.N. et Goodman, K.W. 2016. Precision medicine ethics : selected issues and developments in next-generation sequencing, clinical oncology, and ethics. *Current Opinion in Oncology*, 28(1), pp. 83–87. doi :10.1097/CCO.0000000000000247.

Fjeld, J. *et al.* 2020. Principled artificial intelligence : Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication* [Preprint], (2020–1).

Floridi, L. 2019. Translating Principles into Practices of Digital Ethics : Five Risks of Being Unethical. *Philosophy & Technology*, 32(2), pp. 185–193. doi :10.1007/s13347-019-00354-x.

Fusulier, B. et Sirna, F. 2010. Contrer les inégalités du "pouvoir d'agir", augmenter les capabilités. *Les Politiques Sociales*, n° 3-4(2), pp. 33–38.

Gasser, U. *et al.* 2020. Digital tools against COVID-19 : taxonomy, ethical challenges, and navigation aid. *The Lancet Digital Health*, 2(8), pp. e425–e434. doi :10.1016/S2589-7500(20)30137-0.

Ghobakhloo, M. 2020. Industry 4.0, digitization, and opportunities for sustainability. *Journal of Cleaner Production*, 252, p. 119869. doi :10.1016/j.jclepro.2019.119869.

Gibert, M. 2019. Faut-il avoir peur de la peur de l'IA ? *La Quatrième Blessure*, 11 January. https://medium.com/@martin.gibert/faut-il-avoir-peur-de-la-peur-de-lia-1687abc35342.

GIEC (Groupe d'experts intergouvernemental sur l'évolution du climat). 2021. Sixth Assessment Report, Climate Change 2021 : The Physical Science Basis. https://www.ipcc.ch/report/ar6/wg1/downloads/report/IPCC_AR6_WGI_Full_Report.pdf.

Global Observatory for eHealth. 2015. Atlas of eHealth country profiles : the use of eHealth in support of universal health coverage. *World Health Organization.* https://www.who.int/publications/i/item/9789241565219

Global observatory for eHealth. 2016. Global diffusion of eHealth : Making universal health coverage achievable. *World Health Organization*. http://library.health.go.ug/download/file/fid/2620

Gmach, D. *et al.* 2010. Profiling Sustainability of Data Centers in *Proceedings of the 2010 IEEE International Symposium on Sustainable Systems and Technology*. pp. 1–6. doi :10.1109/ISSST.2010.5507750.

Goodman, J. 2009. From Global Justice to Climate Justice ? Justice Ecologism in an Era of Global Warming. *New Political Science*, 31(4), pp. 499–514. doi :10.1080/07393140903322570.

Gunasekeran, D. V., Wei Wen Tseng R. M., Tham Y-C., et Wong T. Y. 2021. Applications of Digital Health for Public Health Responses to COVID-19 : A Systematic Scoping Review of Artificial Intelligence, Telehealth and Related Technologies. *Npj Digital Medicine* 4 (1) : 1-6. https://doi.org/10.1038/s41746-021-00412-9.

Hagendorff, T. 2020. The Ethics of AI Ethics : An Evaluation of Guidelines. *Minds and Machines*, 30(1), pp. 99–120. doi :10.1007/s11023-020-09517-8.

Hager, G.D., Drobnis A., Fang F., Ghani R. *et al.* 2019. Artificial Intelligence for Social Good. *arXiv :1901.05406 [cs]* [Preprint]. http://arxiv.org/abs/1901.05406.

Hand, D.J. 2018. Aspects of Data Ethics in a Changing World : Where Are We Now ?. *Big Data*, 6(3), pp. 176–190. doi :10.1089/big.2018.0083.

Hashiguchi, T. Cravo, O., Oderkirk J., et Slawomirski, L. 2022. Fulfilling the Promise of Artificial Intelligence in the Health Sector : Let's Get Real. *Value in Health* 25 (3) : 368-73. https://doi.org/10.1016/j.jval.2021.11.1369.

Hébert, C. 2021. Un pour tous, tous pour Une seule santé. *Hinnovic*. https://www.hinnovic.org/post/un-pour-tous-tous-pour-une-seule-santé.

Hund, K. Laporta, D. Fabregas T.P. Laing, T. et Drexhage J. 2020. Minerals for Climate Action : The Mineral Intensity of the Clean Energy Transition. *The World Bank*. https://pubdocs.worldbank.org/en/961711588875536384/Minerals-for-Climate-Action-The-Mineral-Intensity-of-the-Clean-Energy-Transition.pdf.

IEA. 2021. The Role of Critical Minerals in Clean Energy Transitions – Analysis. *International Energy Agency*. https://www.iea.org/reports/the-role-of-critical-minerals-in-clean-energy-transitions/executive-summary.

IEEE. 2017. Ethically aligned design – Version 2 – For Public Discussion. *I. of E. and E.E.* https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf.

Ienca, M. et Vayena, E. 2020. On the responsible use of digital data to tackle the COVID-19 pandemic. *Nature Medicine*, 26(4), pp. 463–464. doi:10.1038/s41591-020-0832-5.

Iyengar, A., Kundu, A. et Pallis, G. 2018. Healthcare Informatics and Privacy. *IEEE Internet Computing*, 22(2), pp. 29–31. doi:10.1109/MIC.2018.022021660.

Jackman, M. 2020. Fault Lines : COVID-19, the Charter, and Long-term Care in *Vulnerable : The Law, Policy and Ethics of COVID-19 de Colleen M. Flood et al.* University of Ottawa Press, pp. 339–354. Available at : https://muse.jhu.edu/book/76885.

Jobin, A., Ienca, M. et Vayena, E. 2019. The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence* 1 (9) : 389-99. https://doi.org/10.1038/s42256-019-0088-2.

Kassab, M. et Graciano Neto, V.V. 2021. Digital Surveillance Technologies to Combat COVID-19 : A Contemporary View. *Procedia Computer Science*, 185, pp. 37–44. doi:10.1016/j.procs.2021.05.005.

Kenny, .Nuala P., Sherwin, S.B. et Baylis, F.E. 2010. Re-visioning Public Health Ethics : A Relational Perspective. *Canadian Journal of Public Health*, 101(1), pp. 9–11. doi:10.1007/BF03405552.

Kim, P.T. 2016. Data-Driven Discrimination at Work. *William & Mary Law Review*, 58, pp. 857–936.

Krantz, I., Sachs, L. et Nilstun, T. 2004. Ethics and vaccination. *Scandinavian Journal of Public Health*, 32(3), pp. 172–178. doi:10.1080/14034940310018192.

Kudina, O. 2021. Bridging Privacy and Solidarity in COVID-19 Contact-tracing Apps through the Sociotechnical Systems Perspective. *Glimpse*, 22(2), pp. 43–54. doi:10.5840/glimpse202122224.

Lagacé, M., Garcia, L. et Bélanger-Hardy, L. 2020. COVID-19 et âgisme : crise annoncée dans les centres de soins de longue durée et réponse improvisée ? in *Vulnerable : The Law, Policy and Ethics of COVID-19 de Colleen M. Flood et al.* University of Ottawa Press, pp. 329–338. Available at : https://muse.jhu.edu/book/76885.

Lai, J. et Widmar, N.O. 2021. Revisiting the Digital Divide in the COVID-19 Era. *Applied Economic Perspectives and Policy*, 43(1), pp. 458–464. doi:10.1002/aepp.13104.

Langlois, L. et Régis, C. 2021. Analyzing the Contribution of Ethical Charters to Building the Future of Artificial Intelligence Governance in Braunschweig, B. and Ghallab, M. (eds) *Reflections on Artificial Intelligence for Humanity*. Cham : Springer International Publishing (Lecture Notes in Computer Science), pp. 150–170. doi:10.1007/978-3-030-69128-8_10.

CEST. 2020. Les enjeux éthiques de l'utilisation d'une application mobile de traçage des contacts dans le cadre de la pandémie de COVID-19 au Québec. *Commission de l'éthique en science et en technologie*. https://www.ethique.gouv.qc.ca/fr/publications/l-utilisation-d-une-application-mobile-de-tracage-des-contacts-dans-le-cadre-d-une-pandemie/.

Ligozat, A.-L., Lefèvre, J. Bugeau, A. et Combaz, J. 2021. Unraveling the hidden environmental impacts of AI solutions for environment. *arXiv :2110.11822* [cs], octobre. http://arxiv.org/abs/2110.11822.

Mackenzie, J.S. et Jeggo, M. 2019. The One Health Approach—Why Is It So Important ? *Tropical Medicine and Infectious Disease*, 4(2), p. 88. doi:10.3390/tropicalmed4020088.

Makri, A. 2019. Bridging the digital divide in health care. *The Lancet Digital Health*, 1(5), pp. e204–e205. doi:10.1016/S2589-7500(19)30111-6.

Martins Van Jaarsveld, G. 2020. The Effects of COVID-19 Among the Elderly Population : A Case for Closing the Digital Divide. *Frontiers in Psychiatry*, 11, p. 577427. doi :10.3389/fpsyt.2020.577427.

Massé, R. 2003. Valeurs universelles et relativisme culturel en recherche internationale : les contributions d'un principisme sensible aux contextes socioculturels. *Autrepart*, n° 28(4), pp. 21–35.

Mbunge, E. *et al.* 2021. Ethics for integrating emerging technologies to contain COVID-19 in Zimbabwe. *Human Behavior and Emerging Technologies*. doi :10.1002/hbe2.277.

McCarthy, M.T. 2016. The big data divide and its consequences. *Sociology Compass*, 10(12), pp. 1131–1140. doi :10.1111/soc4.12436.

Mello, M.M. et Wang, C.J. 2020. Ethics and governance for digital disease surveillance. *Science (New York, N.Y.)*, 368(6494), pp. 951–954. doi :10.1126/science.abb9045.

Mittelstadt, B. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1(11), pp. 501–507. doi :10.1038/s42256-019-0114-4.

Mittelstadt, B.D. et Floridi, L. 2016. The Ethics of Big Data : Current and Foreseeable Issues in Biomedical Contexts. *Science and Engineering Ethics*, 22(2), pp. 303–341. doi :10.1007/s11948-015-9652-2.

Mondin, C. et de Marcellis-Warin, N. 2020. Recension des solutions technologiques développées dans le monde afin de limiter la propagation de la COVID-19 et typologie des applications de traçage. *Observatoire international sur les impacts sociétaux de l'IA et du numérique (OBVIA)*. https://www.docdroid.com/VLokunh/recension-des-solutions-technologiques-developpees-dans-le-monde-afin-de-limiter-la-propagation-de-la-covid-19-et-typologie-des-applications-de-tracage-pdf.

Morley, J., Floridi L., Kinsey L. et Elhalal A. 2020. From What to How : An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science and Engineering Ethics* 26 (4) : 2141-68. https://doi.org/10.1007/s11948-019-00165-5.

Murphy, K. *et al.* 2021. Artificial intelligence for good health : a scoping review of the ethics literature. *BMC Medical Ethics*, 22(1), p. 14. doi :10.1186/s12910-021-00577-8.

Nations Unies. 2021. Cadre mondial d'indicateurs relatifs aux objectifs et aux cibles du Programme de développement durable à l'horizon 2030. https://unstats.un.org/sdgs/indicators/Global%20Indicator%20Framework%20after%202021%20refinement_Fre.pdf.

Naudé, W. 2020. Artificial intelligence vs COVID-19 : limitations, constraints and pitfalls. *AI & SOCIETY*, 35(3), pp. 761–765. doi :10.1007/s00146-020-00978-0.

News, L. 2020. *Covid-19 is magnifying the digital divide*, *LaptrinhX/News*. https://laptrinhx.com/news/covid-19-is-magnifying-the-digital-divide-pGZbpeA/.

Noorman, M. 2016. Computing and Moral Responsibility in Zalta, E.N. (ed.) *The Stanford Encyclopedia of Philosophy*. Winter 2016. Metaphysics Research Lab, Stanford University. Available at : https://plato.stanford.edu/archives/win2016/entries/computing-responsibility/.

O'Doherty, K.C., Christofides E., Yen J. , Beate Bentzen H. *et al.* 2016. If you build it, they will come : unintended future uses of organised health data collections. *Bmc Medical Ethics*, 17, p. 54. doi :10.1186/s12910-016-0137-x.

OMS (Organisation Mondiale de la Santé). 2018. Big data and artificial intelligence for achieving universal health coverage : an international consultation on ethics. http://apps.who.int/iris/bitstream/handle/10665/275417/WHO-HMM-IER-REK-2018.2-eng.pdf?ua=1.

OMS (Organisation Mondiale de la Santé). 2021a. Ethics and governance of artificial intelligence for health : WHO guidance. *World Health Organization*. https://www.who.int/publications/i/item/9789240029200.

OMS (Organisation mondiale de la santé). 2021b. WHO HUB FOR PANDEMIC AND EPIDEMIC INTELLIGENCE. Better Data. Better Analytics. Better Decisions. *World Health Organization, Health Emergencies Programme*. https://cdn.who.int/media/docs/default-source/2021-dha-docs/who_hub.pdf?sfvrsn=8dc28ab6_5.

Oosterlaken, I. 2015. *Technology and human development*. Routledge, Taylor&Francis. USA.

Patsavellas, J. et Salonitis, K. 2019. The Carbon Footprint of Manufacturing Digitalization : critical literature review and future research agenda. *Procedia CIRP*, 81, pp. 1354–1359. doi :10.1016/j.procir.2019.04.026.

Patz, J.A. *et al.* 2014. Climate Change : Challenges and Opportunities for Global Health. *JAMA*, 312(15), pp. 1565–1580. doi :10.1001/jama.2014.13186.

de Pooter, H. 2015. *Le droit international face aux pandémies : vers un système de sécurité sanitaire collective ?* Editions A. Pedone.

Principes de Manhattan (2004) *The Manhattan Principles*. https://oneworldonehealth.wcs.org/About-Us/Mission/The-Manhattan-Principles.aspx.

Ramsetty, A. et Adams, C. 2020. Impact of the digital divide in the age of COVID-19. *Journal of the American Medical Informatics Association*, 27(7), pp. 1147–1148. doi :10.1093/jamia/ocaa078.

Rial-Sebbag, E. 2017. Chapitre 4. La gouvernance des Big data utilisées en santé, un enjeu national et international. *Journal international de bioethique et d'ethique des sciences*, Vol. 28(3), pp. 39–50.

Risse, M. 2019. Human Rights and Artificial Intelligence : An Urgently Needed Agenda. *Human Rights Quarterly*, 41(1), pp. 1–16. doi :10.1353/hrq.2019.0000.

Scassa, T., Millar, J. et Bronson, K. 2020. Privacy, Ethics, and Contact-tracing Apps. *SSRN Scholarly Paper* ID 3651457. Rochester, NY : Social Science Research Network. doi :10.2139/ssrn.3651457.

Sen, A. 1983. Poor, Relatively Speaking. *Oxford Economic Papers*, 35(2), pp. 153–169.

Sen, A. 2013. The Ends and Means of Sustainability. *Journal of Human Development and Capabilities*, 14(1), pp. 6–20. doi :10.1080/19452829.2012.747492.

Shachar, C., Gerke, S. et Adashi, E.Y. 2020. AI Surveillance during Pandemics : Ethical Implementation Imperatives. *The Hastings Center Report*, 50(3), pp. 18–21. doi :10.1002/hast.1125.

Shneiderman, B. 2020. Bridging the Gap Between Ethics and Practice : Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. *ACM Transactions on Interactive Intelligent Systems*, 10(4), p. 26 :1-26 :31. doi :10.1145/3419764.

Siau, K. et Wang, W. 2020. Artificial Intelligence (AI) Ethics : Ethics of AI and Ethical AI. *Journal of Database Management (JDM)*, 31(2), pp. 74–87. doi :10.4018/JDM.2020040105.

Sim, F.M. 2017. Individualism and social solidarity in vaccination policy : some further considerations. *Israel Journal of Health Policy Research*, 6(1), p. 21. doi :10.1186/s13584-017-0147-2.

Solomon, C.G. et LaRocque, R.C. 2019. Climate Change — A Health Emergency. *New England Journal of Medicine*, 380(3), pp. 209–211. doi :10.1056/NEJMp1817067.

Solomon, D.H. *et al.* 2020. The "Infodemic" of COVID-19. *Arthritis & Rheumatology*, 72(11), pp. 1806–1808. doi :10.1002/art.41468.

Spiekermann, S., Korunovska, J. et Langheinrich, M. 2018. Inside the Organization : Why Privacy and Security Engineering Is a Challenge for Engineers. *Proceedings of the IEEE*, pp. 1–16. doi :10.1109/JPROC.2018.2866769.

Stahl, B.C. et Wright, D. 2018. Ethics and Privacy in AI and Big Data : Implementing Responsible Research and Innovation. *IEEE Security Privacy*, 16(3), pp. 26–33. doi :10.1109/MSP.2018.2701164.

Stapleton G., Schröder-Bäck P., Laaser U., Meershoek A. et Popa D. 2014. Global health ethics : an introduction to prominent theories and relevant topics. *Global Health Action*, 7(s2), p. 23569. doi :10.3402/gha.v7.23569.

von Struensee, S. 2021. Mapping Artificial Intelligence Applications Deployed Against COVID-19 Alongside Ethics and Human Rights Considerations. *SSRN Scholarly Paper ID 3889441*. Rochester, NY : Social Science Research Network. doi :10.2139/ssrn.3889441.

Tagliabue, F., Galassi, L. et Mariani, P. 2020. The "Pandemic" of Disinformation in COVID-19. *Sn Comprehensive Clinical Medicine*, pp. 1–3. doi :10.1007/s42399-020-00439-1.

Terry, N. et Coughlin, C.N. 2021. A Virtuous Circle : How Health Solidarity Could Prompt Recalibration of Privacy and Improve Data and Research. *SSRN Scholarly Paper ID 3774366. Rochester*, NY : Social Science Research Network. Available at : https://papers.ssrn.com/abstract=3774366.

The Future Society. 2020. Areas for future action in the responsible AI ecosystem. https://thefuturesociety.org/wp-content/uploads/2021/02/Areas-for-Future-Action-in-the-Responsible-AI-Ecosystem.pdf.

The Shift Project. 2020. Déployer la sobriété numérique. https://theshiftproject.org/article/deployer-la-sobriete-numerique-rapport-shift.

The Shift Project. 2021. Impact environnemental du numérique et gouvernance de la 5G. https://theshiftproject.org/article/impact-environnemental-du-numerique-5g-nouvelle-etude-du-shift/.

Tran, C.D. et Nguyen, T.T. 2021. Health vs. privacy ? The risk-risk tradeoff in using COVID-19 contact-tracing apps. *Technology in Society*, 67, p. 101755. doi :10.1016/j.techsoc.2021.101755.

Tzachor, A. *et al.* 2020. Artificial intelligence in a crisis needs ethics with urgency. *Nature Machine Intelligence*, 2(7), pp. 365–366. doi :10.1038/s42256-020-0195-0.

United Nations. n.d. Universal Values, Principle 2 : Leave No One Behind. *United Nations Sustainable Development Group.* https://unsdg.un.org/2030-agenda/universal-values/leave-no-one-behind

UNDP. 2020. *Human Development Report 2020. The next frontier Human development and the Anthropocene*. http://hdr.undp.org/sites/default/files/hdr2020.pdf.

Vaishya, R. *et al.* 2020. Artificial Intelligence (AI) applications for COVID-19 pandemic. *Diabetes & Metabolic Syndrome : Clinical Research & Reviews*, 14(4), pp. 337–339. doi :10.1016/j.dsx.2020.04.012.

Villani, C. 2018. Donner un sens à l'intelligence artificielle. Pour une stratégie nationale et européenne. https://www.aiforhumanity.fr/pdfs/9782111457089_Rapport_Villani_accessible.pdf.

Voarino, N. 2020. Systèmes d'intelligence artificielle et santé : les enjeux d'une innovation responsable. Thèse. https://papyrus.bib.umontreal.ca/xmlui/handle/1866/23526.

Whittlestone, J. *et al.* 2019. The Role and Limits of Principles in AI Ethics : Towards a Focus on Tensions in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA : Association for Computing Machinery (AIES '19), pp. 195–200. doi :10.1145/3306618.3314289.

Wilson, S. L., et Wiysonge, C. 2020. Social Media and Vaccine Hesitancy. *BMJ Global Health* 5 (10) : e004206. https://doi.org/10.1136/bmjgh-2020-004206.

Wynants, L. *et al.* 2020. Prediction models for diagnosis and prognosis of covid-19 : systematic review and critical appraisal. *BMJ*, 369, p. m1328. doi :10.1136/bmj.m1328.

K. Yeung ; A. Howes et G. Pogrebna. AI governance by Human Rights-Centered Design, Deliberation, and Oversight in *Dubber Pasquale and Das (eds*), The Oxford Handbook of Ethics of AI, Oxford University Press, 2020, p. 77-106.

Yu, H. *et al.* 2018. Building Ethics into Artificial Intelligence. *arXiv :1812.02953 [cs]* [Preprint]. Available at : http://arxiv.org/abs/1812.02953.

DONNÉES[39]

KATE CRAWFORD

Chercheuse principale chez Microsoft Research et cofondatrice et directrice de la recherche à l'AI Now Institute à NYU. Auteure de « Atlas of AI » (2021). New Haven, Yale University Press

---

ODD 3 - Bonne santé et bien-être
ODD 9 - Industrie, innovation et infrastructure
ODD 10 - Énergie propre et d'un coût abordable
ODD 11 - Villes et communautés durables
ODD 12 - Consommation et production durables
ODD 16 - Paix, justice et institution efficaces
ODD 17 - Partenariats pour la réalisation des objectifs

39. Ce chapitre est une traduction du chapitre 3 de « Atlas of AI ». Droits d'auteur accordés à Mila – Institut québécois d'intelligence artificielle par Yale University Press le 23 novembre 2021 pour reproduction dans cette publication.

DONNÉES

RÉSUMÉ

Ce chapitre explore la manière dont les données constituent le fondement essentiel des systèmes d'IA. Il expose les politiques sous-jacentes, les questions éthiques, les limites épistémologiques et l'éventail des préjudices qui découlent des logiques d'extraction et d'accumulation des données dans l'industrie de l'IA. Dans la course à la performance algorithmique, le succès dépend de l'accès aux données et, par conséquent, on en collecte toujours plus. Les ensembles de données extrêmement volumineux sont alors considérés comme des infrastructures neutres ; ils sont interprétés comme des « choses » dépourvues de contexte et de sens et ce, malgré les images profondément personnelles et parfois horribles qu'ils contiennent. Or, en excavant les couches de données, on y découvre des histoires : des récits individuels et collectifs d'injustices historiques, de discriminations et d'inégalités structurelles. Cette représentation largement acceptée des données comme autant de ressources à consommer, de flux à contrôler et d'investissements à exploiter a engendré une sorte d'hybris – une idéologie statistique où seule l'échelle compte.

INTRODUCTION

Une jeune femme regarde vers le haut, les yeux fixés sur quelque chose hors du cadre, comme si elle refusait de reconnaître l'appareil photo. Sur la photo suivante, ses yeux sont fixés sur le centre de l'image. Une autre image la montre avec des cheveux ébouriffés et une expression abattue. Sur la séquence de photos, nous la voyons vieillir au fil du temps, les lignes autour de sa bouche s'abaissant et se creusant. Sur la dernière photo, elle apparaît blessée et découragée. Il s'agit de photographies d'identité judiciaire d'une femme ayant fait l'objet de multiples arrestations sur plusieurs années de sa vie. Ces images sont contenues dans une collection connue sous le nom de NIST Special Database 32–Multiple Encounter Dataset qui est mise à disposition sur Internet pour les chercheurs et chercheuses qui souhaitent tester leur logiciel de reconnaissance faciale (NIST, 2010).

Cette base de données est l'une des nombreuses que gère le National Institute of Standards and Technology (NIST), l'un des laboratoires de sciences physiques les plus anciens et les plus respectés des États-Unis qui fait maintenant partie du ministère du Commerce. NIST a été créé en 1901 pour

renforcer l'infrastructure nationale de mesure et créer des normes capables de concurrencer les rivaux économiques du monde industrialisé, comme l'Allemagne et le Royaume-Uni. Tout, des dossiers médicaux électroniques aux gratte-ciel antisismiques en passant par les horloges atomiques, relève de la compétence de NIST. Cet institut est devenu l'agence de mesure du temps, des protocoles de communication, des structures cristallines inorganiques, des nanotechnologies (Russell, 2014). Son objectif est de rendre les systèmes interopérables en définissant et en soutenant des normes, ce qui inclut désormais l'établissement de normes pour l'intelligence artificielle (IA). L'une des infrastructures de test qu'elle gère concerne les données biométriques.

| FIGURE 1 |

Images tirées de la NIST Special Database 32–Multiple Encounter Dataset (MEDS). National Institute of Standards and Technology, ministère du Commerce, États-Unis.

J'ai d'abord découvert la base de données de photographies d'identité judiciaire en 2017 alors que je cherchais dans les archives de données de NIST. Leurs collections biométriques sont vastes. Depuis plus de cinquante ans, NIST collaborent avec le Federal Bureau of Investigation (FBI) sur la reconnaissance automatisée des empreintes digitales et a conçu des méthodes pour évaluer la qualité des lecteurs d'empreintes digitales et des systèmes d'imagerie (Garris et Wilson, 2005, p. 1). Depuis les attaques terroristes du 11 septembre 2001, NIST contribue à la réponse nationale visant à créer des normes biométriques afin de vérifier et de suivre les personnes entrant aux États-Unis (Garris et Wilson, 2005, p. 1). Ce fut un moment décisif dans la recherche en reconnaissance faciale qui s'est étendue de l'application de la loi au contrôle des personnes traversant les frontières nationales.

Les photographies d'identité judiciaire elles-mêmes sont dévastatrices. Certaines personnes ont des blessures visibles, des contusions et des yeux au beurre noir, d'autres sont en détresse et pleurent. D'autres encore fixent l'appareil photo l'air hagard. La Special Dataset 32 contient des milliers de photos de personnes décédées ayant fait l'objet de plusieurs arrestations durant leurs rencontres répétées avec le système de justice pénale. Les personnes figurant dans les ensembles de données de photographies d'identité judiciaire sont présentées comme des points de données. Il n'y a pas d'histoires, de contextes

ou de noms et, comme les photographies d'identité judiciaire sont prises au moment de l'arrestation, nous ne savons pas si ces personnes ont été inculpées, acquittées ou emprisonnées. Elles sont toutes présentées de la même manière.

En raison de l'inclusion de ces images dans la base de données de NIST, leur utilisation est passée de l'identification de personnes dans les systèmes d'application de la loi à la référence technique pour tester les systèmes commerciaux et universitaires de détection des visages fondés sur l'intelligence artificielle. Dans son compte rendu sur la photographie policière, Allan Sekula fait valoir que les photographies d'identité judiciaire s'inscrivent dans une tradition de réalisme technique visant à « fournir une mesure physionomique standard du criminel » (Sekula, 1986, p. 17). Dans l'histoire de la photographie policière, il existe deux approches distinctes, explique-t-il. Les criminologues comme Alphonse Bertillon, inventeur de la photographie d'identité judiciaire, y voyaient une sorte de machine d'identification biographique, nécessaire pour repérer les récidivistes, tandis que Francis Galton, statisticien et figure fondatrice de l'eugénisme, utilisait le portrait composite de prisonniers pour détecter un « type criminel » déterminé biologiquement (Sekula, pp. 18-19). Galton travaillait dans le cadre d'un paradigme physiognomoniste dans lequel l'objectif était de trouver une apparence généralisée qui pourrait être utilisée pour trouver des traits de caractère profonds à partir de l'apparence physique. Lorsque des photographies d'identité judiciaire sont utilisées comme données d'entraînement, elles ne servent plus d'outils d'identification, mais plutôt à peaufiner une forme automatisée de vision. Nous pouvons considérer ceci comme un formalisme galtonien. Elles sont utilisées pour détecter les composantes mathématiques fondamentales des visages dans le but de « réduire la nature à son essence géométrique » (Sekula, 1986, p. 17).

Les photographies d'identité judiciaire constituent une partie des archives qui est utilisée pour tester les algorithmes de reconnaissance faciale. Les visages contenus dans la Multiple Encounter Dataset sont devenus des images standardisées, un substrat technique pour la comparaison de l'exactitude algorithmique. NIST, en collaboration avec l'Intelligence Advanced Research Projects Activity (IARPA), a organisé des compétitions en utilisant ces photographies d'identité judiciaire dans lesquelles les chercheurs et chercheuses compétitionnent pour déterminer quel algorithme est le plus rapide et le plus exact. Les équipes s'efforcent de battre les autres dans des tâches comme la vérification de l'identité des visages ou l'extraction d'un visage à partir d'une image de vidéo de surveillance (Grother *et al.*, 2017). Les gagnants et gagnantes célèbrent ces victoires qui peuvent leur apporter gloire, offres d'emploi et reconnaissance au sein de l'industrie (Ever AI, 2018).

Ni les personnes figurant sur ces photographies ni leur famille ne peuvent dire quoi que ce soit au sujet de leur utilisation et n'ont probablement aucune idée qu'elles font partie de bancs d'essai de l'IA. Les sujets des photographies d'identité judiciaire sont rarement pris en compte et peu d'ingénieurs et ingénieures y portent une très grande attention. Comme le document de NIST les décrit, ces photographies existent purement pour « peaufiner des outils, des techniques et des procédures de reconnaissance faciale en soutien à l'identification de nouvelle génération (NGI), à la comparaison, à la formation et à l'analyse médico-légales ainsi qu'à la conformité des images de visage et aux normes d'échange entre institutions » (Founds *et al.*, 2011). La description de la Multiple Encounter Dataset mentionne que plusieurs personnes montrent des signes de violence comme des cicatrices, des ecchymoses et des bandages. Mais le document conclut que ces signes sont « difficiles à interpréter en raison de l'absence de vérité terrain pour la comparaison avec un échantillon "propre" » (Curry *et al.*, 2009). Ces personnes ne sont pas considérées comme des individus, mais bien comme faisant partie d'une ressource technique commune, soit simplement un autre composant du programme Facial Recognition Verification Test, la référence standard du secteur.

Au fil de nombreuses années de recherche sur la manière dont sont construits les systèmes d'IA, j'ai consulté des centaines de bases de données, mais les bases de données de photographies d'identité judiciaire de NIST sont particulièrement perturbantes puisqu'elles représentent le modèle de ce qui nous attend. Il ne s'agit pas seulement de l'effet dramatique accablant des images elles-mêmes. Ce n'est pas

non plus uniquement l'atteinte à la vie privée qu'elles représentent, puisque les suspects, suspectes, prisonniers et prisonnières n'ont pas le droit de refuser d'être photographiés. Il s'agit plutôt du fait que les bases de données de NIST annoncent l'émergence d'une logique qui a maintenant envahi le secteur des technologies : la croyance inébranlable que tout est « donnée » et que tout est là pour être utilisé. Il importe peu où la photographie a été prise ou si elle reflète un moment de vulnérabilité ou de détresse, ou si elle représente une forme de honte pour la personne qui y figure. Il est devenu normalisé partout au sein de l'industrie de prendre et d'utiliser tout ce qui est disponible et peu s'arrêtent pour remettre en question les politiques sous-jacentes.

En ce sens, les photographies d'identité judiciaire sont l'édition originale de l'approche actuelle pour faire de l'IA. Le contexte et l'exercice du pouvoir que représentent ces images sont considérés comme étant sans importance, puisque ces dernières n'existent plus en tant que choses distinctes. Ces images ne sont pas considérées comme étant porteuses de sens ou d'un poids éthique en tant qu'images de personnes individuelles ou en tant que représentations du pouvoir structurel dans le système carcéral. On imagine que toutes les significations personnelles, sociales et politiques sont neutralisées. Je soutiens que ceci représente un virage d'*image* à *infrastructure* où la signification ou l'attention qui pourraient être accordées à l'image d'une personne ou le contexte derrière la scène sont présumés être effacés au moment où la photographie fait partie d'une masse regroupée qui alimente un système plus large. Tout est traité comme étant des données à utiliser dans des fonctions, soit du matériel à intégrer pour améliorer la performance technique. Il s'agit d'un principe fondamental de l'idéologie de l'extraction de données.

Chaque jour, les systèmes d'apprentissage automatique sont entraînés en utilisant des images comme celles-ci, des images qui ont été prises sur Internet ou d'institutions publiques sans contexte ni consentement. Elles représentent des histoires personnelles, des inégalités structurelles et toutes les injustices qui découlent de l'héritage du maintien de l'ordre et des systèmes carcéraux aux États-Unis. Mais la présomption que ces images peuvent d'une certaine façon servir de matériel apolitique et inerte influence de quelle manière et ce que « voit » un outil d'apprentissage automatique. Un système de vision par ordinateur peut détecter un visage ou un immeuble, mais non la raison pour laquelle une personne se trouvait dans un poste de police ni même le contexte social et historique entourant cet instant. Ultimement, les instances précises des données, une photographie d'un visage par exemple, ne sont pas considérées comme importantes dans l'entraînement d'un modèle d'IA. Tout ce qui compte est un ensemble suffisamment varié. Selon cette vision du monde, il existe toujours plus de données à recueillir du coffre-fort mondial et en croissance que sont Internet et les plateformes de médias sociaux.

Une personne se tenant devant un appareil photo dans une combinaison orange est alors déshumanisée en tant qu'une autre donnée. L'histoire de ces images, la manière dont elles ont été obtenues et leurs contextes institutionnels, personnels et politiques sont considérés comme étant sans importance. Les collections de photographies d'identité judiciaire sont utilisées comme toute autre source pratique d'images gratuites de visages bien éclairées, une référence pour faire en sorte que des outils, comme la reconnaissance faciale, fonctionnent. Et, dans un cycle sans fin, les visages des personnes décédées et des suspects, suspectes, prisonniers et prisonnières sont recueillis pour peaufiner les systèmes de reconnaissance faciale des services de police et de surveillance des frontières qui sont, quant à eux, utilisés pour surveiller et détenir encore plus de personnes.

La dernière décennie a été témoin d'une capture dramatique de matériel numérique pour la production de l'IA. Ces données sont le fondement de la recherche de sens en IA, non pas en tant que représentations classiques du monde ayant une signification individuelle, mais en tant que collection massive de données pour les abstractions et opérations automatiques. Cette capture de données à grande échelle est devenue si fondamentale au secteur de l'IA qu'elle n'est pas remise en question. Alors, comment en sommes-nous arrivés là ? Quelles façons de concevoir les données ont facilité cette évacuation du contexte, du sens et de la spécificité ? De quelle manière sont acquises, comprises et utilisées les données d'entraînement en apprentissage automatique ? De quelles manières les données

d'entraînement limitent-elles *ce* qu'interprète l'IA dans notre monde et *comment* elle l'interprète ? Quelles formes de pouvoir ces approches renforcent-elles et permettent-elles ?

Dans ce chapitre, je montre de quelle manière les données sont devenues le moteur de la réussite de l'IA et de sa mythologie et de quelle manière tout ce qui peut être facilement extrait est recueilli. Or, les implications plus profondes de cette approche courante sont rarement abordées, même si elles engendrent d'autres asymétries de pouvoir. L'industrie de l'IA a favorisé un genre de pragmatisme impitoyable, avec un minimum de contexte, de prudence ou de pratiques de données fondées sur le consentement, tout en promouvant l'idée que la collecte massive de données est nécessaire et justifiée pour créer des systèmes ayant une « intelligence » informatique profitable. Ceci a entraîné une métamorphose profonde où toutes les formes d'images, de textes, de sons et de vidéos ne sont que des données brutes pour les systèmes d'IA et les fins semblent justifier les moyens. Mais nous devrions demander : « Qui a profité le plus de cette transformation et pourquoi ces narratifs dominants sur les données persistent-ils ? » Et, comme nous l'avons constaté dans les chapitres précédents, la logique de l'extraction qui a façonné la relation à la terre et au labeur humain est aussi une caractéristique déterminante de la manière dont les données sont utilisées et comprises en IA. En regardant attentivement les données d'entraînement à titre d'exemple central dans un ensemble d'apprentissage automatique, nous pouvons commencer à prendre conscience de ce qui est en jeu dans cette transformation.

ENTRAÎNER DES MACHINES À VOIR

Il est utile de considérer pourquoi les systèmes d'apprentissage automatique exigent actuellement des quantités massives de données. Un exemple concret du problème est la vision par ordinateur, un sous-secteur de l'IA visant à apprendre aux machines à détecter et à interpréter des images. Pour des raisons qui sont rarement reconnues dans le secteur de l'informatique, le projet d'interpréter des images est une entreprise profondément complexe et relationnelle. Les images contiennent des choses remarquablement insaisissables, chargées de multiples significations potentielles, de questions sans réponse et de contradictions. Or, il est désormais pratique courante dans les premières étapes de la création d'un système de vision par ordinateur de moissonner des milliers, voire des millions, d'images d'Internet, de créer une série de catégories dans lesquelles les classer et d'utiliser ceci comme fondement quant à la manière dont ce système percevra la réalité observable. Ces vastes collections, nommées ensembles de données d'entraînement, constituent ce que les développeurs et développeuses en IA appellent souvent la « vérité terrain » (Jaton, 2017). La vérité a donc moins à voir avec une représentation factuelle ou une réalité convenue qu'avec, plus couramment, un amalgame d'images moissonnées des diverses sources en ligne disponibles.

Pour un apprentissage automatique supervisé, les ingénieures et ingénieurs humains fournissent des données d'entraînement étiquetées à l'ordinateur. Deux types distincts d'algorithmes entrent alors en jeu : les *apprenants* et les *classifieurs*. L'apprenant est l'algorithme qui est entraîné sur ces exemples de données étiquetées. Il informe ensuite le classifieur quant à la meilleure façon d'analyser la relation entre les nouveaux intrants et l'extrant ciblé souhaité (ou la prédiction). L'algorithme peut prédire si un visage est contenu dans une image ou si un courriel est, en fait, un pourriel. Plus le nombre d'exemples de données correctement étiquetées est grand, plus l'algorithme réussira à produire des prédictions précises. Il existe plusieurs types de modèles d'apprentissage automatique, dont les réseaux neuronaux, la régression logistique et les arbres de décision. Les ingénieurs et ingénieures choisissent un modèle en fonction de ce qu'ils et elles construisent, qu'il s'agisse d'un système de reconnaissance ou de moyens de détecter l'humeur dans les médias sociaux, et l'adaptent à leurs ressources informatiques.

Considérez la tâche de bâtir un système d'apprentissage automatique qui peut détecter la différence entre des photographies de pommes et d'oranges. D'abord, un développeur ou une développeuse

doit recueillir, étiqueter et entraîner un réseau neuronal de milliers d'images étiquetées de pommes et d'oranges. En ce qui a trait au logiciel, les algorithmes procèdent à un relevé statistique des images et développent un modèle pour reconnaître la différence entre ces deux classes. Si tout se passe comme prévu, le modèle entraîné sera en mesure de distinguer des images de pommes et d'oranges qu'il n'a jamais vues auparavant.

Or, si dans notre exemple, toutes les images utilisées lors de l'entraînement sont de pommes rouges et qu'aucune image ne représente une pomme verte, le système d'apprentissage automatique pourrait en déduire que « toutes les pommes sont rouges ». C'est ce qui est connu sous le nom de *raisonnement inductif*, une hypothèse ouverte fondée sur les données disponibles, plutôt qu'un *raisonnement déductif* qui, quant à lui, découle logiquement d'une prémisse (Nilsson, 2009, p. 398). Étant donné la manière dont ce système est entraîné, une pomme verte ne serait pas reconnue comme étant une pomme. Les ensembles de données d'entraînement sont donc au cœur de la manière dont les systèmes d'apprentissage automatique font des inférences. Ils servent de source principale de matériel sur lequel se fondent les systèmes d'IA pour faire leurs prédictions.

Les données d'entraînement définissent aussi bien plus que les fonctionnalités des algorithmes d'apprentissage automatique. Elles sont utilisées pour évaluer leur performance au fil du temps. Comme des pur-sang de grande valeur, les algorithmes d'apprentissage automatique sont constamment en compétition les uns contre les autres partout dans le monde afin de déterminer lesquels performent le mieux avec un ensemble de données en particulier. Ces ensembles de données de référence deviennent l'alphabet sur lequel une *lingua franca* est fondée, plusieurs laboratoires de divers pays convergeant vers des ensembles conformes qui tentent d'essayer de se surpasser les uns les autres. L'une des compétitions les plus reconnues est le ImageNet Challenge lors duquel les chercheurs et chercheuses compétitionnent pour voir laquelle des méthodes peut classer et détecter avec la plus grande précision des objets et des scènes[40].

Lorsque des ensembles de données d'entraînement ont été établis comme des références utiles, ils sont couramment adaptés, développés et élargis. Il en émerge une sorte de généalogie des ensembles de données d'entraînement. Ces derniers héritent de la logique apprise d'exemples antérieurs et, par la suite, donnent naissance à d'autres (Crawford, 2021, ch. 4). Par exemple, ImageNet s'appuie sur la taxonomie des mots hérités de WordNet, la base de données lexicale influente des années 1980, tandis que WordNet hérite de plusieurs sources, dont le Brown Corpus d'un million de mots, publié en 1961. Les ensembles de données d'entraînement reposent donc sur des classifications et des collections antérieures. Comme une encyclopédie en expansion, les formes antérieures demeurent et les nouveaux éléments leur sont ajoutés au fil des décennies.

Les données d'entraînement sont donc la fondation sur laquelle sont construits les systèmes d'apprentissage automatique contemporains[41] (Michalski, 1980). Ces ensembles de données façonnent les limites épistémiques gouvernant la manière dont fonctionne l'IA et, en ce sens, établissent les limites de la façon dont l'IA peut « voir » le monde. Mais les données d'entraînement sont une forme friable de la vérité terrain, et même les plus grandes collections de données ne peuvent échapper aux dérapages fondamentaux qui se produisent lorsqu'un monde infiniment complexe est simplifié et divisé en catégories.

40. Pour plus d'information, consultez « ImageNet Large Scale Visual Recognition Competition (ILSVRC) ». http://image-net.org/challenges/LSVRC/.

41. À la fin des années 1970, Ryszard Michalski a écrit un algorithme fondé sur des variables symboliques et des règles logiques. Ce langage a été populaire dans les années 1980 et 1990, mais, au fur et à mesure que les règles de prise de décisions et de qualification devenaient plus complexes, le langage a perdu de son utilité. Simultanément, la possibilité d'utiliser d'importants ensembles d'entraînement a déclenché un virage de ce regroupement conceptuel vers les approches d'apprentissage automatique contemporaines.

UNE BRÈVE HISTOIRE DE LA DEMANDE POUR LES DONNÉES

« Le monde est arrivé à une époque où des dispositifs complexes bon marché d'une grande fiabilité existent. Et quelque chose va forcément en ressortir. » Voilà les mots de Vannevar Bush, inventeur et administrateur, qui, à titre de directeur du Office of Scientific Research and Development, a supervisé le projet Manhattan et qui, plus tard, a joué un rôle essentiel dans la création de la National Science Foundation. C'était juillet 1945. Les bombes n'avaient pas encore été larguées sur Hiroshima et Nagasaki et Bush avait une théorie au sujet d'un nouveau type de système de connexion de données qui n'avait pas encore vu le jour. Il imaginait les « machines arithmétiques avancées de l'avenir » qui exécuteraient à très grande vitesse et qui « choisiraient leurs propres données et les manipuleraient conformément à des instructions ». Mais Bush était convaincu que les machines nécessiteraient de gigantesques quantités de données, disant que « de telles machines auront d'énormes appétits. L'une d'elles prendra les instructions et les données d'une salle entière de jeunes femmes munies de simples cartes perforées et fournira des feuilles de résultats calculés toutes les quelques minutes. Il y aura toujours beaucoup de choses à calculer dans les affaires détaillées de millions de personnes faisant des choses complexes » (Bush, 1945).

Les « jeunes femmes » auxquelles Bush faisait référence étaient les opératrices des perforatrices à clavier exécutant le travail de calcul quotidien. En tant qu'historiennes, Jennifer Light et Mar Hicks, ont démontré que ces femmes étaient souvent réduites à des dispositifs d'entrée de données pour des enregistrements de données intelligibles. Dans les faits, leur rôle était aussi important à la création de données et au fonctionnement des systèmes que celui des ingénieurs et ingénieures qui concevaient les ordinateurs numériques à cette époque (Light, 1999). Mais la relation entre les données et la machinerie qui les traiterait était déjà imaginée comme l'une de consommation sans fin. Les machines seraient de grandes consommatrices de données et il y aurait certainement un vaste horizon de matériel à extraire de millions de personnes.

Dans les années 1970, les chercheurs et chercheuses en IA exploraient principalement ce que l'on appelle une approche de systèmes experts, soit une programmation fondée sur des règles visant à réduire le champ d'actions possibles en articulant les formes du raisonnement logique. Mais la fragilité et le manque de sens pratique de cette approche sont rapidement devenus évidents dans un cadre réel où une règle établie pouvait rarement gérer l'incertitude et la complexité (Russell et Norvig, 2010, p. 546). De nouvelles approches étaient donc nécessaires. Dès le milieu des années 1980, les laboratoires de recherche se sont tournés vers des approches probabilistes et de force brute. En bref, ils utilisaient plusieurs cycles de calcul afin de calculer autant d'options que possible pour trouver le résultat optimal.

Un exemple important fut le groupe de reconnaissance de la voix d'IBM Research. Le problème de la reconnaissance de la voix avait principalement été abordé à l'aide de méthodes linguistiques, mais les théoriciens de l'information de l'époque, Fred Jelinek et Lalit Bahl, avaient créé un nouveau groupe qui comptait dans ses rangs Peter Brown et Robert Mercer (bien avant que ce dernier devienne un milliardaire associé au financement de Cambridge Analytica, de Breitbart News et de la campagne présidentielle de Donald Trump de 2016). Optant pour quelque chose de différent, leurs techniques ont ultimement produit des précurseurs pour les systèmes de reconnaissance de la voix sous-tendant Siri et Dragon Dictate ainsi que les systèmes de traduction automatique comme Google Translate et Microsoft Translator.

Ils ont commencé à utiliser des méthodes statistiques centrées davantage sur la fréquence d'apparition des mots en relation les uns aux autres plutôt que de tenter d'enseigner aux ordinateurs une approche fondée sur des règles à l'aide de principes grammaticaux ou de caractéristiques linguistiques. Pour que cette approche statistique fonctionne, il fallait une quantité énorme de vraies données de textes et de paroles, ou données d'entraînement. Le résultat, comme l'a écrit Xiaochang Li, érudit des médias, fut « qu'une réduction radicale de la parole à de simples données qui pouvaient être modelées

et interprétées sans connaissance ni compréhension de la linguistique a été nécessaire. La parole, *en tant que telle*, a cessé de compter ». Ce virage a été incroyablement important et est devenu un modèle répété pendant des décennies : la réduction du contexte aux données, de la signification à la reconnaissance des formes statistique. Li explique :

> Le fait de compter sur les données plutôt que les principes linguistiques présentait cependant un nouveau lot de défis, car cela signifiait que les modèles statistiques étaient nécessairement déterminés par les caractéristiques des données d'entraînement. Par conséquent, la taille des ensembles de données est devenue une préoccupation cruciale… De plus grands ensembles de données de résultats observés amélioraient non seulement la fiabilité des estimations pour un processus aléatoire, mais ils augmentaient aussi la probabilité que les données capturent un plus grand nombre de résultats se produisant rarement. La taille des données d'entraînement, en fait, était si cruciale à l'approche d'IBM qu'en 1985, Robert Mercer expliquait la perspective du groupe en déclarant simplement « qu'il n'y a pas de données comme plus de données » (Li, 2017, p. 143).

Pendant plusieurs décennies, il fut remarquablement difficile de mettre la main sur des données. Dans un entretien avec Li, Lalit Bahl disait « qu'à cette époque… vous ne pouviez même pas trouver facilement un million de mots en texte lisible par ordinateur. Et nous cherchions partout pour du texte » (Li, 2017, p. 144). L'équipe a essayé les manuels techniques d'IBM, les livres pour enfants, les brevets de technologie laser, les livres pour les personnes aveugles et même la correspondance écrite à la machine à écrire de Dick Garwin, fellow d'IBM, qui a créé la première conception de bombe à l'hydrogène (Brown et Mercer, 2013). Leur méthode faisait étrangement écho à une nouvelle de Stanislaw Lem, auteur de science-fiction, dans laquelle un homme appelé Trurl décide de construire une machine capable d'écrire de la poésie. Il commence par « huit cent vingt tonnes de livres sur la cybernétique et douze mille tonnes de la meilleure poésie » (Lem, 2003, p. 199). Or, Turl réalise que pour programmer une machine de poésie autonome, il est nécessaire de « répéter la totalité de l'Univers depuis le début, ou du moins une partie importante » (Lem, 2003, p. 199).

En fin de compte, le groupe de reconnaissance continue de la voix d'IBM a trouvé sa « bonne pièce » de l'univers d'une source improbable. En 1969, un important procès antitrust fédéral a été intenté contre IBM. La procédure a duré treize ans et près de mille témoins ont été cités. IBM a employé un personnel important juste pour numériser toutes les transcriptions des dépositions sur des cartes perforées Hollerith. Ceci a fini par créer un corpus de cent millions de mots dans le milieu des années1980. Mercer, notoirement antigouvernemental, l'a appelé « un cas d'utilité accidentellement créé par le gouvernement malgré lui » (Brown et Mercer, 2013).

IBM n'était pas le seul groupe à réunir des mots à la tonne. Entre 1989 et 1992, une équipe de linguistes, d'informaticiens et d'informaticiennes de l'Université de la Pennsylvanie travaillaient sur le projet Penn Treebank, une base de données annotée de textes. Ce groupe a recueilli quatre millions et demi de mots d'anglais américain à des fins d'entraînement de systèmes de traitement du langage naturel. Parmi ses sources figuraient des résumés du ministère de l'Énergie, des articles de Dow Jones Newswires et des rapports du Federal News Service sur des « activités terroristes » en Amérique du Sud (Marcus *et al.*, 1993). Les collections de textes émergentes s'appuyaient sur des collections antérieures auxquelles s'ajoutaient de nouvelles sources. Des généalogies de collections de données sont apparues, chacune fondée sur la dernière, et souvent important des particularités, des problèmes ou des omissions en gros.

Un autre corpus de textes classique est venu des enquêtes sur la fraude d'Enron Corporation après que la société a déclaré la plus importante faillite dans l'histoire des États-Unis. La Federal Energy Regulatory Commission avait saisi les courriels de 158 membres du personnel à des fins d'investigation juridique (Klimt et Yang, 2004) qu'elle avait décidé de publier en ligne parce que « le droit du public à la divulgation l'emporte sur le droit à la vie privée de la personne » (Wood III *et al.*, 2003, p. 12). Ce corpus est devenu une collection extraordinaire. Plus d'un demi-million d'échanges en langage courant pouvaient maintenant être utilisés comme une mine linguistique, une collection représentant néanmoins

les distorsions relatives au genre, à la race et aux antécédents professionnels de ces 158 travailleurs et travailleuses. Le corpus d'Enron a été cité dans des milliers d'articles universitaires. Malgré sa popularité, on s'y attarde attentivement rarement. Le *New Yorker* le décrit comme « un texte de recherche de confiance que personne n'a vraiment lu » (Heller, 2017). Cette construction de données d'entraînement et cette dépendance à elles annonçaient une nouvelle façon de faire. Ce corpus a transformé le domaine du traitement du langage naturel et a jeté les bases de ce qui deviendrait pratique courante en apprentissage automatique.

C'est à ce moment que les graines des éventuels problèmes ont été plantées. Les archives de textes étaient considérées comme des collections neutres de langage, comme s'il existait une équivalence générale entre les mots d'un manuel technique et ceux utilisés pour envoyer des courriels à des collègues. Tout texte pouvait être réorienté et était interchangeable, dans la mesure où il y en avait assez pour entraîner un modèle de langage pour prédire avec un haut degré de réussite quel mot pourrait en suivre un autre. Comme les images, les corpus de textes s'appuient sur la supposition que toutes les données d'entraînement sont interchangeables. Mais le langage n'est pas une substance inerte qui fonctionne de la même manière, quel que soit l'endroit d'où il provient. Des phrases tirées de Reddit seront différentes de celles composées par des cadres chez Enron. Distorsions, écarts et préjugés dans les textes recueillis sont intégrés dans le système plus large et, si un modèle de langage est fondé sur les types de mots qui sont regroupés ensemble, l'endroit d'où proviennent ces mots est important. Il n'existe pas de terrain neutre pour le langage et toutes les collections de textes reflètent aussi le temps, l'endroit, la culture et les politiques. Par ailleurs, les langues pour lesquelles les données sont moins disponibles ne sont pas servies par ces approches et sont souvent négligées (Baker *et al.*, 2009).

Évidemment, plusieurs histoires et contextes sont combinés dans les données d'entraînement d'IBM, dans les archives d'Enron et dans le Penn Treebank. Comment y distinguer ce qui, dans ces ensembles de données, est important ou non à comprendre ? Comment peut-on communiquer des avertissements comme « Cet ensemble de données reflète probablement des distorsions découlant du fait qu'il dépend d'articles de nouvelle sur des terroristes sud-américains des années 1980. » ? Dans un système, les origines des données sous-jacentes peuvent être incroyablement importantes et, malgré cela, trente ans plus tard, il n'existe pas de pratique normalisée pour noter d'où proviennent toutes ces données et de quelle manière elles ont été acquises, sans parler des préjugés et des politiques de classification que contiennent ces ensembles de données qui influenceront tous les systèmes qui dépendront d'eux (Gebru *et al.*, 2021 ; Mitchell *et al.*, 2019 ; Raji et Buolamwini, 2019).

SAISIR LE VISAGE

Alors que le texte lisible par ordinateur devenait très utile pour la reconnaissance vocale, le visage humain était au cœur des préoccupations pour la construction de systèmes de reconnaissance faciale. Un exemple important, financé par le Bureau de développement des programmes technologiques de lutte contre le trafic de drogue du ministère de la Défense des États-Unis, est émergé dans la dernière décennie du vingtième siècle. Ce Bureau avait commandité le programme de technologie de reconnaissance faciale connu sous l'acronyme FERET dans le but de mettre au point la reconnaissance faciale à des fins de renseignements et de mise en application des lois. Avant FERET, peu de données d'entraînement de visages humains étaient disponibles, à l'exception de quelques collections d'une cinquantaine de visages, ce qui est insuffisant pour effectuer de la reconnaissance faciale à l'échelle. Le laboratoire de recherche de l'armée américaine a mené le projet technique consistant à créer un ensemble d'entraînement de portraits de plus de mille personnes, en plusieurs poses. Finalement, l'ensemble comptait 14 126 images. Comme les collections de photographies d'identité judiciaire de NIST, FERET est devenu une référence standard, un outil de mesure commun permettant de comparer les approches à la détection de visages.

L'infrastructure FERET a été créée pour appuyer, une fois de plus, des tâches comme la recherche automatisée de photographies d'identité judiciaire, en plus de la surveillance dans les aéroports et les postes frontaliers, ainsi que la recherche dans les bases de données de permis de conduire à des fins de « détection des fraudes » (le fait qu'une personne soumette plusieurs demandes d'aide sociale était un exemple en particulier mentionné dans les documents de recherche du programme FERET) (Phillips *et al.*, 1996, p. 9). Deux scénarios de test principaux ont été employés. Dans le premier, un livre électronique de photographies d'identité judiciaire de personnes connues était présenté à un algorithme qui, ensuite, devait trouver les meilleures correspondances parmi un vaste ensemble. Le deuxième scénario portait plus particulièrement sur le contrôle aux aéroports et aux postes frontaliers et visait à reconnaître une personne connue, « trafiquant, terroriste ou autres criminels », parmi une grande population de personnes inconnues.

Dans leur conception, ces photographies sont lisibles par ordinateur et ne sont pas destinées à l'œil humain, bien qu'elles constituent un spectacle remarquable. Les images sont étonnamment belles, des photographies haute résolution saisies dans le style formel d'un portrait. Pris à l'aide d'appareils photo 35 mm à l'Université George Mason, les portraits aux cadrages serrés représentent un large éventail de personnes, certaines d'entre elles semblant s'être habillées pour l'occasion, avec leur coiffure soignée, des bijoux et du maquillage. La première série de photographies, prises entre 1993 et 1994, ressemble à une capsule témoin des coiffures et de la mode du début des années 1990. Les personnes devaient tourner leur tête pour adopter plusieurs positions. En feuilletant les images, on voit des images de profil, des images de face, divers degrés d'illumination et, parfois, différentes tenues. Certaines personnes ont été photographiées sur plusieurs années afin de commencer à étudier de quelle manière suivre des personnes au fur et à mesure qu'elles vieillissent. Chaque personne recevait de l'information sur le projet et devait signer une décharge de responsabilité approuvée par le comité d'éthique de l'Université. Les personnes savaient à quoi elles participaient et fournissaient leur plein consentement (Phillips *et al.*, 1996, p. 61). Ce degré de consentement deviendra une rareté au cours des années suivantes.

FERET était la référence d'un style formel de « création de données » avant que l'Internet ne commence à ouvrir la voie à une extraction massive, sans permission ni travail soigné de photographie. Par ailleurs, même à ce stade peu avancé, des problèmes liés au manque de diversité dans les visages recueillis existaient. Le document de recherche de FERET de 1996 admet que « certaines questions ont été soulevées concernant la distribution relative à l'âge, à la race et au genre de la base de données », mais « qu'à ce stade du programme, la question clé était la performance de l'algorithme sur une base de données contenant un nombre plus important de personnes » (Phillips *et al.*, 1996, p. 12). En effet, FERET était extraordinairement utile à cette fin. Avec l'intensification de la détection des terroristes et l'augmentation dramatique du financement de la reconnaissance faciale qui ont suivi les événements du 11 septembre, FERET est devenu la référence la plus couramment utilisée. À partir de ce moment, l'échelle et l'ambition des systèmes biométriques de suivi et des systèmes de vision automatisés ont connu une expansion rapide.

D'INTERNET À IMAGENET

À bien des égards, l'Internet a tout changé. Dans le domaine de la recherche en IA, il est apparu comme quelque chose de comparable à une ressource naturelle, là pour être utilisé. Au fur et à mesure qu'un plus grand nombre de personnes téléversaient leurs images sur des sites Web, sur des services de partage de photographies et, ultimement, sur des plateformes de médias sociaux, le pillage a commencé à être sérieux. Soudainement, les ensembles d'entraînement pouvaient atteindre la taille que les scientifiques des années 1980 n'auraient jamais pu imaginer. Plus besoin d'organiser des séances de prises de photos sous diverses conditions de lumière, de paramètres contrôlés et de dispositifs pour positionner le visage. Il y avait maintenant des millions d'égoportraits dans toutes les conditions

d'éclairage, les positions et les profondeurs de champ possibles. Les gens ont commencé à partager les photos de leurs enfants, des photos de famille et des images de ce à quoi ils ressemblaient il y a dix ans, une ressource idéale pour suivre les ressemblances génétiques et le vieillissement du visage. Chaque jour, des milliers de milliards de lignes de texte, contenant des formes de parole formelles et informelles, étaient publiées. C'était de l'eau au moulin pour l'apprentissage automatique. Et c'était vaste. Par exemple, lors d'une journée moyenne en 2019, environ 350 millions de photos étaient téléversées sur Facebook et 500 millions de tweets étaient publiés (Aslam, 2020). Et cela n'était que deux plateformes américaines. Tout et n'importe quoi pouvaient devenir un ensemble d'entraînement pour l'IA.

Les titans de l'industrie des technologies occupaient dorénavant une position de force : ils disposaient d'une source d'images et de textes indéfiniment renouvelée et, plus les gens partageaient leur contenu, plus le pouvoir de l'industrie des technologies augmentait. Les gens étiquetaient volontiers leurs photos en y ajoutant des noms et des emplacements, sans frais, et cette main-d'œuvre non payée fournissait des données étiquetées plus précises pour les modèles de vision par ordinateur et de langage automatique. Au sein de l'industrie, la valeur de ces collections est très grande. Il s'agit de trésors exclusifs qui sont rarement partagés en raison de questions de confidentialité et de l'avantage concurrentiel qu'ils représentent. Tant ceux au sein de l'industrie que ceux à l'extérieur d'elle, comme dans les laboratoires d'informatique de pointe en milieu universitaire, voulaient les mêmes avantages. Comment pouvaient-ils se permettre de recueillir les données des gens et s'assurer qu'elles étaient étiquetées à la main par des participants humains volontaires ? C'est de là que de nouvelles idées ont commencé à émerger : combiner des images et du texte extraits d'Internet avec la main-d'œuvre des travailleurs et travailleuses du nuage (« *crowdworkers* ») à bas salaire.

L'un des plus importants ensembles de données en IA est ImageNet, d'abord conceptualisé en 2006 lorsque la professeure Fei-Fei Li a décidé de bâtir un immense ensemble de données pour la reconnaissance d'objets. « Nous avons décidé que nous souhaitions faire quelque chose qui n'avait aucun précédent dans l'histoire. Nous allions cartographier l'ensemble du monde des objets », a dit Li (Gershgorn, 2017). L'affiche de cette recherche révolutionnaire a été publiée par l'équipe d'ImageNet lors d'une conférence sur la vision par ordinateur en 2009. Elle était présentée comme suit :

> L'ère numérique a engendré une énorme explosion de données. Les plus récentes estimations chiffrent le nombre de photos sur Flickr à trois milliards, ce qui correspond environ au nombre de vidéos sur YouTube et qui excède le nombre d'images dans la base de données Google Image Search. L'exploitation de ces images permet de proposer des modèles et des algorithmes plus sophistiqués et plus robustes menant à de meilleures applications pour que les utilisateurs et utilisatrices puissent indexer, récupérer, organiser et interagir avec ces données (Deng *et al.*, 2009).

Dès le départ, les données étaient caractérisées comme quelque chose de volumineux, de désorganisé, d'impersonnel et de prêt à être exploité. Selon les auteurs et auteures, « exactement comment de telles données pouvaient être utilisées et organisées était un problème qui restait à résoudre ». En extrayant des millions d'images d'Internet, principalement de moteurs de recherche en utilisant l'option de recherche d'images, l'équipe a produit une « ontologie d'images à grande échelle » qui devait servir de ressources pour « fournir les données d'entraînement et de référence cruciales » aux algorithmes de reconnaissance d'objets et d'images. Grâce à cette approche, ImageNet est devenu énorme. L'équipe a moissonné en masse plus de quatorze millions d'images d'Internet à être organisées en plus de vingt mille catégories. Des préoccupations éthiques concernant l'utilisation de données des gens n'ont été mentionnées dans aucun des documents de recherche de l'équipe, même si plusieurs milliers d'images étaient très personnelles et de nature compromettante.

Une fois les images moissonnées d'Internet, une préoccupation majeure a été soulevée : qui les étiquetterait toutes et les placerait dans des catégories intelligibles ? Comme Li le décrit, le plan initial de l'équipe était d'embaucher des étudiants et étudiantes de premier cycle à dix dollars l'heure pour qu'ils et elles trouvent manuellement les images et les ajoutent à l'ensemble de données (Gershgorn, 2017).

Compte tenu de son budget, elle a réalisé qu'il faudrait quatre-vingt-dix ans pour réaliser le projet. La réponse est venue lorsqu'un étudiant de Li lui a parlé d'un nouveau service : Amazon Mechanical Turk. Comme nous l'avons constaté dans le chapitre 2, cette plateforme distribuée signifiait qu'il était soudainement possible d'avoir accès à une main-d'œuvre distribuée pour effectuer des tâches en ligne, comme étiqueter et trier des images, à grande échelle et à peu de coûts. « Il m'a montré le site Web, et je peux vous dire littéralement que ce jour-là, j'ai su que le projet ImageNet allait se réaliser. Soudainement, nous avions trouvé un outil qui pourrait évoluer, ce qui était inimaginable dans l'embauche d'étudiants et d'étudiantes de premier cycle de Princeton », a dit Li (Gershgorn, 2017). Sans surprise, ces étudiants et étudiantes n'ont pas eu d'emploi.

Plutôt, ImageNet deviendrait, pour un temps, le plus important utilisateur universitaire de Mechanical Turk d'Amazon, déployant une armée de microtravailleurs et microtravailleuses pour trier en moyenne cinquante images par minute dans des milliers de catégories (Markoff, 2016). Des catégories existaient pour les pommes et les avions, les plongeurs et les sumos. Mais des étiquettes cruelles, offensantes et racistes avaient aussi été créées. Des photos de gens étaient classées dans des catégories comme « alcooliques », « hommes-singes et femmes-singes », « fous et folles », « prostitués et prostituées » et « yeux bridés ». Tous ces termes étaient importés de la base de données WordNet et fournis aux microtravailleurs et microtravailleuses pour qu'ils et elles les associent aux images. Pendant dix ans, ImageNet a grossi pour devenir un mastodonte pour la reconnaissance d'objets pour l'apprentissage automatique et une référence très puissante pour le secteur. L'approche d'extraction en masse de données sans consentement et d'étiquetage par des microtravailleurs et microtravailleuses sous-payés deviendrait pratique courante et des centaines de nouveaux ensembles de données d'entraînement suivraient dans les pas d'ImageNet. Comme nous le verrons dans le prochain chapitre, ces pratiques, et les données étiquetées qu'elles généraient, reviendraient éventuellement hanter le projet.

LA FIN DU CONSENTEMENT

Les premières années du vingt-et-unième siècle ont marqué la fin de la collecte de données fondée sur le consentement. En plus d'éliminer la nécessité d'organiser des séances photos, les personnes responsables de créer des ensembles de données ont présumé qu'elles pouvaient utiliser les contenus sur Internet, sans l'obtention d'un consentement, la signature de décharges de responsabilité ou d'examens éthiques. Des pratiques d'extraction encore plus troublantes ont alors commencé à émerger. Sur le campus Colorado Springs de l'Université du Colorado, par exemple, un professeur avait installé un appareil photo dans le couloir principal du campus, photographiant secrètement plus de mille sept cents étudiants, étudiantes et membres du corps professoral, le tout dans le but d'entraîner son système de reconnaissance faciale (Hernandez, 2019). Un projet semblable à l'Université Duke avait permis de recueillir à leur insu des enregistrements vidéo de plus de deux mille étudiants et étudiantes alors qu'ils et elles se déplaçaient entre leurs cours et les résultats ont été publiés sur Internet. L'ensemble de données, appelé DukeMTMC (pour reconnaissance faciale multicible, multicaméra), était financé par le Bureau de recherche de l'armée américaine (Zhang *et al.*, 2017).

Le projet DukeMTMC a été sévèrement critiqué après qu'un projet des artistes et chercheurs Adam Harvey et Jules LaPlace a montré que le gouvernement chinois utilisait ces images pour entraîner des systèmes de surveillance des minorités ethniques. Ceci a déclenché une enquête du comité d'éthique indépendant de Duke qui a déterminé qu'il s'agissait « d'une déviation importante » aux pratiques acceptables. L'ensemble de données a été retiré d'Internet (Satisky, 2019).

Mais, ce qui s'était produit à l'Université du Colorado et à Duke n'était certainement pas des cas isolés. À l'Université Stanford, des chercheurs et chercheuses avaient réquisitionné une webcaméra d'un café populaire de San Franscico pour en extraire près de douze mille images de « la vie de tous les jours dans

un café très fréquenté du centre-ville » sans le consentement de quiconque (Harvey et LaPlace, 2015). Encore et encore, des données extraites sans permission ni consentement étaient téléversées pour les chercheurs et chercheuses en apprentissage automatique qui les utilisaient ensuite comme infrastructure pour des systèmes automatisés d'imagerie.

Un autre exemple est MS-Celeb, l'emblématique ensemble de données d'entraînement de Microsoft qui, en 2016, a moissonné d'Internet environ dix millions de photos de centaines de milliers de célébrités. À cette époque, il s'agissait du plus grand ensemble de données public de reconnaissance faciale au monde et les gens qui y figuraient n'étaient pas uniquement des acteurs, actrices, politiciens et politiciennes de renom. On y trouvait aussi des journalistes, des activistes, des décideurs et décideuses, des universitaires et des artistes (Locker, 2019). Ironiquement, plusieurs personnes incluses dans l'ensemble de données sans leur consentement sont connues pour leur travail critiquant la surveillance et la reconnaissance faciale en soi, dont la documentariste Laura Poitras, l'activiste en droits numériques Jillian York, le critique Evgeny Morozov et l'auteure de *Surveillance Capitalism,* Shoshana Zuboff (Murgia et Harlow, 2019 ; Locker, 2019).[42]

Même lorsque les renseignements personnels sont éliminés des bases de données et que ces dernières sont publiées avec grande prudence, des personnes ont été tout de même identifiées ou des détails hautement sensibles à leur sujet ont été révélés. En 2013, par exemple, la Commission de service de taxi et de limousine de la ville de New York a publié un ensemble de données comprenant 173 millions de déplacements individuels en taxi, incluant l'endroit et l'heure à laquelle les personnes avaient été ramassées, puis déposées, les tarifs et le montant des pourboires. Le numéro du médaillon des chauffeurs de taxi avait été rendu anonyme, mais cela a rapidement été annulé, permettant ainsi aux chercheurs et chercheuses d'en déduire des renseignements sensibles comme les salaires annuels et les adresses de domicile (Franceschi-Bicchierai, 2015). Ces données, une fois combinées à des renseignements publics de sources comme des blogues de célébrités, ont permis d'identifier certains acteurs, actrices, politiciens et politiciennes et de déduire les adresses de personnes fréquentant des bars à strip-tease (Tockar, 2014). Au-delà des préjudices individuels, de tels ensembles de données engendrent aussi des « préjudices prédictifs à la vie privée » pour des groupes ou des communautés dans leur ensemble (Crawford and Schultz, 2019). Par exemple, ce même ensemble de données sur les taxis de New York a été utilisé pour suggérer quels chauffeurs de taxi étaient des musulmans pratiquants en observant quand ils s'arrêtaient aux heures de prière (Franceschi-Bicchierai, 2015).

De cet ensemble de données qui, à prime à bord, semble anonyme et anodin peut être extraites des formes d'information très personnelle et inattendue, mais cette réalité n'a pas freiné la collecte d'images et de textes. Puisque la réussite en apprentissage automatique en est venue à dépendre d'ensembles de données de plus en plus grands, un plus grand nombre de personnes tentent de les acquérir. Mais pourquoi le domaine de l'IA plus large accepte-t-il cette pratique, malgré les problèmes éthiques, politiques et épistémologiques et les préjudices potentiels qu'elle engendre ? Quels croyances, justifications et incitatifs économiques ont normalisé cette extraction massive et cette équivalence générale des données ?

42. Lorsque le *Financial Times* a exposé les contenus de cet ensemble de données, Microsoft l'a retiré d'Internet et un porte-parole de Microsoft a déclaré qu'il avait simplement été retiré « parce que le défi de recherche avait été atteint » (Murgia et Harlow, 2019).

MYTHES ET MÉTAPHORES DES DONNÉES

Souvent citée, l'histoire de l'intelligence artificielle écrite par le professeur Nils Nilsson décrit plusieurs des mythes fondateurs concernant les données en apprentissage automatique. Ce dernier illustre clairement de quelle manière les données sont généralement décrites dans les disciplines techniques : « L'impressionnant volume de données brutes exige des techniques "d'exploration des données" efficaces afin de classer, de quantifier et d'extraire des informations utiles. Les méthodes d'apprentissage automatique jouent un rôle de plus en plus grand dans l'analyse des données, puisqu'elles peuvent traiter d'énormes quantités de données. En fait, plus il y a de données, mieux c'est. » (Nilsson, 2009, p. 495)

Faisant écho à Robert Mercer qui l'avait déclaré plusieurs décennies avant lui, Nilsson percevait que les données étaient partout et prêtes à être utilisées et que cela était parfait pour une classification massive par des algorithmes d'apprentissage automatique (Bowker, 2005, 184-85)[43]. Il s'agissait d'une croyance si commune qu'elle est devenue axiomatique : les données étaient là pour être acquises, épurés et valorisées.

Mais, au fil du temps, des intérêts particuliers ont fabriqué et appuyé avec soin cette croyance. Comme le notent les sociologues Marion Fourcade et Kieran Healy, l'injonction de toujours recueillir des données ne venait pas uniquement des professions dans le domaine des données, mais aussi de leurs institutions et des technologies qu'elles mettaient en œuvre :

> L'ordre institutionnel émanant de la technologie est le plus puissant de tous : nous faisons ces choses *parce que nous le pouvons*… Les professionnels et professionnelles recommandent, l'environnement institutionnel exige et la technologie permet aux organisations de ramasser autant de données individuelles que possible. Peu importe que les quantités recueillies excèdent largement la portée imaginative ou la compréhension analytique d'une firme. La supposition est qu'elles seront éventuellement utiles, c'est-à-dire qu'elles auront de la valeur… Les organisations contemporaines sont à la fois interpellées par l'impératif des données et puissamment équipées de nouveaux outils pour les utiliser (Fourcade et Healy, 2016).

Ceci a produit un genre d'impératif moral pour recueillir des données afin d'améliorer les systèmes, quels que soient les impacts négatifs que la collecte de données puisse éventuellement avoir. Derrière la croyance discutable voulant que « plus, c'est mieux » se cache l'idée que les personnes peuvent entièrement être connues une fois qu'un nombre suffisant de pièces de données sont recueillies (Meyer et Jepperson, 2000). Mais qu'est-ce qui compte comme donnée ? L'historienne Lisa Gitelman note que chaque discipline et institution « dispose de ses propres normes et standards pour l'imagination des données » (Gitelman, 2013, p. 3). Au vingt-et-unième siècle, les données sont devenues tout ce qui peut être saisi.

Des termes comme « exploration des données » et des phrases comme « les données sont le nouveau pétrole » font partie d'un changement de rhétorique qui a transformé la notion de données de quelque chose de personnel, d'intime ou d'assujetti à une propriété et à un contrôle individuels en quelque chose d'inerte et de non humain. Les données ont commencé à être décrites comme une ressource

43. Et, comme Geoff Bowker nous le rappelle notoirement, « *Raw data is both an oxymoron and a bad idea ; to the contrary, data should be cooked with care.* » (Les données brutes sont un oxymore et une mauvaise idée : au contraire, les données devraient être utilisées avec soins. – traduction libre)

à consommer, un flux à contrôler ou un investissement à saisir[44]. L'expression « les données comme du pétrole » s'est répandue et, bien qu'elle dépeigne les données comme une matière à extraire, elle a rarement été utilisée pour souligner le coût associé aux industries de l'extraction pétrolière et minière : le travail sous contrat, les conflits géopolitiques, l'épuisement des ressources et les conséquences dépassant l'échelle de temps humaine.

Ultimement, le mot « données » est devenu froid, recelant à la fois ses origines matérielles et ses fins. Et, si les données sont considérées comme abstraites et immatérielles, elles peuvent donc plus facilement être exclues de la compréhension et des responsabilités traditionnelles en matière d'attention, de consentement ou de risque. Comme Luke Stark et Anna Lauren Hoffman, chercheur et chercheuse, en conviennent, les métaphores des données en tant que « ressource naturelle » là attendant à être découverte sont un truc rhétorique bien établi utilisé par les puissances coloniales (Stark et Hoffmann, 2019). L'extraction est justifiée si elle provient d'une source primitive et « brute »[45]. Si les données sont dépeintes comme du pétrole, attendant d'être extraites, l'apprentissage automatique est désormais considéré comme son processus de raffinage nécessaire.

On a aussi commencé à considérer les données comme du capital, conformément aux visions néolibérales plus larges des marchés en tant que formes principales d'organisation de la valeur. Une fois que les activités humaines sont exprimées par le biais de traces numériques, puis comptabilisées et classées selon des critères de notation, elles fonctionnent en tant que moyen d'extraire de la valeur. Comme Fourcade et Healy le notent, ceux et celles qui disposent des bons signaux de données obtiennent un avantage comme un rabais sur une assurance et un rang plus élevé dans les marchés (Fourcade et Healy, 2016, p. 19)[46]. Les personnes performantes dans l'économie générale ont tendance aussi à réussir dans une économie qui note les données, alors que ceux et celles qui sont les plus démunis deviennent la cible des formes les plus nuisibles de la surveillance et de l'extraction des données. Lorsque les données sont considérées comme une forme de capital, tout est justifié si cela signifie d'en recueillir davantage. Le sociologue Jathan Sadowski soutient aussi que les données opèrent dorénavant comme une forme de capital. Il suggère qu'une fois que tout est compris sous l'angle des données, cela justifie un cycle d'extraction des données toujours plus grand : « La collecte de données est donc stimulée par le cycle perpétuel de l'accumulation du capital qui, à son tour, encourage le capital à construire un monde dans lequel tout est fait de données et à en dépendre. La supposée universalité des données recadre tout comme relevant du domaine du capitalisme des données. Tous les espaces doivent être soumis à la mise en données. Si l'on conçoit l'univers comme une réserve potentiellement infinie de données, cela signifie donc que l'accumulation et la circulation de données peuvent être soutenues pour toujours. » (Sadowski, 2019, p. 8)

Ce désir d'accumuler et de faire circuler est la puissante idéologie sous-tendant les données. L'extraction massive de données est la « nouvelle frontière de l'accumulation et la prochaine étape dans le capitalisme », suggère Sadowski et est la couche fondamentale qui permet à l'IA de fonctionner

44. Plusieurs universitaires ont regardé de près le travail que ces métaphores font. Cornelius Puschmann et Jean Burgess, professeurs en études des médias, ont analysé les métaphores courantes sur les données et ont noté deux catégories généralisées : les données « comme force naturelle à contrôler et [les données] comme ressource à consommer » (Puschmann et Burgess, 2014). Tim Hwang et Karen Levy, chercheur et chercheuse, suggèrent que la description des données comme « le nouveau pétrole » laisse sous-entendre un coût élevé d'acquisition, mais suggère aussi la possibilité de « gros bénéfices pour ceux et celles qui ont les moyens de les extraire » (Hwang et Levy, 2015).

45. Spécialistes des médias, Nick Couldry et Ulises Mejías appellent cela le « colonialisme des données » qui est ancré dans les pratiques prédatrices historiques du colonialisme, mais associé aux méthodes informatiques contemporaines (et embrouillé par elles). Cependant, comme d'autres spécialistes l'ont montré, cette terminologie est à double tranchant, puisqu'elle cache les préjudices réels et persistants du colonialisme (Couldry et Mejías, 2019a ; 2019b ; Segura et Waisbord, 2019).

46. Ils font référence à cette forme de capital sous le nom de « ubercapital ».

(Sadowski, 2019, p. 9). Par conséquent, il existe donc des industries entières, des institutions et des personnes qui ne souhaitent pas que cette frontière, où les données sont là prêtes à être extraites, soit remise en question ou déstabilisée.

Les modèles d'apprentissage automatique nécessitent des flux de données continus pour devenir plus précis. Or, les machines sont asymptotiques, n'atteignant jamais une précision complète, ce qui nourrit la justification pour une plus grande extraction d'un plus grand nombre possible de personnes pour alimenter les raffineries de l'IA. Ceci a créé un mouvement d'abandon des idées comme les « sujets humains », un concept ayant émergé des débats sur l'éthique du vingtième siècle, et la création de « personnes concernées », des agglomérations de points de données sans subjectivité, ni contexte, ni droits clairement définis.

L'ÉTHIQUE À DISTANCE

La très grande majorité des travaux de recherche universitaire sur l'IA est faite sans processus d'évaluation éthique. Mais si les techniques d'apprentissage automatique sont utilisées pour éclairer les décisions dans des domaines délicats comme l'éducation et les soins de santé, alors pourquoi ne sont-elles pas soumises à une évaluation plus rigoureuse ? Pour le comprendre, nous devons nous pencher sur les disciplines précurseurs de l'intelligence artificielle. Avant l'émergence de l'apprentissage automatique et de la science des données, les domaines des mathématiques appliquées, de la statistique et de l'informatique n'avaient historiquement pas été considérés comme des formes de recherche sur des sujets humains.

Durant les premières décennies de l'IA, le risque associé aux travaux de recherche utilisant les données humaines était généralement considéré comme étant minimal[47]. Même si les ensembles de données provenaient souvent de personnes et représentaient leurs vies, les travaux de recherche utilisant ces ensembles de données étaient plutôt considérés comme une forme de mathématiques appliquées ayant peu de conséquences pour les sujets humains. Les infrastructures de protection éthique, comme les comités d'éthique indépendants dans les universités, avaient accepté cette position depuis des années (Federal Register, 2015). Au départ, cela était sensé. Les comités d'éthique indépendants s'étaient principalement concentrés sur les méthodes courantes utilisées dans les expériences biomédicales et psychologiques dans lesquelles les interventions représentaient des risques clairs pour les personnes. L'informatique était considérée comme étant beaucoup plus abstraite.

Lorsque, dans les années 1980 et 1990, l'IA a quitté les contextes de laboratoire et a fait son entrée dans des situations concrètes, comme tenter de prédire quels criminels récidiveront ou qui devrait recevoir des prestations sociales, les préjudices potentiels sont devenus plus nombreux. En outre, en plus des personnes, ces préjudices touchent des communautés entières. Mais il existe toujours une forte présomption que des ensembles de données disponibles publiquement présentent des risques minimaux et, par conséquent, devraient être exempts d'évaluation éthique (Metcalf et Crawford, 2016). Cette idée a été héritée d'une époque antérieure, lorsqu'il était plus difficile de déplacer des données d'un endroit à un autre et très coûteux de les stocker sur de longues périodes. Ces suppositions antérieures ne sont pas en phase avec ce qui se passe actuellement en apprentissage automatique. Aujourd'hui, il est plus facile de relier les ensembles de données, ils sont infiniment réutilisables, constamment mis à jour et fréquemment retirés du contexte de la collecte de données.

47. Je m'appuie ici sur l'histoire de l'examen des sujets humains et des études de données à grande échelle coécrite avec Jake Metcalf. Consultez Metcalf et Crawford (2016).

Le profil de risque de l'IA change rapidement, au fur et à mesure que les outils deviennent plus invasifs et que les chercheurs et chercheuses sont de plus en plus capables d'accéder à des données sans interagir avec leurs sujets. Par exemple, un groupe de chercheurs et chercheuses en apprentissage automatique ont publié un article dans lequel ils soutiennent avoir conçu un « système automatique de classification des crimes » (Seo *et al.*, 2018). Plus particulièrement, ils ont cherché à savoir si un crime violent était lié à un gang, soutenant que leur réseau neuronal pouvait le prédire en utilisant seulement quatre informations : l'arme, le nombre de suspects, le quartier et l'emplacement. Ils et elles y sont arrivés en utilisant un ensemble de données sur les crimes du Service de police de Los Angeles qui contenait des milliers de crimes ayant été étiquetés par la police comme étant liés aux gangs.

Les données sur les gangs sont notoirement biaisées et truffées d'erreurs et, malgré cela, les chercheurs et chercheuses ont utilisé cette base de données et d'autres semblables comme source définitive pour entraîner des systèmes d'IA prédictifs. On a démontré, par exemple, que la base de données CalGang, qui est largement utilisée par la police en Californie, contient des inexactitudes majeures. La vérificatrice de l'État a découvert que vingt-trois pour cent des centaines d'entrées qu'elle a examinées n'étaient pas étayées adéquatement pour être incluses dans la base de données. Cette dernière contenait aussi quarante-deux enfants en bas âge, dont vingt-huit ayant « admis être membre d'un gang » (California State Auditor, 2016). La plupart des adultes figurant dans la liste n'avaient jamais été inculpés, mais une fois inclus dans la base de données, il est impossible de retirer leur nom. Les raisons justifiant leur inclusion pouvaient être aussi simples que d'avoir parlé avec un voisin en portant un chandail rouge. En utilisant de telles justifications insignifiantes, les personnes noires et latino-américaines ont été ajoutées de manière disproportionnée à la liste (Libby, 2016).

Lorsque les chercheurs et chercheuses ont présenté leur projet de prédiction des crimes liés aux gangs à une conférence, certaines personnes présentes ont été troublées. Comme rapporté par *Science,* les questions du public comprenaient : « Comment l'équipe peut-elle être certaine que les données d'entraînement ne sont pas biaisées ? » et « Que se passe-t-il lorsqu'une personne est mal étiquetée comme membre d'un gang ? » Hau Chan, informaticien maintenant à l'Université Harvard ayant présenté les travaux, avait répondu qu'il ne pouvait pas savoir à quoi servirait l'outil. « [Il s'agit du] genre de questions éthiques auxquelles je ne sais répondre adéquatement », avait-il répondu, ajoutant qu'il n'était « qu'un chercheur ». Un membre du public lui a répondu en citant les paroles de la chanson satirique de Tom Lehrer au sujet de Wernher von Braun, un scientifique concepteur de fusées de la Deuxième Guerre mondiale : « *Once the rockets are up, who cares where they come down ?* » (Une fois les fusées lancées, qui se soucie d'où elles retombent ? – traduction libre) (Hutson, 2018)

Cette séparation des questions éthiques des aspects techniques reflète un problème plus large dans le domaine, où la responsabilité des préjudices n'est pas reconnue ou est considérée comme étant au-delà de la portée des travaux de recherche. Comme Anna Lauren Hoffman l'écrit : « Le problème ici ne se limite pas aux bases de données biaisées ni aux algorithmes injustes et aux conséquences accidentelles. Il reflète aussi un problème plus persistant, soit le fait que des chercheurs et chercheuses reproduisent activement des idées qui nuisent aux communautés vulnérables et qui renforcent les injustices actuelles. Même si le système proposé par l'équipe de Harvard pour caractériser la violence liée aux gangs n'est jamais mis en œuvre, des dommages n'ont-ils pas déjà été causés ? Leur projet n'était-il pas un acte de violence culturelle en soi ? » (Hoffmann, 2018). La mise de côté des questions éthiques est nuisible en elle-même et elle perpétue l'idée fausse que les travaux de recherche sont menés dans un vacuum, sans responsabilité pour les idées qu'ils propagent.

La reproduction d'idées nuisibles est particulièrement dangereuse maintenant que l'IA n'est plus une discipline expérimentale utilisée uniquement dans des laboratoires, mais bien testée sur des millions de personnes. Les approches techniques peuvent évoluées rapidement d'articles présentés dans une conférence à une mise en œuvre dans des systèmes de production où les suppositions nuisibles peuvent devenir enracinées et difficiles à inverser.

L'apprentissage automatique et les méthodes en science des données peuvent créer une relation abstraite entre les chercheurs et chercheuses et les sujets dans laquelle les travaux sont faits à distance, éloignés des communautés et des personnes à risque de préjudices. Cette relation de distance entre les chercheurs et chercheuses en IA et les personnes dont les vies sont reflétées dans les ensembles de données est une pratique de longue date. En 1976, lorsque Joseph Weizenbaum, scientifique en IA, a écrit sa critique cinglante sur le domaine, il avait observé que l'informatique cherchait déjà à échapper aux contextes humains (Weizenbaum, 1976, p. 266). Il faisait valoir que les systèmes de données avaient permis aux scientifiques durant la guerre d'opérer une distance psychologique par rapport aux gens « qui seraient mutilés et tués par les systèmes d'armement qui découleraient des idées qu'ils communiquaient » (Weizenbaum, 1976, p. 275-76). La réponse, selon Weizenbaum, était de s'attaquer directement à ce que les données représentent réellement : « Par conséquent, la leçon est que le ou la scientifique ou technologue doit, par des actes de volonté et d'imagination, s'efforcer activement de réduire de telles distances physiques afin de contrer les forces qui tendent à l'isoler des conséquences de ses actes. Il ou elle doit, c'est aussi simple que cela, penser à ce qu'il ou elle est vraiment en train de faire. » (Weizenbaum, 1976, p. 276)

Weizenbaum espérait que les scientifiques et technologues penseraient plus attentivement aux conséquences de leurs travaux et quelles personnes ou quels groupes pourraient être à risque. Mais cela ne deviendrait pas la norme dans le domaine de l'IA. Plutôt, les données sont plus couramment considérées comme étant quelque chose pris à sa guise, utilisé sans restriction et interprété sans contexte. Il existe une culture internationale avide en matière de collecte de données qui peut être exploiteuse et invasive et qui peut produire des formes durables de préjudices[48]. Et il existe plusieurs industries, institutions et personnes qui ont de très bonnes raisons pour maintenir cette attitude de colonisation, c'est-à-dire que les données sont là pour être recueillies, et elles ne souhaitent pas remettre en question ou réglementer cette attitude.

LA SAISIE DES COMMUNS

La culture répandue actuelle de l'extraction des données continue à prendre de l'expansion malgré des préoccupations concernant la vie privée, l'éthique et la sécurité. En consultant les milliers d'ensembles de données qui sont librement accessibles pour le développement de l'IA, j'ai eu un aperçu de ce que les systèmes techniques sont destinés à reconnaître et de comment le monde est présenté pour les ordinateurs d'une manière que les humains voient rarement. Il existe des ensembles de données gigantesques, remplies d'égoportraits de personnes, de tatouages, de parents se promenant avec leurs enfants, de gestes de la main, de personnes conduisant leur voiture, de personnes commettant des crimes sur des télévisions en circuit fermé et des centaines d'activités humaines quotidiennes, comme s'assoir, faire un signe de la main, lever un verre ou pleurer. Toutes les formes de données biographiques, dont les données médicolégales, biométriques, sociométriques et psychométriques, sont saisies et ajoutées à des bases de données pour que des systèmes d'IA puissent trouver des tendances et mener des évaluations.

Les ensembles de données d'entraînement soulèvent des questions complexes du point de vue éthique, méthodologique et épistémologique. Plusieurs d'entre eux ont été créés à l'insu des personnes et sans leur consentement, les données étant recueillies de sources en ligne comme Flickr, des recherches d'images sur Google et YouTube ou offertes par des agences gouvernementales comme le FBI. Ces données sont désormais utilisées pour élargir les systèmes de reconnaissance faciale, moduler les taux

48. Pour en savoir plus sur l'histoire de l'extraction des données et des points de vue des communautés marginalisées, consultez Costanza-Chock (2020) et D'Ignazio et Klein (2020).

d'assurance de soins de santé, pénaliser les chauffeurs distraits et alimenter les outils prédictifs de la police. Mais les pratiques d'extraction des données s'étendent même plus profondément dans des aspects de la vie humaine qui étaient autrefois interdits ou trop coûteux à atteindre. Les compagnies technologiques ont tiré profit d'une variété d'approches pour gagner du terrain. Les données vocales sont recueillies d'appareils posés sur les comptoirs de cuisine ou les tables de nuit. Les données physiques proviennent de montres sur les poignets et de téléphones dans les poches. Les données au sujet des livres et des journaux qui sont lus proviennent de tablettes et d'ordinateurs portables. Les gestes et les expressions faciales sont compilés et évalués dans les milieux de travail et les salles de classe.

La collecte des données des gens pour construire des systèmes d'IA soulève des préoccupations claires en matière de vie privée. Prenez, par exemple, l'entente conclue entre le Royal Free National Health Service Foundation Trust du Royaume-Uni et DeepMind, la filiale de Google, visant le partage des dossiers de patient de 1,6 million de personnes. Au Royaume-Uni, le National Health Service est une institution révérée, chargée d'offrir à tous et à toutes des soins de santé qui sont principalement gratuits, tout en assurant la sécurité des données des patients. Mais, lorsque l'entente avec DeepMind a été étudiée, il a été démontré que l'entreprise avait violé les lois en matière de protection de la vie privée en n'informant pas adéquatement les patients (Revell, 2017). Dans ses conclusions, la commissionnaire à l'information a observé que « le prix de l'innovation ne doit pas être l'érosion des droits fondamentaux à la vie privée » (Information Commissioner's Office, 2017).

Or, d'autres questions sérieuses reçoivent moins d'attention que le respect de la vie privée. Les pratiques d'extraction des données et de construction d'ensembles de données d'entraînement sont fondées sur une saisie commercialisée de ce qui faisait autrefois partie des communs. Cette forme particulière d'érosion est la privatisation à la dérobée, une extraction de la valeur des connaissances de biens collectifs. Un ensemble de données peut toujours être publiquement disponible, mais la métavaleur des données, le modèle qui est créé à partir d'elle, est détenue par des intérêts privés. Bien sûr, plusieurs bonnes choses peuvent être faites avec des données publiques. Mais il y a eu une attente sociale, et dans une certaine mesure, technique voulant que la valeur des données partagées par les institutions publiques et les espaces publics en ligne doive être retournée au bien collectif dans d'autres formes de communs. Plutôt, nous voyons quelques entreprises privées qui disposent dorénavant d'un pouvoir énorme leur permettant d'extraire des informations et des profits de ces sources. La nouvelle ruée vers l'or de l'IA consiste à englober différents champs de connaissances, de sentiments et d'actions humains, tous les types de données disponibles, tous saisis dans une logique d'expansion de collecte sans fin. Nous en sommes rendus à un pillage de l'espace public.

Fondamentalement, les pratiques d'accumulation de données sur plusieurs années ont contribué à une puissante logique extractive, une logique qui est désormais une caractéristique essentielle de la manière dont fonctionne le domaine de l'IA. Cette logique a enrichi les compagnies technologiques disposant des bassins de données les plus importants tandis que les espaces sans collecte de donnée sont devenus dramatiquement plus rares. Comme l'avait prévu Vannevar Bush, l'appétit des machines est énorme. Mais comment elles sont nourries et ce qui leur est fourni a un impact énorme sur la façon dont elles interpréteront le monde et les priorités de leurs maîtres façonneront toujours la manière dont cette vision sera monétisée. En regardant les couches des données d'entraînement utilisées pour façonner et informer les modèles et les algorithmes d'IA, nous constatons que la collecte et l'étiquetage des données sur le monde sont des interventions sociales et politiques, même si elles se font passer pour des interventions purement techniques.

La façon dont les données sont comprises, saisies, classées et nommées est fondamentalement un acte de création et de confinement du monde. Ses ramifications quant à la façon dont l'intelligence artificielle fonctionne dans le monde et à l'égard des communautés les plus touchées sont énormes. Le mythe de la collecte de données en tant que pratique bienveillante en informatique a embrouillé ses opérations de pouvoir, protégeant ceux et celles qui en profitent le plus et évitant la responsabilité de ses conséquences.

RÉFÉRENCES

Aslam, S. 2020. Facebook by the Numbers (2019) : Stats, Demographics & Fun Facts, Omnicore. https://www.omnicoreagency.com/facebook-statistics/

Baker, J.M. *et al.* 2009. Research Developments and Directions in Speech Recognition and Understanding, Part 1, *IEEE Signal Processing Magazine*, 26(3), pp. 75-80.

Bowker, G.C. 2005. Memory Practices in the Sciences. Cambridge, MA : MIT Press.

Brown, P. et Mercer, R. 2013. Oh, Yes, Everything's Right on Schedule, Fred. *Twenty Years of Bitext Workshop, Empirical Methods in Natural Language Processing Conference*, Seattle, Washington, Octobre. http://cs.jhu.edu/~post/bitext

Bush, V. 1945. "As We May Think.", juillet. https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/.

California State Auditor. 2016. The CalGang Criminal Intelligence System. 2015-130. Sacramento, CA. https://www.auditor.ca.gov/pdfs/reports/2015-130.pdf

Costanza-Chock, S. 2020. Design Justice : Community – Led Practices to Build the World We Need. Cambridge, MA : MIT Press.

Couldry, N. et Mejías, U.A. 2019a. Data Colonialism : Rethinking Big Data's Relation to the Con-temporary Subject, *Television and New Media*, 20(4), pp. 336-349. https://doi.org/10.1177/1527476418796632

Couldry, N. et Mejías, U.A. 2019b. The Costs of Connection : How Data Is Colonizing Human Life and Appropriating It for Capitalism. Stanford, CA : Stanford University Press.

Crawford, K. 2021. The Atlas of AI : Power, Politics, and the Planetary Costs of Artificial Intelligence. New Haven : Yale University Press.

Crawford, K. et Schultz, J. 2019. AI Systems as State Actors, *Columbia Law Review*, 119(7). https://columbialawreview.org/content/ai-systems-as-state-actors/

Curry, S. *et al.* 2009. NIST Special Database 32 : Multiple Encounter Dataset I (MEDS-I). NISTIR 7679. National Institute of Standards and Technology. https://nvlpubs.nist.gov/nistpubs/Legacy/IR/nistir7679.pdf

Deng, J. *et al.* 2009. ImageNet : A Large – Scale Hierarchical Image Database, in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 248-255. doi :https://doi.org/10.1109/CVPR.2009.5206848

D'Ignazio, C. et Klein, L.F. 2020. Data Feminism. Cambridge, MA : MIT Press.

Ever AI. 2018. Ever AI Leads All US Companies on NIST's Prestigious Facial Recognition Vendor Test, *Globe News wire*, 27 novembre. https://www.globenewswire.com/news-release/2018/11/27/1657221/0/en/Ever-AI-Leads-All-US-Companies-on-NIST-s-Prestigious-Facial-Recognition-Vendor-Test.html

Federal Register. 2015. Federal Policy for the Protection of Human Subjects. https://www.federalregister.gov/documents/2015/09/08/2015-21756/federal-policy-for-the-protection-of-human-subjects

Founds, A.P. *et al.* 2011. *NIST Special Database 32 : Multiple Encounter Dataset II (MEDS-II)*. NISTIR 7807. National Institute of Standards and Technology. https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=908383

Fourcade, M. et Healy, K. 2016. Seeing Like a Market, *Socio-Economic Review*, 15(1), pp. 9-29. https://doi.org/10.1093/ser/mww033

Franceschi-Bicchierai, L. 2015. Redditor Cracks Anonymous Data Trove to Pinpoint Muslim Cab Drivers, *Mashable*, 28 janvier. https://mashable.com/2015/01/28/redditor-muslim-cab-drivers/

Garris, M.D. et Wilson, C.L. 2005. NIST Biometrics Evaluations and Developments. NISTIR 7204. National Institute of Standards and Technology (NIST). https://www.govinfo.gov/content/pkg/GOVPUB-C13-1ba4778e3b87bdd6ce660349317d3263/pdf/GOVPUB-C13-1ba4778e3b87bdd6ce660349317d3263.pdf

Gebru, T. *et al.* 2021. Datasheets for Datasets' *arXiv:1803.09010* [Preprint]. https://arxiv.org/abs/1803.09010

Gershgorn, D. 2017. The Data That Transformed AI Research—and Possibly the World, *Quartz*, 26 juillet. https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/

Gitelman, L. 2013. "Raw Data" Is an Oxymoron. Cambridge, MA : MIT Press.

Grother, P. *et al.* 2017. The 2017 IARPA Face Recognition Prize Challenge (FRPC). NISTIR 8197. National Institute of Standards and Technology (NIST), p. 26. https://nvlpubs.nist.gov/nistpubs/ir/2017/NIST.IR.8197.pdf

Harvey, A. and LaPlace, J. 2015. Brainwash Dataset, *MegaPixels*. https://megapixels.cc/brainwash/

Heller, N. 2017. What the Enron Emails Say about Us, *New Yorker*, 17 juillet. https://www.newyorker.com/magazine/2017/07/24/what-the-enron-e-mails-say-about-us

Hernandez, E. 2019. CU Colorado Springs Students Secretly Photo-graphed for Government – Backed Facial – Recognition Research, *Denver Post*, 27 mai. https://www.denverpost.com/2019/05/27/cu-colorado-springs-facial-recognition-research/

Hoffmann, A.L. 2018. Data Violence and How Bad Engineering Choices Can Damage Society, *Medium*, 30 avril. https://medium.com/s/story/data-violence-and-how-bad-engineering-choices-can-damage-society-39e44150e1d4

Hutson, M. 2018. Artificial Intelligence Could Identify Gang Crimes—and Ignite an Ethical Firestorm, *Science*, 28 février. https://www.sciencemag.org/news/2018/02/artificial-intelligence-could-identify-gang-crimes-and-ignite-ethical-firestorm

Hwang, T. et Levy, K. 2015. « The Cloud » and Other Dangerous Meta-phors, *Atlantic*, 20 janvier. https://www.theatlantic.com/technology/archive/2015/01/the-cloud-and-other-dangerous-metaphors/384518/

ImageNet Large Scale Visual Recognition Competition (ILSVRC) (no date). https://www.image-net.org/challenges/LSVRC/

Information Commissioner's Office. 2017. Royal Free– Google DeepMind Trial Failed to Comply with Data Protection Law. *Information Commissioner's Office*, 3 juillet. https://ico.org.uk/about-the-ico/news-and-events/news-and-blogs/2017/07/royal-free-google-deepmind-trial-failed-to-comply-with-data-protection-law/

Jaton, F. 2017. We Get the Algorithms of Our Ground Truths : Designing Referential Databases in Digital Image Processing. *Social Studies of Science*, 47(6), pp. 811-840. https://doi.org/10.1177/0306312717730428

Klimt, B. et Yang, Y. 2004. The Enron Corpus : A New Dataset for Email Classification Research, in Boulicat, J.-F. et al. (eds) *Machine Learning: ECML 2004*. Berlin : Springer, pp. 217-226.

Lem, S. 2003. The First Sally (A), or Trurl's Electronic Bard, in Gunn, J. (ed.) *The Road to Science Fiction*. Lanham.

Li, X. 2017. Divination Engines : A Media History of Text Prediction. Ph.D. New York University.

Libby, S. 2016. Scathing Audit Bolsters Critics' Fears about Secretive State Gang Database, *Voice of San Diego*, 11 août. https://www.voiceofsandiego.org/topics/public-safety/scathing-audit-bolsters-critics-fears-secretive-state-gang-database/

Light, J.S. 1999. When Computers Were Women. *Technology and Culture*, 40(3), pp. 455-483.

Locker, M. 2019. Microsoft, Duke, and Stanford Quietly Delete Databases with Millions of Faces, *Fast Company*, 6 juin. https://www.fastcompany.com/90360490/ms-celeb-microsoft-deletes-10m-faces-from-face-database

Marcus, M.P., Marcinkiewicz, M.A. et Santorini, B. 1993. Building a Large Annotated Corpus of English : The Penn Treebank, *Computational Linguistics*, 19(2), pp. 313-330.

Markoff, J. 2016. Pentagon Turns to Silicon Valley for Edge in Artificial Intelligence, *New York Times*, 11 mai. https://www.nytimes.com/2016/05/12/technology/artificial-intelligence-as-the-pentagons-latest-weapon.html

Metcalf, J. et Crawford, K. 2016. Where Are Human Subjects in Big Data Research ? The Emerging Ethics Divide, *Big Data and Society*, 3(1), pp. 1-14. https://doi.org/10.1177/2053951716650211

Meyer, J.W. et Jepperson, R.L. 2000. The "Actors" of Modern Society : The Cultural Construction of Social Agency, *Sociological Theory*, 18(1), pp. 100-120. https://doi.org/10.1111/0735–2751.00090

Michalski, R.S. 1980. Pattern Recognition as Rule – Guided Inductive Infer-ence, *IEEE Transactions on Pattern Analysis Machine Intelligence*, 2(4), pp. 349-361. https://doi.org/10.1109/TPAMI.1980.4767034

Mitchell, M. *et al.* 2019. Model Cards for Model Reporting, in. *FAT* '19 : Proceedings of the Conference on Fairness, Accountability, and Transparency*, Atlanta : ACM Press, pp. 220-229. https://doi.org/10.1145/3287560.3287596

Murgia, M. et Harlow, M. 2019. Who's Using Your Face ? The Ugly Truth about Facial Recognition, *Financial Times*, 19 avril. https://www.ft.com/content/cf19b956-60a2-11e9-b285-3acd5d43599e

National Institute of Standards and Technology (NIST). 2010. Special Database 32—Multiple Encounter Dataset (MEDS). https://www.nist.gov/itl/iad/image-group/special-database-32-multiple-encounter-dataset-meds

Nilsson, N.J. 2009. The Quest for Artificial Intelligence : A History of Ideas and Achievements. New York : Cambridge University Press.

Phillips, J.P., Rauss, P.J. et Der, S.Z. 1996. FERET (Face Recognition Technology) Recognition Algorithm Development and Test Results. ARL-TL-995. Adelphi, M.D. : Army Research Laboratory. https://apps.dtic.mil/dtic/tr/fulltext/u2/a315841.pdf

Puschmann, C. et Burgess, J. 2014. Big Data, Big Questions : Metaphors of Big Data, *International Journal of Communication*, 8, pp. 1690-1709.

Raji, I.D. et Buolamwini, J. 2019. Actionable Auditing : Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products, in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. *AAAI/ACM Conference on AI, Ethics, and Society*, pp. 429-435.

Revell, T. 2017. Google DeepMind's NHS Data Deal "Failed to Comply" with Law", *New Scientist*, 3 juillet. https://www.newscientist.com/article/2139395-google-deepminds-nhs-data-deal-failed-to-comply-with-law/

Russell, A. 2014. Open Standards and the Digital Age : History, Ideology, and Networks. New York : Cambridge University Press.

Russell, S.J. et Norvig, P. 2010. Artificial Intelligence : A Modern Approach. 3e édition. Upper Saddle River, N.J. : Pearson.

Sadowski, J. 2019. When Data Is Capital : Datafication, Accumulation, and Extraction, *Big Data and Society*, 6(1), pp. 1-12. https://doi.org/10.1177/2053951718820549

Satisky, J. 2019. A Duke study recorded thousands of students' faces. Now they're being used all over the world, *Duke Chronicle*, 12 juin. http://shorturl.at/gmq09

Segura, M.S. et Waisbord, S. 2019. Between Data Capitalism and Data Citizenship, *Television and New Media*, 20 (4), pp. 412-419.

Sekula, A. 1986. The Body and the Archive, MIT Press, 39(octobre), pp. 3-64. https://doi.org/10.2307/778312

Seo, S. *et al.* 2018. Partially Generative Neural Networks for Gang Crime Classification with Partial Information, in *proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. *roceAAAI/ACM Conference on AI, Ethics, and Society*, pp. 257-263.

Stark, L. et Hoffmann, A.L. 2019. Data Is the New What ? Popular Metaphors and Professional Ethics in Emerging Data Culture, *Journal of Cultural Analytics*, 1(1). https://doi.org/10.22148/16.036

Tockar, A. 2014. Riding with the Stars : Passenger Privacy in the NYC Taxi-cab Dataset, 15 septembre. https://agkn.wordpress.com/2014/09/15/riding-with-the-stars-passenger-privacy-in-the-nyc-taxicab-dataset/

Weizenbaum, J. 1976. Computer Power and Human Reason : From Judgment to Calculation. San Francisco, CA : W. H. Freeman.

Wood III, P., Massey, W.M. et Brownell, N.M. 2003. *F*ERC Order Directing Release of Information. Federal Energy Regulatory Commission. https://www.caiso.com/Documents/FERCOrderDirectingRelease–InformationinDocketNos_PA02–2–000_etal__Manipulation–ElectricandGasPrices_.pdf

Zhang, Z. *et al.* 2017. Multi-Target, Multi-Camera Tracking by Hierarchical Clustering : Recent Progress on DukeMTMC Project, *arXiv :1712.09531* [Preprint]. https://arxiv.org/abs/1712.09531

ÉCOSYSTÈMES D'INNOVATION POUR UNE IA BÉNÉFIQUE SUR LE PLAN SOCIAL

YOSHUA BENGIO

Professeur à l'Université de Montréal, fondateur et directeur scientifique de Mila – Institut québécois d'intelligence artificielle, directeur scientifique d'IVADO – Institut pour la valorisation des données, codirecteur du programme Apprentissage automatique, apprentissage biologique de CIFAR et membre de l'OBVIA.

ALLISON COHEN

Responsable des projets d'IA appliquée, IA pour l'humanité, Mila – Institut québécois d'intelligence artificielle.

BENJAMIN PRUD'HOMME

Directeur exécutif, IA pour l'humanité, Mila – Institut québécois d'intelligence artificielle.

AMANDA LEAL DE LIMA ALVES

Chercheuse, IA pour l'humanité, Mila – Institut québécois d'intelligence artificielle.

NOAH ODER

Candidat à la maîtrise à l'Université McGill et Sciences Po Paris

---

ODD 4 - Education de qualité

ODD 5 - Egalité entre les sexes

ODD 9 - Industrie, innovation et infrastructure

ODD 10 - Inégalités réduites

ODS 17 - Partenariats pour la réalisation des objectifs

ÉCOSYSTÈMES D'INNOVATION POUR UNE IA BÉNÉFIQUE SUR LE PLAN SOCIAL

RÉSUMÉ

Le potentiel perturbateur que confère la technologie de l'intelligence artificielle (IA) préoccupe les gouvernements, ces derniers souhaitant maximiser la création de croissance économique ou minimiser le risque de violations des droits. Les gouvernements et les institutions internationales ont, par conséquent, concentré leurs efforts sur le financement de l'industrie de l'IA en général ou sur la lutte contre les applications nuisibles. Cependant, ces approches ne permettent pas d'accorder suffisamment d'attention à la manière dont l'IA peut contribuer à des découvertes bénéfiques sur le plan social dans des domaines aussi cruciaux que la découverte de médicaments, la lutte aux changements climatiques et l'éducation. Miser sur l'impact social au moment d'investir et de développer des écosystèmes d'innovation reste un chaînon manquant dans le paysage du développement et de la gouvernance de l'IA et empêche les gouvernements de mettre en œuvre des politiques publiques qui, autrement, favoriseraient une innovation en IA qui serait importante sur le plan social.

Ce chapitre a pour but de promouvoir le potentiel de l'IA à contribuer à un changement social important. Des recommandations visant à soutenir un écosystème d'innovation favorisant les projets d'IA pour le bien social sont aussi présentées. Les sept recommandations proposées ont pour objectifs principaux : i) de permettre un engagement éclairé et hautement compétent dans le domaine de l'IA, ii) de promouvoir une collaboration multidisciplinaire dans la chaîne de valeur du développement de l'IA et iii) de récompenser les acteurs qui contribuent à cet écosystème d'innovation.

INTRODUCTION

L'avènement de l'intelligence artificielle (IA) offre à la société de nombreuses possibilités bénéfiques sur le plan social, notamment en augmentant la rapidité des décisions dérivées des données et en réduisant leur coût (sans oublier qu'elle ouvre la voie à des produits et services entièrement nouveaux et fondés sur les données). L'écosystème dans lequel se développe actuellement l'IA comporte, cependant, plusieurs obstacles qui limitent la réalisation de ce potentiel. Dans le cadre économique actuel, les décisions des acteurs sont guidées par la recherche de profits, ce qui se traduit par une innovation insuffisante dans les domaines à fort impact social positif, mais à faible valeur économique, un phénomène que l'on peut comparer à une « tragédie des communs » (Llyod, 1833) ou à un scénario d'échec du marché.

Comme le suggère ce chapitre, les gouvernements devraient encourager la création d'écosystèmes qui comblent les lacunes en matière d'IA bénéfique sur le plan social. Ces écosystèmes devraient être conçus pour améliorer le bien-être des citoyens et citoyennes et réduire la pression sur les services sociaux, des soins de santé à l'éducation. Les gouvernements peuvent développer ces écosystèmes en utilisant des lignes directrices, des normes et des cadres incitatifs qui influencent le développement de l'IA en catalysant son application de manière à ce qu'elle profite plus largement et durablement à la société. Mais, pour réaliser ce potentiel, le secteur public devra jouer un rôle plus actif dans l'orientation du développement de l'IA.

Comme nous l'explorerons tout au long de ce chapitre, les principaux aspects dans lesquels peut s'impliquer le secteur public comprennent des mesures incitatives stratégiques qui favorisent la recherche et développement d'une IA bénéfique sur le plan éthique et social. Pour bénéficier de ces mesures incitatives, les parties prenantes devraient se soumettre à certaines conditions, notamment la science ouverte, qui sont conçues pour accélérer le progrès des technologies bénéfiques et leur mise en œuvre. Cette stratégie devrait permettre de réorienter l'industrie de l'IA afin d'améliorer la probabilité de création d'outils fondés sur l'IA répondant à des besoins sociaux urgents et d'accroître leur prévalence. Qui plus est, cette stratégie peut remédier au chaînon manquant actuel dans l'approche qu'adoptent les gouvernements en matière d'innovation, à savoir, le fait de miser sur des retombées sociales positives.

Les idées exprimées dans ce chapitre devraient être appliquées de manière à respecter le contexte unique de chaque pays. Les recommandations présentées visent à guider le secteur public afin que toutes les parties prenantes, y compris le secteur privé et la société civile, puissent s'efforcer de créer un écosystème dans lequel l'IA bénéfique sur le plan social est activement encouragée. Compte tenu de la théorie économique de la tragédie des communs et des tendances actuelles au sein de l'industrie de l'IA, nous soutenons qu'un tel développement dépend d'un engagement important du secteur public. Les sections suivantes présentent la signification de « l'IA pour le bien social », la pertinence de l'implication des gouvernements dans les premières étapes du développement de l'IA dans le monde et des recommandations pour fournir le chaînon manquant actuel en matière de leadership des gouvernements dans le domaine de l'IA, un domaine qui connaît une croissance rapide.

LA SIGNIFICATION DE « L'IA POUR LE BIEN COMMUN »

Afin de clarifier les types d'applications promues dans ce chapitre, il est important, en premier lieu, de définir ce que l'on entend par IA pour le bien social. La définition qui articule le mieux notre notion du concept stipule que les projets d'IA pour le bien social impliquent des systèmes d'IA qui sont conçus, développés, mis en œuvre, surveillés et évalués afin de « (i) prévenir, atténuer ou résoudre des problèmes ayant un impact négatif sur la vie humaine ou le bien-être du monde naturel ou (ii) permettre des développements socialement préférables ou écologiquement durables, tout en (iii) n'introduisant pas de nouvelles formes de préjudice ou en n'amplifiant pas les disparités et les inégalités existantes »

(Cowls *et al.*, 2021). Dans ce contexte, nous pouvons nous tourner vers les 17 objectifs de développement durable (ODD) des Nations Unies comme cadre utile pour catégoriser les domaines qui sont considérés comme bénéfiques sur le plan social à l'échelle mondiale. Ces objectifs ont été approuvés par les 193 États membres des Nations Unies et ont été établis en tant que plan directeur pour catalyser le progrès économique, social et environnemental.

LA TRAJECTOIRE QUE SUIT ACTUELLEMENT L'IA

Le cadre économique actuel, caractérisé par la prédominance des forces du marché et la dépendance des États envers des investisseurs privés pour trouver des projets de recherche et développement qui valent la peine d'être réalisés, a, jusqu'à présent, joué un rôle déterminant dans la définition de la structure d'incitation qui guide le développement de l'IA. Malheureusement, ce cadre priorise la création d'outils d'IA qui sont principalement alignés sur les rendements économiques, ne répondant aux besoins sociaux que lorsqu'il est rentable de le faire.

Alors que les dernières décennies ont vu une tendance croissante de l'engagement du secteur privé dans des initiatives sociales (Porter et Kramer, 2006, p.3), cette programmation a largement lieu en marge des activités commerciales principales et n'atteint généralement pas l'échelle nécessaire pour relever plusieurs défis à long terme importants pour nos sociétés. Les outils d'IA actuellement développés présentent trois lacunes communes, à savoir : i) des échecs imprévus, ii) des occasions ratées et iii) des interventions lourdes (Cowls *et al.*, 2021). Ces lacunes sont des manifestations de la structure incitative du développement de l'IA, une structure qui a orienté la technologie dans le sens du profit, et ce, au détriment du bien social.

Pour ce qui est des échecs imprévus, des technologies qui ne sont pas fondées sur le « bien social » sont mises en œuvre avec des conséquences sociales imprévues et parfois néfastes. Un bon exemple est l'agent conversationnel « Tay » de Microsoft lancé sur Twitter en 2016. Tay avait été conçu pour apprendre des utilisatrices et utilisateurs humains et générer son propre contenu dans le style d'une adolescente. L'objectif des développeurs était d'apprendre à Tay à avoir des conversations avec des humains sur presque tous les sujets. Cependant, dans les 24 heures suivant son lancement, l'agent conversationnel a été retiré d'Internet parce qu'il avait commencé à publier des gazouillis racistes, misogynes, homophobes et autrement offensants (Schwartz, 2019). Cet exemple s'inscrit dans une longue liste d'applications d'IA qui créent des préjudices qui ne sont pas atténués de manière proactive par les développeurs en IA et qui représentent donc des échecs imprévus.

Outre les résultats négatifs, il existe un nombre important d'occasions ratées liées à l'utilisation et à la mise en œuvre de l'IA dans des contextes où les retombées sociales ne sont pas la priorité. Par exemple, des outils d'IA sont en cours de développement dans le secteur de la santé pour détecter le cancer de la peau chez les patients. Or, compte tenu de l'asymétrie des ressources, des données et des mesures incitatives commerciales, l'un de ces outils a donné de bons résultats pour les personnes à la peau claire, mais de mauvais résultats pour les personnes à la peau plus foncée (Adamson et Smith, 2018). En créant ces outils sans accorder une importance centrale à la diversité et à l'inclusion, nous ratons des occasions d'améliorer la qualité des soins fournis aux communautés marginalisées et vulnérables qui, compte tenu de la disparité qui existe déjà dans les soins de santé, sont souvent les plus à même de bénéficier de cette nouvelle technologie.

Enfin, en ce qui concerne les interventions lourdes, les outils d'IA sont conçus avec des objectifs et des résultats qui ne génèrent aucune retombée positive claire à la société. En fait, ces outils peuvent parfois être à l'origine de graves préjudices sociaux. Par exemple, un outil d'IA a été développé pour détecter, avec grande précision, si une personne était homosexuelle (Wang et Kosinski, 2017). Ce type d'outil

permet de surveiller les gens en utilisant des informations très personnelles qu'ils peuvent choisir ou non de divulguer publiquement. Ces informations peuvent être divulguées pour maltraiter, ostraciser ou nuire de toute autre manière aux membres de la communauté LGBTI. Les interventions lourdes sont particulièrement préoccupantes dans un scénario où les acteurs développant l'IA peuvent, à leur propre discrétion, décider de créer, de diffuser et de maintenir des interventions d'IA problématiques.

Afin de s'attaquer aux structures de marché qui permettent à ces types d'outils d'IA problématiques de proliférer, plusieurs personnes au sein de l'industrie se tournent, et à juste titre, vers la réglementation. Les risques qu'entraînent les outils d'IA problématiques non réglementés et sous-réglementés sont préoccupants. Les politiques, les programmes et les initiatives qui sont créés en réponse à cette situation sont d'une importance cruciale et nous permettent d'espérer que l'industrie sera moins susceptible de se développer de manière nuisible à l'avenir. Toutefois, ce chapitre ne se veut pas une contribution à cet important corpus de travaux. Son intention est plutôt de sensibiliser les gens à une nouvelle approche que peuvent adopter les gouvernements pour guider le développement de l'IA. Il s'agit d'une approche dans laquelle l'impact social positif est priorisé pour permettre à certaines applications d'IA de proliférer, en particulier celles qui profitent à la société et qui ne sont actuellement pas attrayantes pour les acteurs du marché.

L'innovation en IA pour le bien social n'est pas actuellement exploitée à son plein potentiel. En fait, comme l'a démontré l'approche actuelle de laissez-faire dans l'industrie de l'IA, il est peu probable que le marché développe des technologies bénéfiques sur le plan social ayant la taille et la portée nécessaires pour contribuer à relever certains de nos défis mondiaux les plus difficiles, allant des changements climatiques à l'éducation et à la santé. Ainsi, pour développer l'industrie de l'IA plus rapidement et d'une manière plus robuste et bénéfique sur le plan social, les gouvernements doivent s'impliquer plus activement pour orienter le développement de l'IA vers la résolution de défis sociaux importants.

Pour illustrer comment le système actuel n'est pas conçu pour maximiser les retombées positives pour la société, nous pouvons prendre l'exemple de la résistance aux antimicrobiens (RAM) et du pipeline de découverte de médicaments qui y est associé. Le cas de la RAM démontre comment le capitalisme de laissez-faire ne parvient pas à générer de la recherche et développement dans des domaines d'importance critique pour la société, des domaines où l'innovation pourrait autrement sauver le monde de défis complexes et importants.

Depuis les années 1950, les antibiotiques ont révolutionné les soins de santé, sauvant directement et indirectement d'innombrables vies (en permettant des interventions chirurgicales sûres, par exemple). Cependant, les bactéries que les antibiotiques sont censés combattre ont commencé à se défendre. En effet, par un processus évolutif, les bactéries finissent par muter en souches résistantes aux médicaments antimicrobiens. Ces souches résistantes peuvent proliférer en raison de l'inefficacité des antibiotiques. En 2019, la résistance aux antimicrobiens a été associée au décès de 4,95 millions de personnes. Parmi ces cas, la RAM était la cause directe du décès de 1,27 million de personnes, un nombre qui devrait augmenter (Murray *et al*., 2022 ; Organisation mondiale de la santé, 2019b). Si aucun changement n'est apporté à notre utilisation des antibiotiques ni au pipeline actuel de découverte de médicaments, on prévoit que les bactéries résistantes aux antibiotiques causeront 10 millions de décès par an d'ici 2050 (Review on Antimicrobial Resistance, 2014). À titre de comparaison, la pandémie de COVID-19 est, jusqu'à présent, responsable de la mort d'environ 5,6 millions de personnes (Organisation mondiale de la santé, 2022).

Compte tenu du coût humain associé à la prolifération de bactéries résistantes aux antibiotiques et de l'impact économique qui en découlerait, la valeur sociale des technologies visant à prévenir et à atténuer la RAM devrait également être énorme. Si elle n'est pas atténuée, on estime que la RAM entraînera une baisse du PIB mondial de 2 à 3,5 % par an, ce qui, cumulé jusqu'en 2050, représente une diminution de 60 à 100 billions de dollars dans les échanges de biens et de services à l'échelle mondiale (Review on

Antimicrobial Resistance, 2014). Gardez à l'esprit que ces estimations pourraient être prudentes. Il est possible qu'une souche soit si mortelle et transmissible qu'elle pourrait menacer l'ensemble de l'espèce humaine, sans parler de l'ordre économique et de l'organisation sociale.

Une question évidente se pose donc : pourquoi les entreprises pharmaceutiques n'investissent-elles pas simplement dans la recherche et développement d'antibiotiques efficaces contre les nouvelles souches de bactéries actuelles et futures ? Bien que ces travaux de recherche puissent sauver un nombre incalculable de vies, d'importantes sommes d'argent et peut-être même l'ordre social tel que nous le connaissons, ils ne sont pas rentables dans les conditions actuelles du marché. En fait, pour que les entreprises pharmaceutiques rentabilisent leurs investissements, la demande pour leurs médicaments doit être forte. Or, dans le cas de la résistance aux antimicrobiens, le nombre de personnes initialement infectées par une souche mutée ne représente souvent qu'un faible pourcentage de la population infectée (Plackett, 2020). Ainsi, même si de nouveaux antibiotiques empêchent la propagation d'une nouvelle souche, ils freinent simultanément la demande pour ce nouveau médicament en empêchant cette souche de se multiplier.

En outre, les médecins prescrivent à juste titre les antibiotiques existants comme première ligne de défense pour retarder l'apparition de mutations entraînant une résistance aux nouveaux antibiotiques, ce qui réduit encore la taille du marché des nouveaux antibiotiques. Par conséquent, ce scénario de marché n'est pas suffisamment intéressant pour les entreprises développant des médicaments dont les profits sont normalement directement proportionnels au nombre de consommatrices et consommateurs potentiels, ce nombre devenant donc un critère important dans l'évaluation de la rentabilité potentielle d'un médicament. Ceci est particulièrement le cas dans le contexte du développement d'antibiotiques qui, contrairement à d'autres catégories de médicaments, sont vendus à de très bas prix. En fait, historiquement, il n'y a eu de place sur le marché que pour un seul médicament rentable par infection bactérienne (McKenna, 2020). Par conséquent, il n'y a pas assez de recherche et développement sur des médicaments qui seraient efficaces contre les mutations mortelles avant qu'il ne soit trop tard (Organisation mondiale de la santé, 2019a). Puisque le développement d'un nouveau médicament peut prendre une décennie, l'ironie est que ces médicaments *finiraient* par être développés et distribués à grande échelle après que la souche mutée aurait proliféré, mais seulement une fois que le coût humain, social, politique et économique aurait été important. Ainsi, pour relever de façon préventive ce défi sociétal, l'intervention des gouvernements dans la recherche et développement menant à la découverte de médicaments en vue d'atteindre ces objectifs socialement importants est essentielle[49].

Le cas de la résistance aux antimicrobiens démontre de quelle manière un décalage important peut exister entre les besoins sociaux et les rendements financiers. Dans ce cas, les marchés peuvent échouer à générer les produits dont la société a si désespérément besoin, tant sur le plan économique que sur le plan humain. Il ne faut donc pas compter uniquement sur les marchés pour inventer la technologie dont nous avons besoin en tant que société. Plutôt, le secteur public doit être responsable de stimuler la recherche et développement de manière très efficace du point de vue de l'impact social.

En ce qui concerne les nouvelles applications d'IA, leur potentiel peut être exploré dans le contexte de la découverte de médicaments, étant donné leur capacité prometteuse à accélérer le processus de recherche et développement, qui prend actuellement 10 ans en moyenne, et à réduire le coût de la découverte de médicaments, qui se chiffre actuellement en milliards de dollars (PhRMA, 2015). Puisque l'IA peut explorer un volume beaucoup plus important de médicaments candidats dans l'espace moléculaire, cette technologie peut, en outre, contribuer à la découverte de médicaments plus efficaces.

49. Les gouvernements ont investi dans des organismes de recherche, notamment le Fonds multipartenaire contre la résistance aux antimicrobiens (AMR MPTF), le Partenariat mondial sur la recherche-développement en matière d'antibiotiques (GARDP) et le Fonds d'action contre la RAM. En outre, des gouvernements comme ceux de la Suède, de l'Allemagne et des États-Unis pilotent des modèles de remboursement pour financer l'innovation dans la recherche sur la RAM.

Pourtant, il existe au moins trois obstacles majeurs à la réalisation du potentiel de l'IA. Premièrement, il y a la question de la disponibilité des données. Les ensembles de données ont souvent une portée limitée et les entreprises ne les mettent pas à la disposition du public, principalement dans le but de protéger leurs investissements de la concurrence. Deuxièmement, l'accès à l'expertise en IA est encore insuffisant, notamment au sein des entreprises en démarrage et dans les pays du Sud, où un éventail plus large d'études et d'applications novatrices pourraient être autrement explorées. Troisièmement, la taille et la portée limitées de la plupart des laboratoires de recherche universitaires constituent un obstacle, étant donné qu'ils pourraient potentiellement apporter des contributions importantes à l'écosystème de la découverte de médicaments en créant des ensembles de données en interne, entre autres capacités. Contrairement aux entreprises pharmaceutiques, les laboratoires universitaires fonctionnent de manière ascendante, les étudiants et étudiantes diplômés et les professeurs et professeures disposant d'une grande liberté pour entreprendre les recherches de leur choix, ce qui est excellent pour la recherche exploratoire de base, mais moins efficace lorsqu'il s'agit de recherche et développement orientée vers une mission. D'autre part, le processus de recherche et développement industrielle est plus descendant afin de s'adapter aux objectifs stratégiques des entreprises, une approche qui s'est avérée efficace pour convertir les idées de départ en produits. En gardant ces exemples à l'esprit, les sections suivantes donnent un aperçu de la manière dont les gouvernements peuvent s'engager dans le développement de l'IA afin d'éliminer chacun des obstacles qui empêchent l'adoption de l'IA pour une utilisation bénéfique sur le plan social.

LA TRAGÉDIE DU CADRE DES COMMUNS

Établie en 1833 par l'économiste britannique William Forster Lloyd, la tragédie des communs est un concept qui peut mettre en lumière la nécessité d'une implication stratégique des gouvernements pour stimuler des résultats bénéfiques sur le plan social au sein du marché. Lloyd s'est demandé ce qui se passerait si chaque agriculteur, agissant dans son propre intérêt, autorisait son bétail à paître sur une parcelle de prairie commune. En l'absence de règles convenues collectivement sur la manière dont les agriculteurs collaboreraient pour maintenir la prairie au fil du temps, la parcelle de terre s'épuiserait rapidement. En effet, en l'absence de règles communes, la consommation d'herbe deviendrait un jeu à somme nulle (Llyod, 1833). En conséquence, les agriculteurs seraient incités à continuer d'envoyer leur bétail paître jusqu'à épuisement, même si l'épuisement de la prairie commune conduisait ultimement à la destruction de la ressource et, au bout du compte, n'était dans l'intérêt de personne. La tragédie des communs pourrait être résolue grâce à des règles convenues collectivement qui influenceraient les décisions des personnes dans la mesure où elles commenceraient à agir d'une manière plus alignée sur les intérêts du groupe.

Sur le marché de l'IA, où les règles du jeu ne sont pas bien définies (LaCroix et Mohseni, 2021; Benkler, 2019), les intérêts des entreprises, intrinsèquement axées sur les profits, prévalent. En conséquence, certains biens publics sont non seulement épuisés, leur création et leur maintien sont aussi découragés, ce qui entraîne des résultats sociaux sous-optimaux. Cette situation est particulièrement préoccupante compte tenu de l'incroyable potentiel de l'IA en matière d'utilité sociale. Pour que les principales parties prenantes commencent à produire une IA plus bénéfique sur le plan social, les gouvernements doivent réécrire la structure d'incitation qui régit la prise de décisions des parties prenantes dans ce domaine. En fin de compte, l'État est le seul acteur ayant une influence suffisante pour influer sur les pratiques de l'industrie au rythme et à l'échelle nécessaires.

La création de nouvelles mesures incitatives pour stimuler l'innovation n'est pas un concept nouveau. En fait, le succès d'industries telles que les technologies de l'information, les biotechnologies et les nanotechnologies a compté sur des investissements publics bien avant que les acteurs privés n'entrent dans ces domaines, ces investissements soutenant la recherche et développement initiale nécessaire

pour amener ces technologies à maturité et à la rentabilité. Ce n'est donc qu'après que les risques initiaux ont été assumés par le gouvernement, grâce à des investissements substantiels dans la recherche fondamentale et les infrastructures, qu'un marché a été créé pour ces innovations et d'autres innovations révolutionnaires (Mazzucato, 2013).

S'il est vrai que de nombreux gouvernements ont déjà investi massivement dans le marché de l'IA, ces investissements ont généralement été orientés vers des applications d'IA commercialement viables. Les gouvernements ont, par exemple, financé la recherche au sein de l'industrie de l'IA en payant une partie des coûts de recherche et développement (gouvernement du Canada, 2018 ; Wiggers, 2021). Cette structure incitative exige que la recherche soit suffisamment attrayante sur le plan commercial pour que les entreprises soient motivées à assumer les coûts de l'autre partie de ces travaux. Par conséquent, pour changer la trajectoire actuelle de l'IA, les gouvernements doivent reconnaître leur rôle dans l'investissement et la promotion de l'adoption d'une IA responsable sur le plan social puisque, comme nous l'avons vu, cela ne se fera pas uniquement par le biais des marchés.

Il est recommandé aux gouvernements d'adopter une approche à long terme lorsqu'ils conçoivent les mesures incitatives qui régiront le domaine, et ce, parce que les calculs de profits impliquant des rendements à court terme réduisent souvent de manière exponentielle les résultats attendus à plus long terme. En pratique, cette approche à long terme consiste à rediriger stratégiquement le rendement du capital investi vers de futures initiatives de bien commun plutôt que de se concentrer sur des cycles à court terme axés sur le profit, qui s'avèrent actuellement la direction suivie par de nombreux acteurs du secteur privé (Lazonick et Mazzucato, 2013). En effet, les financements apportés par « les fonds de capital-risque ont tendance à se concentrer dans les domaines à fort potentiel de croissance, à faible intensité de capital et à faible complexité technologique, puisque cette dernière augmente considérablement les coûts » (Mazzucato, 2013, p. 55). Malheureusement, cette structure incitative est mal alignée sur les investissements à grande échelle qui sont nécessaires pour développer le domaine de l'IA pour le bien social. L'innovation dans cet espace devrait plutôt être considérée comme un processus cumulatif qui ne débouchera sur des produits de meilleure qualité et moins coûteux qu'après des *années* de recherche et développement industrielle (Lazonick et Mazzucato, 2013).

MESURES INCITATIVES DES GOUVERNEMENTS

Les gouvernements doivent jouer un rôle de premier plan dans la création de mesures incitatives régissant le domaine de l'IA afin d'opérer un changement positif parmi les types de projets d'IA en cours de développement. Les principaux leviers de changement peuvent être mis en pratique par des mesures incitatives positives et négatives. Les mesures incitatives négatives peuvent inclure des pénalités financières et non financières pour les développements socialement problématiques, tandis que les mesures incitatives positives peuvent inclure des récompenses financières et non financières pour les comportements bénéfiques sur le plan social. Que ce soit par des moyens positifs ou négatifs, les mesures incitatives doivent être appropriées afin d'influencer adéquatement les décisions des acteurs du secteur privé.

Les recommandations ci-dessous sont classées en fonction du groupe de parties prenantes qu'elles visent (de la personne à la société, en passant par les institutions). Chacune de ces parties prenantes joue un rôle unique et précieux dans l'écosystème de l'innovation et doit être mobilisée pour générer le changement souhaitable. Les points d'intervention recommandés sont établis pour atteindre trois objectifs principaux que les gouvernements devraient poursuivre :

- Permettre un engagement informé et hautement compétent dans le domaine de l'IA (*recommandations 1-3*) ;
- Promouvoir la collaboration multidisciplinaire dans l'ensemble de la chaîne de valeur (*recommandations 4-5*) ;
- Récompenser les acteurs qui contribuent à un écosystème favorisant une IA bénéfique sur le plan social (*recommandations 6-7*).

- **Recommandation 1:** Former des talents et créer une expertise à tous les niveaux, de la littératie numérique de base au personnel hautement compétent en IA dans les universités, en combinant à la fois sensibilisation sociale et formation technique.

À l'échelle mondiale, l'expertise en IA est rare. Sans compter que ceux et celles qui sont compétents en IA sont concentrés dans des pays, des secteurs, des industries et des groupes démographiques particuliers, ce qui a des répercussions sur les types d'applications d'IA qui sont développées (Groupe de la Banque mondiale, 2021). Un point de départ pour remédier à la pénurie de talents, et pour favoriser une communauté mondiale capable de mettre l'IA au service du développement social, consiste à investir en littératie numérique. Selon la définition de l'UNESCO (2018), « la littératie numérique est la capacité d'accéder à l'information, de la gérer, de la comprendre, de l'intégrer, de la communiquer, de l'évaluer et de la créer en toute sécurité et de manière appropriée au moyen de technologies numériques pour l'emploi, d'emplois décents et d'entrepreneuriat. Elle comprend des compétences qui sont diversement désignées sous le nom de connaissances en informatique, connaissances des TIC, connaissances en information et connaissances des médias » (traduction libre). Des efforts ont été déployés pour améliorer la littératie en IA de divers groupes de la population, à savoir ceux qui n'ont pas de formation technique (Kong *et al.*, 2021), qui font partie de groupes sous-représentés dans l'industrie (Office for Students, 2020) et qui, autrement, n'apprendraient pas l'IA dans le cadre du programme d'études standard (Lee *et al.*, 2021). Chacune de ces initiatives a donné des résultats encourageants qui devraient être explorés davantage.

Les compétences en IA sont également fortement concentrées le long de lignes géographiques. On peut observer ce phénomène en analysant la concentration des résultats en IA dans le petit nombre de pays qui accueillent l'écrasante majorité des talents en IA (Groupe de la Banque mondiale, 2021). Qui plus est, il arrive souvent que les travailleuses et travailleurs compétents du Sud se déplacent pour trouver du travail dans le Nord, ce qui entraîne un exode des cerveaux qui s'est fait grandement sentir dans les pays dépourvus de pôles de recherche et d'industrie en IA (McKinsey Global Institute, 2020). Les tendances indiquent aussi que les compétences en IA sont très majoritairement détenues par des hommes. Selon l'indice mondial de l'écart entre les sexes du Forum économique mondial (2020), les femmes ne représentent que 26 % de la main-d'œuvre dans les domaines des données et de l'IA à l'échelle mondiale. Au Canada, la répartition entre les sexes parmi les professionnels des données et de l'IA est de 70 % d'hommes et de 30 % de femmes. Dans le milieu universitaire, l'écart est encore plus grand. En fait, selon le Global AI Talent Report (Hudson et Mantha, 2020), les femmes n'ont rédigé que 15 % des articles sur l'IA publiés sur arXiv, une archive en libre accès largement utilisée par la communauté de l'IA.

La rareté et le déséquilibre des talents en IA se traduisent par des applications d'IA conçues par et pour certaines personnes et pas d'autres ainsi que par un écart croissant dans la concentration de la richesse et du pouvoir (Crawford, 2021). Selon l'International Data Corporation (IDC), les revenus mondiaux du marché de l'IA devraient dépasser les 300 milliards de dollars en 2024 (Savage, 2020). Pour qu'un plus grand nombre de personnes puisse bénéficier de cette opportunité économique, le développement des compétences est essentiel (OCDE, 2015).

Il est recommandé aux gouvernements de rendre la littératie numérique et la formation en IA plus largement disponibles et accessibles à tous les groupes démographiques *et* de veiller à ce que les gens possèdent les atouts nécessaires pour faire face aux conséquences éthiques et sociales en aval des

outils qu'ils créent. Il est important, lors de la formation des talents en IA, de les sensibiliser aux conséquences en aval afin de dissuader les applications problématiques et de promouvoir celles qui sont bénéfiques sur le plan social. En outre, cette formation permettrait aux praticiens et praticiennes de l'IA de considérer plus systématiquement les utilisations abusives potentielles de leurs outils et de prendre des mesures pour atténuer ces risques à l'avance[50].

- **Recommandation 2 :** Alimenter le pipeline de découverte de connaissances dans des applications importantes sur le plan social, de la recherche fondamentale exploratoire en IA à son adoption dans l'industrie.

Bien que la possibilité d'une croissance alimentée par l'IA soit largement comprise au sein de l'industrie, les taux d'adoption sont relativement lents en raison de préoccupations concernant le manque de maturité de la technologie et son développement rapide (Deloitte, 2019). Cependant, en intégrant le pipeline de découverte de connaissances, de l'exploration à la mise en œuvre, les gouvernements peuvent catalyser les investissements industriels dans des capacités internes en IA et, par le fait même, soutenir l'adoption de la technologie.

Afin d'intégrer les groupes de parties prenantes, les gouvernements devraient encourager l'utilisation de solutions d'IA tout au long de la chaîne de valeur (c.-à-d. dans la recherche, le développement et la mise en œuvre). Cela peut se faire en créant un écosystème dans lequel des chercheuses et chercheurs indépendants reçoivent des subventions pour explorer, tester et développer de nouveaux algorithmes bénéfiques sur le plan social dont pourrait tirer parti l'industrie.

Par exemple, afin d'intégrer les parties prenantes tout au long de la chaîne de valeur de l'IA, le Canada a établi la Stratégie pancanadienne en matière d'IA, une initiative de 125 millions de dollars qui vise à stimuler le leadership canadien dans le domaine de l'IA. Le levier de changement principal consiste à créer des écosystèmes d'IA locaux et régionaux qui soutiennent les talents en IA, favorisent l'adoption par l'industrie et permettent de mieux comprendre les implications sociales de l'IA dans toute la chaîne de valeur (CIFAR, 2017). Mila – Institut québécois d'intelligence artificielle est un exemple de pôle de recherche qui contribue à un écosystème d'IA plus large à Montréal et au-delà en développant la recherche fondamentale et appliquée en IA grâce à plus de 900 chercheuses et chercheurs et un grand nombre de partenaires industriels, des entreprises en démarrage aux entreprises technologiques bien établies[51].

- **Recommandation 3 :** Renforcer l'écosystème de l'IA grâce à la création de pôles d'excellence en recherche et formation en IA.

Pour mettre en place des écosystèmes d'innovation, les pays doivent attirer des talents de grande qualité et leur fournir les ressources et le soutien nécessaires pour qu'ils puissent se maintenir à long terme. La Commission européenne (2020) a établi une stratégie à cet égard en énonçant une série de mesures recommandées pour en arriver à un « écosystème d'excellence » tout au long de la chaîne de connaissances et de valeur. La Commission affirme que, contrairement au paysage fragmenté qui caractérise actuellement les centres d'excellence européens, une approche paneuropéenne permettrait d'atteindre l'échelle requise pour concurrencer les instituts de pointe à l'échelle mondiale. Selon la Commission, une approche centralisée permettrait de renforcer la formation et d'attirer les chercheuses et chercheurs, ce qui mènerait au développement de technologies de grande qualité et débloquerait

50. L'outil d'évaluation de l'incidence algorithmique du Canada est un outil d'évaluation des risques utile qui peut aider les développeuses et développeurs en IA à découvrir les domaines de risque qui doivent être atténués avant la mise en œuvre. https://www.canada.ca/fr/gouvernement/systeme/gouvernement-numerique/innovations-gouvernementales-numeriques/utilisation-responsable-ai/evaluation-incidence-algorithmique.html

51. Pour en savoir plus sur Mila, visitez le site Web : https://mila.quebec/.

d'importants investissements en IA (Commission européenne, 2020). Un tel modèle pourrait inspirer d'autres régions du monde, où les ressources peuvent être rares et où un effort transnational pourrait permettre de renforcer un écosystème qui développe une IA bénéfique sur le plan social afin de relever des défis communs.

- **Recommandation 4 :** Financer et intégrer les différentes parties du pipeline afin de minimiser l'exode tout au long de la chaîne de valeur, en particulier dans la poursuite d'objectifs sociétaux pertinents.

Si les gouvernements veulent réussir à créer des entreprises génératrices de marchés, leur financement doit aller au-delà de la recherche initiale. Tout au long de l'histoire, les investissements publics qui ont généré des révolutions technologiques (dans des domaines tels que les technologies de l'information, les biotechnologies et les nanotechnologies) ont visé les parties prenantes de l'ensemble de la chaîne d'innovation. Or, d'une manière générale, il est encore nécessaire de s'attaquer à l'exode des cerveaux, des idées inexploitées et des entreprises en démarrage prometteuses, par exemple, tout au long de la chaîne de valeur (Spicer *et al.*, 2018). Un exemple de processus intégré peut être observé aux États-Unis où la National Science Foundation (NSF) a mené des recherches fondamentales qui ont été transmises à la DARPA et au National Institutes of Health (NIH) pour être développées, testées et appliquées. Ensuite, par le biais d'agences telles que Small Business Innovation Research (SBIR), les entreprises cherchant à mettre la recherche en production recevaient le financement de démarrage nécessaire pour le faire (Block et Keller, 2011). Afin que les gouvernements parviennent à favoriser de nouveaux marchés, l'innovation doit ainsi être encouragée de manière stratégique tout au long du cycle de vie de la recherche et développement.

Cependant, sans mesures ni suivis appropriés, il est très difficile de s'assurer que l'engagement des gouvernements a eu un impact. À ce titre, les gouvernements devraient mesurer le succès de leurs interventions en fonction des retombées économiques, sociales et environnementales qu'elles génèrent. Bien qu'il puisse être difficile sur le plan politique de mettre l'accent sur les retombées à long terme d'un investissement, il s'agit d'un exercice utile pour s'assurer que les marchés créés sont ceux qui génèrent le plus de retombées positives pour la société.

- **Recommandation 5 :** Faciliter la croissance de l'écosystème des entreprises en démarrage en IA pour le bien social et favoriser son lien avec l'industrie [52].

Le secteur privé joue un rôle de premier plan dans l'écosystème d'innovation en IA. Cependant, les acteurs du secteur privé pourraient en faire plus pour obtenir de meilleurs résultats dans ce domaine. Plus précisément, ces derniers devraient s'engager dans une relation symbiotique les uns avec les autres, tirant parti de leurs forces respectives pour une croissance sectorielle optimale. Les forces des grandes entreprises comprennent leur capacité à s'offrir les meilleurs talents, installations informatiques et laboratoires d'expérimentation, qui produisent tous des données de grande qualité et qui génèrent des produits d'IA intéressants. Les entreprises en démarrage sont aussi essentielles, puisqu'elles se concentrent sur des domaines niches du développement de l'IA qui contiennent souvent les technologies d'IA les plus poussées. En outre, elles sont incroyablement dynamiques et peuvent s'adapter rapidement à l'évolution des besoins, qu'il s'agisse de processus, de talents ou d'opérations.

Des organisations comme le European AI Startup Landscape sont un exemple intéressant quant à la manière de favoriser les liens entre les entreprises en démarrage, les grandes entreprises et les sociétés de capital de risque (European AI Startup Landscape, n.d.). La promesse de ces partenariats est

52. Le European Startup Landscape est un exemple intéressant de l'établissement d'un réseau d'entreprises en démarrage spécialisées en IA et de mise en place d'un écosystème dynamique avec d'autres parties prenantes telles que l'industrie. Consultez le https://www.ai-startups-europe.eu/.

qu'ils peuvent renforcer l'écosystème pour l'innovation et stimuler la diffusion technologique dans de nouveaux domaines (Groupe de la Banque mondiale, 2021). En plus de contribuer au développement général de l'écosystème des entreprises en démarrage en IA, les gouvernements pourraient tirer parti de leurs contributions financières pour encourager l'innovation bénéfique sur le plan social en fonction de leurs priorités gouvernementales.

- **Recommandation 6 :** Stimuler la recherche et l'innovation dans les domaines où la valeur sociétale est grande, mais où la valeur commerciale est trop faible pour les entreprises.

Comme nous l'avons mentionné précédemment, certains domaines de recherche en IA sont sous-développés, non pas parce qu'ils n'offrent pas de retombées sociétales évidentes, mais parce qu'ils ne sont pas suffisamment attrayants sur le plan commercial pour les entreprises. Pour s'assurer que l'écosystème de l'IA pour le bien social est robuste, les gouvernements doivent cibler ces domaines de recherche et inciter les parties prenantes concernées à s'y engager.

Souvent, pour être en mesure de trouver des voies de recherche et développement prometteuses, l'expertise de personnes spécialisées en IA et dans le domaine des applications est requise. Il est donc recommandé aux gouvernements de s'associer à des centres d'excellence dont les équipes d'expertes et d'experts internes peuvent trouver les domaines de recherche sous-explorés ayant des retombées sociales pertinentes. Il est aussi recommandé aux gouvernements de faire appel à ces équipes pour évaluer les mérites des demandes de subvention[53]. Ce type de partenariat peut stimuler une croissance et un développement économiques plus importants que ceux que les gouvernements pourraient autrement réaliser seuls.

Pour les domaines de recherche qui nécessitent des investissements particulièrement importants, il est recommandé que des organisations financées par les gouvernements soient créées à l'échelle internationale pour déterminer de manière indépendante les orientations les plus stratégiques de l'innovation en IA. Ces organisations seraient chargées de rédiger les contrats d'approvisionnement en innovations et de définir les indicateurs de réussite. L'indépendance de ces organisations leur donnerait la liberté de planifier sur un horizon temporel à plus long terme et de tenir compte de moins de demandes politiques. Néanmoins, ces organisations seraient responsables de rendre compte de leurs activités de manière cohérente et transparente afin de veiller à ce qu'elles soient tenues responsables de leurs décisions.

- **Recommandation 7 :** Établir un cadre pour promouvoir le partage des connaissances et des données entre les acteurs, tout en préservant la confidentialité des données.

Même si les gouvernements parvenaient à mettre en œuvre les six premières recommandations, il y aurait encore toute une série d'occasions ratées en raison de l'accès limité à des ensembles de données importants. En effet, sans accès aux données appropriées, il peut s'avérer impossible d'entraîner des modèles d'apprentissage automatique performants.

Malheureusement, les acteurs qui recueillent des données tentent souvent de les garder secrètes ou d'en conserver le monopole. Par conséquent, ces derniers « créent une rareté artificielle de connaissances exactement de la même manière qu'un cartel des boulangers crée une rareté artificielle de pain » (Maurer, 2003, p. 175). Lorsque le propriétaire d'une propriété intellectuelle restreint la manière dont elle peut être utilisée, une série d'inefficacités en matière d'innovation peuvent survenir. Par exemple, sans le partage des données, les parties prenantes disposant de l'expertise, de l'imagination et des installations matérielles nécessaires pour créer des produits d'IA novateurs

53. Le National Institutes of Health (NIH) (l'agence de recherche médicale aux États-Unis) a créé un centre d'excellence national dans le cadre de son programme « Bridge2AI » afin de l'aider à catalyser des recherches prometteuses.

pourraient être incapables de le faire ou d'obtenir des résultats optimaux sans avoir accès aux ensembles de données nécessaires. Ce goulot d'étranglement freine les travaux ultérieurs et peut avoir des répercussions sur l'écosystème.

Il est essentiel d'améliorer l'accès aux données ainsi que leur gestion pour permettre le développement de l'IA et d'autres applications numériques. En Europe, par exemple, un rapport a étudié l'ampleur du coût d'opportunité lié au manque de données en format ouvert. En examinant sept indicateurs, soit le temps consacré, le coût du stockage, le coût des licences, la rétraction de la recherche, le financement en double, l'interdisciplinarité et la croissance économique potentielle, l'étude a révélé que le coût estimé du non-partage des données atteignait 10 milliards d'euros par an (Commission européenne, 2020). C'est pour cette raison que le Comité de la politique scientifique et technologique de l'OCDE a fait valoir que l'accès aux données devrait devenir une priorité politique majeure au sein de l'OCDE (OCDE, 2021).

Les demandes de propositions sont l'un des leviers que peuvent utiliser les gouvernements pour favoriser le partage des données. En demandant des demandes de propositions aux entreprises œuvrant en IA, les gouvernements peuvent exiger, comme condition d'obtention d'un contrat, que les entreprises choisies rendent ouvertement disponibles tous les ensembles de données générés dans le cadre d'un projet. Bien que cette condition puisse entraîner la nécessité de mieux rémunérer l'entreprise à qui est octroyé le contrat, les avantages à long terme de ces politiques sont susceptibles de l'emporter sur les coûts. On estime que les données ouvertes peuvent débloquer annuellement 3 billions de dollars à l'échelle mondiale en valeur économique en contribuant à l'innovation dans tous les secteurs de l'économie (McKinsey, 2014, p.10). Cela s'explique par le fait qu'un plus grand accès aux données abaisse les barrières pour travailler en IA et augmente la concurrence avec les nouveaux entrants sur le marché qui se joignent à l'industrie. Les avantages financiers de cette politique se manifestent dans de nouvelles sources de revenus, des économies et un surplus économique dans des domaines allant de l'éducation au transport. Avec de nouveaux entrants sur le marché et une plus grande innovation, les gouvernements créent un environnement favorable au développement de nouveaux produits d'IA bénéfiques sur le plan social.

CONCLUSION

L'IA est un outil incroyablement puissant qui a le potentiel de générer des découvertes bénéfiques sur le plan social dans des domaines d'une importance cruciale, de l'éducation à l'environnement en passant par les soins de santé. Cependant, pour que ces possibilités se concrétisent, les gouvernements doivent activement façonner la trajectoire de la recherche et développement en IA en impliquant toutes les parties prenantes au sein de l'écosystème de l'IA. Sinon, les acteurs sectoriels sont laissés à eux-mêmes pour décider du développement du domaine de l'IA, ce qui tend à marginaliser les innovations qui ont une grande valeur sociétale lorsqu'elles ne sont pas suffisamment attrayantes sur le plan commercial. Les gouvernements devraient agir pour réorienter cette industrie. Ils devraient investir dans la littératie et l'éducation en IA, mettre en place un écosystème multipartite bien intégré, créer des mesures incitatives suffisantes le long du pipeline pour engager et maintenir les talents, inspirer l'IA pour des applications de bien social et promouvoir le partage des données. En adoptant une telle approche, la société pourra commencer à exploiter les promesses de l'IA en tant qu'outil de développement social *et* économique.

RÉFÉRENCES

Adamson, A. et Smith, A. 2018. Machine Learning and Health Care Disparities in Dermatology. *JAMA Dermatology*. vol. 154, nº 11, pp. 1247–1248. DOI :10.1001/jamadermatol.2018.2348

Benkler, Y. 2019. *Don't let industry write the rules for AI*. https://www.nature.com/articles/d41586-019-01413-1

Block, F. L., et M. R. Keller. 2011. *State of innovation : The U.S. government's role in technology development*. Boulder, CO : Paradigm Publishers.

CIFAR. 2017. *Pan-Canadian AI Strategy*. https://cifar.ca/ai/

Cowls, J., Tsamados, A., Taddeo, M., & Floridi, L. 2021. A definition, benchmark and database of AI for social good initiatives. *Nature Machine Intelligence*, *3*(2), 111-115.

Deloitte. 2019. *Future in the balance ? How countries are pursuing an AI advantage*. https://www2.deloitte.com/us/en/insights/focus/cognitive-technologies/ai-investment-by-country.html

European AI Startup Landscape. n.d. *Motivation*. https://www.ai-startups-europe.eu/

European Commission. 2020. *On artificial intelligence – A European approach to excellence and trust*. https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

European Economic and Social Committee. 2021. How the Digital Transformation can put Humans at the Centre of Robotics and Automation : collaboration between humans and machines for better quality products and services. April 2021. doi : 10.2864/733324

Government of Canada. 2018. *Government of Canada invests in artificial intelligence and start-up innovation across Canada*. Ottawa, Innovation, Science and Economic Development Canada. https://www.canada.ca/en/innovation-science-economic-development/news/2018/10/government-of-canada-invests-in-artificial-intelligence-and-start-up-innovation-across-canada.html

Hudson, S. et Mantha, Y. 2020. *Global AI Talent Report 2020*. https://jfgagne.ai/global-ai-talent-report-2020/

Kong, S. C., Cheung, W. M. Y. et Zhang, G. 2021. Evaluation of an artificial intelligence literacy course for university students with diverse study backgrounds. *Computers and Education : Artificial Intelligence*, 100026.

LaCroix, T., et Mohseni, A. 2020. *The Tragedy of the AI Commons*. arXiv preprint arXiv :2006.05203.

Lazonick, W. et Mazzucato, M. 2013. The risk-reward nexus in the innovation-inequality relationship : who takes the risks ? Who gets the rewards ?. *Industrial and Corporate Change*, vol. 22, nº 4, pp. 1093-1128.

Lee, I. *et al*. 2021. Developing Middle School Students' AI Literacy. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, pp. 191-197.

Lloyd, W. F. 1833. *Two lectures on the checks to population*. JH Parker.

Maurer, S. 2003. Designing Public–Private Transactions that Foster Innovation. Esanu, J.M. et Uhlir, P.F. (eds). *The Role of Scientific and Technical Data and Information in the Public Domain : Proceedings of a Symposium*. National Academies Press, pp. 175-79.

Mazzucato, M. 2013. *The entrepreneurial state : Debunking the public vs. private myth in risk and innovation*. London : Anthem Press.

McKenna, M. 2020. *The antibiotic paradox : why companies can't afford to create life-saving drugs*. https://www.nature.com/articles/d41586-020-02418-x

McKinsey Global Institute. 2020. *How to Ensure Artificial Intelligence Benefits Society : A Conversation with Stuart Russell and James Manyika*. https://www.mckinsey.com/featured-insights/artificial-intelligence/how-to-ensure-artificial-intelligence-benefits-society-a-conversation-with-stuart-russell-and-james-manyika

Murray, C. J. L., *et al*. 2022. Global burden of bacterial antimicrobial resistance in 2019 : a systematic analysis. *The Lancet*, vol. 399, nº 10325, pp. 625 – 655. DOI :https://doi.org/10.1016/S0140-6736(21)02724-0

OECD. 2015. Making Open Science a Reality. OECD Science, Technology and Industry Policy Papers, nº 25, OECD Publishing, Paris. http://dx.doi.org/10.1787/5jrs2f963zs1-en

OECD. 2021. *Recommendation of the OECD Council concerning Access to Research Data from Public Funding*. https://www.oecd.org/sti/recommendation-access-to-research-data-from-public-funding.htm

Office for Students. 2020. *Apply now – new courses in artificial intelligence and data science*. https://www.officeforstudents.org.uk/news-blog-and-events/press-and-media/apply-now-new-courses-in-artificial-intelligence-and-data-science/

PhRMA. 2015. *Biopharmaceutical Research & Development : The Process Behind New Medicines*. http://phrma-docs.phrma.org/sites/default/files/pdf/rd_brochure_022307.pdf

Plackett, Benjamin. 2020. Why Big Pharma Has Abandoned Antibiotics. *Nature Outlook : Antimicrobial Resistance*, 21 October.https://www.nature.com/articles/d41586-020-02884-3.

Porter, M. E. et Kramer, M. R. 2006. Strategy & Society : The Link between Competitive Advantage and Corporate Social Responsibility. *Harvard Business Review*, December 2006. https://hazrevista.org/wp-content/uploads/strategy-society.pdf.

Review on Antimicrobial Resistance. 2014. *Antimicrobial resistance : tackling a crisis for the health and wealth of nations*. Review on Antimicrobial Resistance.

Savage, N. 2020. The Race to the Top among the World's Leaders in Artificial Intelligence. *Nature*, December. https://www.nature.com/articles/d41586-020-03409-8.

Schwartz, O. 2019. In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation. *IEEE Spectrum*, 25 November. https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation.

Spicer, Z. *et al*. 2018. *Reversing the Brain Drain : Where is Canadian STEM Talent Going ?*. https://brocku.ca/social-sciences/political-science/wp-content/uploads/sites/153/Reversing-the-Brain-Drain.pdf

UN General Assembly. 2015. *Transforming our world : the 2030 Agenda for Sustainable Development*. https://www.refworld.org/docid/57b6e3e44.html

Wang, Y. et Kosinski, M. 2022. *Deep neural networks are more accurate than humans at detecting sexual orientation from facial images*. OSF, 23 June. https://doi.org/10.17605/OSF.IO/ZN79K

Wiggers. K. 2021. *U.S. agencies are increasing their AI investments*. AI Weekly. San Francisco, VentureBeat. https://venturebeat.com/2021/09/11/ai-weekly-u-s-agencies-are-increasing-their-investments-in-ai/#:~:text=R%26D%20spending%20reached%20%241.2%20billion,by%20a%20combined%20%2481%20million.

World Bank Group. 2021. *Harnessing Artificial Intelligence for Development in the Post-COVID-19 Era. A Review of National AI Strategies and Policies*. https://thedocs.worldbank.org/en/doc/2e658ef2144a05f30e254221ccaf7a42-0200022021/original/DD-Analytical-Insights-Note-4.pdf

World Economic Forum. 2020. *Data Explorer : Global Gender Gap Index*. http://reports.weforum.org/global-gender-gap-report-2020/dataexplorer/

World Health Organization. *2019a. 2019 Antibacterial agents in clinical development : an analysis of the antibacterial clinical development pipeline*. https://apps.who.int/iris/bitstream/handle/10665/330420/9789240000193-eng.pdf

——. 2019*b*. *New report calls for urgent action to avert antimicrobial resistance crisis.* https://www.who.int/news/item/29-04-2019-new-report-calls-for-urgent-action-to-avert-antimicrobial-resistance-crisis

——. 2021*a*. *Cancer*. https://www.who.int/news-room/fact-sheets/detail/cancer

——. 2021*b*. Road traffic injuries. https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

——. 2022. *WHO Coronavirus (COVID-19) Dashboard*. https://covid19.who.int/

UN MANIFESTE EN FAVEUR DE L'INTELLIGENCE ARTIFICIELLE POUR LE SUIVI DU DÉVELOPPEMENT DURABLE : L'ANGLE MORT ENTRE LES ODD, LES INVESTISSEMENTS ET LA CONFIANCE

JOHN SHAWE-TAYLOR

Professeur d'informatique statistique et d'apprentissage automatique et détenteur de la chaire UNESCO en intelligence artificielle à University College London et directeur du Centre international de recherche en intelligence artificielle (IRCAI) sous l'égide de l'UNESCO au Jozef Stefan Institute en Slovénie.

DANIEL MIODOVNIK

Directeur chez Social Finance et cofondateur de leur Digital Labs. Il est conseiller au Centre international de recherche en intelligence artificielle (IRCAI) sous l'égide de l'UNESCO.

DAVOR ORLIC

Adjoint à la recherche honoraire au UCL Centre for Artificial Intelligence et directeur de l'exploitation au Centre international de recherche en intelligence artificielle (IRCAI) sous l'égide de l'UNESCO au Jozef Stefan Institute en Slovénie.

- **ODD 6** - Eau propre et assainissement
- **ODD 7** - Énergie propre et d'un coût abordable
- **ODD 9** - Industrie, innovation et infrastructure
- **ODD 10** - Inégalités réduites
- **ODD 11** - Villes et communautés durables
- **ODD 12** - Consommation et production responsables
- **ODD 13** - Mesures relatives à la lutte contre les changements climatiques
- **ODD 15** - Vie terrestre
- **ODD 16** - Paix, justice et institutions efficaces
- **ODD 17** - Partenariats pour la réalisation des objectifs

UN MANIFESTE EN FAVEUR DE L'INTELLIGENCE ARTIFICIELLE POUR LE SUIVI DU DÉVELOPPEMENT DURABLE : L'ANGLE MORT ENTRE LES ODD, LES INVESTISSEMENTS ET LA CONFIANCE

RÉSUMÉ

L'atteinte des objectifs de développement durable (ODD) des Nations Unies suscite un très grand intérêt. Partout dans le monde, des pays, des entreprises et des investisseurs et investisseuses sont déterminés à lutter contre la crise économique, sociale et environnementale mondiale. Les investisseurs et investisseuses ont déjà consacré 89 billions $ US en actifs à des investissements ciblant des résultats soutenant les ODD dans le cadre du programme Principes pour l'investissement responsable (PRI).

Or, sans moyens fiables et objectifs pour évaluer les progrès, cet élan pourrait bien s'essouffler. Nous constatons une érosion de la confiance des populations envers les gouvernements, les sociétés technologiques et l'industrie. L'absence d'un cadre uniforme et la subjectivité actuelle des données et des notations nous empêchent d'aller de l'avant.

Nous sommes d'avis que l'intelligence artificielle (IA) et les données sont essentielles pour renforcer la confiance et fournir des éléments de preuve étayant les progrès mesurables réalisés envers l'atteinte des ODD. Nous pouvons déjà nous appuyer sur des exemples montrant la façon dont des résultats clairement définis et mesurables peuvent débloquer des investissements pour réaliser les ODD. Notamment, des indicateurs de résultats clairs et une collecte de données efficace sont à l'origine de contrats fondés sur les résultats totalisant 10 millions $ octroyés pour répondre à des besoins en matière de services d'assainissement en milieu rural au Cambodge (Objectif 6 : Garantir l'accès à tous à des services d'alimentation en eau et d'assainissement gérés de façon durable).

> Nous avons donc préparé un manifeste qui appelle les ONG, l'ONU, les entreprises, les investisseurs et investisseuses et les pays à bâtir de manière collaborative un système robuste, accessible et transparent pour mesurer et certifier l'atteinte des ODD. Ensemble, nous pouvons mettre en place l'IA et l'écosystème de données nécessaires pour créer de la confiance et permettre aux investisseurs et investisseuses, aux entreprises et aux gouvernements de démontrer les progrès réalisés, d'obtenir des investissements et, ultimement, de changer le monde.

INTRODUCTION

Nous n'avons pas, à ce moment, une compréhension objective d'où nous en sommes exactement quant aux progrès réalisés envers l'atteinte des objectifs de développement durable (ODD). Nous sommes d'avis qu'un mécanisme technique novateur universellement reconnu ainsi qu'un mécanisme pour les aspects financiers et les investissements permettraient de créer de la confiance entre toutes les parties prenantes. Le chaînon manquant est la convergence entre le recours à l'intelligence artificielle (IA) pour mesurer les progrès réalisés envers l'atteinte des ODD et aux obligations à impact social qui, ensemble, peuvent être utilisées par les gouvernements pour financer de telles entreprises techniques. Dans ce chapitre, nous proposons un narratif qui pourrait possiblement éliminer ce chaînon manquant.

OBLIGATIONS À IMPACT SOCIAL

De piètres installations sanitaires, particulièrement dans des endroits où la défécation à l'air libre est courante, sont associées à des risques pour la santé allant de la propagation de maladies à la contamination de l'eau potable. Pour aider le gouvernement cambodgien à mettre en place des installations sanitaires dans certains des foyers les plus pauvres et les plus vulnérables du pays, Social Finance s'est associée à la Stone Family Foundation, à International Development Enterprises (iDE) et à l'Agence américaine pour le développement international (USAID) afin de créer la première obligation à impact social pour l'assainissement.

L'objectif de l'obligation à impact social était d'atteindre, d'ici 2023, une couverture sanitaire de 85 % en milieu rural dans les régions ciblées, permettant ainsi à 1600 villages de devenir des collectivités sans défécation à l'air libre. L'atteinte de ce jalon accélérerait les efforts du Cambodge dans le but d'en arriver à des services d'assainissement universels avant les cibles des objectifs de développement durable 2030 (Social Finance, 2021).

Après une décennie de croissance impressionnante en matière de couverture sanitaire en milieu rural au Cambodge, les foyers restants avaient tendance à être situés dans les régions les plus pauvres et les plus difficiles d'accès. Afin d'aider le gouvernement cambodgien à atteindre cette cible ambitieuse d'ici 2023, iDE, un fournisseur de services d'assainissement en milieu rural de premier plan, cherchait à avoir accès à du financement pour innover. L'obligation à impact social a répondu à ce besoin. La Stone Family Foundation a fourni le financement initial à iDE et, par le fait même, les ressources nécessaires pour

concevoir et réaliser un programme de services d'assainissement en milieu rural pour atteindre les foyers les plus pauvres et les plus vulnérables. USAID a, quant à elle, accepté de fournir jusqu'à 10 millions £ de financement fondé sur les résultats à la Stone Family Foundation si le programme d'iDE permettait à ces villages de devenir des collectivités sans défécation à l'air libre.

En novembre 2019, l'obligation à impact social a été lancée. Aux dernières nouvelles, USAID rapportait que 500 villages étaient devenus des collectivités sans défécation à l'air libre, 88 738 foyers ayant maintenant un accès confirmé à des installations sanitaires conformes aux lignes directrices à cet égard du gouvernement royal du Cambodge. À ce jour, USAID a versé 3 125 000 $ fondés sur des résultats (USAID, 2019). Il s'agit donc d'un exemple de la manière dont les données peuvent permettre d'obtenir un financement novateur qui stimule des progrès envers l'atteinte des ODD.

Défis et importance de la vérification

L'exemple présenté plus haut confirme notre croyance voulant que les investisseurs et investisseuses souhaitent fortement engager leurs capitaux auprès d'entreprises qui sont en mesure de contribuer de manière positive au développement durable. En d'autres mots, de tels investisseurs et de telles investisseuses acceptent de possiblement obtenir un rendement financier du capital investi plus faible ou à plus long terme dans la mesure où ils et elles ont l'assurance que leurs capitaux seront utilisés pour faire progresser des objectifs de développement durable généraux ou particuliers. Cela ne devrait pas surprendre, compte tenu de l'intérêt qu'ont suscité au cours des dernières années les investissements éthiques, un intérêt qui a même poussé certains investisseurs et certaines investisseuses à cesser de soutenir des entreprises dont les actions étaient contraires à l'éthique, notamment des entreprises promouvant l'utilisation du tabac ou ayant recours à une main-d'œuvre bon marché dans des ateliers de misères. La différence principale entre les contraintes liées à un investissement dans des entreprises dont les activités sont contraires à l'éthique et à un investissement en développement durable réside dans le fait qu'il est possible de fournir des éléments de preuve, relativement facilement vérifiables, démontrant qu'une entreprise a agi d'une manière précise contraire à l'éthique, alors qu'un tout autre degré de preuve est requis pour démontrer qu'une entreprise contribue de manière soutenue au développement durable.

L'exemple d'un article récent analysant comment l'émission d'obligations vertes dans le secteur forestier du Brésil (Ferrando *et al.*, 2021) a mené à la transformation de territoires et d'éléments naturels en « solutions temporelles et spatiales » pour les besoins du capital financier mondial décrit très bien cette difficulté. Voici un extrait du résumé :

> Through the study of recent green bond issuances realized by private companies active in the forestry sector in Brazil, we discuss how green bonds as a "new" form of "green" debt put nature at work and transform the territories and natural elements in the global south into "temporal and spatial fixes" for the needs of global financial capital (page 410).

Il ne s'agit que d'un exemple de la difficulté à démontrer, à l'aide de données objectives et vérifiables, son bilan écologique. Une étude scientifique récente s'est penchée sur l'ampleur avec laquelle les compensations carbone génèrent les effets escomptés et a trouvé des éléments de preuve suggérant une surestimation : « Les résultats suggèrent que les méthodologies acceptées pour quantifier les crédits carbone surestiment les impacts sur la déforestation évitée et sur l'atténuation des changements climatiques » (West *et al.*, 2020).

La question concernant en qui nous pouvons avoir confiance pour fournir des informations objectives et exactes est au cœur de cette difficulté. En effet, Tariq Fancy, le premier responsable des placements durables de BlackRock, a remis en question l'ensemble de l'initiative ESG : « Mais d'autres problèmes relatifs aux placements ESG existent, notamment leur subjectivité et le manque de fiabilité des données et des notations » (Amaro, 2021).

La question clé réside dans le fait que les personnes générant les notations et les données sont celles qui profiteront potentiellement d'une évaluation positive, créant ainsi un conflit d'intérêts et l'érosion subséquente de la confiance qui est au cœur de l'initiative.

Des rapports très encourageants existent certainement, comme le rapport décrivant les travaux au Costa Rica ayant reçu le prix environnemental Earthshot du prince William et la récente soumission à la liste des 100 meilleurs projets de l'IRCAI[54] portant sur l'utilisation de la vision par ordinateur pour détecter les émissions de carbone dans les forêts de la Zambie décrite ci-dessous :

> Our project is based on detecting and reducing carbon emission in forest using computer vision. We intended to collect data using satellite and also data scientist, machine learning and artificial intelligence. After collecting the data, we are going to preprocess it, and it will be ready for training and metrics and performance evaluation using keras software for analysis. The impact of this project is for about 300 people within and near the national parks near the forest that will benefit from this project (Zamculture, 2019).

L'ampleur du soutien et de la volonté de travailler envers l'atteinte des objectifs de développement durable des Nations Unies atteint, aujourd'hui, des niveaux inégalés. Bien que cet élan de soutien soit extrêmement positif, il y a un réel danger de désillusion si les entreprises et les pays ne disent pas toute la vérité, comme certains éléments de preuve le suggèrent. Sans moyens objectifs et fiables pour évaluer les progrès, il y a aussi un danger que les médias sociaux, par exemple, soient utilisés pour ternir l'image d'une entreprise en répandant des rumeurs sans fondement voulant que ses réalisations soient fausses. De telles situations nuiraient considérablement à l'intérêt et au soutien aux investissements envers les ODD.

L'ampleur du soutien aux investissements envers les ODD se voit, entre autres, dans le succès du programme de Principes pour l'investissement responsable (PRI), qui consiste à investir avec les ODD (UNPRI, 2022), dans le cadre duquel les investisseurs et investisseuses ont engagé, ensemble, 89 billions $ US en actifs sous gestion. Le cadre de référence proposé est résumé dans le programme (UNPRI, 2020) et comprend les parties suivantes :

1. Identification des résultats
2. Définition des politiques et des objectifs
3. Les investisseurs façonnent les résultats
4. Les systèmes financiers façonnent les résultats collectifs
5. Les parties prenantes internationales collaborent pour obtenir des résultats alignés avec les ODD

Le cadre de référence est bien construit et établit les visées du programme, soit de diriger les investissements pour soutenir les ODD des Nations Unies. Au cœur de cette approche est le besoin des « investisseurs et investisseuses à chercher individuellement à accroître les résultats positifs, à diminuer les résultats négatifs et à mesurer les progrès réalisés envers les cibles établies ».

Bien que la question de la mesure soit soulevée, la question plus large de la confiance est également importante à saisir. Comme indiqué ci-dessous, le rapport souligne qu'une évaluation plus objective des indicateurs de performance clés encourage les parties prenantes à appuyer les initiatives :

> With more objective assessment of SDG KPIs there is greater opportunity for stakeholders to support initiatives that are making verifiable impact : these could be individual investors, governments, other companies making informed choices about collaboration, etc.

54. L'IRCAI est le Centre international de recherche en intelligence artificielle sous l'égide de l'UNESCO. Site Web : ircai.org.

Or, nous vivons à une époque où il existe une méfiance généralisée envers les institutions et les leaders qui entraîne une majorité de personnes à croire que les gouvernements et les leaders du milieu des affaires cherchent à les tromper (UNESCO, 2020). Dans ce contexte d'érosion de la confiance, nous croyons que l'IA peut jouer un rôle essentiel dans le développement durable en fournissant cette pièce manquante au casse-tête et, par conséquent, nous proposons le manifeste suivant :

Il existe un besoin urgent de créer un système robuste pour mesurer et certifier l'atteinte des indicateurs de performance clés (IPC) des ODD et, lorsque cela est possible, pour fournir des éléments de preuve quant aux interventions qui sont à l'origine de tout changement (positif ou négatif). Ce système et son fonctionnement doivent gagner la confiance de toutes les parties prenantes : citoyens et citoyennes, gouvernements, compagnies technologiques et industrie.

RÉALISER LE MANIFESTE

Nous portons maintenant notre attention à la question de savoir comment donner vie à ce manifeste. Nous soutiendrons ici que la confiance peut être créée lorsque les conclusions sont fondées sur des données recueillies et vérifiables et que les forces et les faiblesses des déductions tirées des données sont présentées de manière impartiale.

Le rôle des données

Tous les types d'ensembles de données peuvent être utiliser pour évaluer les divers aspects de la réalisation des différents IPC des ODD. Une telle base a le potentiel de :

- mesurer si un résultat a été obtenu ;
- consigner ce résultat d'une manière qui est fiable pour tous et toutes ;
- garantir qu'un résultat est vérifiable et attribuable à un service ou à un produit ;
- utiliser ces données pour effectuer un paiement et analyser la manière d'améliorer les services, puisque nous ne devrions pas être satisfaits et satisfaites tant que les ODD n'auront pas été entièrement réalisés.

Des données sont recueillies à une vitesse sans précédent à l'aide de capteurs locaux et à distance. Un mouvement bien établi soutient aussi que de telles collectes de données et, de manière plus générale la science, doivent être ouvertes. L'UNESCO a, par exemple, établi l'initiative Science ouverte qui s'appuie sur l'idée que la science ouverte permet aux informations, aux données et aux résultats scientifiques d'être plus largement accessibles (libre accès) et utilisés de manière plus fiable (données ouvertes) avec la participation active de toutes les parties prenantes (ouverture vers la société) :

> The idea behind Open Science is to allow scientific information, data and outputs to be more widely accessible (Open Access) and more reliably harnessed (Open Data) with the active engagement of all the stakeholders (Open to Society) (Masakhane, 2022).

La science ouverte saisit parfaitement bien le rôle potentiel et l'approche qui peuvent engendrer une confiance dans les données, mais qui peuvent aussi encourager une participation plus large à l'exploration scientifique. Il s'agit là d'un élément important de l'instauration de la confiance, à savoir que tous les groupes doivent sentir qu'ils peuvent participer, tant à la collecte des données qu'à leur vérification et à leur analyse. Dans ce cas, les groupes peuvent faire référence à différentes régions du monde, à différentes tranches de la société, à différentes disciplines scientifiques ou à différents gouvernements, ONG ou entreprises. Le modèle des pays développés qui consiste à apporter des

solutions toutes faites à des problèmes éloignés peut très facilement mener à la résolution du mauvais problème ou à l'omission de conditions locales essentielles, ce qui se traduit en une piètre solution ou, pire encore, à aucune solution, mais qui contribue à l'érosion de la confiance, tant en ce qui a trait à la collaboration qu'à la science en général.

Une partie importante de la science ouverte et des données ouvertes est la reconnaissance que les défis locaux exigent une participation locale pour définir le défi, recueillir les données et collaborer à la recherche de solutions. L'initiative Masakhane est un excellent exemple d'organisation qui tente, avec beaucoup de succès, de faire justement cela pour les langues africaines :

> Masakhane is a grassroots organization whose mission is to strengthen and spur NLP research in African languages, for Africans, by Africans. Despite the fact that two thousand of the world's languages are African, African languages are barely represented in technology. The tragic past of colonialism has been devastating for African languages in terms of their support, preservation and integration. This has resulted in technological space that does not understand our names, our cultures, our places, our history (Masakhane, 2023).

Les technologies nécessaires pour certifier la validité des données sont bien étudiées et sont de plus en plus mises en œuvre. Dans certains cas, cela peut se faire relativement simplement, comme dans le cas de données recueillies à distance par des satellites. La marque Fairtrade fait face au problème plus difficile de surveiller ses produits et ses producteurs pour veiller à ce que ses normes soient maintenues, mais elle est un exemple de marque de confiance qui a réussi à gérer cette tâche complexe en faisant appel à FLOCERT, une organisation indépendante :

> FLOCERT, an independent organization, checks that the Fairtrade Standards have been met by the farmers, workers and companies that are part of the product supply chains. In order to reassure consumers that this has happened, we license the use of the FAIRTRADE Mark on products and packaging to signal the standards have been met (VideoLectures.NET, 2020).

Or, bien que nous ne voulions pas sous-estimer ce défi, nous croyons qu'il y a lieu d'être optimiste quant au fait que l'initiative « Science ouverte » puisse fournir un cadre dans lequel la tâche de collecte et de certification des données pertinentes peut être conçue et réalisée. Cependant, recueillir et certifier des données en soi n'est pas suffisant pour attester de l'atteinte des IPC, sans parler de l'attribution de la responsabilité. Pour cela, nous devons extraire des informations et des connaissances des données et c'est là que l'IA peut jouer un rôle vital.

Le rôle de l'IA

L'IA et l'apprentissage automatique sont des technologies qui peuvent être utilisées pour extraire des informations utiles de données, et ce, de manière vérifiable et transparente. C'est pourquoi ces technologies ont de plus en plus un rôle à jouer. Par exemple, Aidan O'Sullivan a eu recours à l'IA pour analyser des images satellitaires multispectrales dans le but d'évaluer la qualité de l'eau des lacs partout dans le monde (Schölkopf, 2019). À première vue, bien que cela puisse paraître ne nécessiter qu'un accès aux images satellitaires, certaines données « vérité terrain » relatives à la qualité de l'eau recueillies dans différents lacs jouent un rôle essentiel qui fournissent les données d'apprentissage permettant à l'IA d'évaluer correctement la qualité de l'eau à partir de plusieurs mesures multispectrales et d'en arriver à formuler des généralisations, et ce, à partir d'un petit nombre de mesures vérité terrain. Il s'agit d'un exemple du besoin de recueillir des données locales nécessitant une validation et une certification appropriées. Or, dans des cas particuliers, il peut aussi être nécessaire de peaufiner les méthodes d'IA afin de quantifier la précision des prévisions.

Une fois de plus, cet exemple illustre les diverses contributions nécessaires et la manière dont une collaboration de bonne volonté peut créer un écosystème qui inspirera confiance en raison

de sa transparence, de son ouverture et de sa connectivité. Nous reviendrons sur ce thème plus bas, mais nous devons d'abord discuter d'une composante technologique essentielle requise, mais qui, à ce jour, n'a pas atteint la maturité nécessaire : les jumeaux numériques et la modélisation mathématique pour l'IA qui permettent aux modèles complexes de surveiller les IPC et fournissent des éléments de preuve soutenant un lien de causalité entre les actions et les résultats.

Le défi consiste à attribuer le mérite ou la responsabilité des changements dans les IPC aux divers acteurs impliqués. Il peut s'agir de preuves de l'exploitation continue d'une ressource, comme la déforestation, ou de preuves d'interventions visant à résoudre les problèmes à l'origine d'une tendance négative, comme les interventions visant à améliorer la qualité de l'eau. L'analyse de la causalité en apprentissage automatique est bien établie (Schölkopf, 2019), mais nécessite d'être adaptée à ce que l'on appelle désormais les jumeaux numériques. Ces derniers sont des modèles informatiques d'un phénomène ou d'un écosystème en particulier qui peuvent être utilisés pour tester la façon dont les diverses interventions ont influencé les différents IPC ou pourraient les influencer. Par conséquent, en construisant un modèle complexe d'un environnement particulier, nous sommes en mesure de répondre aux questions de type « et si » et d'attribuer la responsabilité des changements observés et documentés. Comme susmentionné, un modèle complexe nécessitera des avancées en IA et en modélisation mathématique, des avancées qui s'appuieront plus particulièrement sur de récentes avancées comme le programme d'ingénierie centrée sur les données du Alan Turing Institute (ATI, 2021).

QUELS SONT LES OBSTACLES À LA RÉALISATION DU MANIFESTE ?

Plusieurs enjeux peuvent entraver la mise en œuvre du manifeste et il est sage d'évaluer les risques que ces enjeux pourraient poser pour sa réalisation. En voici une courte liste.

La première préoccupation est le manque de définitions communes des résultats et de moyens pour les mesurer dans lesquels le public, les entreprises et les ONG ont confiance. Les IPC des ODD établis par les Nations Unies se veulent un point de départ, mais, afin de bâtir l'entente et la confiance requises, cette question devra faire l'objet d'une attention particulière, jumelée à un engagement technique et public.

Ceci mène naturellement à la deuxième préoccupation, à savoir qu'il existe un problème d'action collective concernant qui est, et qui devrait être, responsable d'établir les définitions de résultats et les solutions technologiques qui les capturent et les consignent. Le sujet de bâtir des solutions pour mesurer et vérifier les résultats n'a pas d'attrait évident pour le financement et, compte tenu de sa nature, nous ne savons pas quel serait l'organe de financement idéal pour ce type d'initiative ni le calendrier pour le rendement du capital investi. Notre manifeste est conçu pour plaider en faveur de ce financement en faisant valoir qu'il est judicieux d'investir dans de telles initiatives, mais laisse sans réponse la question relative aux sources potentielles de ce financement.

La troisième préoccupation concerne l'éthique et la confidentialité des données ainsi que ce qui est éthique de recueillir et de stocker. Cette question doit être considérée en collaboration avec les personnes et les collectivités concernées afin de bâtir la confiance nécessaire en ce qui a trait à la façon dont sont utilisées les données, et ce, en suivant les lignes directrices de la Recommandation sur l'éthique de l'IA de l'UNESCO.

La dernière préoccupation que nous souhaitons soulever est celle de la gouvernance, soit la question à savoir qui est responsable « d'approuver » une définition de résultat ou l'utilisation de l'IA pour mesurer les ODD. Bien entendu, il est essentiel que la gouvernance soit responsable et transparente afin d'engendrer la confiance requise. Cette dernière composante s'ajoute aux précédentes et est essentiellement l'élément clé pour que le manifeste soit crédible et efficace.

Si nous souhaitons exploiter le potentiel des données et de l'IA pour mesurer les progrès réalisés envers l'atteinte des ODD, créer de la crédibilité et permettre des investissements dans des entreprises qui se concentrent sur ces objectifs, nous devons éliminer tous ces obstacles.

COMMENT POUVONS-NOUS ÉLIMINER LES OBSTACLES ?

La directive la plus importante pour éliminer tous les obstacles et les facteurs de risque est peut-être de travailler en collaboration : le public, les investisseurs et investisseuses, les entreprises, les gouvernements, les organisations internationales et les ONG doivent s'unir pour établir des normes relatives à la définition des résultats, à la collecte des données, à la vérification, à la désignation de la responsabilité, etc. Ce n'est qu'en assurant un consensus que la méthodologie acceptée ne pourra être discréditée par les critiques d'une ou de plusieurs parties prenantes.

Le deuxième principe directeur est de commencer à petite échelle pour instaurer la confiance, la confiance ne pouvant être bâtie du jour au lendemain. Plutôt que de cibler dès le départ les 17 ODD, nous devrions commencer par un nombre limité d'ODD afin de démontrer le potentiel des données et de l'IA dans l'élimination des obstacles à la confiance et le renforcement de la crédibilité de l'approche. Cela permettra de vérifier les hypothèses clés concernant l'instauration de la confiance et la perception des risques et de constater si cela débloque davantage d'investissements pour réaliser les ODD.

Le troisième principe directeur est de tirer parti d'applications et d'outils existants qui peuvent être adaptés. Plusieurs solutions d'IA émergentes pourraient appuyer cette ambition. Pour ne pas réinventer la roue, nous devons comprendre ce qui existe et ce qui peut être utilisé et adapté.

Nous avons déjà souligné l'importance de mettre en œuvre une gouvernance transparente. Selon nous, cela peut être atteint en établissant des méthodologies acceptées pour déterminer les définitions des résultats, les approches à la collecte de données et autres mécanismes qui sont acceptables à toutes les parties prenantes et en déterminant de quelle manière ce processus d'approbation se réalise. Nous avons besoin d'un cadre de responsabilité pour les résultats relatifs aux ODD qui permet aux organisations de vérifier ce qui a été réalisé. Le cadre de responsabilité doit être conçu en partenariat avec le public, les investisseurs et investisseuses, les entreprises, les gouvernements et les ONG afin de bâtir la confiance et l'utilité. Cet engagement et ce partenariat doivent inclure les personnes et les collectivités concernées pour que ces derniers tiennent compte de leurs expériences et de leurs attentes. Nous ne pouvons pas laisser des entreprises ou des ONG détachées des expériences quotidiennes des personnes établir ce qu'est pour elles un résultat et la façon dont ce dernier devrait être mesuré.

En somme, il serait essentiel que les entreprises s'impliquent, mais nous devons du même coup veiller à ce que la méthodologie soit définie par un groupe de parties prenantes élargi et représente l'intérêt supérieur de toutes les sociétés à l'échelle internationale. Il est naturel de supposer que les organisations internationales assument un rôle de premier plan dans ce processus avec les Nations Unies et l'UNESCO, les centres de catégorie 2, comme l'IRCAI, fournissant une assistance technique.

VERS UN PARTENARIAT MONDIAL

L'éventail des compétences et l'étendue des zones géographiques concernées font de la mesure des ODD un défi réellement mondial qui nécessite, dans chacune des régions, l'engagement d'équipes de recherche locales pouvant répondre à cet appel à l'action. Cela signifie un financement distribué ne provenant pas d'une source ou d'un organe de financement, mais de diverses sources, et la portée,

dont la taille et la quantité de contributions, allant d'appels ouverts pour des solutions techniques à des microprojets dans des établissements de recherche en IA. Pour réussir, la clé est d'instaurer la confiance dans l'approche et ceci ne peut être imposé au monde, mais se réalise plutôt en créant une large coalition incluant des ONG, des entreprises, des gouvernements et des organisations internationales. Ce n'est qu'en nous assurant que les populations partout se sentent représentées que cette approche pourra obtenir le soutien et la confiance de toutes les sociétés.

Pour que cette démarche réussisse, un élément essentiel sera la transparence en ce qui a trait à ce qu'une technologie ou une solution peut fournir, c'est-à-dire une description de ses avantages et de ses désavantages potentiels, afin que des critiques ne puissent créer un narratif suggérant « qu'on nous trompe ». Il est aussi primordial de créer un langage de données commun pour faciliter les discussions et les ententes multipartisanes relativement à des stratégies appropriées ou, en d'autres mots, pour dépolitiser la discussion. Il pourrait être plus facile d'atteindre tous ces desiderata si l'attention est d'abord portée sur un seul ODD ou sur un petit sous-ensemble d'ODD pour lesquels les points de vue sont moins polarisés. En renforçant la confiance dans ce contexte, l'occasion serait créée d'élargir la démarche à d'autres ODD qui représentent de plus grands défis.

Ce partenariat mondial pourrait initialement être piloté en bâtissant une communauté de recherche en développement durable et IA grâce à un réseau renforçant les centres d'excellence et de recherche en IA partout dans le monde et facilitant la collaboration et le réseautage. Ce nouveau réseau mondial de centres d'excellence en IA en développement durable aurait pour objectif de mettre en place un réseau permettant de renforcer grandement les capacités de recherche dans ce domaine et de le rendre attrayant aux scientifiques, aux nouveaux talents, aux investisseurs et investisseuses, tant à impact social qu'à capital de risque, et aux décideurs et décideuses. Cette initiative devrait aussi contribuer au développement d'une IA éthique et digne de confiance, comme décrite dans la recommandation de l'UNESCO.

CONCLUSION

> Plus tôt dans ce chapitre, nous avons plaidé en faveur de ce manifeste, mais il est utile d'explorer quelles autres retombées pourraient découler de sa réalisation réussie. Le fait que les marchés financiers constituent une machinerie très sophistiquée pour garantir que les ressources investies produisent le meilleur rendement financier possible est une analogie efficace. Le développement durable remet en question la croyance voulant que le rendement financier doive être la seule façon de mesurer les retombées des investissements et nous avons fait valoir que ce point de vue bénéficie d'un soutien de plus en plus important. Toutefois, il n'existe pas de mécanisme correspondant pour mesurer la performance des entreprises par rapport à ces nouveaux critères. Si nous voulons « joindre l'acte à la parole », il est urgent de créer des mécanismes tels que le préconise notre manifeste. Ce n'est que par une utilisation plus efficace des données et de l'IA que nous pourrons éviter l'effet « d'écoblanchiment » par lequel les entreprises, par le biais du marketing et des relations publiques, déclarent au public et à leurs clients et clientes qu'elles réalisent les ODD, alors que ce n'est pas le cas. Plus important encore, cela ouvrirait une voie robuste et vérifiable aux investisseurs et investisseuses d'appuyer le développement durable et aux entreprises de convaincre des investisseurs et investisseuses et les gouvernements de la valeur de leurs produits pour la société. Cela permettrait aussi aux entreprises de montrer la valeur ajoutée et les économies potentielles que leurs produits pourraient contribuer aux gouvernements en matière d'améliorations quantifiables aux ODD, fournissant de ce fait de l'information utile pour les obligations à impact social.

RÉFÉRENCES

Amaro, S. 2021. Blackrock's former sustainable investing chief says ESG is a dangerous placebo. 24 août. CNBC. https://www.cnbc.com/2021/08/24/blackrocks-former-sustainable-investing-chief-says-esg-is-a-dangerous-placebo.html

ATI, 2021. Data-centric engineering. The Alan Turing Institute. https://www.turing.ac.uk/research/research-programmes/data-centric-engineering

Edelman. 2020. *21st Annual Edelman Trust Barometer*. https://www.edelman.com/sites/g/files/aatuss191/files/2021-01/2021-edelman-trust-barometer.pdf

Fairtrade Foundation. n.d. What Fairtrade does. https://www.fairtrade.org.uk/what-is-fairtrade/what-fairtrade-does/

Ferrando, T., Miola, I., Junqueira, G. O., Prol, G. M., Vecchione-Goncalves, M. et Herrera, H. 2021. Capitalizing on green debt, Journal of World-Systems Research, vol. 27, nº 2, pp. 410-437.

Schölkopf, B. 2019. Causality for machine learning. https://arxiv.org/abs/1911.10500

Social Finance. 2021. Cambodia rural sanitation : Bringing safe sanitation to rural communities in Cambodia. https://www.socialfinance.org.uk/projects/cambodia-rural-sanitation

Masakhane. 2022. Page d'accueil de Masakhabe. https://www.masakhane.io/

UNESCO. 2020. Page d'accueil de Open Science. https://www.unesco.org/en/natural-sciences/open-science

UNPRI. 2020. Investing with SDG Outcomes : A Five-Part Framework. https://www.unpri.org/sustainable-development-goals/investing-with-sdg-outcomes-a-five-part-framework/5895.article (consulté le 15 février 2022).

UNPRI. 2022. About the PRI. https://www.unpri.org/about-us/about-the-pri

USAID. 2019. The Cambodia Rural Sanitation DIB : Lessons Learnt from the First Year. https://www.thesff.com/system/wp-content/uploads/2021/03/Development-Impact-Bond-lessons-learnt-March-2021.pdf

VideoLectures.NET. 2020. AI and Climate : Water Quality Measurement System. http://videolectures.net/IRCAILaunch2021_sullivan_ai_climate/

West, T. A. P., Börner, J., Sills, E. O. et Kontoleon, A. 2020. Overstated carbon emission reductions from voluntary REDD+ projects in the Brazilian Amazon, PNAS, vol. 117, nº 39. pp. 24188-24194. https://doi.org/10.1073/pnas.2004334117

Zamculture. 2019. Zambia school mapping project. https://github.com/Zamculture/Zambia-School-Mapping-Project–

L'IA AU SERVICE DES ODD – ET APRÈS ? VERS UNE CULTURE HUMAINE DE L'IA POUR LE DÉVELOPPEMENT ET LA DÉMOCRATIE

EMMANUEL LETOUZÉ

Marie-Curie Fellow, Universitat Pompeu Fabra et Directeur, Data-Pop Alliance.

NURIA OLIVER

Directrice Scientifique, ELLIS Unit Alicante Foundation et Chief Data Scientist, Data-Pop Alliance.

BRUNO LEPRI

Chercheur, Fondazione Bruno Kessler et Data-Pop Alliance.

PATRICK VINCK

Professeur adjoint, Université Harvard et co-Directeur, Data-Pop Alliance.

- **ODD 1** - Pas de pauvreté
- **ODD 2** - Faim « zéro »
- **ODD 3** - Bonne santé et bien-être
- **ODD 5** - Égalité entre les sexes
- **ODD 8** - Travail décent et croissance économique
- **ODD 9** - Industrie, innovation et infrastructure
- **ODD 10** - Inégalités réduites
- **ODD 11** - Villes et communautés durables
- **ODD 12** - Consommation et production durables
- **ODD 16** - Paix, justice et institutions efficaces
- **ODD 17** - Partenariats pour la réalisation des objectifs

L'IA AU SERVICE DES ODD – ET APRÈS ? VERS UNE CULTURE HUMAINE DE L'IA POUR LE DÉVELOPPEMENT ET LA DÉMOCRATIE

RÉSUMÉ

L'intelligence artificielle (IA) peut contribuer aux objectifs de développement durable (ODD) et au Programme de développement durable à horizon 2030 des Nations Unies qui visent entre autres à éliminer l'extrême pauvreté, à réduire les inégalités entre les sexes, à protéger les écosystèmes naturels et à promouvoir des sociétés inclusives. Une des façons pour y arriver est d'utiliser l'IA et les nouvelles « miettes d'information numérique » pour estimer des indicateurs des ODD afin de prendre des décisions plus éclairées. Or, dans un monde où la démocratie est de plus en plus mise à l'épreuve, notamment en raison de l'influence qu'exerce l'IA sur les inégalités et la polarisation, l'utilisation de l'IA au service du progrès humain et des ODD exige des changements plus profonds que de simplement fournir un meilleur carburant à un vieux moteur. Les principaux écueils et le potentiel de l'IA ne sont pas technologiques, mais bien politiques et culturels.

Notre chapitre évalue de manière critique les principes et les lacunes clés du narratif et des initiatives « d'IA au service des ODD » et discute des limites et des conditions d'une vision de la « culture humaine de l'IA » grâce à laquelle les sociétés apprennent à utiliser l'IA en tant qu'inspiration et instrument contrôlé par les humains pour devenir meilleures. Pour ce faire, il est nécessaire d'accroître la sensibilisation et d'acquérir des compétences et des systèmes afin d'assurer un suivi de tous les ODD, y compris les plus sensibles sur le plan politique qui concernent la liberté de la presse, en considérant de nouveaux objectifs et en favorisant la participation et la collaboration de tous les citoyens et de toutes les citoyennes concernés aux initiatives utilisant l'IA et inspirées par elle. À cette fin, nous appelons les citoyens et citoyennes, les décideurs et décideuses, les scientifiques, les éducateurs et éducatrices, les donateurs et donatrices, les journalistes, les membres de la société civile et les employés et employées à lire ce chapitre et à réfléchir aux perspectives qui y sont présentées dans l'espoir qu'ils et elles façonneront l'IA et en tireront parti dans le but de promouvoir et de protéger le développement humain et la démocratie d'ici 2030 et après.

INTRODUCTION

En septembre 2021, le magazine *Wired* publiait un article intitulé « How Valencia crushed COVID with AI » (Marx, 2021). Décrivant une initiative primée menée par Nuria Oliver, l'une des personnes ayant participé à la rédaction de ce chapitre, l'article décrit une instance où l'intelligence artificielle (IA), utilisant des métadonnées de téléphones cellulaires et des données de sondages épidémiologiques et en ligne, a été utilisée par le gouvernement pour prendre des décisions de politiques éclairées ayant des conséquences directes sur la santé publique et les activités économiques. Cette initiative est un exemple de vision positive où l'IA, le nouvel épicentre de la « Révolution des données », peut aider l'humanité à progresser vers l'atteinte d'objectifs communs, dont les 17 objectifs de développement durable (ODD) des Nations Unies (ONU) et le programme qui les sous-tend, ces derniers ayant été adoptés officiellement par 193 États membres en septembre 2015.

Dans sa version simplifiée, l'argument sur lequel s'appuie le discours dominant en faveur de l'utilisation de l'IA au service des ODD repose sur le fait que l'explosion de la quantité et de la diversité des données relatives aux actions et aux interactions humaines recueillies par des appareils et services numériques (c.-à-d. les mégadonnées) ainsi que les améliorations parallèles des systèmes algorithmiques capables d'apprendre de ces données (p. ex., l'apprentissage automatique) pourraient aider les décideurs et décideuses, les chercheurs et chercheuses, les organisations non gouvernementales (ONG), les entreprises et d'autres groupes concernés à mieux mesurer et, par la suite, influencer les processus et les résultats reflétés dans les ODD ou pertinents à ceux-ci. Plusieurs initiatives et publications, notamment de notre groupe, suggèrent qu'il existe une vérité partielle dans cette proposition de valeur : les indicateurs, les informations et les initiatives alimentés par l'IA peuvent, bien entendu, éclairer des décisions et des mesures qui contribuent aux ODD. Mais il est temps de reconnaître que cet argument et une grande partie des discussions qui l'entourent n'abordent pas certains éléments précis, les nuances, les écueils et les zones grises (Letouzé, 2015b).

Par exemple, un problème important de telles discussions est la supposition que les bonnes intentions des décideurs et décideuses ou des dirigeants et dirigeantes à l'échelle mondiale sont limitées par des informations insuffisantes ou inadéquates et que le fait d'éliminer cette contrainte, grâce à des méthodes d'IA, aurait un impact majeur. En réalité, les principaux obstacles empêchant les données et l'IA d'être utiles pour les ODD et le développement humain ne sont généralement pas fondamentalement technologiques. Ces obstacles sont plutôt les incitatifs, les dynamiques de pouvoir et les iniquités qui déterminent qui contrôle et utilise les ressources clés. Par conséquent, pour cette raison et d'autres encore, nous sommes d'avis que la vision de l'IA au service des ODD nécessite une théorie du changement plus claire et plus audacieuse ainsi qu'un meilleur plan, tous deux fondés sur une base conceptuelle et contextuelle solide.

Cette contribution porte particulièrement sur deux sujets : (1) l'absence de discussion sur le rôle que jouent la politique, le pouvoir et, ultimement, la culture dans le contexte de l'utilisation de l'IA dans les efforts visant l'atteinte des ODD et (2) les changements pragmatiques et les ingrédients qui, selon nous, sont nécessaires pour que l'IA satisfasse aux attentes à son égard et échappe aux prévisions les plus inquiétantes.

Une proposition clé est de créer les conditions nécessaires pour instaurer une culture humaine de l'IA dans laquelle l'IA sera utilisée comme instrument contrôlé par des humains et comme source d'inspiration pour favoriser l'émergence de sociétés d'apprentissage.

Pour ce faire, nous utilisons un cadre analytique que l'on appelle « les 4 C pour l'IA » qui aide à décrire d'une manière systématique et structurée les exigences et les éléments constitutifs essentiels de l'IA et à en discuter. Nous proposons également une taxonomie des canaux de contributions, y compris le « canal de mesure », considérant des cas d'utilisation courants afin de révéler la théorie du changement liant de manière explicite les applications d'IA et les résultats de développement. Nous résumons

ensuite les principaux obstacles et risques auxquels nous faisons actuellement face en utilisant les 4 C comme cadre. Enfin, compte tenu de la résistance politique et économique au changement, nous décrivons les caractéristiques d'une nouvelle théorie du changement et d'une nouvelle vision que nous appelons la « culture humaine de l'IA » qui, selon nous, pourrait appuyer les programmes d'ODD et les programmes démocratiques au cours de la prochaine décennie et après, y compris les cibles des ODD les plus délicates sur le plan politique et d'autres objectifs.

L'IA ET LES ODD : CLARIFICATIONS CONCEPTUELLES ET CONTEXTUELLES

L'IA est une discipline de l'informatique ou de l'ingénierie qui comprend une variété de méthodes et de domaines (Vinuesa *et al.*, 2020), comme l'apprentissage automatique, la vision par ordinateur, le traitement du langage naturel et la parole, appliqués à un large éventail de secteurs ayant divers degrés d'impacts sociétaux. Bien que l'IA en tant que discipline existe depuis les années 1950, plusieurs facteurs interreliés lui ont, au cours des quinze dernières années, donné un élan et un renouveau (Lazer *et al.*, 2009). Premièrement, la disponibilité de riches ensembles de données numériques d'envergure qui fournissent le carburant aux méthodes alimentées par l'IA fondée sur les données. Deuxièmement, l'amélioration des capacités de calcul et le développement d'algorithmes d'apprentissage automatique sophistiqués, nommés apprentissage profond, qui peuvent apprendre de données à grande échelle en tirant parti de la grande performance des calculs (King, 2013). Troisièmement, l'émergence et la croissance d'écosystèmes d'entreprises, de groupes de recherche, d'organisations publiques et internationales et de citoyens-consommateurs et de citoyennes-consommatrices. Enfin, le quatrième facteur qui a insufflé un élan à l'IA est l'avènement d'un état d'esprit et d'une culture qui valorisent l'efficacité et la prévisibilité, mais aussi, dans une certaine mesure, la responsabilité, la rentabilité et la mesure, des caractéristiques ancrées dans l'adage qui veut que « vous ne puissiez pas gérer ce que vous ne pouvez pas mesurer » (Weigend, 2013). Un bon exemple de la puissance de ces facteurs interreliés est la performance améliorée des systèmes de traduction de langage en temps réel. Par conséquent, en nous appuyant sur des travaux existants (King, 2013 ; Weigend, 2013 ; Letouzé 2014 ; Letouzé 2015a), nous proposons que l'IA doive être conceptualisée et discutée non pas en tant que simple discipline technologique, mais bien en tant que phénomène sociotechnologique composé de quatre éléments clés (figure 1).

1. *Crumbs* (les miettes d'information numérique) : les données numériques que les humains « laissent derrière » (Pentland, 2012) comme sous-produits de leurs actions et de leurs interactions impliquant des appareils et des services numériques (Letouzé *et al.*, 2013) (consulter le tableau 1 de l'annexe). Ces dernières constituent l'intrant brut des méthodes d'IA fondées sur les données.

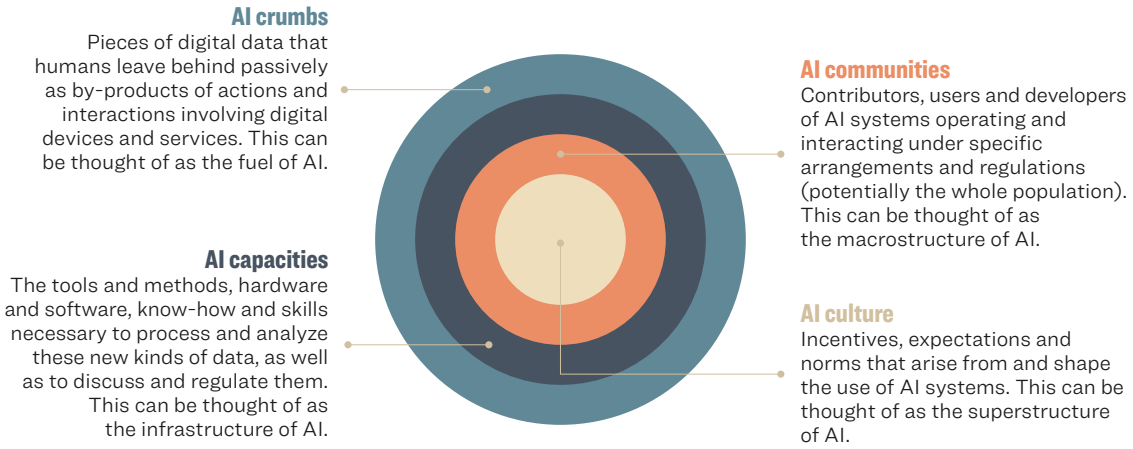
2. *Capacities* (les capacités) : les outils et les méthodes, le matériel et les logiciels, le savoir-faire et les compétences nécessaires pour traiter et analyser ce nouveau type de données. Les capacités peuvent être considérées comme étant l'infrastructure de l'IA.

3. *Communities* (les communautés) : les contributeurs et contributrices, les utilisateurs et utilisatrices et les développeurs et développeuses des systèmes d'IA exploitant et interagissant dans le respect d'arrangements et de règlements précis, dont les agences des Nations Unies et autres parties prenantes au sein du mouvement élargi de la « Révolution des données ». Les communautés peuvent être considérées comme étant la macrostructure de l'IA.

4. *Culture* (la culture) : l'ensemble des incitatifs, des attentes, des idéologies et des normes qui façonnent l'utilisation des systèmes d'IA et qui en découlent, p. ex., la superstructure de l'IA dans un sens marxiste.

| FIGURE 1 |

Les 4 C de l'IA en tant que phénomène sociotechnologique, basée sur Letouzé (2015).



Nous croyons que ce cadre conceptuel aide à évaluer et à discuter de manière structurée et holistique les caractéristiques et les exigences actuelles et futures de l'IA en tant que partie d'un écosystème complexe.

Il est aussi utile pour décrire la genèse et le contexte de « l'IA au service des ODD » et des narratifs et initiatives de la « Révolution des données ». Dans les faits, l'un des premiers rapports portant sur le nexus de l'IA et des ODD est antérieur aux deux. En 2012, UN Global Pulse a publié le papier blanc intitulé « Big Data for Development : Challenges and Opportunities » (UN Global Pulse, 2012) qui jetait les bases de plusieurs discussions qui ont, depuis, eu lieu. En 2013, le Groupe de personnalités de haut niveau chargé d'étudier le programme de développement pour l'après-2015 a appelé à « une révolution des données pour le développement durable » (consulter la figure 2). Un an plus tard, un groupe consultatif d'experts indépendants nommé par le secrétaire des Nations Unies a publié le rapport intitulé « A World that Counts : Mobilizing the data revolution for sustainable development » (IEAG, 2014). L'attente était, et demeure, que l'IA puisse contribuer à lutter contre la pénurie de statistiques officielles dans les pays en développement (Letouzé et Jütting, 2015), une situation nommée une « tragédie statistique » (Devarajan, 2013) ou un « tarissement de données » (*The Economist*, 2014), ce qui améliorerait alors les résultats de développement, comme le suggèrent les phrases suivantes : « better data for better decisions and better lives » (Melamed, 2018) et « [d]ata are the lifeblood of decision-making and the raw material for accountability » (IEAG, 2014).

| FIGURE 2 |

(Groupe de personnalités de haut niveau chargé d'étudier le programme de développement pour l'après-2015, 2013).

> "Too often, development efforts have been hampered by a lack of the most basic data about the social and economic circumstances in which people live… Stronger monitoring and evaluation at all levels, and in all processes of development (from planning to implementation) will help guide decision making, update priorities and ensure accountability. This will require substantial investments in building capacity in advance of 2015. A regularly updated registry of commitments is one idea to ensure accountability and monitor delivery gaps. We must also take advantage of new technologies and access to open data for all people."
>
> *Bali Communiqué of the High-Level Panel, March 28, 2013*

Depuis, plusieurs groupes ont soutenu tirer parti de l'IA au service des ODD et de nombreux efforts ont été déployés en ce sens (Vinuesa *et al.*, 2020 ; Tomašev *et al.*, 2020)[55]. Or, la question fondamentale, à savoir quelle incidence l'IA a exactement sur les ODD ou peut avoir sur eux, c.-à-d. la théorie (ou les théories) sous-jacente du changement concerné, n'a pas été suffisamment investiguée ni articulée. Les auteurs et autrices de cette contribution ont proposé d'examiner les diverses « fonctions de l'IA », comme la prévision, la prescription, etc. (Letouzé *et al.*, 2013), tandis que d'autres ont suggéré de structurer l'analyse par secteur d'impact (Vinuesa *et al.*, 2020). Dans cette contribution, nous utilisons une taxonomie bâtie autour de quatre canaux et modalités de contribution pour rendre explicites les relations causales possibles entre les applications d'IA et les résultats concrets : mesure et suivi, exactitude et ingéniosité, conception, suivi et évaluation des politiques et autres activités.

L'IA au service des ODD : quatre canaux de contribution

Nous avons établi les quatre canaux principaux suivants :

1. Un canal « Mesure et suivi » qui vise à combler les « lacunes dans les données » et à améliorer la conscience situationnelle relativement à des indicateurs d'ODD particuliers ou à des indicateurs qui leur sont étroitement associés.
2. Un canal « Exactitude et ingéniosité » grâce auquel sont explicitement conçus les produits et les services fondés sur l'IA afin qu'ils aient un impact sur au moins un des domaines ciblés par les ODD.
3. Un canal « Conception, suivi et évaluation des politiques » avec la conception d'approches émergentes fondées sur l'IA qui cherchent à concevoir et à mettre en œuvre des politiques et des programmes fondés sur des faits.
4. Un canal « Autres activités » qui comprend tout autre système d'IA qui n'a pas été conçu spécifiquement en tenant compte des ODD, parce que les développeurs et développeuses pourraient ne jamais en avoir entendu parler, mais qui aura une incidence sur les ODD dans le futur.

La liste est loin d'être exhaustive, mais vise à fournir de manière structurée un résumé de l'état de la situation.

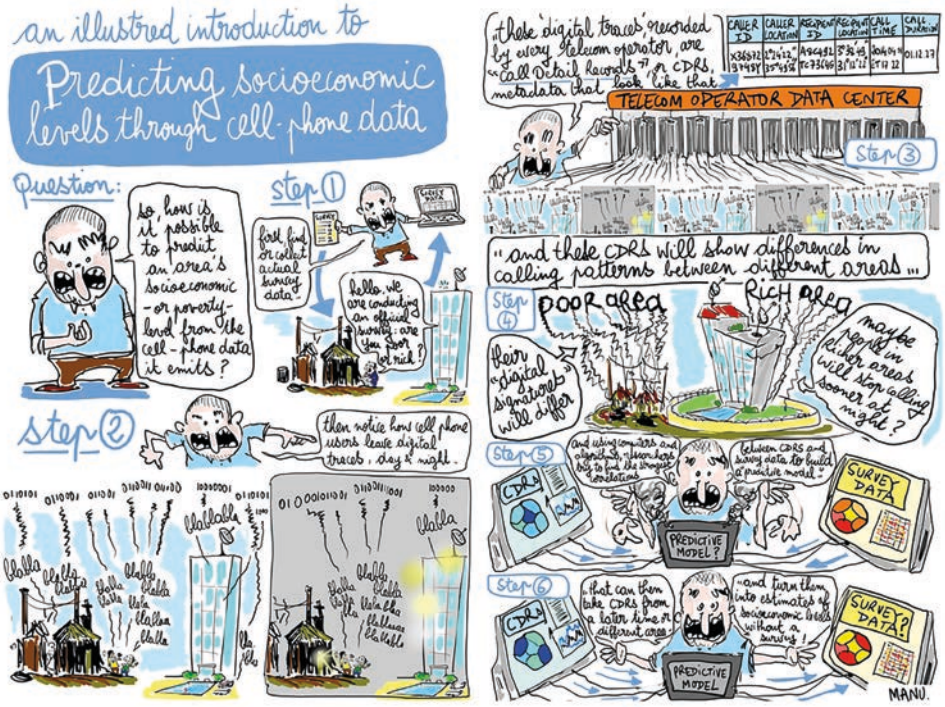
55. Des listes des efforts pertinents visant à tirer parti de l'IA au service des ODD ont été établies dans plusieurs référentiels. Par exemple, le SDG AI Repository (2021) de l'ITU, la base de données du AI4SDGs Think Tank (2021) et la base de données de l'initiative de recherche AIxSDGs (Saïd Business School, 2021) de l'Université d'Oxford, qui contient plus de 100 projets.

Canal de contribution et efforts : « Mesure et suivi »

Comme suggéré plus haut, on soutient depuis longtemps que l'IA pourrait contribuer à la promotion des ODD en aidant à les mesurer et à les suivre. Les buts et les indicateurs associés aux ODD qui ont été mesurés ou estimés par des approches d'IA sont généralement ceux qui se voient dans les miettes d'information numérique (p. ex., la consommation d'électricité nous en dit long sur le statut socio-économique) et sont actuellement suivis par des données courantes qui fournissent une vérité terrain. Les principes et les étapes de base de ces approches sont décrits dans la figure 3.

| FIGURE 3 |

Prédire les niveaux socio-économiques à l'aide de données d'utilisation de téléphones cellulaires (Emmanuel Letouzé, 2013).



Plusieurs problèmes relatifs au canal « Mesure et suivi » peuvent être soulevés. Le premier est le risque découlant d'une surveillance de l'État ou d'une entreprise. Un second est la validité scientifique de certaines mesures. Par exemple, il est concevable de créer des systèmes de suivi de la cohésion sociale fondés sur la fréquence des contacts physiques et numériques, qui peut être obtenue de registres d'appels détaillés, mais il reste à déterminer si ces interactions constituent une mesure significative et valable de la cohésion sociale. En outre, de telles mesures sont limitées par les préjugés et les iniquités structurales et en sont souvent le reflet, comme cela sera discuté dans les prochaines sections. Enfin, la question clé, à savoir si et comment de meilleures mesures des résultats de développement, comme dans le cas des ODD, influencent ces résultats, sera discuté plus loin.

La section suivante montre des exemples choisis tirés de plusieurs études et projets pilotes ayant eu recours à l'IA pour estimer des indicateurs compris dans les 17 ODD (Letouzé, 2015a ; Oliver, 2021).

Exemples d'efforts de mesure et de suivi selon l'ODD

L'**ODD 1** a fait l'objet de nombreux efforts tirant parti des données d'observation de la Terre, comme les émissions de lumière et les caractéristiques des toits (Jean *et al.*, 2016), les métadonnées de téléphones cellulaires (Sundsoy *et al.*, 2016 ; Soto *et al.*, 2011), les transactions bancaires numériques et les publicités en ligne (Cruz *et al.*, 2019).

L'**ODD 2** a été ciblé par des techniques d'IA qui analysent des données météorologiques (USAID, 2010), satellitaires, démographiques (Quinn *et al.*, 2010) et socio-économiques (Okori et Obua, 2011) pour détecter la faim et le rendement des cultures dans les pays en développement (Zhu *et al.*, 2018 ; Ghandi et Armstrong, 2016).

L'**ODD 3** a été ciblé par des méthodes d'IA qui suivent des données des médias sociaux pour trouver des éclosions et des épidémies de diverses maladies et des préoccupations concernant les vaccins (Letouzé, 2015b). Des objets personnels connectés abordables ont aussi permis la collecte de données longitudinales à grande échelle (Clifton *et al.*, 2014).

L'**ODD 4** a été ciblé par l'IA par l'entremise de méthodes d'apprentissage automatique visant à mesurer la présence et la performance d'étudiants et d'étudiantes. Par exemple, l'utilisation de données socio-économiques et d'Internet pour prédire les taux de décrochage (Freitas *et al.*, 2020).

L'**ODD 5** a été ciblé par l'IA par le biais de l'utilisation de données de médias sociaux pour révéler des points chauds de violence conjugale. D'autres méthodes d'IA, comme la reconnaissance vocale et l'analyse de conversations, ont été utilisées pour mettre en lumière des préjugés fondés sur le genre et la participation des femmes à des réunions (Fedor *et al.*, 2009).

L'**ODD 6** a été cartographié à l'aide de l'IA par l'entremise de différentes mesures pour détecter et suivre les sources importantes de contamination de l'eau (Wu *et al.*, 2021), y compris les réseaux d'eau potable (Dogo *et al.*, 2019), et pour estimer la consommation d'eau dans les régions urbaines et rurales (Brentan *et al.*, 2017).

L'**ODD 7** a été ciblé par l'IA par le biais de techniques qui peuvent estimer l'accès à de l'énergie pour l'électrification et à du carburant de cuisine propre grâce à une observation très fréquente de la Terre (Pokhriyal *et al.*, 2021).

L'**ODD 8** a été cartographié à l'aide de l'IA par l'entremise de l'utilisation de données satellitaires pour estimer le PNB à l'échelle nationale et infranationale et de données d'Internet pour estimer les taux d'inflation (Letouzé, 2015b).

L'**ODD 9** a été ciblé par l'IA par le biais de techniques permettant de suivre les infrastructures existantes en analysant les images aériennes (Bao *et al.*, 2019 ; Ren *et al.*, 2020 ; Xu *et al.*, 2019) et de détecter la construction d'infrastructures, la production de polluants industriels (Xu *et al.*, 2015) et les anomalies dans la consommation d'énergie.

L'IA a tenté de cartographier l'**ODD 10** en utilisant des données de crédit de temps d'antenne et de téléphones cellulaires pour établir le statut socio-économique.

L'**ODD 11** a été ciblé par des techniques d'IA axées sur la planification urbaine, l'estimation de la densité urbaine à partir d'images aériennes (Lu *et al.*, 2010), l'étude de l'utilisation des transports par l'entremise de données de cartes de transport et l'établissement de points chauds pour les crimes (Bogomoloy, 2014) et le trafic de drogues illicites (Li *et al.*, 2019).

L'**ODD 12** a été ciblé par l'IA par le biais de la création de cartes sur l'utilisation du territoire afin de fournir un portrait précis de l'état de l'utilisation des ressources naturelles (Talkudar *et al.*, 2020) et d'inférer des comportements de consommation et d'élimination socialement responsables (Talkudar *et al.*, 2020).

L'**ODD 13** a été cartographié par l'IA par l'entremise de données satellitaires pour mesurer la production nette principale de méthane et faire des observations sur le méthane ainsi que pour suivre les émissions de gaz à effet de serre provenant des populations et de l'énergie (Letouzé, 2015b).

L'**ODD 14** a été ciblé par l'IA par le biais de projets qui suivent la qualité des océans en utilisant des méthodes d'apprentissage profond, l'analyse d'images aériennes et satellitaires et la classification et qui ont permis d'estimer le volume de débris de plastique (Martin *et al.*, 2018) et le flux de CO_2 (Chen *et al.*, 2019) ainsi que détecter des fuites de pétroles (Jiao *et al.*, 2019).

L'**ODD 15** a été cartographié par des méthodes d'IA qui assurent un suivi de la déforestation (de Bem *et al.*, 2020), de la qualité des forêts (Zhao *et al.*, 2019) et de la biomasse aérienne (Madhab Ghosh and Behera, 2018) ainsi que de la classification de la faune sauvage (Tabak *et al.*, 2018) et de la détection de commerce illégal d'espèces sauvages (Di Minin *et al.*, 2019).

L'**ODD 16** a été ciblé par l'IA axée sur la corruption et l'extrémiste en appliquant des algorithmes d'IA à la corruption au sein des gouvernements (Adam and Fazekas, 2018) et aux transactions financières (West and Bhattacharya, 2016) et, aussi, par l'entremise du traitement du langage dans le contenu des médias sociaux (Johansson *et al.*, 2017).

Canal de contribution et efforts : « Exactitude et ingéniosité »

Les efforts déployés dans ce canal utilisant l'IA ne cherchent pas à mesurer un ODD, mais à optimiser les systèmes et les processus qui éclairent la prise de décisions dans des domaines ciblés par au moins un des 17 ODD. Ils sont généralement décrits comme étant « exacts » ou « ingénieux » et appliqués à des domaines comme l'agriculture, la médecine et les soins de santé, le développement urbain, etc. Un tel exemple est le Famine Action Mechanism (FAM) qui prend en charge l'analyse de risque, le financement et la programmation pour lutter contre la famine (ODD 2) (Badr *et al.*, 2016). L'IA peut également améliorer le bien-être d'un enfant par la détection précoce des besoins (Schwartz *et al.*, 2017), ce qui a une incidence sur les inégalités (ODD 10). D'autres initiatives aident à la prise de décisions cliniques et de santé publique en produisant, par exemple, des prévisions de cancer (Esteva *et al.*, 2017), de tuberculose (Doshi, 2017), de la probabilité de soins intensifs (Kaji *et al.*, 2019) et des besoins de soutien en santé mentale (Walsh *et al.*, 2017).

D'autres systèmes pertinents pour les ODD 9 et 11 visent à optimiser la collecte des déchets et le recyclage et à prédire des modèles de matières résiduelles solides (Kannangara, 2018). Les efforts promouvant la production et la consommation responsables et l'action pour le climat (ODD 12 et 13) portent plus particulièrement sur l'optimisation des systèmes de production, comme l'estimation de l'impact de l'exploitation forestière (Hethcoat *et al.*, 2019) et la prévision de l'occurrence et des impacts des événements météorologiques extrêmes (Lee *et al.*, 2020 ; Radke *et al.*, 2019 ; Wong *et al.*, 2020, Pastor-Escuredo *et al.*, 2014), comme le projet Artificial Intelligence for Disaster Response qui utilise des données des médias sociaux (Ong *et al.*, 2020). D'autres comprennent des systèmes tutoriels intelligents (STI) et des interfaces éducatives pour aider à concevoir des outils d'apprentissage pour les étudiants et étudiantes en situation de handicap (Abdul Hamid, 2018) qui, quant à eux, sont pertinents pour les ODD 4 et 10. Un autre exemple est *Bob Emploi* (Marion, 2018) qui promettait de faire plus efficacement le lien entre les chercheurs et chercheuses d'emploi et les possibilités d'emploi (ODD 8). Les inquiétudes associées à ce canal sont souvent liées à l'équité et à la gouvernance de systèmes automatisés (Lepri *et al.*, 2017).

Canal de contribution et efforts : « Conception, suivi et évaluation de politiques »

La possibilité d'utiliser l'IA pour améliorer les politiques et les programmes tout au long de leur cycle de vie, de leur conception à leur évaluation, a fait l'objet de beaucoup d'attention au cours des dernières années (Bamberger *et al.*, 2016 ; Letouzé *et al.*, 2019). Un des arguments soulevés est que l'IA et les nouvelles sources de données offrent la possibilité de capturer les réponses comportementales et les perceptions de la population ciblée dans les médias sociaux et d'autres sources de données, et ce, quasiment en temps réel. Cette capacité aide à obtenir une réponse à la question ultime en élaboration de politique : « cette intervention a-t-elle fonctionné ? » ou, mieux encore, « fonctionne-t-elle maintenant ? », ce qui permet une correction plus rapide de la trajectoire. Cette façon de penser se résume par le virage de « prouver » à « améliorer » dans le domaine du suivi et de l'évaluation (Letouzé *et al.*, 2019). Cependant, il existe encore peu d'applications concrètes. Un exemple est l'utilisation de l'IA pour mieux cibler l'assistance sociale (Noriega-Campero *et al.*, 2020) en prédisant les faux positifs (c.-à-d. les personnes qui reçoivent des prestations alors que, selon les règles, elles ne devraient pas) et les faux négatifs (c.-à-d. les personnes qui ne reçoivent pas de prestations alors que, selon les règles, elles devraient). Un autre exemple est l'utilisation de l'IA pour aider à détecter la fraude gouvernementale (West, 2021).

Or, l'IA a des effets contradictoires sur le « défi de l'évaluabilité ». Par exemple, il est difficile de savoir dans quelle mesure la *causalité* peut être établie entre les interventions et les résultats (Bamberger *et al.*, 2016) parce que l'IA peut créer plusieurs boucles de rétroaction et échos qui compliquent davantage l'inférence causale et le pouvoir prédictif, comme dans le cas de l'exemple célèbre de « *l'échec lamentable* » de Google Flu Trends (Lazer *et al.*, 2014). Dans l'avenir, l'IA est susceptible d'influencer

de manière fondamentale l'élaboration de politiques, notamment en contribuant à cerner de nouvelles préoccupations et questions d'intérêt. Mais cela ne devrait pas signifier qu'il est possible d'éviter une conception scientifique rigoureuse fondée sur plusieurs méthodes, comme des lignes directrices formulées à cet effet l'ont soulevé (Bamberger *et al.*, 2016), et que l'IA peut se substituer à des systèmes démocratiques efficaces.

Canal de contribution et efforts : « Autres activités »

Le dernier canal comprend toutes les approches d'IA qui sont utilisées et qui ont un impact sur les ODD au quotidien, de manière positive ou négative, sans avoir été conçues dans cette optique (ou de manière très éloignée). Bien qu'il s'agisse peut-être du moyen le plus puissant par lequel l'IA a une incidence sur les ODD, il est impossible de dire si globalement, et pour qui, l'impact net est positif ou négatif, et ce, à la fois en raison de la multitude des effets sur des personnes et des groupes différents et du fait que ces systèmes sont encore très récents (Vinuesa *et al.*, 2021). Google Maps peut, par exemple, réduire la pollution et le stress en encourageant les gens à éviter de prendre leur voiture lorsqu'il y a du trafic, mais il peut aussi entraîner le décès de chauffeurs qui manipulent leur téléphone en conduisant. Les plateformes de médias sociaux nous aident peut-être à communiquer et à recevoir des informations en temps réel, mais elles renforcent probablement la polarisation politique (Barret, Hendrix et Sims, 2021) et créent des comportements de dépendance (Zheng et Lee, 2016). La question de savoir si les services alimentés par l'IA fournis par Amazon sont globalement positifs ou négatifs pour les gens et la planète peut être débattue sans fin dans un sens ou dans l'autre, selon les perspectives et les indicateurs. Il est donc nécessaire de retenir que les effets de l'IA doivent être évalués et discutés de manière beaucoup plus approfondie, transparente et respectueuse sur la base des données disponibles afin de maximiser les impacts positifs (Vinuesa *et al.*, 2021), tout en gardant à l'esprit qu'il n'y a presque jamais de vérité définitive.

Principaux défis et limites en ce qui a trait aux données, aux capacités, aux communautés et à la culture

Les défis et les limites actuellement associés aux initiatives « d'IA au service des ODD » ont fait l'objet de plusieurs études (Letouzé et Oliver, 2019). Nous les résumons ci-dessous en utilisant comme cadre les 4 C de l'IA : *crumbs* (les miettes d'information numérique, soit les données), *capacities* (les capacités), *communities* (les communautés) et *culture* (la culture).

Miettes d'information numérique : verrouillées, biaisées, désordonnées et sensibles

Bien que nous « baignions dans les données », accéder et utiliser systématiquement et de manière sûre ces miettes d'information numérique pour entraîner les systèmes d'IA demeure un défi de taille. Les sociétés privées contrôlent, de manière juridique ou pratique, ou les deux, la plupart des miettes d'information numérique d'IA et ces dernières hésitent souvent à les partager ou à faciliter leur accès et recueillent fréquemment de telles données avec le consentement ou le contrôle limités des personnes auprès de qui elles sont recueillies. Cette situation se produit en partie en raison de considérations commerciales : certaines entreprises conçoivent, ou concevront prochainement, leurs propres services commerciaux fondés sur des données dans le cadre de stratégies de « monétisation des données » et craignent que le partage de données fournisse des renseignements à leurs concurrents. En outre, certains de ces ensembles de données contiennent des renseignements personnels, ce qui entraîne des risques pour la réputation et des risques juridiques importants que les entreprises ne sont pas prêtes à prendre. Compte tenu de ce que nous savons maintenant des limites de l'anonymisation des données (de Montjoye *et al.*, 2013 ; 2015) et même de la confidentialité différentielle dans la pratique (de Montjoye *et al.*, 2019), ces préoccupations sont particulièrement importantes pour les entreprises soumises au Règlement général sur la protection des données (RGPD) de l'Union européenne. Certaines plateformes

de médias sociaux ont conçu des interfaces API permettant le partage et la standardisation automatisés des données. Cependant, plusieurs permettent uniquement des recherches des messages archivés. Bien que les données satellitaires, comme celles fournies gratuitement par la NASA (la National Aeronautics and Space Administration des États-Unis) et par l'Agence spatiale européenne (ESA), soient généralement moins coûteuses que la cartographie sur le terrain, certains produits de télédétection sont coûteux, ce qui crée une barrière à l'accès.

La stabilité et la prévisibilité de l'accès aux données s'avèrent un autre défi, compte tenu du fait que plusieurs projets et projets pilotes sont, pour l'instant, ponctuels, ce qui limite la faisabilité et la désirabilité de recourir à long terme à des mesures et à des suivis fondés sur l'IA des indicateurs de développement humain. Quelle que soit la taille ou la richesse d'un ensemble de données, et peut-être particulièrement dans le cas des grands ensembles complexes, il est nécessaire de demander quelles informations ils contiennent vraiment et ce qu'elles transmettent. Les miettes d'information numérique d'IA ne sont généralement pas représentatives de l'ensemble de la population d'intérêt et peuvent refléter et exacerber des inégalités structurelles et des préjugés existants (Bradley *et al.*, 2021). Comme discuté dans d'autres chapitres de ce volume, les modèles entraînés en utilisant de telles données sont généralement non pertinents et, dans certains cas, injustes ou dangereux pour les segments de la population n'étant pas représentés dans les ensembles de données d'entraînement. Ces préjugés auront tendance à être plus importants dans le cas de technologies ayant des taux de pénétration plus faibles. Un manque de représentativité nuit à l'interprétation et à l'utilisabilité comme cela est défini par les concepts de validité interne et externe ainsi qu'à la légitimité de ces systèmes (Flashcard Machines, 2011).

Alors que toutes les statistiques circonscrivent l'expérience humaine en négligeant plusieurs de ses aspects, les miettes d'information numérique proviennent de processus de collecte beaucoup moins contrôlés que ceux employés pour recueillir des statistiques officielles. Plusieurs consistent en textes non structurés et générés par les utilisateurs et utilisatrices de sorte que l'information peut avoir été produite dans de faux profils ou par de vraie personne partageant de l'information qui ne reflète pas exactement leurs propres perceptions ou agissements. La nécessité de combiner, dans plusieurs cas, des miettes d'information numérique avec des statistiques officielles à des fins d'entraînement ou de vérification terrain constitue un dernier défi. Pour y arriver, les statistiques doivent être facilement disponibles et accessibles, ce qui entre souvent en conflit avec les niveaux techniques et de confiance (Letouzé et Jütting, 2015).

Capacités : les nantis et les démunis

L'étendue actuelle des capacités en IA, qu'elles soient humaines, technologiques, scientifiques ou financières, représente le deuxième groupe de défis et de limites aux ODD. Sans conteste, les capacités en IA sont *très* inégalement réparties dans le monde et cette situation a des répercussions que nous ne comprenons pas encore pleinement et que nous avons, encore moins, abordées. Plusieurs nations, institutions et communautés ne disposent pas des ressources humaines ni financières nécessaires pour créer et exploiter les types de systèmes d'IA conçus et utilisés dans les universités et les sociétés mondiales de premier plan. Malgré les progrès réalisés au cours de la dernière décennie, les pays du Sud sont encore loin derrière les pays riches dans tout ce qui a trait aux capacités technologiques et nous ne savons pas si le fossé se réduit ou s'élargit en raison de la pandémie de COVID-19 (UNCTAD, 2021).

Les capacités humaines sont un autre facteur limitant évident qui se manifeste, par exemple, dans le manque ou la pénurie de personnel compétent dans les bureaux de statistiques des pays du Sud, où les jeunes diplômés en informatique ont bien plus de chances de travailler dans une entreprise technologique locale ou mondiale que dans une agence gouvernementale sous-financée. Les logiciels d'analyse populaires, tels Python et R peuvent bien être gratuits, le personnel local peut ne pas disposer des outils nécessaires pour les utiliser ou peut tout simplement ne pas être encouragé à le faire. Étant donné

la diversité des sources de données et des techniques utilisées pour concevoir ou utiliser l'IA, les besoins de formation et de recyclage sont généralement grands (Dondi *et al.*, 2021 ; Brown *et al.*, 2019).

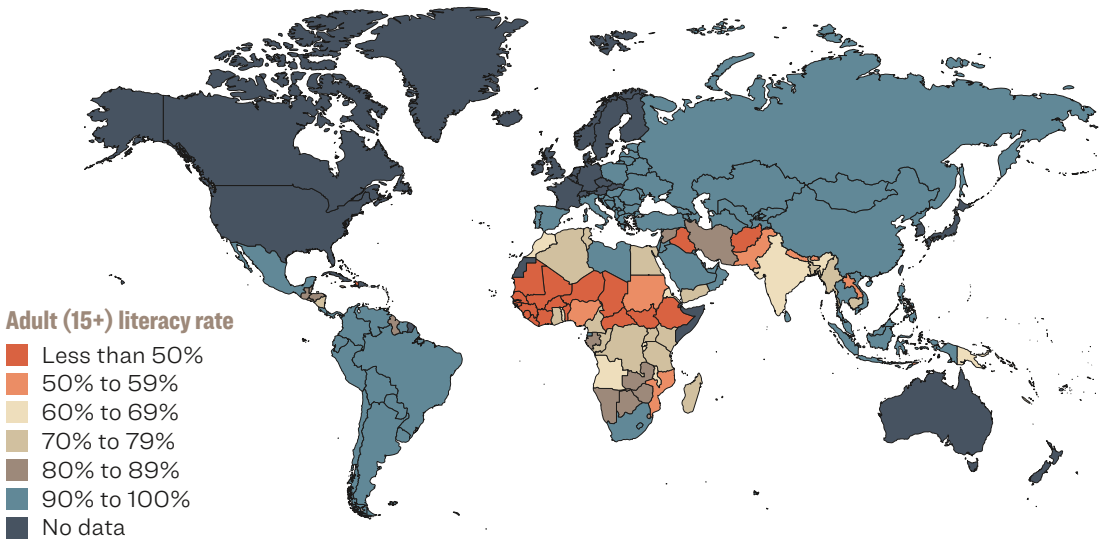
Outre les capacités technoscientifiques avancées, les principales parties prenantes ne disposent généralement pas des compétences pertinentes, particulièrement dans les pays en développement, une situation qui peut être représentée par les niveaux d'alphabétisation des adultes (figure 3). Les appels à la promotion de la littéracie en matière de données sont les bienvenus, mais ces efforts doivent aller au-delà de la simple formation en codage des étudiants, étudiantes, professionnels et professionnelles (Letouzé, Bhargava *et al.*, 2015). Les contraintes en matière de capacité comprennent également la normalisation limitée des méthodologies et des technologies permettant d'accéder aux données dans le respect de la vie privée[56] (Dwork et Roth, 2014) et aux tentatives comme le projet Open Algorithm (OPAL) (Roca et Letouzé, 2016). Des techniques permettant de corriger les préjugés découlant de l'échantillon utilisé en recourant à des techniques et à des sources statistiques standards sont en cours de développement (Zagheni et Weber, 2012 ; Letouzé, Pestre et Zagheni, 2019), mais il faut en faire davantage pour s'assurer que les préjugés sont systématiquement évalués et traités dans les ensembles de données originaux.

Un autre enjeu en matière de capacités concerne les très grandes exigences énergétiques et l'empreinte carbone du traitement et du stockage des données d'IA. Selon une étude, la consommation d'énergie des centres de données en Europe pourrait augmenter de 28 % entre 2018 et 2030 (Montevecchi *et al.*, 2020), alors qu'une autre estime que les émissions de dioxyde de carbone découlant de l'entraînement d'un modèle d'apprentissage profond pour le traitement automatique des langues sont équivalentes à celles produites par une personne moyenne vivant aux États-Unis en *deux ans* (Strubell *et al.,* 2019). Sur une note positive, des infrastructures écoénergétiques sont en développement (Lei et Masanet, 2020), l'IA pourrait aider à optimiser la consommation d'énergie (Gao, 2014) et des recherches sont menées pour mesurer plus efficacement les émissions de carbone dues à l'IA, (Lacoste *et al.*, 2019 ; Henderson *et al.*, 2020 ; Cowls *et al.*, 2021), mais ces tendances pourraient tout de même ne pas être durables.

56. La confidentialité différentielle consiste à effectuer une analyse statistique des ensembles de données qui peuvent contenir des renseignements personnels, de sorte qu'en observant le résultat de l'analyse des données, il est impossible de déterminer si les données d'une personne en particulier ont été incluses ou non dans l'ensemble de données original.

| FIGURE 4 |

Taux de littéracie chez les adultes par pays (UNESCO, 2017).



Communautés : piètres liens et inclusion

Comme dans le cas de l'initiative valencienne, pour que les efforts en IA portent leur fruit, la participation de plusieurs parties prenantes issues du secteur privé, des gouvernements, des universités, des organisations internationales et des organisations de la société civile est nécessaire, même si leurs motivations, contraintes et priorités peuvent souvent différer (Letouzé et Oliver, 2019). Des progrès ont été réalisés au cours des dernières années pour renforcer les liens et la confiance entre les parties prenantes, notamment par les défis « data for good », comme le défi « Data for Refugees », ainsi que d'autres projets pilotes et initiatives, dont la mise en place récente par l'Union européenne d'un groupe d'experts sur la « facilitation de l'utilisation de nouvelles sources de données pour les statistiques officielles » après que d'autres initiatives semblables aient été mises en place (Salah et al., 2018 ; Skibinski, 2020 ; Commission Européenne, 2022). Des modalités de collaboration pour aider à créer des projets au sein de la « communauté de l'IA » ont été proposées, telles que Data Collaboratives, tout comme l'ont été des lignes directrices et des objectifs possibles de collaboration (Tomašev *et al.*, 2020). Mais des obstacles importants à de telles initiatives demeurent, comme l'absence de modèles d'affaires clairs pour le partage de données ainsi que les incertitudes réglementaires, les préoccupations éthiques et le calcul politique (Letouzé *et al.*, 2015 ; Letouzé et Oliver, 2019).

L'inclusion et la participation nettement inadéquates des groupes de personnes marginalisées, vulnérables et issues de minorités, non pas simplement dans les ensembles de données, mais même (ou plus particulièrement) dans les différentes étapes des processus et projets d'IA demeurent une limite majeure à l'utilisation de l'IA au service des ODD. Les données et les systèmes d'IA ne sont ni neutre ni objectifs. Ils reflètent les questions et les préférences des groupes qui disposent du pouvoir de les mettre en œuvre. Assurer la protection des données et la protection des renseignements personnels de chacun et de chacune dans le but d'atténuer les préjudices potentiels est d'une importance capitale,

mais la confidentialité devrait également être conceptualisée pour inclure la confidentialité collective (Kammourieh *et al.*, 2017). La confidentialité devrait aussi comprendre la faculté d'agir, soit la capacité des personnes représentées dans les systèmes d'IA ou touchées par ceux-ci d'avoir leur mot à dire, au-delà du simple fait de donner leur consentement lorsqu'on leur demande (Letouzé *et al.*, 2015). Une tentative visant à offrir un médium pour une représentation et une inclusion locales accrues est le Council for the Orientation of Development and Ethics (CODE) mis en place par la Data-Pop Alliance pour tous ces projets (Letouzé et Yáñez, 2021). Mais il est nécessaire d'en faire bien plus pour promouvoir une participation et une inclusion adéquates des personnes concernées par les systèmes d'IA.

Culture : quand les peurs, la méfiance et la cupidité entrent en jeu

Malgré l'enthousiasme que suscite l'IA dans certains cercles, le sentiment général dans l'espace public et, dans une certaine mesure, dans « la communauté de l'IA au service du bien » est marqué par la méfiance et la peur (Ford, 2015 ; Ikkatai *et al.*, 2022 ; Schmelzer, 2019). La méfiance envers l'IA ou les partenaires d'IA peut limiter l'impact positif qu'elle peut avoir sur les ODD et représente un défi de taille puisque cette méfiance est enracinée dans des préoccupations légitimes alimentées par des échecs répétés, des scandales publics et une concurrence entre les États. Parallèlement, la limitation des pires excès des applications de l'IA peut entraîner des mesures juridiques et réglementaires trop restrictives susceptibles d'entraver l'innovation.

Au-delà des préoccupations et des récriminations, la résistance au changement est alimentée par des habitudes et des intérêts bien perçus. Par exemple, les premières tentatives visant à tirer parti des données « non traditionnelles » ont été accueillies avec scepticisme par la communauté statistique officielle et les milieux gouvernementaux pour des raisons scientifiques, mais aussi par crainte de perdre leur pertinence (Letouzé et Jütting, 2015). En même temps, certains décideurs et certaines décideuses ont peu de raisons de faire pression pour des changements fondamentaux et des investissements en IA. Même en supposant la mise en place d'un système d'IA hautement performant, les décideurs et décideuses peuvent choisir d'ignorer les informations produites. Ce « déficit de décision », bien connu dans le secteur humanitaire, désigne le décalage entre l'information et l'action et découle en partie d'un manque d'habitude à utiliser les données pour prendre rapidement des décisions ou d'une méfiance à l'égard de ces données, ainsi que d'autres facteurs politiques, comme nous le verrons plus en détail dans la section suivante.

Cette insignifiance apparente des faits pourrait être en partie attribuée à une surcharge de données qui ont « tué les faits et la vérité » (Lepore, 2020). Aussi, comme la psychologie l'a démontré, il est très difficile pour les humains de changer leurs idées et leurs actions, lorsque le changement va à l'encontre de déterminants religieux, politiques, économiques, culturels ou autres profondément ancrés dans nos identités ou lorsque le comportement découle d'une dépendance (Kolbert, 2017). Par exemple, la science a démontré, il y a des décennies, les effets nuisibles de nos modes de vie sur les émissions de carbone et la biodiversité et ceux de la consommation d'alcool sur notre santé, mais changer des croyances et des comportements profondément ancrés est très difficile.

La confiance est une exigence clé pour que les projets d'IA fonctionnent et pour que les gens acceptent lentement les faits soutenus par la science et elle est généralement mieux servie par des discussions rationnelles et respectueuses. Or, la confiance n'est souvent pas assez forte entre les parties prenantes clés. L'expérience et plusieurs études ont montré que des facteurs « intangibles » sans lien avec les données, la technologie, les compétences ou les réglementations ont un impact important sur les possibilités et la manière dont l'IA est utilisée à des fins d'intérêt public (West, 2021).

Vers une culture humaine de l'IA pour le développement, l'apprentissage et la démocratie au 21e siècle

Dans cette dernière section, nous souhaitons proposer une vision possiblement légèrement différente et à plus long terme de la manière dont l'IA pourrait contribuer aux objectifs de développement humain, y compris tous les ODD et d'autres, ainsi qu'aux principes et processus démocratiques. Nous nous penchons de manière plus approfondie sur certains des principes fondamentaux du programme des ODD standard et sur le discours qui l'accompagne à une époque de méfiance grandissante et d'inégalités en partie alimentées par l'omniprésence de l'IA dans nos vies. De ce fait, nous établissons les contours et les exigences d'une vision d'une culture humaine de l'IA proposée dans d'autres contributions et explorons ce qui peut être fait à ce sujet.

Reformuler nos problèmes en utilisant le narratif standard relativement à l'IA au service des ODD

Comme susmentionné, l'argument voulant que l'IA puisse contribuer à la promotion du développement humain par l'entremise des ODD faiblit, et ce, en raison de plusieurs dures réalités mondiales, dont deux sont abordées ci-après.

La nature et le fonctionnement des régimes politiques partout dans le monde est la première. En effet, l'argument et le discours courant de la communauté de « l'IA au service du bien » tient fortement à l'hypothèse que ceux et celles qui prennent des décisions lourdes de conséquences se soucient du bien-être des citoyens et citoyennes et qu'ils et qu'elles n'ont pas en temps opportun les données pertinentes et de grande qualité pour prendre de meilleures décisions. Dans ce contexte, il s'ensuit que des mesures sont importantes de la même façon que nous croyons que le sont les institutions, c.-à-d. qu'elles ont un effet causal sur les résultats (Przeworski, 2004 ; Acemoglu et Robinson, 2012 ; Letouzé, 2018). Or, concrètement, nous soutenons que certains dirigeants et dirigeantes sont peu enclins à mettre en œuvre des politiques fondées sur des données probantes, surtout lorsque ces dernières suggèrent qu'ils et qu'elles devraient mettre en œuvre des politiques qui vont à l'encontre de leurs intérêts politiques ou qu'ils et qu'elles devraient quitter leur poste. Parallèlement, ils et elles ont tout intérêt à tirer parti de ces nouvelles technologies, dont l'IA, pour la surveillance et le contrôle des populations (Lillis, 2021).

Le fait que les 193 chefs et cheffes de gouvernements des États membres de l'ONU aient signé les ODD à l'époque où ils ont été créés s'avère à la fois leur plus grande force et leur plus grande faiblesse. Leur force, parce que, bien qu'ils ne soient pas juridiquement contraignants, les ODD aident les sociétés à tenir ces signataires responsables des objectifs de développement fixés collectivement et clairement énoncés. Leur faiblesse, car le caractère des régimes de nombreux signataires sont tel que si l'un des ODD ou l'ensemble de l'entreprise avait représenté une menace pour le statut quo, ils ne les auraient très probablement pas signés. On a même soutenu que les ODD « nuisaient à la démocratie » en « faisant pression en faveur d'un programme judicieusement calibré pour éviter de contrarier les dictateurs, les kleptocrates et les auteurs et autrices de violations des droits fondamentaux dans le monde » (Smith et Gladstein, 2019). Bien que cette déclaration puisse sembler choquante, elle n'est pas entièrement sans fondement. Lorsque la démocratie semble reculer, les autocrates ont été enhardis par la pandémie de COVID-19. Selon le Economist Intelligence Unit (2021), « en 2020, partout dans le monde, les citoyens et citoyennes ont vécu la plus importante réduction des libertés individuelles jamais entreprises par des gouvernements en temps de paix (et peut-être même en temps de guerre) » et « la démocratie mondiale à continuer son déclin brutal en 2021 ». Les iniquités, notamment en ce qui a trait au revenu, continuent à s'accroître (Ferreira, 2021 ; Oxfam, 2022) et, au moment de la rédaction de ce chapitre, le scandale des Pandora Papers venait d'éclater (ICIJ, 2021). En présence de tous ces événements combinés, il semble naïf de soutenir que l'obstacle principal à l'éradication de la pauvreté, à l'élimination des inégalités des genres et à la préservation de l'environnement, entre autres, est un manque de disponibilité de données ou d'algorithmes d'IA opportuns et pertinents aux dirigeants et dirigeantes politiques et économiques.

Les intérêts politiques et économiques l'emportent généralement sur les preuves scientifiques et les statistiques officielles lorsqu'il s'agit de déterminer les priorités et les politiques qui influencent des résultats concrets (figure 5). Dans ce contexte, le narratif courant de « l'IA ou les données au service du bien » et de la « Révolution des données » peut non seulement être inopérant, mais aussi contre-productif, en fournissant des arguments aux praticiens et praticiennes du développement et aux politiciens et politiciennes pour échapper à leur responsabilité. En mettant l'accent sur le manque de données et les merveilles que de meilleures informations obtenues grâce à l'IA permettraient de réaliser, il est facile pour eux et elles, particulièrement ceux et celles qui sont corrompus, incompétents ou les deux, de dire haut et fort qu'ils et qu'elles n'ont pas été en mesure d'améliorer X parce qu'ils et qu'elles n'avaient pas les bonnes données concernant X. Pour être clair, à notre avis, les pays et les communautés pauvres ne sont pas pauvres parce que leurs dirigeants et dirigeantes ne disposent pas de bonnes données sur la pauvreté à leur sujet. Ils sont pauvres et leur pauvreté n'est pas correctement prise en compte parce qu'ils ne comptent pas. Lorsqu'un moteur est brisé, améliorer son carburant ne fera pas l'affaire. La question est de savoir comment le réparer.

| FIGURE 5 |

La Révolution des données est arrivée (Améliorera-t-elle toutes les vies ?), illustration tirée de Emmanuel Letouzé, illustration à l'événement Eurostat NTTS, 13 mars 2019.

Dans cette entreprise, l'IA peut certainement aider, bien qu'elle soit accompagnée de certains défis. En effet, le deuxième enjeu majeur que nous souhaitons aborder est de réitérer le rôle de l'IA dans le bris de confiance envers les experts et expertes, les institutions, les voisins et voisines et, ultimement, les faits. De plus en plus d'études suggèrent que les plateformes de médias sociaux et les géants technologiques, qui sont en fait des entreprises de données dominant presque totalement le marché, contribuent à la polarisation politique et certains craignent qu'ils ne menacent la survie même des pratiques et des systèmes démocratiques. (Helbing *et al.*, 2017, Bergstrom et West, 2020, Risse, 2021). Ceci signifie également que les bienfaits objectifs de l'IA, comme la capacité à détecter le cancer ou la fraude, peuvent être considérés comme suspicieux. Dans un contexte de crises écologiques et sociopolitiques multiples aggravé par les conditions actuelles, il en résulte que l'on peut difficilement s'attendre à ce que l'IA contribue de manière transparente à « reconstruire en mieux » après la pandémie de COVID-19, sans un changement fondamental dans la manière dont les systèmes d'IA sont conçus, utilisés et réglementés et surtout pour qui et dans quels objectifs.

De nouveaux cadres juridiques et réglementaires apparaissent dans le monde pour guider l'utilisation des données et de l'IA. Toutefois, ces développements se limitent largement à une certaine région ou à un certain pays en particulier et n'arrivent pas à créer efficacement de nouveaux droits mondiaux. Citons par exemple le « droit à l'oubli » et le Règlement général sur la protection des données (RGPD), qui ne sont pas des normes mondiales et entraînent en fait un traitement numérique inégal des personnes. À mesure que nos vies physiques et numériques s'entremêlent, il est peut-être plus fondamental de repenser nos droits humains et tout aussi fondamental de formaliser les droits et responsabilités des systèmes d'IA. Les principes d'Asilomar pour l'IA[57] sont une première étape importante dans cette direction, mais ils se limitent à la recherche et au développement en IA et ne sont pas des règles et des normes globales convenues à l'échelle internationale, soumises à l'application et à la responsabilité, qui sont nécessaires de toute urgence pour réduire le risque d'un avenir dystopique de l'IA, y compris le potentiel de guerre alimentée par l'IA.

Une question qui retient de plus en plus l'attention est de savoir si les réglementations en matière d'IA doivent se concentrer sur les exigences ex ante ou sur la responsabilité ex post. Bien que l'accent soit actuellement mis sur les premières, les secondes pourraient être plus réalistes étant donné la nature distribuée des systèmes d'IA.

Caractéristiques, exigences et retombées attendues d'une vision et d'une culture humaines de l'IA pour promouvoir le développement humain et réchapper les idéaux et les processus démocratiques

Malgré ces tendances croissantes préoccupantes, nous sommes d'avis que l'IA peut contribuer à la promotion des objectifs démocratiques et de développement. Fondamentalement, les systèmes d'IA ne sont pas que de puissants outils pouvant réaliser des tâches précises. Ils démontrent également comment des nodes travaillant ensemble sur la base de données et de rétroactions peuvent apprendre collectivement à atteindre plus efficacement un ensemble d'objectifs communs. Paradoxalement, bien que le cerveau humain ait servi d'inspiration à l'IA, nous soutenons que l'IA pourrait et devrait dorénavant agir comme une analogie source d'inspiration pour de meilleurs systèmes et sociétés fondés sur l'apprentissage, dans la mesure où les bons ingrédients sont disponibles, cultivés et utilisés.

Comme suite à des contributions antérieures, cette idée de considérer l'IA et de l'utiliser comme un instrument (systèmes d'IA étroits qui excellent à des tâches précises) et une source d'inspiration pour les sociétés humaines en s'appuyant sur une volonté et une capacité renouvelées pour un apprentissage collectif se nomme « culture humaine de l'IA » (Pentland, 2017, Letouzé et Pentland, 2018). Il s'agit d'une vision dans laquelle les diverses parties (nodes) composant les sociétés humaines collaborent

57. Consulter https://futureoflife.org/ai-principles/.

pour apprendre et renforcer ce qui contribue à la progression vers des objectifs communs, y compris en utilisant des outils d'IA. Une culture humaine de l'IA, par exemple, se demanderait si l'objectif de construire une société plus sûre et plus pacifique est mieux servi par la « guerre contre la drogue » et les politiques d'incarcération de masse qui y sont associées et qui ont été utilisées dans certaines parties de l'Amérique du Nord au cours des dernières décennies que par d'autres moyens (Pearl, 2018). Ce faisant, cette culture pourrait tirer parti de l'IA pour aider à proposer d'autres approches et à les tester, mais elle pourrait aussi opter pour des solutions simples.

Une culture humaine de l'IA s'appuie également sur une vision en vertu de laquelle la désirabilité et la légitimité de certains objectifs, comme la croissance du PIB ou l'optimisation des profits, seraient réévaluées d'une manière systématique et continuelle en fonction de leurs effets, comme dans un système d'apprentissage. Les principaux ingrédients et exigences d'une telle culture sont assez bien connus. Par exemple, il est nécessaire d'encourager une culture de discussions raisonnées et rationnelles, de coopération et, par conséquent, de confiance entre les nodes bien au-delà de ce que l'on observe aujourd'hui entre les groupes afin que la mesure ait la possibilité de compter comme elle compte en IA. De plus, une telle culture requiert de disposer de données d'entrée et d'informations de retour précises et opportunes pour que le système puisse constamment apprendre. En outre, elle nécessite de vastes sociétés possédant de bonnes connaissances des données, un plus grand contrôle de la part des sujets concernés sur les données à leur sujet, notamment grâce à la création de « coopératives de données » ou d'autres mécanismes d'accès et de partage de données (Pentland et Hardjono, 2020) ainsi qu'une presse libre (UNESCO, 2022).

Le voie vers une culture humaine de l'IA impliquerait de faire revivre ou de réinventer les principes démocratiques de participation, d'autogestion et de gouvernement par le biais de discussions fondées sur la compassion rationnelle (Bloom, 2016), y compris, et de plus en plus, à l'échelle locale. Il nécessiterait aussi de créer des incitatifs, des moyens et des habitudes de la part de toutes les parties prenantes pour demander une évaluation systématique de toutes les décisions collectives. Cette évaluation devrait être menée en utilisant les meilleures données et méthodologies disponibles afin d'ajuster de futures itérations et de contribuer à un ensemble de preuves sur les actions qui entraînent des résultats donnés. En ce sens, afin d'éviter d'accentuer les inégalités que semble engendrer l'économie numérique, de tels incitatifs, moyens et habitudes devraient impliquer une reconsidération de la manière dont les différentes formes de capital, y compris le capital numérique, sont partagées (Gardels, 2022).

Il ne sera pas facile d'instaurer une culture humaine de l'IA qui place les discussions rationnelles et respectueuses fondées sur la confiance et les faits au cœur d'un nouveau contrat social parmi les humains et entre les humains et les machines dans les sociétés du 21e siècle. Ceci est le cas surtout parce qu'il est nécessaire d'aborder les excès et les abus perpétrés par de puissants acteurs à l'origine de la plupart des maux de l'humanité et de considérer les voix dissidentes et les complexités des réalités humaines. Comme cela a été suggéré précédemment, il ne s'agit pas simplement d'utiliser l'IA pour fournir un meilleur carburant à de vieilles machines. Il s'agit, et requiert, plutôt de mettre à niveau ces systèmes en utilisant l'IA comme un instrument quand et au besoin, et comme source d'inspiration.

Nouveaux indicateurs et le prochain programme des ODD ?

Une façon concrète de commencer à établir de nouveaux indicateurs et le prochain programme des ODD est d'encourager des efforts en IA qui visent à suivre toutes les cibles des ODD, notamment les indicateurs de niveau 3 sensibles sur le plan politique de l'ODD 16 qui vise à « Promouvoir des sociétés justes, pacifiques et inclusives ». Ces derniers comprennent l'indicateur 16.6.2. « Proportion de la population dont la dernière expérience avec les services publics a été satisfaisante par l'analyse de données des médias sociaux » (Data-Pop Alliance, 2018) et l'indicateur 16.10.1 « Nombre de cas avérés de meurtres, d'enlèvements, de disparitions forcées, de détentions arbitraires et d'actes de torture dont ont été victimes des journalistes, des personnes travaillant dans les médias, des syndicalistes et des défenseurs

des droits de l'homme au cours des 12 mois précédents » (Muñoz *et al.*, 2021). Ces efforts pourraient obtenir le soutien d'organisations de recherche et de défense des intérêts internationales ainsi que d'entreprises partageant une vision semblable et prêtes à faire pression auprès des gouvernements qui sont réticents à discuter et à aborder ces phénomènes.

De nouveaux objectifs qui reflètent de nouvelles réalisations et priorités sociétales devraient aussi être considérés. Certains groupes suggèrent déjà de nouvelles priorités, notamment la santé, le bien-être et les droits des animaux (Visseren-Hamakers, 2020), l'espace durable (ITU News, 2021), l'espace pour tous et toutes (National Space Society, 2020), une vue numérique signifiante et sûre (Jespersen, n.d.), l'assurance que l'âge numérique appuie les gens, la planète, la prospérité et la paix (Luers, 2020), le développement et les handicaps (Le Marrec, 2016).

L'IA pourrait également aider à déterminer quels ODD devraient être priorisés en fonction de l'intérêt public exprimé et d'études de faisabilité. De tels efforts devraient être déployés sous une supervision humaine par le biais d'une conception de participation rigoureuse afin de veiller à ce qu'ils ne reflètent pas les préjugés structurels présents dans les bases de données. La façon d'atténuer les préjugés structurels pourrait être semblable à celle utilisée pour établir les priorités de recherche en AI (Vinuesa *et al.*, 2020) ou pour intégrer des valeurs éthiques dans les systèmes d'IA (Rahwan, 2017).

CONCLUSIONS : LA RÉVOLUTION DES DONNÉES ET DE L'IA DOIT ÊTRE POLITISÉE

La pandémie de COVID-19 a exposé et exacerbé des lignes de faille structurelles préexistantes dans notre société. Notre monde est de plus en plus numérique et inéquitable. Alors que la numérisation ne cesse de progresser, la démocratie et l'égalité semblent reculer. Dans ce contexte, l'essor de l'IA apparaît comme le cas parfait du feu de Prométhée. Elle peut certainement aider à mesurer et à promouvoir plus efficacement les ODD et d'autres objectifs de développement humain, malgré les défis et les obstacles qui se dressent sur la route et qui peuvent être résolus grâce à des investissements appropriés dans les données, les capacités, les collaborations et les initiatives. Mais l'IA peut aussi alimenter davantage les inégalités, la polarisation et l'effondrement de la confiance.

Nous soutenons fondamentalement que les problèmes à régler ne sont pas intrinsèquement technologiques. Ils sont principalement politiques, économiques et culturels et enracinés dans l'avidité personnelle, l'accaparement par les élites, la soif de pouvoir et la méfiance sociétale. Il s'ensuit que les solutions doivent être essentiellement politiques et culturelles. Par conséquent, à moins qu'il n'y ait une reconnaissance que le discours actuel sur l'IA au service des ODD, selon lequel la contrainte principale est le manque d'indicateurs sur les tableaux de bord des dirigeantes et dirigeants mondiaux, penche du côté de la complaisance ou de la naïveté, l'IA ne tiendra jamais sa promesse. Dans le scénario « comme d'habitude », dans lequel l'IA demeure contrôlée par des personnes et des groupes motivés par le pouvoir et les profits, l'IA est plus susceptible de mener à un contrôle technologique des citoyens et citoyennes et de l'alimenter et, ce faisant, de réduire les choix et les libertés et d'abaisser le niveau de vie des perdants et perdantes de l'accroissement des inégalités économiques, sociales, politiques et environnementales.

Mais nous ne perdons pas espoir en l'IA. Bien que l'IA imite le cerveau humain, nous soutenons paradoxalement que les sociétés humaines pourraient maintenant tenter d'imiter les systèmes d'IA en valorisant et en encourageant les capacités d'apprentissage et la coopération. Nous appelons ceci une culture humaine de l'IA et nous décrivons cette culture comme utilisant l'IA en tant qu'analogie inspirante et ensemble d'instruments pour mesurer, surveiller et atteindre des objectifs établis collectivement. L'objectif le plus important est de faire respecter et de protéger les principes et les processus démocratiques, plus particulièrement en assurant à tous et à toutes un contrôle et une

transparence accrus sur la conception et l'utilisation des systèmes d'IA qui ont un impact sur leur vie. Cela doit être associé à des mécanismes clairs et fermes de responsabilité et de conformité en ce qui a trait à la conception et à l'utilisation de ces systèmes. Peut-être le cas de Valence, en Espagne, mentionné au début de ce chapitre, montre qu'une culture humaine de l'IA est possible.

RÉFÉRENCES

Acemoglu, D. et Robinson, J. 2012. *Why Nations Fail : The Origins of Power, Prosperity and Poverty*. New York : Crown Business.

Avendano, R., Jütting, J. et Kuhm, M. 2020. *The Palgrave Handbook of Development Cooperation for Achieving the 2030 Agenda*. London : Palgrave Macmillan.

Barret, P., Hendrix, J. et Sims, G. 2021. Fueling the fire : How social media intensifies U.S. political polarization – and what can be done about it. *NYU Stern Center for Business and Human Rights*, 21 septembre.

Big Data UN Global Working Group. 2019. Training, Skills and Capacity-Building — UN GWG for Big Data.

Bloom, P. 2016. *Against Empathy : The Case for Rational Compassion*. London : The Bodley Head.

Bradley, V. C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X. et Flaxman, S. 2021. UnrePrésentative big surveys significantly overestimated US vaccine uptake. *Nature*, nº 600, pp. 695–700.

Commission Européenne. 2022. High-Level Expert Group on Artificial Intelligence : shaping Europe's digital future. 13 juin. https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai

Cowls, J., Tsamados, A., Taddeo, M. et Floridi, L. 2021. The AI Gambit — Leveraging Artificial Intelligence to Combat Climate Change : Opportunities, Challenges, and Recommendations. https://ssrn.com/abstract=3804983

Devarajan, S. 2013. Africa's statistical tragedy. *Review of Income and Wealth*, nº 59, pp. 9-15. https://doi.org/10.1111/roiw.12013

Dickinson, E. 2011. GDP : A Brief History. *Foreign Policy*. 3 janvier. https://foreignpolicy.com/2011/01/03/gdp-a-brief-history/

Economist. 2015. How to catch the overfishermen. 22 janvier. https://www.economist.com/leaders/2015/01/22/how-to-catch-the-overfishermen

Economist. 2022. Daily chart : A new low for global democracy. https://www.economist.com/graphic-detail/2022/02/09/a-new-low-for-global-democracy

Ferreira, F. 2021. Inequality in the times of COVID-19. IMF Finance and Development. https://www.imf.org/external/pubs/ft/fandd/2021/06/inequality-and-covid-19-ferreira.htm

Flashcard Machine. 2011. Internal v. External Validity.13 juin. https://www.flashcardmachine.com/internal-vsexternal-validity.html

Ford, P. 2015. Our fear of artificial intelligence. MIT Technology Review. 11 février. https://www.technologyreview.com/2015/02/11/169210/our-fear-of-artificial-intelligence/

Grameen Foundation. 2011. Lessons learned from AppLab's first three years in Uganda. 21 janvier. https://grameenfoundation.org/blog/lessons-learned-applab%E2%80%99s-first-three-years-uganda#.XehL5OhKhPZ

Griswold, W. 2008. *Cultures and societies in a changing world*. 3rd edition. Thousand Oaks, CA : Pine Forge.

Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D. et Pineau, J. 2020. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, nº 21, pp. 1-43.

United Nations. 2013. *A New Global Partnership : Eradicate Poverty and Transform Economies through Sustainable Development. : The Report of the High-Level Panel of Eminent Persons on the Post-2015 Development Agenda.* https://sustainabledevelopment.un.org/content/documents/8932013-05%20–%20HLP%20Report%20–%20A%20New%20Global%20Partnership.pdf

ICIJ. 2021. Pandora Papers. International Consortium of Investigative Journalists. https://www.icij.org/investigations/pandora-papers/

Ikkatai, Y., Hartwig, T., Takanashi, N. et Yokoyama, H. M. 2022. Octagon measurement : public attitudes toward AI ethics. *International Journal of Human–Computer Interaction,* pp. 1–18. 10 janvier. https://doi.org/10.1080/10447318.2021.2009669

Independent Expert Advisory Group (IEAG). 2014. A World that Counts : Mobilizing the Data Revolution for Sustainable Development. https://www.undatarevolution.org/wp-content/uploads/2014/12/A-World-That-Counts2.pdf

Jespersen, S. n.d. Advocating for an 18th Sustainable Development Goal : A meaningful and safe digital life. VERTIC. https://www.vertic.com/our-thinking/advocating-for-an-18th-sustainable-development-Objectif-a-meaningful-and-safe-digital-life

King, G. 2013. *Big Data is not about the data !* Allocution présentée au Golden Seeds Innovation Summit, New York City. Institute for Quantitative Social Science, 30 janvier. http://gking.harvard.edu/files/gking/files/evbase-gs.pdf

Kolbert, E. 2017. Why facts don't change Our minds, *New Yorker Magazine*. https://www.newyorker.com/magazine/2017/02/27/why-facts-dont-change-our-minds

Lacoste, A., Luccioni, A., Schmidt, V. and Dandres, T. 2019.Quantifying the carbon emissions of machine learning. arXiv :1910.09700 [cs.CY]

Le Marrec, J. 2016. Where is Sustainable Development Objectif Goal 18 ? Blogue sur le développement et les handicaps http://developmentanddisability.blogspot.com/2016/04/where-is-sustainable-development-Objectif-18.html

Lepore, J. 2020. The End of Knowledge : How Data Killed Facts. Lecture given at the Fox Center for Humanistic Inquiry, Emory University. 8 avril.

Letouzé, E. 2013. Could Big Data provide alternative measures of poverty and welfare ? *Development Progress* blogue, 11 juin. http://www.developmentprogress.org/blog/2013/06/11/could-big-data-provide-alternative-measures-poverty-and-welfare

Letouzé, E. 2014. Big Data for development : Facts and figures. SciDev.net, 15 avril. http://www.scidev.net/global/data/feature/big-data-for-development-facts-and-figures.html.

Letouzé, E. 2015a. *Big Data and Development : An Overview*. Data-Pop Alliance. http://datapopalliance.org/item/white-paper-series-official-statistics-big-data-and-human-development/

Letouzé, E. 2015b. Thoughts on Big Data and the SDGs. Data-Pop Alliance. 18 février. https://datapopalliance.org/wp-content/uploads/2020/09/7798BigData-Data-Pop-Alliance-Emmanuel-Letouze.pdf

Letouzé, E., del Villar Z., Molina, R. L., Nieto B. F., Romero, G., Ricard, J., Vazquez, D. et Maya, L. A. C. 2022. Parallel Worlds : Revealing the Inequity of Access to Urban Spaces in Mexico City Through Mobility Data. In : *Measuring the City : The Power of Urban Metrics*. Ahn, C., Ignaccolo, C. and Salazar-Miranda, A (éditeurs). https://projections.pubpub.org/pub/01kebgos/release/1

Letouzé, E., Meier, P. et Vinck, P. 2013. Big Data for conflict prevention : New oil and old fires. Dans *New Technology and the Prevention of Violence and Conflict*, pp. 4-27. International Peace Institute. https://www.ipinst.org/images/pdfs/IPI_Epub-New_Technology-final.pdf

Letouzé, E. Noonan, A., Bhargava, R., Deahl, E., Sangokoya, D. and Shoup, N. 2015. Beyond data literacy : reinventing community engagement and empowerment in the age of data. MIT Media Lab, September. https://www.media.mit.edu/publications/beyond-data-literacy-reinventing-community-engagement-and-empowerment-in-the-age-of-data/

Letouzé, E. et Pentland, A. 2018. Human AI for human development. *ITU Journal : ICT Discoveries* Special Issue, nº 2 (Décembre). https://www.itu.int/en/journal/002/Pages/15.aspx.

Letouzé, E., Pestre, G. et Zagheni, E. 2019. The ABCDE of Big Data : Assessing biases in call-detail records for development estimates. *The World Bank Economic Review*, 3 décembre. https://doi.org/10.1093/wber/lhz039.

Letouzé, E. et Oliver, N. 2019. Paper sharing is caring : Four key requirements for sustainable private data sharing and use for public good. Data-Pop Alliance*;* Vodafone Institute for Society and Communications, novembre. https://datapopalliance.org/paper-sharing-is-caring-four-key-requirements-for-sustainable-private-data-sharing-and-use-for-public-good/.

Letouzé, E., Vinck, P. et Kammourieh, L. 2015. The law, politics and ethics of cell phone data analytics. Data-Pop Alliance Big Data and Development Primer Series, avril. http://datapopalliance.org/item/white-paper-the-law-politics-and-ethics-of-cell-phone-data-analytics/

Letouzé, E. et Yáñez Soria, I. 2021. The CODE for building participatory and ethical data projects. Data-Pop Alliance. https://datapopalliance.org/the-code-for-building-participatory-and-ethical-data-projects/

Lillis, K. B. 2022. NSA watchdog finds "concerns" with searches of Americans' communications. *CNN*. 1er février. https://edition.cnn.com/2022/01/31/politics/nsa-watchdog-concerns-searches-american-communications/index.html

Luers, A. 2020. The missing SDG : Ensure the digital age supports people, planet, prosperity & peace. Inter Press Service News Agency, 6 juillet. http://www.ipsnews.net/2020/07/missing-sdg-ensure-digital-age-supports-people-planet-prosperity-peace/

Marx, W. 2021. How Valencia crushed COVID with AI. *Wired*. 9 septembre. https://www.wired.co.uk/article/valencia-ai-covid-data

Montjoye, Y.-A., L. Radaelli, V. K. Singh, et A. S. Pentland. "Unique in the shopping mall : On the reidentifiability of credit card metadata." *Science* 347, nº 6221, pp. 536-39. https://doi.org/10.1126/science.1256297

Montjoye, Y-A, Hidalgo, C. A., Verleysen, M. et Blondel, V. D. 2013. Unique in the crowd : The privacy bounds of human mobility. *Scientific Reports*, vol. 3, nº 1, pp. 1-5. https://doi.org/10.1038/srep01376

Nations Unies. 2013. *A New Global Partnership : Eradicate Poverty and Transform Economies Through Sustainable Development : The Report of the High-Level Panel of Eminent Persons on the Post-2015 Development Agenda.* https://sustainabledevelopment.un.org/content/documents/8932013-05%20–%20HLP%20Report%20–%20A%20New%20Global%20Partnership.pdf

Oliver, N. 2021. *Artificial Intelligence for Social Good : The Way Forward. European Commission* (à paraître).

Pearl, B. 2018. Ending the War on Drugs : By the numbers. Fiche de données. Center for American Progress. https://www.americanprogress.org/issues/criminal-justice/reports/2018/06/27/452819/ending-war-drugs-numbers/

Pentland, A. 2012. Reinventing society in the wake of Big Data : A conversation with Alex (Sandy) Pentland. *Edge*. 30 août. https://www.edge.org/conversation/alex_sandy_pentland-reinventing-society-in-the-wake-of-big-data

Pentland, A. 2014. *Social Physics : How Good Ideas Spread-The Lessons from a New Science*. New York : Penguin Press.

Przeworski, A. 2004. Institutions matter ? *Gouvernement and Opposition*, vol. 39, nº 4,pp. 527-40. https://doi.org/DOI:10.1111/j.1477-7053.2004.00134.x

Rasmussen, K. et McArthur, J. 2017. How successful were the millennium development Goals ? Blogue *Brookings*, 11 janvier. https://www.brookings.edu/blog/future-development/2017/01/11/how-successful-were-the-millennium-development-Objectifs/

Roca, T. et Letouzé, E. 2016. Open algorithms : A new paradigm for using private data for social good. Data-Pop Alliance. https://datapopalliance.org/open-algorithms-a-new-paradigm-for-using-private-data-for-social-good/

Salah, A., Pentland, A., Lepri, B., Letouzé, E., Vinck, P., de Montjoye, Y-A., Dong, X. et Dagdelen, O. 2018. Data for refugees : the D4R challenge on mobility of Syrian refugees in Turkey." *arXiv*. 14 octobre. http://arxiv.org/abs/1807.00523.

Schmelzer, R. 2019. Should we be afraid of AI ? Forbes. 31 octobre. https://www.forbes.com/sites/cognitiveworld/2019/10/31/should-we-be-afraid-of-ai/.

Skibinski, A. 2020. Expert Group on Facilitating the Use of New Data Sources for Official Statistics. *CROS – European Commission*. 4 décembre. https://ec.europa.eu/eurostat/cros/content/expert-group-facilitating-use-new-data-sources-official-statistics_en.

Smith, J. et Gladstein, A. 2018. How the UN's Sustainable Development Goals undermine democracy. *Quartz Africa*, 7 juin. https://qz.com/africa/1299149/how-the-uns-sustainable-development-Objectifs-undermine-democracy/

Tomašev, N., Cornebise, J., Hutter, F. et al. 2020. AI for social good : Unlocking the opportunity for positive impact. *Nature Communications*, vol. 11, nº 2468. https://doi.org/10.1038/s41467-020-15871-z

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M. et Nerini, F. F. 2020. The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications*, vol. 11, nº 233. https://doi.org/10.1038/s41467-019-14108-y

UNCTAD. 2021. *Technology and Innovation Report 2021. Catching technological waves Innovation with Equity*. https://unctad.org/system/files/official-document/tir2020_en.pdf

UN Global Pulse. 2012. *Big Data for Development : Challenges and Opportunities.* https://www.unglobalpulse.org/wp-content/uploads/2012/05/BigDataforDevelopment-UNGlobalPulseMay2012.pdf

UN Global Pulse. 2015. Big Data and Development : An Overview. https://www.unglobalpulse.org/document/big-data-for-development-in-action-un-global-pulse-project-series/

UNESCO. 2017. *Literacy Continues to Rise from One Generation to the Next*. UNESCO Fact sheet No. 45. Septembre. https://uis.unesco.org/sites/default/files/documents/fs45-literacy-rates-continue-rise-generation-to-next-en-2017_0.pdf

United Nations Development Programme. 2019. Human Development Index. http://hdr.undp.org/en/content/human-development-index-hdi.

Visseren-Hamakers, I. J. 2020. The 18th Sustainable Development Objectif. Earth System Governance, vol. 3. https://doi.org/10.1016/j.esg.2020.100047. https://www.sciencedirect.com/science/article/pii/S2589811620300069

West, D.M. 2021. Using AI and machine learning to reduce Gouvernement fraud. Brookings. 10 septembre.https://www.brookings.edu/research/using-ai-and-machine-learning-to-reduce-Gouvernement-fraud/

Zagheni, E. et Weber I. 2012. You are where you e-mail : Using e-mail data to estimate international migration rates. Dans *WebSci '12 : Proceedings of the 4th Annual ACM Web Science Conference*, pp. 348-351. ACM Digital Library. https://doi.org/10.1145/2380718.2380764

Zheng, X. et Lee, M.K.O. 2016. Excessive use of mobile social networking sites : Negative consequences on individuals, *Computers in Human Behavior*, vol. 65, pp. 65-76, https://doi.org/10.1016/j.chb.2016.08.011

| ANNEXE | Taxonomie et exemples de sources de mégadonnées

Types	Exemples	Possibilités
CATÉGORIE 1 : DONNÉES SUR LES GAZ D'ÉCHAPPEMENT		
Appareils mobiles	Registres détaillés des appels GPS (surveillance de parc de véhicules, LAV autobus)	Estimer la répartition de la population et le statut socio-économique dans des endroits aussi divers que le Royaume-Uni et le Rwanda.
Transactions financières	Identification électronique Permis électroniques (p. ex., assurances) Cartes de transport (dont les cartes de fidélité des compagnies aériennes) Cartes de crédit/débit	Fournir des informations essentielles sur les mouvements de population et les réactions comportementales après une catastrophe.
Transport	GPS (surveillance de parc de véhicules, LAV autobus) ; EZ passes	Fournir une évaluation précoce des dommages causés par les ouragans et les tremblements de terre.
Traces en ligne	Témoins Adresses IP	Atténuer l'impact des maladies infectieuses par une surveillance plus opportune grâce aux journaux d'accès de l'encyclopédie en ligne Wikipédia.
CATÉGORIE 2 : CONTENU NUMÉRIQUE		
Médias sociaux	Tweets (Twitter API) « *Check-ins* » (Foursquare) Contenu Facebook Vidéos YouTube	Fournir une alerte précoce en cas de menaces allant des épidémies à l'insécurité alimentaire.
Contenu en ligne/enrichi par les utilisateurs et utilisatrices	Cartographie (OpenStreetMap, Google Maps, Yelp) Surveillance/ production de rapports (uReport)	Permettre aux utilisateurs et utilisatrices d'ajouter des données terrain qui sont utiles notamment à des fins de vérification.
CATÉGORIE 3 : DONNÉES DE TÉLÉDÉTECTION		
Physique	Compteurs intelligents Capteurs de vitesse/poids Sismographes de USGS	Des capteurs ont été utilisés pour évaluer la demande d'utiliser des capteurs pour estimer la demande de cuisinières à haut rendement à différents prix en Ouganda ou la volonté de payer pour des distributeurs de chlore au Kenya.
À distance	Imagerie satellitaire (NASA TRMM, LandSat) Véhicules aériens sans pilote	Des images satellitaires révélant des changements, par exemple, dans la qualité des sols ou la disponibilité de l'eau, ont été utilisées pour éclairer les interventions agricoles dans les pays en développement.

Un aperçu des initiatives portant sur les ODD

ODD/champ d'impact	Projet ou initiative	Organisation	Sources de données et outils	Qu'est-ce qui est suivi ou étudié ?	Description	Pays ou région	Conséquences de l'utilisation d'approches fondées sur les données	Années	Niveau	Type d'organisation
Objectif 16 : Paix, justice et institutions efficaces	FollowTheMoney.org	National Institute on Money in Politics	Rapports de financement de campagnes	Financement de campagnes	Compiler et catégoriser des **rapports de financement de campagnes** rendus publics	É.-U.	Promouvoir la transparence du financement des campagnes électorales et le libre accès à un grand nombre de rapports transnationaux	2010-présent	Niveau III	Gouvernement
Objectif 12 : Consommation et production responsables, **Objectif 8 :** Travail décent et croissance économique	Des données obtenues de la lecture de prix dans l'indice des prix à la consommation en Suisse : une alternative à la collecte de prix sur le terrain	Bureau fédéral de la statistique de la Suisse (FSO)	Données obtenues de la lecture de prix	Indice de prix à la consommation	Utiliser des **données obtenues par la lecture de prix** pour calculer l'indice des prix à la consommation pour les aliments et les groupes connexes	Suisse	Améliorer la collecte des prix servant à l'établissement de l'indice des prix à la consommation : améliorer la qualité et réduire les coûts et le fardeau administratif	2018-présent	Non classifié	Gouvernement
Objectif 11 : Villes et communautés durables, **Objectif 12 :** Consommation et production responsables	Utilisation de l'imagerie satellitaire et de données géospatiales pour le recensement agricole et le recensement d'immeubles et de logements	Bureau de la statistique de la Mongolie	Imagerie satellitaire, données géospatiales	Production agricole	Utiliser des **images satellitaires et des données géospatiales** pour déterminer les types de cultures et estimer la production afin de créer un premier recensement agricole	Mongolie	Compléter les données existantes avec des images satellitaires	2017	Non classifié	Gouvernement
Objectif 3 : Bonne santé et bien-être	Évaluation du potentiel de dissémination internationale du virus Ebola par les voyages aériens commerciaux pendant l'épidémie ouest-africaine de 2014	Flowminder	Données de l'Association internationale du transport aérien, itinéraires historiques des vols des voyageurs	Épidémie d'Ebola	Modéliser le nombre attendu d'infections au virus Ebola exportées à l'échelle internationale, l'effet potentiel des restrictions des voyages aériens et l'efficacité du contrôle des voyageurs dans les aéroports aux points d'entrée et de sortie internationaux en utilisant les **données du transport aérien international et les itinéraires historiques des voyageurs**	Guinée, Liberia et Sierra Leone	Éclairer les décideurs et décideuses sur les préjudices potentiels des restrictions de voyage et sur les sites de dépistage les plus efficaces	2014	Non classifié	Université
Objectif 2 : Faim « zéro », **Objectif 3 :** Bonne santé et bien-être, **Objectif 5 :** Égalité entre les sexes	Mégadonnées et infonuagiques – piloter la « télésanté » pour la production de rapports sur le financement communautaire basé sur la performance au Ghana	Banque mondiale	Sondages en ligne	Efficacité du projet d'amélioration de la nutrition en santé maternelle et infantile	Rendre compte des performances des équipes de santé communautaire en utilisant des **outils d'enquête logiciels sur Android**	Ghana	Contourner les délais, les contraintes de capacité et les problèmes de qualité des données associés aux rapports papier	S.O.	Niveau III	International, gouvernement

ODD/champ d'impact	Projet ou initiative	Organisation	Sources de données et outils	Qu'est-ce qui est suivi ou étudié ?	Description	Pays ou région	Conséquences de l'utilisation d'approches fondées sur les données	Années	Niveau	Type d'organisation
Objectif 1 : Pas de pauvreté	Prévision de la pauvreté et de la prospérité partagée à l'aide de données de téléphones cellulaires	Banque mondiale	Registres détaillés d'appels	Estimer et prévoir la pauvreté et la prospérité partagée	Mesurer les « empreintes numériques » de la population en analysant les **registres d'appels de téléphones cellulaires** à l'aide de techniques d'exploration de données et d'apprentissage par ordinateur afin d'estimer et de prévoir la pauvreté et la prospérité partagée	Guatemala	Fournir une solution abordable, pratique et évolutive pour cartographier la pauvreté	2019	Niveau III	International, gouvernement, organisation privée
Objectif 1 : Pas de pauvreté **Objectif 2 :** Faim « zéro », **Objectif 11 :** Ville et communautés durables, **Objectif 13 :** Mesures relatives à la lutter contre les changements climatiques	Prévision de la vulnérabilité aux inondations et renforcement de la résilience grâce aux mégadonnées	Banque mondiale	Données du nuage Google (élévation, imagerie satellitaire, données de recensement)	Risque d'inondation	Utiliser les données du nuage **Google et des données de recensement et d'imagerie satellitaire** pour affiner les prévisions de risque d'inondation en surface au Bangladesh	Bangladesh	Déterminer et définir les populations à risque et améliorer la planification de la gestion des risques de catastrophes	2019	Niveau III	International
Objectif 11 : Villes et communautés durables	Villes fragiles	Igarape Institute	Sources structurées et non structurées	Fragilité	Noter les villes selon un indice de fragilité en utilisant des **sources structurées et non structurées**	À l'échelle mondiale	Comprendre les dimensions de la fragilité des villes grâce à une plateforme de visualisation des données	2010 – 2017	Niveau I	Université, ONG, international
Objectif 5 : Égalité entre les sexes	Chega de FiuFiu	Chega de FiuFiu	Rapports enrichis par les utilisateurs et utilisatrices sur le harcèlement et la discrimination fondée sur le genre	Discrimination fondée sur le genre, violence contre les femmes	Géolocaliser les rapports des citoyens et citoyennes pour créer une carte qui rapporte les points chauds des lieux dangereux et inconfortables pour les femmes en utilisant les **rapports d'incidents de harcèlement provenant d'utilisateurs et d'utilisatrices et géolocalisés**	Brésil	Rendre visibles les points chauds du harcèlement de rue fondé sur le genre	2013-présent	Non classifié	ONG

ODD/champ d'impact	Projet ou initiative	Organisation	Sources de données et outils	Qu'est-ce qui est suivi ou étudié ?	Description	Pays ou région	Conséquences de l'utilisation d'approches fondées sur les données	Années	Niveau	Type d'organisation
Objectif 5 : Égalité entre les sexes	Cartographier la violence électronique à l'égard des femmes (eVAW)	Hamara Internet	Rapports enrichis par les utilisateurs et utilisatrices sur le harcèlement électronique	Discrimination de genre, violence contre les femmes	Géolocaliser les **rapports des citoyens et citoyennes** sur la violence électronique à l'égard des femmes (eVAW) pour cartographier les incidents de violence à l'égard des femmes dans différentes villes du Pakistan	Pakistan	Rendre visibles les points chauds du harcèlement de rue fondé sur le genre	2014-2016	Non classifié	ONG, international
Objectif 16 : Paix, justice et institutions efficaces	Indice Ibrahim de gouvernance africaine	Mo Ibrahim Foundation	Informations provenant d'agences internationales, de projets de données, d'enquêtes	Performance en matière de gouvernance	Mesurer et suivre les performances en matière de gouvernance en utilisant des **données agrégées, regroupées et pondérées provenant de plusieurs sources**, notamment d'agences internationales, de projets de données et d'enquêtes	Afrique	Améliorer la transparence et la responsabilité de la gouvernance en réunissant plusieurs sources de données	2016-présent	Niveau II	International
Objectif 5 : Égalité entre les sexes	Hollaback !	Knight Foundation	Rapports enrichis par les utilisateurs et utilisatrices sur le harcèlement	Harcèlement	Recueillir et suivre les **signalements de harcèlement en ligne, dans la rue et sous d'autres formes provenant des utilisateurs et utilisatrices**	E.-U., Bosnie-Herzégovine, Canada, Colombie, + 12 autres pays	Rendre visible un harcèlement rarement signalé et culturellement accepté	2019	Non classifié	ONG

L'INFLUENCE DU PARLEMENT DE WESTMINSTER SUR LA STRATÉGIE DU ROYAUME-UNI EN MATIÈRE D'IA

LORD CLEMENT-JONES

Lord Clement-Jones, CBE, a été président du comité spécial sur l'IA de la Chambre des lords (Royaume-Uni), coprésident du groupe parlementaire multipartite sur l'IA, membre fondateur du Groupe parlementaire de l'OCDE sur l'IA, membre du Comité ad hoc sur l'IA du Conseil de l'Europe, et porte-parole des libéraux-démocrates en matière de numérique.

ODD 8 - Travail décent et croissance économique

ODD 9 - Industrie, innovation et infrastructure

ODD 16 - Paix, justice et institutions efficaces

L'INFLUENCE DU PARLEMENT DE WESTMINSTER SUR LA STRATÉGIE DU ROYAUME-UNI EN MATIÈRE D'IA

RÉSUMÉ

Au Royaume-Uni (R.-U.), les assises appropriées sont en place pour accueillir une stratégie et un cadre de gouvernance en matière d'intelligence artificielle (IA). Pour ce qui est de percevoir l'incidence de l'IA et d'évaluer les répercussions qu'elle a sur la société, le pays s'est montré à l'avant-garde à divers égards. Tant le Parlement que le gouvernement ont montré qu'ils saisissaient les enjeux en cause. Le premier l'a fait par l'entremise de comités spéciaux et de groupes multipartites, tandis que le second a mis en place une série d'initiatives stratégiques réunissant un certain nombre d'organisations clés, notamment l'Office for AI, le Centre for Data Ethics and Innovation, l'AI Council et l'Alan Turing Institute. La tâche consiste maintenant à coordonner les efforts des nombreuses parties prenantes qui se penchent sur l'avenir de l'IA au R.-U. afin de convenir d'une approche fondée sur le risque quant à la gouvernance de l'IA. Cette approche devra généralement se conformer à des principes établis par l'Union européenne, le Conseil de l'Europe et l'OCDE ainsi qu'à un ensemble de normes communes visant divers outils d'audit et d'évaluation du risque. Ainsi, ceux et celles qui mettent au point, fournissent ou déploient des systèmes d'IA pourront suivre les balises réglementaires qui leur font cruellement défaut à l'heure actuelle, et le pays restera au premier plan du développement de l'IA.

INTRODUCTION

> [...] en cette ère numérique mondiale, il faut que les normes et les règles que nous établissons soient communes.
>
> Cela signifie notamment d'établir des normes et des règles permettant de tirer le maximum de l'intelligence artificielle d'une manière responsable, par exemple en garantissant que les algorithmes ne reproduisent pas les préjugés des personnes qui les conçoivent.
>
> Ainsi, nous voulons que notre Centre for Data Ethics and Innovation, chef de file mondial, travaille en étroite collaboration avec des partenaires internationaux afin qu'ils réfléchissent ensemble aux façons d'assurer un déploiement sûr, éthique et innovant de l'intelligence artificielle (May, 2018).

C'est en ces mots que Theresa May s'est exprimée à l'occasion de la rencontre du Forum économique mondial tenue à Davos en janvier 2018. Celle qui était alors première ministre a axé son discours principal sur la stratégie du Royaume-Uni (R.-U.) quant au développement de l'intelligence artificielle (IA) et sur le fait qu'elle souhaitait que son pays devienne un leader mondial dans les décisions visant à orienter le déploiement de l'IA de manière sûre et éthique. Elle n'aurait pu démontrer plus vigoureusement l'importance que le R.-U. accordait et accorde toujours à la mise au point d'une stratégie efficace, voire exemplaire, à l'échelle internationale en matière d'IA.

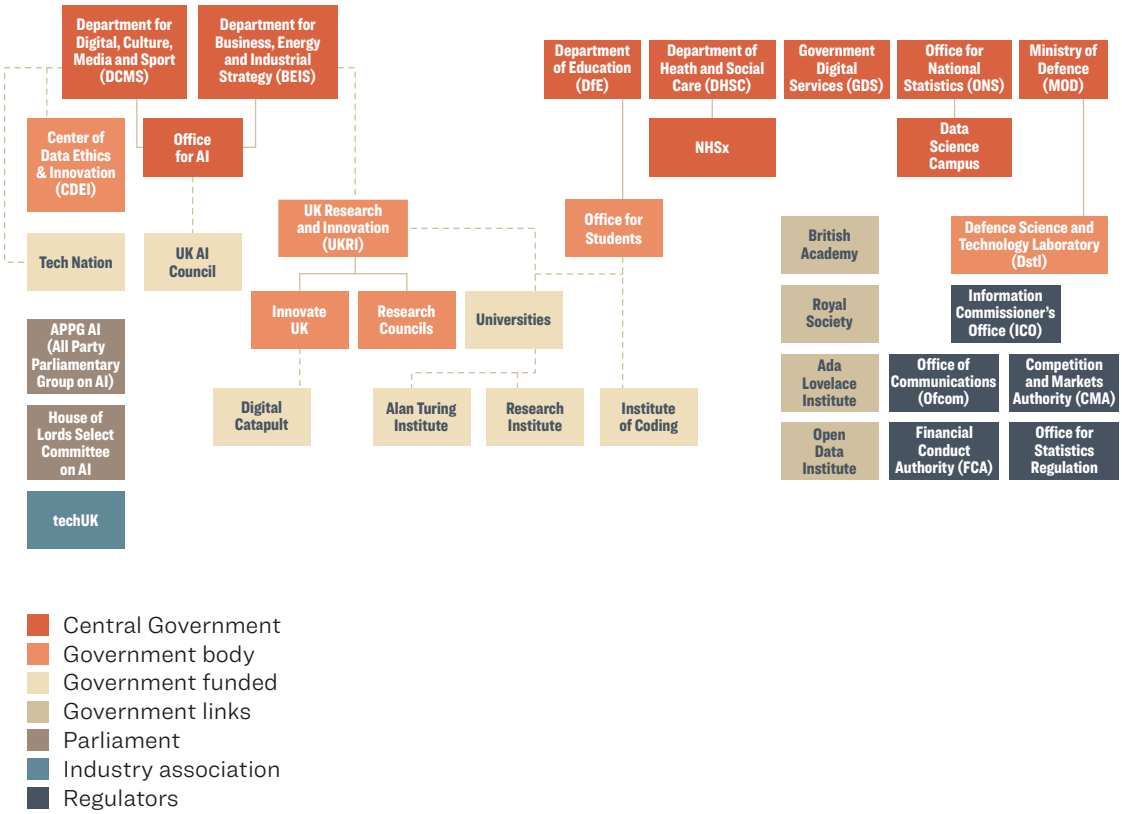
L'ÉCOSYSTÈME DE L'IA AU R.-U.

Bien que la stratégie nationale soit devenue une priorité, l'environnement entourant l'élaboration de politiques en matière d'IA au R.-U. s'est complexifié au fil du temps en ce sens qu'aucune entité en particulier ne s'est vu attribuer la responsabilité de concevoir et de mettre en œuvre cette stratégie. Portrait à ce jour (voir également la figure 1) :

- Au gouvernement, il y a un bureau spécial, l'Office for AI, qui relève à la fois du ministère des Affaires, de l'Énergie et de la Stratégie industrielle ainsi que du ministère du Numérique, de la Culture, des Médias et du Sport. Le déploiement de l'IA au sein du gouvernement est du ressort d'un nouveau bureau chargé du numérique et des données, le Central Digital and Data Office, et de son pendant technique, le Government Digital Service. Dans le système de santé, le service NHSX a le mandat de concevoir des solutions numériques, notamment grâce à l'IA.
- Sur le plan réglementaire, l'Information Commissioner's Office (ICO) supervise l'activité capitale de la gouvernance des données, le Centre for Data Ethics and Innovation (CDEI) fournit des conseils quant à l'éthique liée à l'usage et à la mise au point de l'IA, et une foule d'organismes de surveillance s'intéressent à l'application des algorithmes dans leur domaine, que ce soit par exemple les télécommunications (Ofcom), les marchés (Competition and Markets Authority [CMA]) ou la bourse (Financial Conduct Authority [FCA]).
- Du côté de la recherche et de l'innovation, l'UK Research and Innovation constitue l'organisme public de surveillance. L'Alan Turing Institute est par ailleurs un centre d'excellence voué à la recherche en IA. Grâce à ses membres, il entretient des relations avec de nombreuses organisations – universités, organismes et centres du réseau Catapult – par exemple avec le conseil de recherche EPSRC, de même que les organismes publics Innovate UK et Digital Catapult, qui ont le mandat de faciliter la recherche et le développement dans le domaine.
- En ce qui concerne les organismes non gouvernementaux, la figure de proue est l'AI Council, un comité réunissant des spécialistes qui innovent, développent l'IA et l'utilisent, et qui représentent le monde des affaires, celui de la recherche, le secteur tertiaire ou la fonction publique.
- D'autres organisations influent également sur le cours des travaux, notamment la Royal Society, la British Academy, l'Open Data Institute, la fondation NESTA, l'Ada Lovelace Institute, ainsi que le festival CogX et l'extraordinaire communauté qu'il réunit. Les organismes Big Brother Watch et Liberty too ont pour leur part mené des campagnes sur les répercussions d'une surveillance intrusive par l'IA.

| FIGURE 1 |

'AI public policy and regulation in the UK',
© 2021PricewaterhouseCoopers LLP
(reproduite avec permission, reprise dans Axente 2021).



Quand on en vient aux relations avec des organismes internationaux tels que les Nations Unies, l'Organisation de coopération et de développement économiques (OCDE), le Conseil de l'Europe ou le Partenariat mondial sur l'intelligence artificielle en ce qui concerne l'élaboration d'une politique d'IA à l'échelle internationale, l'expertise de l'Alan Turing Institute, du CDEI et de l'Office for AI entrent en jeu à divers égards.

Voilà qui contraste avec le paysage de l'IA dans plusieurs pays – l'environnement observé au Canada ou en Allemagne, par exemple, est nettement moins complexe – et qui constitue à la fois une force et une faiblesse, comme j'espère bien l'expliquer dans ce chapitre. Même s'il existe ce qu'on a appelé un « éthos de la collaboration », qui « relie les ministères du gouvernement au milieu universitaire, aux groupes de réflexion et au monde des affaires, et qui façonne peu à peu l'évolution de l'IA au R.-U. » (Axente, 2021), l'axe et la cadence d'élaboration d'une politique et d'une stratégie ont fait l'objet

de critiques à plusieurs égards, par exemple en ce qui a trait à la prise de décisions algorithmique dans le secteur public ou au déploiement d'une technologie de reconnaissance faciale en direct dans des endroits publics.

Ma tâche et celle de mes collègues parlementaires consistait et continue de consister à démêler ces intrications et le rôle des nombreuses entités, de faire pression pour une coordination stratégique de la politique en matière d'IA, et d'influer sur la conception et la mise en œuvre de cette politique, ne serait-ce qu'en évaluant les possibilités et les risques que l'IA présente pour la société. Nous avons fait cela au moyen d'une multitude de rapports – du comité des sciences et de la technologie de la Chambre des communes, du comité spécial sur l'IA de la Chambre des lords et d'un groupe parlementaire multipartite sur l'IA –, de débats parlementaires et de questions, et grâce au travail de notre comité sur les normes de la vie publique.

Que ce soit dans le comité spécial ou le groupe parlementaire multipartite sur l'IA, nous avons repéré des solutions ou des systèmes d'IA concernant une multitude de domaines, notamment l'éducation, les villes intelligentes, la santé et la gestion de l'énergie. Nous avons examiné le potentiel de certaines applications d'IA en particulier. À titre de parlementaires, nous avions également le souci d'atténuer les risques que comporte l'IA sur le plan éthique et sociétal.

La conception d'une première stratégie britannique en matière d'IA : l'examen de Hall et Pesenti et la stratégie industrielle

Pour en revenir à notre propos, le discours de la première ministre May, la politique dont ce discours fait état et une partie de la structure organisationnelle trouvent leurs sources dans un examen indépendant. En mars 2017, le gouvernement britannique a commandé un tel examen à Dame Wendy Hall, professeure de sciences informatiques à l'Université de Southampton, et à Jérôme Pesenti, alors chef de la direction de Benevolent Tech, et il leur a demandé de se pencher sur les répercussions économiques de l'IA au R.-U. Au mois d'octobre suivant, ils ont publié leur rapport intitulé *Growing the Artificial Intelligence Industry in the UK* (Hall et Pesenti, 2017).

Le rapport de Hall et Pesenti (2017) contient un certain nombre de recommandations clés traçant une voie bien définie pour élaborer la stratégie du R.-U. en matière d'IA :

- Étant donné l'importance des ensembles de données servant à l'entraînement et au fonctionnement des systèmes d'IA, mettre en place des fiducies de données (les Data Trusts), soit des structures fiables et éprouvées qui facilitent le partage de données entre des organisations qui en détiennent et d'autres qui cherchent à en utiliser pour mettre au point l'IA.
- Améliorer l'offre de compétences et reconnaître la valeur ainsi que l'importance de la diversité de la main-d'œuvre en IA en établissant un important programme de cours universitaires (de niveau maîtrise) liés à l'IA pouvant accueillir une première cohorte de 300 étudiants et étudiantes, en offrant à des diplômés et diplômées souhaitant entreprendre une maîtrise en IA, mais n'ayant pas étudié en informatique ou en science des données, la possibilité de faire une propédeutique d'un an et en admettant dans les principales universités 200 doctorants et doctorantes supplémentaires faisant de la recherche sur l'IA.

- Faire de l'Alan Turing Institute l'institut national voué à l'IA et à la science des données, et créer au R.-U. un programme international de bourses Turing de recherche en IA.
- Établir un conseil britannique sur l'IA pour faciliter la coordination et la croissance dans le domaine de l'IA à l'échelle nationale.

Le gouvernement du R.-U. a par ailleurs publié, en novembre 2017, son livre blanc *Industrial Strategy : Building a Britain Fit for the Future*, selon lequel mettre l'IA « à l'avant-plan de la révolution britannique relative à l'IA et aux données » constitue l'un des quatre grands enjeux fondamentaux pour l'avenir du pays. Cette stratégie industrielle a reconnu le rôle primordial que joue l'éthique dans l'adoption de l'IA au pays, ce qui a entraîné la mise sur pied, à la fin de 2018, du CDEI dans l'objectif « de garantir que les données et l'IA offrent les meilleures retombées possibles à la société en soutenant un usage éthique et innovant de celle-ci » (United Kingdom Government, 2017).

La stratégie industrielle du R.-U. a par la suite mené, au début de 2018, à une entente pour financer le domaine de l'IA à hauteur de 950 millions de livres sterling de manière à appliquer presque toutes les recommandations formulées par Hall et Pesenti. Elle a également conduit à l'établissement, au sein du gouvernement, de l'Office for AI en vue de coordonner la mise en œuvre des recommandations (United Kingdom Government, 2019a).

L'activité parlementaire au R.-U. : le groupe parlementaire sur l'IA et le rapport de la Chambre des lords

Le mois où l'examen de Hall et Pesenti a été annoncé, le député Stephen Metcalfe et moi avons organisé la première rencontre du nouveau groupe parlementaire multipartite sur l'IA mis sur pied avec l'aide du Big Innovation Centre et de Justin Anderson, alors membre de l'Hypercat Alliance, en vue d'exposer nos préoccupations concernant le manque d'attention du parlement quant à l'avenir de l'IA. Le groupe allait aider les députés et députées ainsi que mes pairs de la Chambre des lords à nouer le dialogue avec la communauté de l'IA. Il devait contribuer à façonner une politique en matière d'IA pour le R.-U. et, particulièrement, à se pencher sur les enjeux éthiques, moraux et sociétaux ainsi que sur la gouvernance et les incidences réglementaires.

J'ai alors abordé de telles questions éthiques et morales en faisant référence à Tay, le robot de clavardage de Microsoft en interaction avec le public, qui avait été désactivé moins d'une semaine après sa mise en service, en mars 2016, parce qu'il avait tenu des propos racistes et sexistes (Taylor, 2017). Le nouveau groupe parlementaire a discuté de ce sujet :

> Allons-nous vraiment inculquer nos valeurs à l'IA ? Le souhaitons-nous ? [...] Si nous souhaitons inculquer les pires aspects du comportement humain, tel que nous semblons pouvoir le faire dans des cas comme Tay, ou même inculquer une conduite violente à des robots militaires [...], nous devrions réfléchir aux valeurs de manière fort différente.

L'éthique et la réglementation liées à l'IA ne faisaient pas partie du mandat d'examen de Hall et Pesenti. Cependant, peu après sa mise sur pied, en juin 2017, le comité spécial sur l'IA de la Chambre des lords, que j'ai été appelé à présider, a été chargé « d'évaluer les répercussions économiques, éthiques et sociales découlant des avancées de l'intelligence artificielle ». Dès le début de nos travaux, nous nous sommes posé cinq questions et les avons également posées aux personnes venues témoigner devant nous (United Kingdom Parliament, 2018a, p. 153) :

1. De quelle manière l'IA intervient-elle dans le quotidien des gens et comment est-ce appelé à changer ?
2. Quelles possibilités l'IA offre-t-elle au R.-U. ? Comment en profiter ?
3. Quels sont les risques et les effets potentiels de l'IA ? Comment les éviter ?
4. Comment devrait-on mobiliser le public par rapport à la responsabilisation en matière d'IA ?
5. Quels enjeux éthiques la mise au point et l'usage de l'IA soulèvent-ils ?

Les principales structures ont été mises en place et le but de chacune a été fixé, puis le comité spécial sur l'IA de la Chambre des lords a publié, en avril 2018, un rapport intitulé *AI in the UK : Ready, Willing and Able ?* Ce rapport en disait beaucoup sur la stratégie en matière d'IA qui prenait alors forme. Le comité s'y demandait si l'approche était la bonne et mentionnait le besoin de coordination pour accoucher de la stratégie. Nous avons également insisté sur le besoin de déployer et d'utiliser les systèmes d'IA de manière éthique.

Nos travaux ont conclu que le R.-U. bénéficiait d'une position avantageuse pour devenir un leader mondial relativement à l'évolution de l'IA, étant donné qu'on y trouvait, dans un faible rayon, d'importantes entreprises du domaine, une dynamique culture de recherche universitaire, un vigoureux écosystème d'entreprises émergentes, ainsi qu'une constellation de forces juridiques, éthiques, financières et linguistiques. Si on la gérait soigneusement, l'IA pouvait offrir de grandes possibilités à l'économie britannique (United Kingdom Parliament, 2018a).

Nous avons formulé des recommandations afin d'aider le gouvernement et le pays à constater le potentiel de l'IA pour notre économie et notre société, et de protéger celle-ci contre d'éventuels risques ou menaces. Nous avons signalé qu'une mauvaise gestion pouvait miner la confiance du public dans l'IA. Le R.-U. n'avait qu'une chance de façonner son rôle distinctif de pionnier de l'IA éthique.

Nous avons notamment indiqué dans nos recommandations que le gouvernement devait établir un cadre stratégique national – qui s'harmonise à la stratégie industrielle –, entre autres pour coordonner l'élaboration de la politique du R.-U. en matière d'IA et en assurer la mise en application. « Le R.-U. doit chercher à façonner l'évolution de l'IA et son utilisation de manière active ou il risque d'en accepter passivement de nombreuses conséquences (United Kingdom Parliament, 2018a). »

En attendant les principes de l'OCDE sur l'intelligence artificielle (OCDE, 2019), qui allaient venir plus tard, nous avons proposé cinq principes qui pourraient servir de base à un code régissant le domaine de l'IA à l'échelle nationale et internationale.

À ce moment-là, nous n'avions pas recommandé la mise sur pied d'un nouvel organisme de surveillance propre à l'IA, mais avions mentionné qu'un tel cadre de principes pouvait éventuellement étayer la réglementation, si cela s'avérait nécessaire, et que les organismes de surveillance alors en place étaient mieux à même d'encadrer l'IA dans leurs domaines respectifs.

Nous avions le souci particulier d'éviter que des préjudices déjà observés se répètent, sans qu'on le veuille, dans les systèmes automatisés à venir, et de voir à ce que ces systèmes soient soigneusement conçus dès le départ. Comme l'avaient recommandé Hall et Pesenti (2017), il nous fallait veiller à établir de nouveaux mécanismes et de nouvelles structures, par exemple les fiducies de données. Pour nous assurer que notre usage de l'IA n'allait pas par mégarde nuire à certains groupes au sein de la société, nous avons demandé au gouvernement d'encourager l'élaboration de nouvelles approches pour auditer les ensembles de données employés en IA ainsi que d'accroître la diversité en ce qui a trait à la formation

et au recrutement de spécialistes dans le domaine. Étant donné la forte probabilité qu'il soit difficile d'embaucher du personnel, nous avons également recommandé au gouvernement de prévoir un financement notable pour soutenir la formation et l'acquisition de compétences. Le perfectionnement professionnel continu deviendrait une nécessité. Toutes ces mesures mises ensemble allaient selon nous faire en sorte que le R.-U. demeurerait concurrentiel dans le domaine et qu'il bénéficierait de la confiance du public. Notre rapport reste un document qui a eu de l'influence sur la politique publique, car il a adopté une approche holistique pour encadrer l'IA en abordant à la fois les possibilités qu'elle offre, ses répercussions sur la société, les risques et les enjeux éthiques qu'elle pose, ainsi que la mobilisation publique à cet égard.

La réaction du gouvernement, prise 1

Tout rapport soumis au Parlement s'évalue en fait selon que le gouvernement en accepte ou non les recommandations. À ce propos, nous avons dressé un bilan contrasté (United Kingdom Government, 2018).

Parmi les éléments positifs, soulignons :

- La reconnaissance par le gouvernement du besoin de maintenir et d'accroître la confiance du public en adoptant une approche éthique à l'échelle nationale et internationale.
- La nomination d'un nouveau président à la tête du CDEI ainsi que le début de consultations sur le rôle et les objectifs de ce centre, notamment au sujet d'ententes relatives à la gestion des fiducies de données et de l'accès à des ensembles de données publics.
- La reconnaissance, par la CMA, d'enjeux liés à la concurrence concernant le monopole sur les données.
- La reconnaissance du besoin de « multiples perspectives [...] durant les phases de conception, de déploiement et de mise en service d'algorithmes » ainsi que de diversité au sein de la main-d'œuvre en IA.
- L'engagement à soutenir un plan national de perfectionnement professionnel.

D'un autre côté :

- L'entente sur le financement accordé au domaine de l'IA constituait un premier pas, mais n'allait pas suffire à élaborer un cadre stratégique national.
- On ne savait pas au juste si le nouvel Office for AI au sein du gouvernement allait assurer une coordination accrue avec le nouveau conseil sur l'IA et si le CDEI allait avoir les ressources et le statut nécessaires pour parvenir à un cadre éthique national.
- Le ministère de la Santé n'admettait que de manière mitigée le besoin de transparence, particulièrement pour les applications employées dans le réseau de la santé.
- Le ministère de l'Éducation montrait une attitude défensive par rapport au dossier des stages et au besoin de remanier la taxe d'apprentissage, et il semblait peu comprendre le besoin d'une capacité de raisonnement critique et créatif de même que de compétences en informatique.
- Le ministère de la Défense a continué de s'appuyer, dans ses commentaires, sur une définition de « autonome » que plus aucun pays n'utilise quand il est question de systèmes d'armes létaux autonomes (SALA).

Le gouvernement a donc eu des réactions qui auguraient bien, d'autres moins. À cette étape, nous reconnaissions toutefois que la politique en matière d'IA en était à ses balbutiements au R.-U. et que le gouvernement avait entrepris de manière satisfaisante l'élaboration de la politique.

Un autre rapport : pas de place au laisser-aller

À l'automne 2020, le comité de liaison de la Chambre des lords, qui coordonne le travail des comités spéciaux, nous a demandé, à certaines personnes ayant été membres du comité spécial sur l'IA et à moi, d'effectuer un suivi et d'examiner comment la situation avait progressé depuis notre dernier rapport.

En décembre 2020, après avoir recueilli le point de vue de ministres, d'organismes de surveillance ainsi que d'intervenantes et intervenants clés, nous avons publié un nouveau rapport, *AI in the UK : No Room for Complacency*, qui présentait les progrès réalisés par le gouvernement britannique (United Kingdom Parliament, 2020a). Il contenait également les principales recommandations suivantes.

Confiance du public et gouvernance des données

Une sensibilisation accrue du public est essentielle pour faire adopter davantage l'IA ainsi que pour remettre en cause toute organisation qui déploie l'IA d'une manière semblant dangereuse d'un point de vue éthique. Le gouvernement devait concrètement prendre des moyens pour expliquer au grand public comment l'IA se sert des renseignements personnels. En outre, il fallait accélérer la cadence d'élaboration d'une politique en vue de protéger les données, par exemple au moyen des fiducies de données. En n'agissant pas ainsi, on se ferait doubler par les avancées technologiques.

Code d'éthique

Dès notre premier rapport, il y avait eu un consensus manifeste selon lequel l'IA éthique était la seule façon viable d'aller de l'avant. Depuis, le R.-U. était devenu signataire de la recommandation de l'OCDE sur l'IA, qui englobait cinq principes d'une « approche responsable en soutien d'une IA digne de confiance » (OCDE, 2019) ainsi que les principes sur l'IA du G20, qui s'avéraient peu contraignants. Nous avons alors signalé que le temps était venu pour le gouvernement du R.-U. de cesser de définir des règles éthiques pour plutôt se pencher sur la façon de les instaurer dans la mise au point et le déploiement de systèmes d'IA. Le gouvernement devait montrer la voie à suivre pour faire de l'IA éthique une réalité. Ne pas agir ainsi revenait à dissiper les progrès réalisés jusque-là et à ne pas profiter des possibilités que présente l'IA pour toute la population du pays. Nous avons donc demandé au CDEI d'établir et de publier des normes éthiques nationales concernant la mise au point et le déploiement de l'IA.

Risques et réglementation

En matière de risques et de réglementation, nous avons indiqué que les utilisateurs et utilisatrices ainsi que les responsables des orientations politiques devaient mieux comprendre les risques que pose l'utilisation concrète de l'IA ainsi que la manière de les évaluer et de les atténuer. Notre rapport recommandait que l'ICO conçoive – de pair avec le CDEI, l'Office for AI et l'Alan Turing Institute – une formation destinée à celles et ceux qui sont chargés de la surveillance.

Compétences et amélioration des compétences

En ce qui a trait aux compétences, nous jugions très préoccupante l'inertie du gouvernement. Son action n'était pas à la hauteur des défis auxquels faisaient face les travailleurs et travailleuses du pays quant à sa cadence, son ampleur et sa portée. Lorsque la pandémie de COVID-19 allait connaître une accalmie et que le gouvernement britannique allait devoir se pencher sur son incidence économique, le monde du travail aurait changé. L'IA n'allait pas nécessairement remplacer de très nombreuses personnes effectuant un même travail, mais elle créerait d'autres emplois et ferait appel à diverses compétences. Il importait que le gouvernement et l'industrie prennent des mesures pour s'assurer que les compétences numériques soient actualisées au R.-U., et que les gens aient la chance de parfaire leurs compétences et de se perfectionner afin de s'adapter à un marché du travail en pleine évolution. Il fallait concevoir un plan

de formation précis pour aider les gens à travailler dans un contexte d'automatisation et avec l'IA de même qu'à maximiser leur potentiel. Il y avait également un urgent besoin de faire place à la diversité et à l'inclusion au sein de la main-d'œuvre en IA et d'améliorer la littératie numérique.

Coordination stratégique

Nous avions conclu que le gouvernement britannique avait bien fait de mettre sur pied diverses structures afin de formuler des conseils sur l'IA à long terme. Nous avons toutefois lancé une mise en garde contre tout laisser-aller. Il s'avérait essentiel d'établir une coordination entre les différentes organisations qui se préoccupaient de l'évolution de l'IA, y compris les nombreux organismes de surveillance. Le gouvernement britannique devait mieux coordonner sa politique en matière d'IA ainsi que l'usage des données et des technologies que font les gouvernements locaux et national. Ainsi, nous avons recommandé la création d'un comité du Cabinet dont la principale tâche serait de veiller à l'élaboration d'une stratégie quinquennale en matière d'IA, stratégie qui devait amener la société à tirer parti de l'IA avant que l'inverse ne se produise.

Engagement à l'international

L'une des conclusions de notre nouveau rapport était que le R.-U. devait afficher son leadership à l'échelle internationale par rapport à des enjeux communs par l'entremise d'initiatives telles que le Partenariat mondial sur l'intelligence artificielle. Au regard des SALA, nous nous inquiétions tout autant qu'avant du manque d'intervention, malgré la mise sur pied d'un nouveau centre de développement de l'autonomie au sein du ministère de la Défense. Le travail du centre allait selon nous être entravé par l'incapacité d'harmoniser la définition de « armes autonomes » retenue par le R.-U. avec celle de ses partenaires internationaux.

La feuille de route sur l'IA

L'AI Council a fait paraître sa feuille de route sur l'IA en janvier 2021, soit peu après le dernier rapport du comité de la Chambre des lords, dont elle reprend nettement un certain nombre d'éléments (United Kingdom Government, 2021a). Le document *AI Roadmap* recommande en effet de concevoir une stratégie nationale en matière d'IA au R.-U. Il rappelle que le pays devrait piloter l'élaboration de normes appropriées sur la gouvernance des données en plus d'adopter une réglementation « claire et adaptable » en se basant sur les conseils des organismes de surveillance, notamment ceux de l'ICO.

Dans la feuille de route, il est mentionné que le public devrait se sentir rassuré puisque « l'usage de l'IA est sûr, sécuritaire, juste, éthique et dûment encadré par des entités indépendantes » et que les organismes de surveillance ont le pouvoir d'appliquer des sanctions. La feuille de route insiste non seulement sur la constante évolution des normes de l'industrie ainsi que la réglementation et les cadres appropriés en ce qui a trait à la responsabilité algorithmique, mais également sur la nécessité de devenir un leader mondial de la réglementation adaptée et de la gouvernance. On y suggère qu'une entité indépendante fournisse des conseils quant aux « prochaines étapes de l'évolution des mécanismes de gouvernance, y compris l'évaluation des risques et l'étude des effets, les principes soutenant les pratiques exemplaires, les processus éthiques et les mécanismes institutionnels qui amélioreront et maintiendront la confiance du public » (United Kingdom Government, 2021a).

La réaction du gouvernement, prise 2

À la suite du dernier rapport des lords, le gouvernement a publié un document d'orientation en février 2021, laissant encore une fois voir une réaction mitigée, particulièrement en ce qui concerne les compétences, mais il a tenu compte de notre principale suggestion d'élaborer une stratégie nationale en matière d'IA. Au moment où nous écrivons ces lignes, à l'automne 2021, il devait livrer la stratégie pour laquelle

il a certainement bénéficié de l'apport considérable de la feuille de route de l'AI Council et, je l'espère, du concours du Parlement.

Le gouvernement a accueilli favorablement tant les recommandations du rapport que son appel à éviter un laisser-aller. Il a relevé les éléments figurant aussi dans la feuille de route de l'AI Council, notamment le fait qu'il devait axer son approche sur la coordination entre les organisations – que ce soit au gouvernement et dans la fonction publique, entre les organismes de surveillance ou dans le milieu universitaire et l'industrie – afin de « s'assurer que l'élan des dernières années ne [faiblirait] pas, mais qu'il se [raviverait] plutôt pour stimuler la relance et la prospérité économiques aux quatre coins du R.-U., et qu'il nous [permettrait] d'être des leaders dans la résolution d'enjeux mondiaux en matière d'IA » (United Kingdom Government, 2021e).

En ce qui concerne la sensibilisation du public et les données, le gouvernement a absolument admis l'importance capitale de faire avancer ces dossiers en accélérant l'élaboration de cadres juridiques exploitables quant à la gouvernance des données, en particulier des données de santé publique, et en se penchant sur les enjeux de la concurrence relative aux données, tel que le recommandaient dans des rapports un groupe de travail de la CMA et un autre groupe de spécialistes (United Kingdom Government, 2019b, 2020a).

Pour ce qui est de l'éthique et des recommandations concernant l'adoption de normes nationales relatives à la mise au point et au déploiement de l'IA, le gouvernement s'engageait à ce que le Government Digital Service évalue la mise au point d'un mécanisme approprié et efficace pour améliorer la transparence en cas de recours à la prise de décisions assistée par des algorithmes dans le secteur public. Il examinait également les rôles que le CDEI serait appelé à jouer.

En ce qui a trait aux emplois et au plan d'action qui avait été qualifié d'inadéquat pour prévoir les compétences et le perfectionnement qui s'avéreront nécessaires, le gouvernement a fait valoir que *the future of work* – « l'avenir du travail » – allait guider les politiques d'un certain nombre de ministères. C'était prévu que le domaine de l'éducation et de la formation aux adultes prendrait énormément d'ampleur, car il fallait remettre des travailleurs et travailleuses à niveau et les préparer à naviguer dans l'économie postpandémique en leur offrant des compétences « garanties à vie ». Le gouvernement a rappelé l'annonce qu'il avait faite en 2020 à propos de stages en IA. Il s'est montré d'accord avec notre rapport et la feuille de route sur l'IA quant au besoin de diversité. Il a par ailleurs souligné la remise de 1 000 doctorats dans 16 centres de formation doctorale, la tenue de 100 cours de maîtrise financés par l'industrie, ainsi que l'organisation de 2 500 cours et l'offre de 1 000 bourses pour soutenir la reconversion professionnelle (vers l'IA) de personnes représentant des groupes minoritaires.

Au regard des recommandations du comité qui se rapportaient à la confiance du public et à la réglementation, le gouvernement a mentionné la mise sur pied – par la CMA, l'ICO et Ofcom – d'un forum coopératif sur la réglementation du numérique, lequel facilitera la coordination des initiatives de surveillance des marchés numériques ainsi que la coopération dans des domaines jugés importants, ce qui pourrait mener à la conception d'une formation sur la réglementation relative à l'IA. Le gouvernement a aussi signalé son intention de concevoir une stratégie relative à l'éducation aux médias, laquelle « garantirait une approche coordonnée et stratégique relativement à l'éducation aux médias offerte en ligne ainsi qu'à la sensibilisation des enfants, des jeunes et des adultes » (United Kingdom Government, 2020g).

Quant à la coordination stratégique, le gouvernement a confirmé que la politique en matière d'IA et le fait de favoriser l'adoption de l'IA relevaient de la responsabilité partagée des ministres du Numérique, de la Culture, des Médias et du Sport ainsi que des Affaires, de l'Énergie et de la Stratégie industrielle. Il a soutenu que les avantages de l'IA étaient l'affaire des organismes et du gouvernement dans son ensemble.

La réaction qui a sans doute été la plus surprenante et encourageante concernait les SALA :

> Nous convenons que le R.-U. doit pouvoir participer aux débats internationaux sur les armes autonomes et jouer activement un rôle de leader moral et éthique sur la scène mondiale. Nous convenons également de l'importance de nous assurer que des définitions officielles n'affaiblissent pas nos arguments ni qu'elles nous éloignent de nos alliés.

Même si on ne s'entendait pas encore sur une définition valable de « SALA », le R.-U. avait récemment reconnu les dernières définitions retenues par l'OTAN pour « autonome » et « autonomie ». Le gouvernement a signalé que le pays avait du poids dans les discussions portant sur cet enjeu au sein du Groupe d'experts gouvernementaux sur les SALA, établi par les Nations Unies dans le cadre de la Convention sur certaines armes classiques. En outre, le ministère de la Défense préparait alors la publication d'une nouvelle stratégie relative à l'IA dans son champ de compétence et allait « continuer à se montrer proactif en abordant des enjeux éthiques entourant la mise au point et l'utilisation de l'IA à des fins militaires ».

Les rapports parlementaires sur des applications en particulier : la technologie de reconnaissance faciale en direct et la prise de décisions algorithmique

Parallèlement aux interventions des comités parlementaires dans le débat général sur les éventuelles possibilités de l'IA et la gouvernance de celle-ci, d'autres entités gouvernementales se sont prononcées sur l'incidence de certaines utilisations de systèmes d'IA, et particulièrement sur le recours à de telles applications dans le secteur public.

En mai 2018, le comité des sciences et de la technologie de la Chambre des communes, sous la présidence de Sir Norman Lamb, a publié le rapport *Algorithms in Decision-Making* (United Kingdom Parliament, 2018b), qui portait sur l'usage d'algorithmes pour la prise de décisions dans le secteur public et le monde des affaires. Ce document servait une mise en garde quant au besoin de cerner et de combattre les partis pris ainsi qu'au besoin de transparence et d'imputabilité. Il faisait également ressortir des zones grises du Règlement général sur la protection des données concernant la prise de décisions automatisée.

Le comité a ensuite présenté, en juillet 2019, ses constats quant au travail du commissaire à la biométrie et de l'organisme Forensic Science Regulator. Il a recommandé l'imposition d'un moratoire sur l'usage de la technologie de reconnaissance faciale en direct (United Kingdom Parliament, 2019).

Il constatait que la stratégie du gouvernement en matière de biométrie n'était pas à la hauteur après cinq ans d'élaboration :

> Peut-être n'est-ce même pas une « stratégie » du tout. Il y a un manque de cohérence et de prévoyance dans sa perspective, et elle ne parvient pas à combler le vide législatif que le Home Office a laissé se manifester par rapport aux nouvelles technologies de biométrie.

Le comité a également demandé au gouvernement « de décréter un moratoire sur l'usage qui est actuellement fait de la technologie de reconnaissance faciale », précisant qu'il fallait éviter d'en faire l'essai « jusqu'à l'adoption d'un cadre législatif et d'orientations sur les protocoles d'essai, ainsi qu'à l'établissement d'un système de surveillance et d'évaluation ».

Devant chacun de ces rapports, le gouvernement s'est montré extrêmement réticent à promettre un véritable plan d'action. Il n'a d'ailleurs réagi au second rapport que près de deux ans après sa publication.

En mars 2019, l'influent comité sur les normes de la vie publique – soit l'organisme consultatif indépendant chargé de conseiller le premier ministre quant aux normes éthiques qui s'appliquent à l'ensemble de la vie publique au R.-U., lequel était alors sous la présidence de Lord Evans of Weardale – a toutefois lancé un processus d'enquête. Il souhaitait comprendre en quoi l'IA touchait les principes de Nolan, soit les sept principes régissant la vie publique au R.-U. : désintéressement, intégrité, objectivité, imputabilité, transparence, honnêteté et leadership. Il voulait également vérifier si la stratégie du gouvernement réussirait à faire respecter ces principes à mesure que l'IA s'implanterait dans les services publics. Le comité mentionnait des préoccupations d'ordre éthique en lien, par exemple, avec l'usage de données partiales ou la face cachée des algorithmes. Il a fait valoir que l'IA devait être au service du bien public.

En février 2020, le comité sur les normes de la vie publique a publié le rapport *Artificial Intelligence and Public Standards* (United Kingdom Government, 2020c), qui contient un certain nombre de recommandations clés en vue de consolider le cadre éthique du R.-U. lié au déploiement de l'IA dans le secteur public. Il a rappelé au gouvernement que le cadre de réglementation et de gouvernance du R.-U. à ce sujet demeurait en constante évolution et qu'il avait relevé des lacunes. Il y avait un urgent besoin d'orientation et de réglementation en ce qui concerne en particulier les enjeux de la transparence et des données partiales. Le gouvernement devait clarifier quels principes il fallait suivre. Le comité a également mentionné que le respect des principes de la vie publique nécessitait le concours des organismes publics utilisant l'IA dans leurs services de première ligne. Chaque organisme public devrait indiquer en quoi son utilisation de l'IA se conforme à la législation touchant les technologies axées sur les données ainsi que mettre en place une gouvernance claire et fondée sur le risque quant à cette utilisation. Il faudrait par ailleurs qu'un organisme de surveillance (le CDEI, par exemple) s'emploie à repérer les lacunes dans la réglementation et qu'il fournisse des conseils à d'autres organismes de surveillance et au gouvernement au sujet d'enjeux associés à l'IA. Le gouvernement devrait aussi se demander comment imposer l'obligation d'une étude d'impact afin d'évaluer, pour chaque projet, les répercussions potentielles de l'IA sur les principes de la vie publique, et ce, dès l'étape de la conception du projet. De telles études d'impact seraient donc obligatoires et devraient être publiées.

Encore une fois, le gouvernement a réagi tardivement en ne publiant qu'en mai 2021 son document d'orientation (United Kingdom Government, 2021d). Il a toutefois répondu plus favorablement au rapport du comité sur les normes de la vie publique qu'à ceux du comité des sciences et de la technologie. Il a convenu que le nombre de principes liés à l'IA et leur diversité pouvaient semer la confusion au moment de déployer l'IA dans le secteur public. Le gouvernement du R.-U. a fait valoir qu'il avait adhéré aux principes multilatéraux sur l'IA, y compris ceux de l'OCDE, et qu'il s'était engagé à les mettre en œuvre à titre de membre fondateur du Partenariat mondial sur l'intelligence artificielle. Les lignes directrices du Forum économique mondial sur l'approvisionnement relatif à l'IA avaient mené à des lignes directrices propres au R.-U. Celles-ci étaient le fruit du travail de l'Office for AI – mené en collaboration avec le Centre pour la quatrième révolution industrielle du Forum économique mondial, du Government Digital Service, du Government Commercial Function ainsi que du Crown Commercial Service – et visait à ce que les organismes publics puissent s'approvisionner en IA de manière sûre et éthiquement responsable.

Pour clarifier les principes éthiques et les orientations, le gouvernement avait auparavant publié un document en ligne intitulé *Data Ethics and AI Guidance Landscape* (United Kingdom Government, 2020d), qui contenait une liste de ressources liées à l'éthique des données destinée aux fonctionnaires. Il allait envisager l'élaboration d'un mécanisme efficace et approprié pour améliorer la transparence quant à l'usage d'algorithmes assistant la prise de décisions dans le secteur public. La Equality and Human Rights Commission allait orienter les autorités publiques sur la manière de s'assurer que tout système d'IA respecte le devoir d'équité imposé au secteur public.

La position du gouvernement est demeurée insatisfaisante en ce qui concerne le déploiement de certains systèmes d'IA en particulier, notamment :

- Au regard de la technologie de reconnaissance faciale en direct, le gouvernement a promis, dans sa réponse de mars 2021 au rapport du comité des sciences et de la technologie (United Kingdom Parliament, 2021c), qu'il présenterait à l'institut national de police des directives sur l'usage de cette technologie (en conformité avec l'affaire Bridge). Le comité a toutefois pris une initiative inhabituelle en écrivant aux ministres du Home Office (United Kingdom Parliament, 2021b) pour leur exprimer de « sérieuses inquiétudes quant au manque de progrès réalisés par le gouvernement relativement à la pérennité des services de sciences judiciaires, à l'agrément des laboratoires, à la gouvernance en matière de biométrie et à la gestion des photographies d'identité judiciaire ». Ils ont réclamé que les lignes directrices nationales soient mises à jour et ont demandé si le gouvernement avait l'intention d'adopter un cadre législatif clair par rapport à la technologie de reconnaissance faciale automatisée. Il est maintenant possible de consulter un guide provisoire, mais celui-ci fait déjà l'objet de critiques de la part de l'actuel commissaire à la surveillance vidéo et de l'un de ses prédécesseurs.
- En ce qui concerne la prise de décisions algorithmique, le gouvernement a publié, à la suite d'une controverse sur l'usage d'algorithmes dans les domaines de l'éducation, du logement et de l'immigration, un cadre favorisant une prise de décisions automatisée éthique, transparente et responsable au sein du secteur public (United Kingdom Government, 2021c). Ni le Central Digital and Data Office ni le Bureau du Cabinet ne disposent toutefois de normes de conformité satisfaisantes ou d'un mécanisme d'application pour veiller au respect des principes contenus dans ce cadre.
- Au même moment, l'organisme Big Brother Watch publiait son rapport *Poverty Panopticon*, qui expose l'étendue des enjeux de la prise de décisions algorithmique auxquels font face les administrations locales (Big Brother Watch, 2021).

Ainsi, j'ai soutenu, au cours des dernières années, des projets de loi d'initiative parlementaire prévoyant la surveillance de la prise de décisions algorithmique par une autorité publique et demandant un moratoire ainsi qu'un examen sur l'usage de la technologie de reconnaissance faciale automatisée (United Kingdom Parliament, 2020b). Ces projets de loi visaient à offrir un puissant cadre législatif et réglementaire pour protéger les libertés civiles. J'ai aussi abordé des enjeux liés à la réglementation durant des débats ou des périodes de questions[58].

Un nouveau comité de la Chambre des lords s'intéressant à la justice et aux affaires intérieures assure maintenant un suivi à cet égard en plus d'examiner l'emploi de nouvelles technologies par les forces de l'ordre (United Kingdom Parliament, 2021a).

Le bilan de l'influence parlementaire

Dresser une liste des actions du gouvernement et plus généralement des politiques élaborées en matière d'IA fait ressortir une évidence : la situation a progressé. Il reste néanmoins difficile d'évaluer exactement les avancées pour lesquelles le Parlement a eu une influence déterminante. Ce dernier a probablement contribué, la plupart du temps, à maintenir l'élan et à alimenter la réflexion plutôt qu'à changer l'orientation fondamentale des politiques.

58. À titre d'exemple, consulter United Kingdom Parliament (2020c).

Énumérons toute une gamme de progrès observés :

- L'ICO, l'Alan Turing Institute, le CDEI et l'Office for AI se sont mis d'accord pour collaborer à la conception, au déploiement et au suivi d'une formation sur des enjeux d'intérêt pour ceux et celles qui surveillent l'IA.
- L'Office of AI élabore actuellement une stratégie nationale en matière d'IA. Il s'est montré actif en proposant un guide d'utilisation de l'IA ainsi que des balises relatives à l'approvisionnement.
- Le gouvernement a publié un document pour encadrer la prise de décisions algorithmique dans le secteur public.
- Le CDEI a fait ses preuves en publiant plusieurs rapports au sujet, par exemple, des partis pris observés dans la prise de décisions algorithmique (United Kingdom Government, 2020f) ainsi que des techniques de ciblage en ligne utilisées pour la publicité ainsi que la personnalisation ou la diffusion de contenus (United Kingdom Government, 2020h). Il s'est aussi intéressé à d'autres enjeux comme l'hypertrucage et la désinformation par des moyens audiovisuels, les effets de l'IA sur les assurances personnelles, les haut-parleurs et les assistants vocaux intelligents. Il a également publié l'*AI Barometer*, soit une importante analyse examinant les possibilités de l'IA et les risques à envisager de manière urgente, les défis de la gouvernance, et l'utilisation des données au R.-U. (United Kingdom Government, 2020b).
- Une importante tâche a été confiée à l'Open Data Institute en ce qui concerne les fiducies et les organismes gérant des données.
- L'Alan Turing Institute a nettement favorisé la collaboration dans l'écosystème de l'IA à l'échelle nationale et internationale, notamment en faisant participer 400 personnes au projet ExplAIn, mené avec l'ICO, et en élaborant des politiques avec l'OCDE et le Conseil de l'Europe.
- Après des débuts difficiles en pleine pandémie de COVID-19, l'AI Council a conçu son importante feuille de route sur l'IA.
- Certains organismes de surveillance tels que l'autorité responsable de la bourse (la FCA) et le commissariat à l'information (l'ICO) ont mené des expériences en matière de réglementation.

À d'autres égards, l'influence des comités parlementaires s'est avérée moindre :

- Nous devons formellement adopter un ensemble de principes à l'échelle nationale et instaurer, dans le secteur public, des mécanismes clairs d'évaluation des risques et de la conformité. Il faut également des normes éthiques concrètes pour que les responsables au sein des entreprises acceptent que soient évalués la conception et l'usage de systèmes d'IA, ce qui aiderait à déterminer s'il est nécessaire d'imposer des règles impératives plutôt que des directives souples n'ayant pas caractère obligatoire, et dans quels cas les imposer. Par la suite, nous devons élaborer des outils en matière d'audits, d'études d'impact, de certification et de surveillance continue.
- Le CDEI étant maintenant un organisme clé dans l'écosystème de l'IA, il faut lui attribuer une fonction statutaire et définir précisément son rôle.
- Nous devons faire progresser plus rapidement le dossier des fiducies de données et des autres structures de partage des données. L'Open Data Institute a fait du bon travail, mais il n'y a toujours pas de cadre législatif clair à cet égard. Il faut en faire beaucoup plus et créer des outils fiables pour l'échange des données publiques, par exemple celles du système de santé, en nous appuyant sur ce qui a été fait à l'international quand c'est pertinent de le faire.
- Nous devons également faire progresser la littératie en ligne et la littératie numérique, qui permettent elles-mêmes de susciter la confiance du public. La proposition selon laquelle Ofcom se chargera de le faire en vertu de la nouvelle loi sur les préjudices en ligne est insuffisante.

- Les organismes publics ont eu une influence limitée sur le déploiement généralisé de technologies de reconnaissance faciale en direct.
- Nous considérons comme des éléments majeurs l'évaluation des répercussions de l'IA sur l'emploi, l'évaluation des compétences qui seront requises à l'avenir, la diversité de la main-d'œuvre nécessaire, et l'ampleur du perfectionnement professionnel qu'entraîne l'automatisation. Le gouvernement du R.-U. doit reconnaître que l'IA aura incessamment des effets perturbateurs sur l'emploi et qu'il importe donc, dans une certaine mesure, de parfaire les compétences numériques et de revoir les cibles de perfectionnement. Son action ne concorde pas – quant à sa cadence, son ampleur et sa portée – avec les besoins de perfectionnement de nombreux travailleurs et travailleuses au pays. Il faut agir davantage pour accroître la diversité et l'inclusion au sein de la main-d'œuvre du secteur des technologies.

CONCLUSION : LE GOUVERNEMENT À LA CROISÉE DES CHEMINS SUR LE PLAN RÉGLEMENTAIRE

Tel que le font ressortir les propos qui précèdent, les instances parlementaires et gouvernementales se sont largement entendues quant à la voie à suivre pour élaborer une vaste stratégie nationale en matière d'IA, même si les comités parlementaires se sont parfois montrés impatients d'augmenter la cadence et d'élargir la portée des travaux.

Pour concevoir une telle stratégie, il s'est avéré crucial de coordonner les efforts d'organisations clés, notamment de l'Office for AI, de l'AI Council, de l'ICO, du CDEI et de l'Alan Turing Institute, et il demeure crucial de le faire pour parvenir à des stratégies nationales ou internationales. L'IA s'avère complexe et suscite des réactions. La pandémie de COVID-19 a entraîné un recours accru à la technologie, ce qui a mis l'accent sur les possibilités et les risques associés à l'usage de l'IA et, plus particulièrement, des données. Il est d'ailleurs plus évident que jamais que nous devons stimuler la confiance du public pour lui faire adopter l'IA.

Le gouvernement britannique convient manifestement de ce qui précède. Un doute subsiste toutefois à savoir s'il convient également que, pour faire de l'IA éthique une réalité, il faut évaluer les risques concrets de l'IA, particulièrement ses effets sur les droits civils et sociaux, puis, en fonction des résultats obtenus, établir des normes ou imposer des règlements visant la conception, la mise au point et le déploiement éthiques des systèmes d'IA.

Nous devons mieux définir à partir de quel moment il s'avère approprié d'imposer une réglementation ou de réduire le pouvoir des entreprises. C'est en 2021 que la communauté internationale de l'IA a commencé à décider comment elle comptait le faire concrètement, grâce au recours accru, par des organes internationaux tels que l'Union européenne ou le Comité ad hoc sur l'intelligence artificielle du Conseil de l'Europe (CAHAI, de son nom anglais), à une approche « horizontale » et multisectorielle fondée sur le risque en ce qui a trait à la réglementation touchant l'IA et à la gouvernance de celle-ci.

Des initiatives fondamentales ont d'ailleurs déjà été entreprises à cet égard. En avril 2021, l'Union européenne a soumis une proposition de législation sur l'intelligence artificielle (Commission européenne, 2021). Le CAHAI avait quant à lui élaboré une étude de faisabilité, diffusée en décembre 2020, qui envisageait les options pour l'établissement d'un cadre juridique fondé sur des normes du Conseil de l'Europe en ce qui a trait aux droits humains, à la démocratie et à l'État de droit (Conseil de l'Europe, 2020).

Nous avons maintenant franchi une première étape pour ce qui est de déterminer à quels égards nous pouvons et devons nous fier à des codes d'éthique et à quels égards nous devrions plutôt préconiser une gouvernance éthique ou même y aller à fond en imposant une réglementation. Un tel dilemme entre des directives souples et des règles impératives est loin d'être tranché. Il est toutefois évident qu'une mise en commun de l'expertise à l'échelle internationale portera ses fruits. Le R.-U. se trouve donc à la croisée des chemins. Il a hébergé une rencontre des ministres des pays du G7 chargés du numérique et des technologies en avril 2021, puis accueilli un forum sur les technologies de l'avenir (le premier Future Tech Forum) en novembre. Nous devons néanmoins aller au-delà des principes et mettre en œuvre des normes de gouvernance de l'IA et des solutions concrètes. Il semble que les objectifs de concevoir une IA digne de confiance et de faire du R.-U. un leader de l'adoption d'une IA éthique se soient estompés.

À mon sens, pour atténuer les risques de l'IA et susciter la confiance du public, que ce soit dans le secteur public ou le secteur privé, le principe fondamental devrait consister à mettre l'IA à notre service, et non l'inverse. Reste à savoir si les responsables des politiques et de la surveillance adhèrent à ce principe et s'ils respectent l'obligation de prévoir une gamme de solutions réglementaires et de politiques selon le degré de risque que pose l'IA.

Comment le R.-U. devrait-il donc procéder ? Lui faudrait-il adopter une approche « horizontale » semblable à celle qui a guidé l'Union européenne et le Conseil de l'Europe ? Devrait-il plutôt réglementer l'IA à la pièce à mesure que des enjeux se présentent ? La voie à suivre pourrait bien d'abord passer par l'imposition d'études d'impact générales en vue de calculer les risques que pose le recours à des systèmes d'IA en particulier, puis par l'obligation d'auditer les systèmes à risque et d'en assurer la surveillance.

En juillet 2020, l'ICO a publié des lignes directrices afin d'aider les organisations à atténuer les risques associés à l'IA relativement à la protection des données (United Kingdom Government, 2020[e]). Ces orientations proposent un cadre quant à l'audit des systèmes d'IA. À partir d'une approche équilibrée fondée sur le risque, elles présentent une méthodologie d'audit contenant :

1. des outils et des procédures servant à mener des audits et des enquêtes ;
2. des conseils détaillés en matière d'IA et de protection des données ;
3. une panoplie de conseils pratiques pour aider les organisations à auditer la conformité de leurs propres systèmes d'IA.

Après la publication de ces lignes directrices, l'ICO a lancé une version préliminaire de sa « boîte à outils » sur les risques de l'IA et la protection des données (United Kingdom Government, 2021b), qui vise à aider les organisations ayant recours à l'IA à cerner les enjeux liés à la protection des renseignements personnels et présente des pratiques exemplaires en vue d'atténuer les risques. Ainsi, nous avons jeté les assises d'un régime réglementaire éthique et fondé sur le risque, ce qui est également intéressant pour les développeurs et développeuses ainsi que les investisseurs et investisseuses.

Au début de l'année 2022, le gouvernement du R.-U. a promis de publier un livre blanc sur la gouvernance en matière d'IA afin d'exposer ses projets de règlements. Les assises appropriées sont certainement en place. Or, il reste à mesurer l'influence qu'aura le Parlement au moment de déterminer ce qui y prendra appui.

Janvier 2022

RÉFÉRENCES

Axente, M. L. 2021. *How do we ensure the responsible use of AI by governments ?* Digital Tech ITP, 7 avril. https://digitaltechitp.nz/2021/04/07/how-do-we-ensure-the-responsible-use-of-ai-by-governments/

Big Brother Watch. 2021. *Poverty Panopticon – The hidden algorithms shaping Britain's welfare state*. London. https://bigbrotherwatch.org.uk/wp-content/uploads/2021/07/Poverty-Panopticon.pdf

Commission européenne. 2021. *Proposition de règlement du parlement européen et du conseil établissant des règles harmonisées concernant l'intelligence artificielle (législation sur l'intelligence artificielle) et modifiant certains actes législatifs de l'Union.* Bruxelles, la Commission. https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0020.02/DOC_1&format=PDF

Conseil de l'Europe. 2020. *Étude de faisabilité.* Strasbourg, Comité ad hoc sur l'intelligence artificielle (CAHAI). https://rm.coe.int/cahai-2020-23-final-etude-de-faisabilite-fr-2787-2531-2514-v-1/1680a1160f

Hall, W. et Pesenti, J. 2017. *Growing the artificial intelligence industry in the UK.* London, Department for Digital, Culture, Media & Sport et Department for Business, Energy & Industrial Strategy. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/652097/Growing_the_artificial_intelligence_industry_in_the_UK.pdf

May, T. 2018. Allocution principale. 25 janvier, Davos, Forum économique mondial. https://www.weforum.org/agenda/2018/01/theresa-may-davos-address/

OCDE. 2019. *Recommandation du Conseil sur l'intelligence artificielle.* https://legalinstruments.oecd.org/fr/instruments/OECD-LEGAL-0449

Taylor, C. 2017. *Lord Clement-Jones : On regulation and ethical and moral dilemmas in artificial intelligence.* LinkedIn. 17 mai. https://www.linkedin.com/pulse/lord-clement-jones-regulation-ethical-moral-dilemmas-claire-taylor/?trk=read_related_article-card_title

United Kingdom Government. 2017. *Industrial Strategy : Building a Britain fit for the future*. London, Department for Business, Energy and Industrial Strategy. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/730048/industrial-strategy-white-paper-web-ready-a4-version.pdf

———. 2018. *Government Response to the Lords Select Committee on Artificial Intelligence Report*. London, Department for Business, Energy and Industrial Strategy. https://www.gov.uk/government/publications/ai-in-the-uk-ready-willing-and-able-government-response-to-the-select-committee-report

———. 2019a. *AI Sector Deal.* London, Department for Business, Energy and Industrial Strategy. https://www.gov.uk/government/publications/artificial-intelligence-sector-deal/ai-sector-deal

———. 2019b. *Unlocking digital competition : Report of the Digital Competition Expert Panel*. London, Digital Competition Expert Panel. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/785547/unlocking_digital_competition_furman_review_web.pdf

———. 2020a. *A New Pro-competition Regime for Digital Markets : Advice of the Digital Markets Taskforce.* London, Digital Markets Taskforce. https://assets.publishing.service.gov.uk/media/5fce7567e90e07562f98286c/Digital_Taskforce_-_Advice.pdf

———. 2020b. *AI Barometer*. London, Centre for Data Ethics and Innovation. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/894170/CDEI_AI_Barometer.pdf

———. 2020c. *Artificial Intelligence and Public Standards*. London, Committee on Standards on Public Life. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/868284/Web_Version_AI_and_Public_Standards.PDF

———. 2020d. *Data ethics and AI guidance landscape*. Page Web. London, Department for Digital, Culture, Media and Sport. https://www.gov.uk/guidance/data-ethics-and-ai-guidance-landscape

———. 2020e. *Guidance on AI and Data Protection*. London, Information Commissioner's Office. https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-ai-and-data-protection/

———. 2020f. *Review into Bias in Algorithmic Decision-making*. London, Centre for Data Ethics and Innovation. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/957259/Review_into_bias_in_algorithmic_decision-making.pdf

———.2020g. *Online Harms White Paper*. London, Department for Digital, Culture, Media and Sport. https://www.gov.uk/government/consultations/online-harms-white-paper/online-harms-white-paper

———.2020h. *CDEI Review of Online Targeting*. London, Centre for Data Ethics and Innovation. https://www.gov.uk/government/publications/cdei-review-of-online-targeting

———. 2021a. *AI Roadmap*. London, AI Council. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/949539/AI_Council_AI_Roadmap.pdf

———. 2021b. *Blog : New toolkit launched to help organisations using AI to process personal data understand the associated risks and ways of complying with data protection law*. London, Information Commissioner's Office. https://ico.org.uk/about-the-ico/news-and-events/blog-new-toolkit-launched-to-help-organisations-using-ai/

———. 2021c. *Ethics, Transparency and Accountability Framework for Automated Decision-Making*. London, Office for Artificial Intelligence. https://www.gov.uk/government/publications/ethics-transparency-and-accountability-framework-for-automated-decision-making/ethics-transparency-and-accountability-framework-for-automated-decision-making

———. 2021d. *Government Response to the Committee on Standards in Public Life's 2020 Report AI and Public Standards*. London, Department for Digital, Culture, Media and Sport. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/987905/Government_Response_to_the_Committee_on_Standards_in_Public_Life_s_2020_Report_AI_and_Public_Standards__Accessible_version_.pdf

———. 2021e. *Government Response to the House of Lords Select Committee on Artificial Intelligence*. London, Department for Digital, Culture, Media and Sport. https://www.gov.uk/government/publications/government-response-to-the-house-of-lords-select-committee-on-artificial-intelligence/government-response-to-the-house-of-lords-select-committee-on-arti-cial-intelligence

United Kingdom Parliament. 2018a. *AI in the UK : Ready, Willing and Able ?* London, House of Lords Artificial Intelligence Committee. https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10002.htm

———. 2018b. *Algorithms in Decision-making*. London, House of Commons Science and Technology Committee. https://publications.parliament.uk/pa/cm201719/cmselect/cmsctech/351/35102.htm

———. 2019. *The work of the Biometrics Commissioner and the Forensic Science Regulator*. Page Web. London, House of Commons Science and Technology Committee. https://publications.parliament.uk/pa/cm201719/cmselect/cmsctech/1970/197001.htm

———. 2020a. *AI in the UK : No Room for Complacency. Seventh Report of Session*. London, House of Lords Liaison Committee. https://publications.parliament.uk/pa/ld5801/ldselect/ldliaison/196/196.pdf (———. 2020b. A*utomated Facial Recognition Technology (Moratorium and Review)* (HL Bill 87). London. https://publications.parliament.uk/pa/bills/lbill/58-01/087/5801087_en_1.html

———. 2020c. *Facial recognition surveillance*. London, House of Lords Hansard. https://hansard.parliament.uk/Lords/2020-01-27/debates/E1922D0C-2EA8-4ED1-89F3-B7EE2475EDED/FacialRecognitionSurveillance#contribution-A8290171-B9DD-45A6-B2F4-4F87BD55D783

———. 2021a. *Call for evidence launched on new technologies in law enforcement*. London, Justice and Home Affairs Committee. Annonce. 22 juillet. https://committees.parliament.uk/committee/519/justice-and-home-affairs-committee/news/156778/call-for-evidence-launched-on-new-technologies-in-law-enforcement/

———. 2021b. *Forensics and biometrics : Follow-up*. Lettre. 20 juillet. London, House of Commons. https://committees.parliament.uk/publications/6876/documents/72517/default/

———. 2021c. *Government Response to the Work of the Biometrics Commissioner and the Forensic Science Regulator*. London, Department for Digital, Culture, Media and Sport. https://publications.parliament.uk/pa/cm5801/cmselect/cmsctech/1319/131902.htm

INTELLIGENCE ARTIFICIELLE ET DROITS DES PEUPLES AUTOCHTONES

VALMAINE TOKI

Professeure, Ngatiwai Nga Puhi, Te Piringa – Faculté de droit, Université de Waikato, Nouvelle-Zélande.

ANDELKA M. PHILLIPS

Chargée d'enseignement en droit, sciences et technologies, École de droit, Université du Queensland, Australie, et adjointe de recherche, HeLEX Centre, Université d'Oxford, Royaume-Uni.

- ODD 3 - Bonne santé et bien-être
- ODD 7 - Énergie propre et d'un coût abordable
- ODD 9 - Industrie, innovation et infrastructure
- ODD 10 - Inégalités réduites
- ODD 11 - Villes et communautés durables
- ODD 12 - Consommation et production responsables
- ODD 13 - Mesures relatives à la lutte contre les changements climatiques
- ODD 15 - Vie terrestre
- ODD 16 - Paix, justice et institutions efficaces
- ODD 17 - Partenariats pour la réalisation des objectifs

INTELLIGENCE ARTIFICIELLE ET DROITS DES PEUPLES AUTOCHTONES

RÉSUMÉ

Considérant que l'IA sera de plus en plus utilisée dans tous les secteurs et deviendra omniprésente dans la société, il est urgent de discuter de la manière dont elle peut être exploitée au profit de différents groupes sociaux. En particulier, ceux qui font face à des réalités différentes et vivent selon des visions du monde différentes qui pourraient être très utiles pour le développement et l'application de l'IA. Ce chapitre explore le lien entre les droits des peuples autochtones et l'intelligence artificielle (IA) en se penchant tant sur le caractère procédural de ces droits que sur leur essence. Pour ce faire, il présente un survol de l'IA et de la manière d'envisager ce concept. En plus, le chapitre offre une analyse de comment l'IA prend actuellement en considération les droits des peuples autochtones, ce qui sert de contexte pour aborder différentes visions autochtones du monde. A partir de la perspective des auteures, établies en Nouvelle-Zélande, l'exemple du peuple autochtone Māori est soulevé pour montrer la voie à suivre en ce qui a trait aux procédures et à la substance des droits. Par la suite, la question autour de comment l'IA peut expressément reconnaître les droits des peuples autochtones et les intégrer est analysée davantage au moyen d'une étude de cas relative à la mise en place d'un microréseau électrique sur l'île d'Aotea, aussi appelée l'île de la Grande Barrière. L'approche innovatrice de permettre à une collectivité de gérer la distribution électrique peut favoriser à la fois la protection des droits des peuples autochtones, le respect de la vie privée, ainsi que l'autodétermination et la souveraineté sur les données. À l'ère où les données sont d'ailleurs vues comme « le nouveau pétrole », des questions capitales se posent relativement aux effets sur les Autochtones du déploiement d'un large éventail de nouvelles technologies.

Ce chapitre part de la prémisse que les technologies sont loin d'être neutres et continuent de comporter des risques et des bénéfices pour les communautés. Le but poursuivi c'est de mettre en lumière certains de ces enjeux, et d'encourager la poursuite des discussions et la recherche en la matière.

INTRODUCTION

Ce chapitre vise à explorer le lien entre les droits des peuples autochtones et l'intelligence artificielle (IA) en se penchant tant sur le caractère procédural de ces droits que sur leur essence. Pour ce faire, nous commençons par présenter un survol de l'IA et de la manière d'envisager ce concept. Nous expliquons ensuite comment l'IA prend actuellement en considération les droits des peuples autochtones, ce qui sert de contexte pour aborder différentes visions autochtones du monde. Comme nous sommes établies en Nouvelle-Zélande, nous prenons l'exemple des Māori, peuple autochtone d'ici, pour montrer la voie à suivre en ce qui a trait aux procédures et à la substance des droits. Nous examinons par la suite comment l'IA peut expressément reconnaître les droits des peuples autochtones et les intégrer, au moyen d'une étude de cas relative à la mise en place d'un microréseau électrique sur l'île d'Aotea, aussi appelée l'île de la Grande Barrière. Nous avons choisi cet exemple, car il montre comment le fait de permettre à une collectivité de gérer la distribution électrique peut favoriser à la fois la protection des droits des peuples autochtones, le respect de la vie privée, ainsi que l'autodétermination et la souveraineté sur les données. À l'ère où les données sont d'ailleurs vues comme « le nouveau pétrole », des questions capitales se posent relativement aux effets sur les Autochtones du déploiement d'un large éventail de nouvelles technologies.

Il est d'abord particulièrement important de souligner que, au regard des injustices commises tout au long de l'histoire – par exemple, l'exploitation des Autochtones et d'autres groupes marginalisés aux fins de la recherche scientifique –, les technologies sont loin d'être neutres et continuent de comporter des risques et des bénéfices pour les communautés. Notre but est de mettre en lumière certains de ces enjeux, et d'encourager la poursuite des discussions et la recherche en la matière. Nous avons rédigé ce chapitre en 2021, soit en temps de COVID-19, alors que des régions de notre pays étaient de nouveau confinées et que le traçage des cas accentuait davantage l'érosion du droit à la confidentialité pour les collectivités et les individus.

Alors que le Règlement général sur la protection des données (RGPD) (Union européenne, 2016) exerce une influence internationale qui a mené à la mise à jour par l'Aotearoa/Nouvelle-Zélande en 2020 de sa législation en matière de respect de la vie privée – notamment par la promulgation d'une loi sur la protection des renseignements personnels (New Zealand Government, 2020) –, l'épidémie a entraîné de nouveaux défis. Plus précisément, la nécessité du traçage menace le respect de la vie privée et la protection des renseignements personnels, et pourrait même aller jusqu'à exacerber des inégalités qui avaient déjà cours. Nous ne remettons pas en question l'importance du traçage des cas, mais constatons que le règne des téléphones intelligents a plus généralement contribué à accroître la surveillance de la population. Dans un effort visant à freiner la propagation de la COVID-19, le besoin d'accéder aux données concernant les déplacements d'une personne, son statut vaccinal ou l'information relative à ses tests de dépistage multiplie le nombre d'entités ayant accès à ces renseignements sanitaires et à d'autres données de nature délicate. La Nouvelle-Zélande, par exemple, a récemment mis en place des mesures qui imposent aux gens de montrer un passeport vaccinal pour accéder à certains services, notamment les salons de coiffure et les restaurants. Nous devons garder à l'esprit que des risques accrus de cybermenace pèsent sur toutes les organisations, ce qui comprend les risques d'un piratage des bases de données médicales comme l'attaque associée au rançongiciel WannaCry (Landi, 2019). Autre exemple plus récent : en Nouvelle-Zélande, une telle attaque logicielle ayant ciblé le Waikato District Health Board a touché plus de 4 000 personnes (New Zealand Herald, 2021 ; Keall, 2021). Comme Phillips l'a déjà fait remarquer, il importe désormais d'aborder la technologie d'une manière holistique et inclusive, et un faible encadrement ne mènera probablement pas à « un monde plus sûr et plus juste » que le monde actuel (Phillips et Mian, 2019). Pour y parvenir, il faudrait davantage de débats publics et de mobilisation quant aux enjeux de l'évolution, de l'adoption et de la surveillance des technologies. Le large potentiel des technologies de l'IA rend ce besoin d'autant plus criant.

CONTEXTE

L'IA recouvre un large éventail de technologies, qu'elles soient déjà omniprésentes sur le marché, encore en évolution ou à l'état embryonnaire. De manière générale, il existe une distinction entre le concept d'IA générale, où la machine a une intelligence semblable à celle des humains (en anglais *human-level machine intelligence* [HLMI]) et l'IA restreinte (Bostrom, 2014, pp. 1-21 ; Fjelland, 2020 ; Russell, 2021). Un bon nombre des technologies actuellement utilisées à la maison ou au travail constituent des exemples d'IA restreinte. On n'a qu'à penser aux filtres antipourriels des boîtes de messagerie ou aux dispositifs de reconnaissance vocale, ou encore à certaines machines intelligentes comme l'AlphaGo, conçu par Google pour le jeu de go (House of Commons Science and Technology Committee, 2016, p. 5). Ce dernier exemple montre bien en quoi consiste l'IA restreinte, c'est-à-dire une IA appliquée à une tâche précise, voire à une série de tâches, mais qui ne saurait exceller dans d'autres domaines que celui pour lequel elle a été programmée. Il existe d'ailleurs une foule d'applications conçues pour s'amuser, par exemple pour jouer aux échecs.

La spéculation est forte en ce qui a trait à la mise au point d'une machine intelligente ayant un potentiel semblable à celui de l'être humain (donc de type HLMI), ce qui impliquerait qu'un agent intelligent utilise des capacités de raisonnement dans une multitude de situations, à l'instar de ce que ferait une personne (Bostrom, 2014, pp. 3-5 ; Russell, 2021, p. 514). De nombreux investissements sont actuellement consacrés à la recherche en la matière, et une éventuelle « explosion d'intelligence » – aussi appelée « singularité technologique » – fait d'ailleurs l'objet de nombreuses discussions. Le moment où aura lieu cette éventuelle explosion et sa survenue même demeurent incertains (Fjelland, 2020 ; Eliot, 2020, chap. 4). À ce jour, plusieurs avancées ont touché l'IA restreinte (Russell, 2021, p. 514). La mise au point de véhicules autonomes contribue à la recherche sur l'IA générale et, dans ce contexte, les agents intelligents pourraient avoir à composer avec des questions très complexes (Bradshaw-Martin, 2020). L'exemple des véhicules autonomes illustre bien certaines limites actuelles de l'IA (Technology Quarterly, 2020).

L'IA fait l'objet de nombreuses définitions, dont aucune n'a obtenu un consensus. Figure centrale du domaine et premier à employer l'expression « intelligence artificielle », John McCarthy l'a présentée comme « la science et l'ingénierie permettant la fabrication de machines intelligentes, particulièrement de programmes informatiques intelligents », avant de préciser qu'elle « s'apparente au fait d'utiliser des ordinateurs pour comprendre l'intelligence humaine » (McCarthy, 2007). Ce concept pourrait être mis en opposition avec celui de l'« intelligence naturelle » dont bénéficient les organismes vivants (Williams et Shipley, 2021, p. 45). L'idée d'une machine dont l'intelligence gagne en autonomie par rapport à celle des humains a un caractère anthropomorphique. L'autonomie des machines est d'ailleurs envisagée dans une pluralité de contextes allant de la mise au point de véhicules autonomes à la naissance de robots humanoïdes qui pourraient avoir une intelligence semblable à la nôtre (Calo, 2017).

De nombreux rapports internationaux ont également nourri la réflexion au sujet d'une définition élargie de l'IA (House of Commons Science and Technology Committee, 2016 ; Executive Office of the President National Science and Technology Council Committee on Technology, 2016 ; European Parliament Committee on Legal Affairs, 2016). Au Royaume-Uni, un comité parlementaire a fourni une définition intéressante, que nous reprenons pour assurer une bonne compréhension de la suite de ce chapitre :

> AI can be loosely thought of as : a set of statistical tools and algorithms that combine to form, in part, intelligent software that specializes in a single area or task. This type of software is an evolving assemblage of technologies that enable computers to simulate elements of human behaviour such

> as learning, reasoning and classification[59] (House of Commons Science and Technology Committee, 2016, pp. 5-6).

La définition fournie par Kaplan et Haenlein est également utile, puisqu'elle décrit l'IA comme la « capacité d'un système à interpréter des données externes de manière cohérente, à apprendre de ces données, et à utiliser cet apprentissage afin d'atteindre des objectifs et d'accomplir des tâches précises grâce à une faculté d'adaptation flexible » (Kaplan et Haenlein, 2019, p. 15). En revanche, l'apprentissage automatique est considéré comme une « combinaison d'algorithmes pouvant par eux-mêmes apprendre des concepts spécialisés, sans besoin d'avoir été programmés à cet égard au préalable » (House of Commons Science and Technology Committee, 2016, p. 6).

Par ailleurs, la Commission européenne a publié en avril 2021 une proposition de règlement visant l'IA (Commission européenne, 2021). S'il en venait à être adopté, ce règlement aurait vraisemblablement des répercussions au-delà de l'Union européenne, comme cela a été le cas du RGPD (Union européenne, 2016). Par conséquent, il s'avère également utile de faire référence à la définition de « système d'IA » contenue dans la proposition :

> « système d'intelligence artificielle » (système d'IA), un logiciel qui est développé au moyen d'une ou plusieurs des techniques et approches énumérées à l'annexe I et qui peut, pour un ensemble donné d'objectifs définis par l'homme, générer des résultats tels que des contenus, des prédictions, des recommandations ou des décisions influençant les environnements avec lesquels il interagit (Commission européenne, 2021, art. 3.1).

Cette définition ne s'applique pas à un large éventail de technologies basées sur l'IA auxquelles les peuples autochtones pourraient s'intéresser. Étant donné l'influence qu'a eue le RGPD sur la législation internationale en matière de protection des renseignements personnels, il importe que la proposition de règlement fasse l'objet d'un examen de la communauté internationale et que les peuples autochtones prennent part aux discussions. Des modifications sont attendues alors que la proposition a déjà été critiquée. Le réseau EDRi (European Digital Rights), notamment, a publié une déclaration en collaboration avec 119 groupes de la société civile (EDRi, 2021a et 2021b). Cette déclaration en appelle à une transparence accrue, à un ajustement par rapport au risque que posent les systèmes d'IA, ainsi qu'à l'interdiction de certains systèmes posant des risques inacceptables, y compris l'interdiction de tous les systèmes voués à la notation sociale, la reconnaissance des émotions, la catégorisation biométrique discriminatoire, la prédiction de la criminalité, le profilage en contexte de migration (EDRi, 2021a). Les systèmes d'IA comportant des risques inacceptables, particulièrement ceux qui opèrent une analytique prédictive, sont d'un grand intérêt en ce qui concerne les Autochtones ; nous y revenons plus loin.

L'IA a un potentiel d'influence partout dans la société, y compris en matière de justice pénale, de cybersécurité ou de diagnostics médicaux. Bien qu'elle soit souvent présentée comme un moyen de résoudre certains enjeux sociaux, des questions subsistent quant aux violations des droits humains associés à la protection des renseignements personnels et à la discrimination. Par exemple, l'utilisation de logiciels de prédiction de la criminalité, particulièrement ceux qui évaluent le risque de récidive, a montré des biais problématiques (d'Alessandro *et al.*, 2017 ; O'Neil, 2016, pp. 85-89), ce que Heaven ne manque pas de souligner relativement à un outil de prise de décisions sur la mise en liberté conditionnelle :

---

59. Il est possible d'envisager l'IA, grossièrement, comme un ensemble d'outils statistiques et d'algorithmes qui se combinent notamment pour former un logiciel intelligent spécialisé dans un domaine ou une tâche. Un tel logiciel résulte d'un assemblage de technologies en évolution qui permettent aux ordinateurs de simuler certains comportements humains tels que l'apprentissage, le raisonnement ou la classification.

> a tool called COMPAS, used in many jurisdictions to help make decisions about pretrial release and sentencing, issues a statistical score between 1 and 10 to quantify how likely a person is to be rearrested if released. The problem lies with the data the algorithms feed upon. For one thing, predictive algorithms are easily skewed by arrest rates. According to US Department of Justice figures, you are more than twice as likely to be arrested if you are Black than if you are white. A Black person is five times as likely to be stopped without just cause as a white person (Heaven, 2020).

Un autre exemple d'un tel outil est PredPol (aujourd'hui appelé Geolitica), qui s'appuie sur le traitement des données historiques liées à la criminalité dans le but de prédire dans quels lieux des actes criminels pourraient être commis (O'Neil, 2016, p. 85 ; Geolitica, 2021).

L'IA est souvent perçue comme un outil d'analyse des données qui refléterait les valeurs de la personne qui en fait la programmation. C'est le « bon sens » ou le discernement derrière la programmation qui fait la différence. Il importe néanmoins de se rappeler que cette tâche ne relève habituellement pas d'une seule personne, mais d'une équipe de conception. Essayer de mettre au point une IA qui reflète une large gamme de valeurs humaines représente un défi, mais le fait de travailler en équipe favorise l'adoption d'une approche inclusive. Bien sûr, les peuples autochtones ne sont pas homogènes, et ce ne sont pas toutes les communautés ni toutes les entreprises qui comprennent en leur sein des personnes ayant les compétences appropriées en vue de concevoir des systèmes d'IA intégrant les valeurs autochtones. Plusieurs approches pourraient s'avérer nécessaires, et il faut espérer que de nouveaux modèles et outils favoriseront une telle inclusion. Ainsi, on pourrait élaborer des modèles qui permettraient aux communautés et aux entreprises travaillant avec des communautés autochtones de concevoir des systèmes d'IA reconnaissant les valeurs d'un peuple autochtone en particulier dès le début d'un projet, comme on le fait pour la confidentialité programmée (Cavoukian, 2011) et la sécurisation dès la conception (Lovejoy, 2020). Nous espérons que ce chapitre stimulera les discussions à venir relativement à ces enjeux.

Le fait est reconnu : plus on cède d'autonomie à l'IA dans la prise de décisions, plus forte est la menace d'une inversion du pouvoir et, par conséquent, d'une érosion des droits humains et des valeurs (Liu et Zawieska, 2017). Afin de freiner l'érosion et d'assurer une meilleure reconnaissance des droits fondamentaux, il est nécessaire de se demander si le concept d'« alignement des valeurs » (Kim et Mejia, 2019) – qui vise à garantir l'intégration à l'IA de valeurs importantes – est une solution satisfaisante pour veiller à ce que l'IA soit programmée selon des valeurs de respect des droits humains et, si c'est possible, des droits des peuples autochtones (Maitra, 2020, p. 321).

Même si chaque peuple autochtone est unique, les différents peuples partagent une vision du monde semblable à certains égards. De manière générale, ils entretiennent quant à leurs droits une perspective holistique et relationnelle qui s'appuie sur la cosmologie et la relation avec la nature. Toutefois, l'IA n'adopte habituellement pas une perspective empreinte des droits des peuples autochtones. Ainsi, l'enjeu général pressant pour ces derniers ne se résume pas seulement à l'éventuelle érosion de leurs droits ou aux données en cause dans un modèle d'IA typique. La question est aussi de savoir de quelle manière procéder pour qu'une vision autochtone du monde soit intégrée à une IA modèle. Pour y réfléchir, il est nécessaire de se pencher d'abord sur les mesures de protection en place puis, dans un second temps, de se demander, par exemple, comment protéger la propriété intellectuelle relative à l'intégration de cette vision autochtone.

Il a été relevé que l'IA « a évolué dans une chambre d'échos épistémiques et [que] les partis pris dont elle est empreinte sont caractéristiques d'une suprématie blanche, laquelle ressort d'un ensemble de systèmes imbriqués et superposés » (Lewis, 2020). Cette affirmation vient renforcer la question suivante : Comment, si cela est possible, le droit à l'autodétermination des peuples autochtones peut-il être compris et reconnu en contexte d'IA ?

PERSPECTIVE ACTUELLE

La conception et la mise en œuvre de l'IA touche les Autochtones à la fois en tant que membres de collectivités et à titre de sujets qui participent à des projets d'IA (Walker et Hamilton, 2018). Sans grande surprise, ils se demandent si l'IA est une nouvelle (r)évolution ou s'il s'agit d'une nouvelle forme de colonisation (Whaanga, 2020, p. 35). Il est en l'occurrence possible d'établir certains parallèles avec le déploiement de biotechnologies qui utilisent les ressources des peuples autochtones, lequel a été taxé de biocolonialisme (Whitt, 1998 ; Indigenous Peoples Council on Biocolonialism, 2006). Au-delà de l'alignement des valeurs, on tient peu compte actuellement de la façon d'intégrer les perspectives autochtones et de protéger les données qui concernent les Autochtones dans le domaine de l'IA. Il convient de souligner qu'il n'est pas rare que ces données soient examinées et traitées sans une reconnaissance adéquate des droits des peuples autochtones (Maitra, 2020, p. 323).

Les droits des peuples autochtones sont entiers et indivisibles. La Déclaration des Nations Unies sur les droits des peuples autochtones fournit un cadre pour comprendre ces droits fondamentaux, qui comprennent par exemple, les droits à la culture, à l'éducation, au territoire et aux ressources, ainsi qu'au savoir traditionnel (Nations Unies, 2007). La clé de voûte de la Déclaration est le principe d'autodétermination, dont découlent tous les autres droits. L'application de ces droits repose sur une vision autochtone du monde.

L'IA est toutefois loin de refléter une telle conception. Au contraire, elle reproduit plutôt les valeurs et les idéaux de la perspective scientifique occidentale (Williams et Shipley, 2021 ; 2019), et elle ne peut déterminer des normes par elle-même. L'IA n'est pas dotée d'une conscience, elle ne ressent ni joie, ni culpabilité, ni remords, et elle est loin de se soucier des conséquences générales de ses actions ou des personnes qui les subissent (Williams et Shipley, 2021, p. 44).

Si aucune norme ne guide le fonctionnement de l'IA, celle-ci reflète alors les valeurs des gens qui la programment, et ces derniers sont formés et travaillent dans des environnements où prédomine une vision occidentale du monde. Une ontologie réductionniste des données et une épistémologie construite pour les algorithmes visent à maximiser l'efficacité dans l'exécution des tâches de l'IA, et non à se pencher sur le caractère moral de ces tâches (Williams et Shipley 2021, p. 44). Cela dit, l'IA ne peut que suivre des règles qui s'expriment et s'évaluent de manière quantitative (Williams et Shipley, 2021, p. 44).

Étant donné que l'alignement des valeurs présente des lacunes sur les plans tant technologique que philosophique, il convient de se demander si ce concept représente une réelle mesure de protection contre les violations des droits humains et s'il protège adéquatement les droits des peuples autochtones, en particulier ceux qui ont trait à la propriété intellectuelle et aux données personnelles (Maitra, 2020, p. 321).

COMMENT L'IA POURRAIT-ELLE RECONNAÎTRE LES DROITS DES PEUPLES AUTOCHTONES ?

Cette partie examine des éléments nécessaires, en matière de procédure et de substance des droits, pour que les systèmes d'IA reconnaissent les droits des Autochtones. Il importe de reconnaître dès le départ que la programmation de l'IA a jusqu'à maintenant surtout été une affaire d'hommes partageant les valeurs de la communauté scientifique occidentale. Sans surprise, l'IA reflète donc leurs préjugés ainsi que leurs conceptions de l'éthique et du bon sens (Weidenbener, 2019). Par conséquent, il n'est pas garanti que l'intégration de valeurs morales à l'IA, en amont, protégera à elle seule de manière suffisante les droits humains ou ceux des peuples autochtones (Bostrom, 2014, pp. 185-207).

Afin que les droits des Autochtones soient adéquatement « transmis » aux systèmes d'IA, il importe d'en envisager le caractère procédural et la substance, pour éventuellement s'assurer que, à mesure que l'IA devient une entité autonome qui influera sur nos structures sociales ou notre identité en tant qu'être humain, les droits des peuples autochtones seront pris en considération. Pour ce faire, il faut reconnaître l'importance des données, personnelles et communautaires, et des droits qui y sont associés. Quand on recueille des données auprès de communautés autochtones, par exemple, il faut admettre ou reconnaître qu'il s'agit de « données autochtones » recueillies en contexte autochtone.

Droits en matière de procédure

Les droits des peuples autochtones en matière de procédure sont clairement énoncés à l'article 18 de la Déclaration des Nations Unies sur les droits des peuples autochtones (Nations Unies, 2007) :

> Les peuples autochtones ont le droit de **participer à la prise de décisions** sur des questions qui peuvent concerner leurs droits, par l'intermédiaire de représentants qu'ils ont eux-mêmes choisis conformément à leurs propres procédures, ainsi que le droit de conserver et de développer **leurs propres institutions décisionnelles** [caractères gras ajoutés].

La clé de voûte est le droit à l'autodétermination, formulé à l'article 3 de la Déclaration : « les peuples autochtones ont le droit à l'autodétermination. En vertu de ce droit, ils d´´terminent librement leur statut politique et assurent librement leur développement économique, social et culturel ».

La souveraineté des Māori sur leurs données comprend la défense de leurs droits inhérents et de leurs intérêts en matière de collecte, de propriété et d'utilisation de ces données, y compris dans un contexte d'IA (Kukutai et Taylor, 2016 ; Te Mana Raraunga Māori Data Sovereignty Network, s. d.). Toute initiative visant à incorporer des données relatives aux Autochtones au moyen d'un algorithme devrait tenir compte de ces droits non seulement au moment de la conception du programme en lui-même, mais aussi quand il est question de sa gestion ou de son contrôle. L'outil mondial Navigateur autochtone illustre concrètement cette manière de faire. Le Navigateur offre des outils permettant d'évaluer comment sont reconnus les droits des peuples autochtones. Les données qu'il recueille ne représentent pas des données statistiques officielles, mais elles cernent les perceptions et les expériences des Autochtones (IWGIA et ILO, 2021, p. 20). Un tel processus sous-tend le droit à un consentement préalable, donné librement et en connaissance de cause (IWGIA et ILO, 2021, p. 17). Il s'avérerait utile d'élaborer des normes techniques et des modèles en collaboration avec le Navigateur, car une telle mesure aiderait les équipes chargées de la conception et de la programmation à cerner les besoins. Observons par exemple le travail du PCI Security Standards Council, qui a proposé des lignes directrices et mis en œuvre un système de certification en vue d'améliorer dans l'ensemble la sécurité des données de paiement (PCI Security Standards Council, s. d.). Cet organisme a également élaboré une norme technique relative à la sécurité des données de paiement par carte. Nous estimons que l'élaboration de telles normes techniques et de systèmes de certification à l'échelle internationale serait utile à la mise au point de systèmes d'IA ou d'autres technologies dont les Autochtones se servent.

Substance des droits

La Déclaration des Nations Unies sur les droits des peuples autochtones (Nations Unies, 2007) énonce les droits fondamentaux suivants :

> Les peuples autochtones ont le droit de **maintenir et de renforcer leurs institutions** politiques, juridiques, économiques, sociales et **culturelles distinctes**, tout en conservant le droit, si tel est leur choix, de participer pleinement à la vie politique, économique, sociale et culturelle de l'État (art. 5).
>
> Les peuples autochtones ont le droit de revivifier, **d'utiliser, de développer et de transmettre aux générations futures leur histoire, leur langue, leurs traditions orales, leur philosophie, leur**

> **système d'écriture et leur littérature**, ainsi que de choisir et de conserver leurs propres noms pour les communautés, les lieux et les personnes (art. 13(1)).
>
> Les États se concertent et coopèrent de bonne foi avec les peuples autochtones intéressés – par l'intermédiaire de leurs propres institutions représentatives – avant d'adopter et d'appliquer des mesures législatives ou administratives susceptibles de concerner les peuples autochtones, afin d'obtenir **leur consentement préalable, donné librement et en connaissance de cause** (art. 19).
>
> Les peuples autochtones ont le droit de préserver, de contrôler, de protéger et de développer **leur patrimoine culturel, leur savoir traditionnel et leurs expressions culturelles traditionnelles** ainsi que les manifestations de leurs **sciences, techniques** et culture, y compris leurs ressources humaines et génétiques, leurs semences, leur pharmacopée, leur connaissance des propriétés de la faune et de la flore, leurs traditions orales, leur littérature, leur esthétique, leurs sports et leurs jeux traditionnels et leurs arts visuels et du spectacle. Ils ont également le droit de préserver, de contrôler, de protéger et de développer leur **propriété intellectuelle collective de ce patrimoine culturel, de ce savoir traditionnel et de ces expressions culturelles traditionnelles** (par. 31(1) [caractères gras ajoutés]).

Tout programme d'IA qui chercherait à exploiter ou à extraire des données relatives aux Autochtones par le biais d'un algorithme (ou autre) devrait donc reconnaître non seulement le droit à un consentement préalable, donné librement et en connaissance de cause, mais également les droits associés au savoir traditionnel, aux expressions culturelles traditionnelles, et aux manifestations des sciences et techniques. Recueillir des données autochtones, et donc un savoir traditionnel, sans avoir obtenu un tel consentement enfreint manifestement ces droits. C'est le cas, par exemple, quand une entreprise commerciale utilise le savoir traditionnel associé à certaines plantes médicinales sans qu'une communauté ait d'abord donné son consentement, de manière libre et en connaissance de cause. C'est aussi le cas quand une recherche médicale « se sert » d'Autochtones, notamment pour breveter des lignées cellulaires cultivées à partir d'échantillons sanguins prélevés sur une Guayamí du Panama ou un Hagahai (WIPO, 2006 ; Indigenous Peoples Council on Biocolonialism et Harry, 1995).

Les droits fondamentaux respectent une vision autochtone du monde. Reconnaître ces droits comme un ensemble guidé par le droit à l'autodétermination – et en comprendre tant le caractère procédural que la substance – fournit une solide base à toute initiative qui, en matière d'IA, s'intéresse aux données relatives aux Autochtones.

VISION AUTOCHTONE DU MONDE

Bien qu'ils proviennent de partout, les peuples autochtones partagent une vision du monde assez similaire, fondée sur la nature et la cosmologie. L'intégration de savoirs traditionnels aux systèmes d'IA serait souhaitable, par exemple, pour anticiper l'évolution des changements climatiques. Quoiqu'elle reste un défi dans le domaine de l'IA, l'intégration des perspectives autochtones permettrait l'avènement d'un « autre type d'IA » (Kesserwan, 2018), une IA qui refléterait et appuierait une relation éthique et fondée sur la réciprocité. Nous proposons trois brèves hypothèses sur la manière dont les visions du monde autochtones peuvent ajouter de la valeur au développement de l'IA. Pour cela, nous nous inspirons des exemples des peuples Navajo, Lakota et Hawaï. Nous explorons ensuite plus en détail l'exemple du cas du peuple Māori.

Navajo

Le peuple Diné (Navajo) reconnaît l'univers qui l'accueille et honore ses responsabilités envers celui-ci (Haskie, 2002, citant Griffin-Pierce, 1992). Cette vision du monde se trouve dans le concept d'*Hózhó*, une philosophie complexe alliant bien-être et croyances, et respectant certains préceptes qui guideraient les pensées, actions, comportements et discours (Kahn-John et Koithan, 2015).

Tout comme plusieurs peuples autochtones, les Navajo adhèrent à des principes tels que l'harmonie ou l'équilibre. Leurs croyances se centrent sur une interdépendance et une connexion des êtres animés et inanimés. Ils sont également conscients de l'importance du bonheur individuel et relèvent aussi le lien étroit entre bien-être physique, émotionnel, psychologique et spirituel (Haskie, 2002, p. 25, citant Cleary et Peacock, 1998). L'existence d'un cadre éthique régissant l'harmonie et les comportements moraux pourrait éventuellement s'appliquer à une structure d'IA qui, à titre d'exemple, aspirerait à rendre le système de justice pénale plus équitable qu'il ne l'est actuellement.

Lakota

De manière semblable, le peuple des Lakota estime que tout dans l'univers revêt une dimension intérieure (l'âme) et une dimension matérielle (le corps) (Posthumous, 2018). Un respect des êtres inanimés aussi bien que des êtres animés constitue ainsi l'essence de la vie chez les Lakota (Deloria,1998).

Partant de ces constatations, la question qui se pose est de savoir si une éventuelle structure d'IA lakota remplirait différents rôles antagonistes – de l'armement autonome à la surveillance de masse – tout en préservant l'ontologie relationnelle (Lewis *et al.*, 2019). Il a d'ailleurs été proposé de suspendre par intermittence l'évolution de l'IA pour établir une approche relationnelle (Lewis *et al.*, 2019).

De plus, afin de surmonter les différences de formes ou de caractéristiques propres à chaque système d'IA, il est suggéré que des éléments comme la vocation de l'IA, son codage ou même l'équipe qui la conçoit soient considérés comme un ensemble (Lewis *et al.*, 2019). Logiquement, ce serait alors aux responsables de la programmation de veiller à ce recentrage, ce qui présente en soi, une nouvelle série de défis.

Hawaii

Selon la vision du monde des Hawaïens et Hawaïennes (*kānaka maoli*), le concept fondamental du *pono* relève d'une « approche éthique… qui privilégie la multiplicité plutôt que la singularité » afin d'atteindre l'équilibre et l'harmonie (Lewis *et al.*, 2019). Selon cette vision, il n'est pas envisageable de faire fi du *pono* et d'accorder la préséance à l'individu sur la collectivité (Lewis *et al.*, 2019). Dans une relation, le bien-être de chaque membre compte, tandis que les intérêts personnels restent toujours au second plan (Lewis *et al.*, 2019).

L'IA est un outil créé par les humains et pour le progrès. Si la vision hawaïenne (*kānaka maoli*) devait s'appliquer au domaine de l'IA, ce serait, comme le pensent aussi les Lakota et les Navajo, pour retenir un cadre éthique et une forme de partage entre l'IA et les humains. Il faudrait alors redéfinir le concept d'autonomie ou à tout le moins l'appliquer de manière intermittente. Dans tous les cas, un compromis serait de mise, ce qui n'est pas non plus idéal.

S'il y avait combinaison de toutes les considérations précédentes et si les mécanismes d'alignement des valeurs étaient étendus – en fonction du principe directeur accordant une importance capitale aux relations –, certains problèmes s'en trouveraient améliorés. Mais sans une reconnaissance générale préalable du principe de l'autodétermination, il n'est pas certain que ces mesures s'avèrent très efficaces.

Il semble donc que les épistémologies autochtones se fondant sur un système de valeurs existant qui invite au respect mutuel entre humains et machines constituent plutôt l'approche à privilégier pour aller au-delà du simple alignement des valeurs.

Māori

Le point de vue māori est quant à lui centré sur le *tikanga*. Il renvoie à une philosophie tridimensionnelle complexe qui évoque les concepts « de l'intérieur ». *Tikanga* signifie « direct et sans ambiguïté », et fait intervenir les notions morales de la justice et de l'équité (Benton *et al.*, 2013, p. 429). Cette définition varie toutefois en fonction des circonstances et des personnes en cause (Toki, 2018). Le *tikanga* māori est un concept contextuel (New Zealand Law Commission, 2001). *C'*est un concept de plus en plus reconnu par les tribunaux néo-zélandais comme faisant partie intégrante, en matière de common law, du système judiciaire et de la « loi applicable » (*Trans-Tasman Resources Ltd v Taranaki-Whanganui Conservation Board* [2021] NZSC 127 at [169]).

Le *tikanga* māori trouve sa source dans le *Te Ao*, monde auquel appartient le peuple māori (Marsden, 1992, p. 117). La cosmologie et l'élaboration de récits sont intrinsèques au *Te Ao*, établissant des relations, ou *whakapapa*, entre l'animé et l'inanimé, les individus, leur environnement et le monde spirituel (Waitangi Tribunal, 2014, p. 20). L'interdépendance de ces éléments soutient un mécanisme semblable à celui d'une constitution sociale (Toki, 2018). Le principe de *whakapapa* est fondamental dans le *Te Ao* (Toki, 2018). Il s'agit d'un réseau complexe de réalités mettant tout en lien (Waitangi Tribunal, 2014, pp. 22-25). En tant que construction relationnelle, il explique la création de l'univers et la manière dont l'union et la complémentarité ou l'équilibre entre les êtres a entraîné une nouvelle forme de vie (Marsden, 2003). Le *whakapapa* a toujours été au centre de l'identité. L'individu fait partie de la collectivité et, en conséquence, se lie aux autres par le *whakapapa.* Le *whanaungatanga* est la « colle » qui cimente ces « parties » et est souvent décrit comme « l'état ou les circonstances faisant qu'on est un proche, par lien de parenté, de même que les droits, responsabilités, et comportements attendus au sein de ces relations » (Benton *et al.*, 2013, p. 524). En tant que composante du *tikanga*, le *whanaungatanga* « embrasse le *whakapapa* et se concentre sur les rapports » (Mead, 2003, p. 28). Il est indispensable pour les Māori, car le *whanau* (la famille) procure bien-être physique, émotionnel et spirituel. Tout comme les individus s'attendent à être soutenus par la collectivité, la collectivité s'attend à être soutenue par les individus ; il s'agit là d'une obligation et d'un principe fondamental (Mead, 2003, p. 28).

Le *tikanga* est donc une structure guidant des principes ou règles de base (Toki, 2018). Il est assorti d'autres concepts comme le *mana* et le *tapu*, qui encadrent aussi les rapports, ou le *whakapapa*, qui lie les personnes, l'environnement et le monde spirituel (Toki, 2018). Le *tikanga* māori a pour but l'atteinte d'un équilibre et d'une harmonie pour l'individu lui-même, pour sa communauté et même pour un ensemble plus large. Des régulateurs – le *tapu* et le *mana* – contribuent à remédier à tout déséquilibre dans un processus qui repose sur la réciprocité, l'*aroha* (l'amour) et le *manaaki* (l'attention).

L'*aroha* est un concept relatif au champ émotionnel qui s'apparente à une manière instinctive de réagir dans les relations qu'on entretient. Il s'agit d'une composante essentielle de la philosophie māori et elle prend la forme d'un processus de guérison (Benton *et al.*, 2013, p. 47). Pour le *kaumatua* et la *kuia*, le principe de l'*aroha* est à la base du don, du partage et du soutien entre le *whanau*.

Le processus d'équilibrage et réciprocité s'appelle *utu.* Enfin, l'*utu* est « l'exercice d'un droit de réparation » faisant que chaque chose « en implique une en retour : une satisfaction, une rançon, une récompense ou une réaction, comme une forme de paiement ou de réponse », et il est étroitement lié au *mana* (Benton *et al.*, 2013, p. 46). Souvent perçu comme un principe de réciprocité ou d'équivalence, l'*utu* a pour principal objectif de rétablir l'équilibre et l'harmonie tout en maintenant les rapports, donc le *whanaungatanga* (Mead, 2003, p. 31).

Il serait compliqué d'isoler un concept tel que le *mana* et de l'intégrer dans une structure d'IA sans l'accompagner des concepts tels le *tapu*, le *whakapapa* ou l'*utu*. Considéré à part, le *mana* perd son essence ; il risque d'être redéfini si on le prend isolément.Appliquer le principe relationnel et « interconnexionnel » d'une vision autochtone à l'IA pourrait lui apporter des particularités que les approches scientifiques occidentales ne sauraient lui offrir. Étant donné que la vision autochtone n'opère habituellement pas de distinction entre l'animé et l'inanimé, c'est ce système de valeurs s'appliquant aux relations qui devrait s'intégrer au domaine de l'IA et y former un cadre éthique. Grâce à ce savoir, nous pourrions concevoir des cadres relationnels en vue de protéger les droits et de favoriser l'autonomisation des peuples autochtones.

ÉTUDE DE CAS MĀORI

L'étude de cas qui suit s'intéresse à un projet de mise en œuvre pratique de l'IA visant à améliorer les conditions de vie d'une collectivité autochtone māori.

L'île d'Aotea, ou île de la Grande Barrière, est située à environ 100 km au nord-est d'Auckland. D'une étendue de 285 km^2, elle abrite plusieurs petites communautés māori. Il n'existe pas de réseau électrique centralisé sur l'île. Les habitants et habitantes, privés de réseau, utilisent leurs propres systèmes d'énergie solaire ou d'alimentation par piles. À ces sources d'énergie s'ajoutent des générateurs à essence ou au diesel, le gaz naturel et des feux de bois puisque, dans la quasi-totalité des cas, ils sont loin de couvrir l'entièreté des besoins énergétiques des ménages. Ceux-ci dépendent donc largement des générateurs de secours (Aotea Great Barrier Island Local Board Plan, 2020).

Le manque d'infrastructures sur l'île d'Aotea représente une occasion d'améliorer les conditions de vie difficiles des populations māori qui y vivent ainsi que de contribuer à la réduction des émissions de gaz à effet de serre et à l'essor de l'énergie propre en Nouvelle-Zélande.

Une solution envisageable serait d'établir un microréseau électrique intelligent entièrement alimenté par des énergies renouvelables. Cela, par le biais d'une approche fondée sur le *tikanga* et une démarche structurée et progressive qui lierait les systèmes d'énergie de toutes les petites communautés (Apperley, 2019). L'objectif ultime d'un tel réseau serait le partage énergétique entre les ménages et entre les communautés māori de l'île de la Grande Barrière. Ce réseau ne couvrirait que l'île, ce qui optimiserait la protection des droits de sa population en matière de respect de la vie privée. Il est déjà prévu que, même à un stade préliminaire du projet, la structure fractale du microréseau contribuerait à la fois à générer de l'énergie et à en partager dans la collectivité māori. Le *marae* (le lieu où ont traditionnellement lieu les rencontres) pourrait être le noyau du système, le centre de stockage à partir duquel répartir l'énergie.

Ce projet montre comment une approche holistique fondée sur des données relatives à la consommation énergétique et aux besoins des communautés et adoptant une approche relationnelle comme le *tikanga* pourrait résoudre des problèmes concrets liés à la consommation et à l'offre énergétiques. Rendre la collectivité responsable de ce microréseau constituerait une manière d'encourager l'autonomisation.

Un aspect important des solutions basées sur l'IA pour relever les défis communautaires est l'utilisation des données. Il importe de reconnaître que, dans le monde actuel, le droit au respect de la vie privée de tous et toutes – mais surtout des Autochtones et autres personnes marginalisées – reste passablement menacé.

La collecte des données mentionnée dans cette étude de cas suit une convention et emploie une technologie pouvant être associée à l'internet des objets (IdO). L'IdO « fait généralement référence à des scénarios où la connectivité de réseau et le potentiel de l'informatique s'appliquent à des objets, des capteurs ou d'autres appareils de la vie quotidienne qui ne sont normalement pas considérés comme des ordinateurs et leur permettent de générer, d'échanger et d'employer des données sans grande intervention humaine. Il n'existe toutefois pas de définition unique et universelle de l'IdO » (Internet Society, 2015, p. 5)Il faut souligner que les données recueillies par des compteurs électriques intelligents sont parfois des données personnelles sensibles. Elles peuvent servir plus largement à analyser des habitudes de vie :

> behaviors of residents including bathroom activities, cooking, housework, sleep cycles, and meal times can be inferred from seemingly non-sensitive smart meter readings. It has been shown that even the current TV channel and specific audiovisual content displayed on a television can be identified based on the corresponding household's electricity usage profile (Kröger, 2019, page 152).

Les données sur la consommation énergétique peuvent fournir de l'information sensible. Elles pourraient par exemple permettre d'établir l'appartenance religieuse, par une comparaison des données des ménages les jours de festivals religieux (Karwe et Müller, 2015, p. 228 ; Cleemput, 2018, p. 3 ; Reimann, 2019). Les compteurs intelligents permettraient également d'en apprendre sur la santé des personnes en examinant leur usage d'appareils médicaux (Pham et Månsson, 2019) ou encore de deviner leur statut d'emploi (Anderson *et al.*, 2017 ; Murrill, 2012 ; Greveler *et al.*, 2012). C'est pourquoi offrir aux populations locales la technologie dont elles ont besoin et le contrôle de celle-ci constituerait une manière de régler certains problèmes. La mise en œuvre de tels projets devra tenir compte d'enjeux liés à la sécurité des données.

Si une collectivité participe à la conception d'un tel système dès le départ, elle pourrait alors adopter une approche tenant compte de la confidentialité programmée et de la souveraineté sur les données (Data Sovereignty Now, 2020 ; Nagel et Lycklama, 2021). Selon l'approche préconisée par Nagel et Lycklama, les collectivités autochtones devraient bénéficier d'une pleine souveraineté sur leurs données. De plus, vu la quantité de renseignements que glanent les compteurs intelligents, il serait judicieux d'en désactiver certaines fonctionnalités afin de se conformer aux principes de confidentialité des renseignements en vigueur en Nouvelle-Zélande, selon lesquels on ne doit recueillir que les données nécessaires (il s'agit d'ailleurs du premier principe énoncé). De telles recommandations abondent également dans le sens d'une limitation quant à la quantité de données à caractère personnel recueillies, tel que le prévoit l'article 5 du RGPD (Union européenne, 2016). Étant donné les éventuelles menaces à la sécurité et à la protection des renseignements personnels, il serait judicieux de mettre en œuvre une approche de sécurisation dès la conception, de manière conforme au cinquième principe de confidentialité en vigueur en Nouvelle-Zélande, aux exigences prévues à l'article 32 du RGPD, et au principe d'intégrité et de confidentialité énoncé à l'article 5 du RGPD (Union européenne, 2016). Pour développer davantage les considérations concernant la protection des données et ses implications pour les communautés autochtones, les sous-sections suivantes analyseront les différences entre une approche conventionnelle de collecte de données dans le projet de réseau intelligent et une approche autochtone.

Approche conventionnelle

Le microréseau intelligent évoqué précédemment collige et partage des données. Il pourrait s'apparenter au système nerveux numérique d'un corps dont l'IA serait le cerveau. En utilisant ces deux composantes dans un réseau alimenté par l'énergie solaire, l'IA servira à gérer la consommation électrique non seulement dans un ménage, mais aussi au sein d'un regroupement de ménages raccordés au réseau. Un tel réseau intelligent vise les principaux objectifs suivants :

1. Une consommation électrique optimale au sein d'un regroupement de ménages, en fonction de la demande de chacun ainsi que des appareils qui s'y trouvent ;
2. L'optimisation de la consommation de divers appareils (afin de réduire la demande énergétique, en période de pointe, d'appareils gourmands : réfrigérateurs, cuisinières, machines à laver, etc.) ;
3. Le réacheminement de l'énergie solaire excédentaire vers un système de stockage, ce qui permettrait son utilisation même après le coucher du soleil ;
4. L'affichage, la notification et le contrôle à distance, au moyen d'un tableau de bord intégré à des dispositifs comme un téléphone intelligent, un ordinateur, une tablette électronique ou l'appareil intelligent lui-même.

L'IA conventionnelle analysera les données qui lui permettront d'accomplir les tâches requises et prédéfinies. Ces données, une fois compilées, amélioreront la prévision des besoins énergétiques et conduiront à une sorte de profil associé à un ménage en particulier ou à un regroupement de ménages. La mise en place d'un tel système devra néanmoins tenir compte des principes de la confidentialité programmée (Cavoukian, 2011) et de la souveraineté dès la conception (Data Sovereignty Now, 2020). Cette approche conventionnelle offre de nets avantages et apporte une certaine structure. Toutefois, étant donné que le projet de microréseau vise une collectivité māori, adopter le *tikanga* ou une approche autochtone est un critère particulier et même essentiel.

Approche autochtone

Le projet dont il a été question précédemment permettrait d'élargir la portée des données habituellement recueillies, comme les données socioéconomiques, afin qu'elles reflètent davantage le *Te Ao* et la vision du monde des Māori. Des valeurs qui ne sont normalement pas considérées comme importantes dans l'univers conventionnel de l'IA, par exemple la façon dont les *kuia* et *kaumatua* (les aînées et aînés) consomment de l'énergie, pourront être prises en considération afin de mieux répondre aux besoins de ces personnes. Les principes *tikanga* que sont le *whanaungatanga* et le *whakapapa*, qui mettent l'accent sur les rapports, renforcent cette approche. Les concepts associés de *manaaki* ou de l'*aroha* l'enrichiraient aussi.Cette approche n'est pas nouvelle ; elle a été utilisée dans le programme du Navigateur autochtone, mentionné précédemment. Ce programme procure aux communautés autochtones des outils pour évaluer de quelle manière leurs droits sont respectés.

Il serait avantageux de concevoir des outils complémentaires dont ces communautés pourraient se servir au moment de mettre en œuvre leurs propres technologies, ce qui favoriserait l'autodétermination et la souveraineté sur les données à l'échelle locale. Pareille mesure aiderait également les équipes chargées de la programmation à comprendre ce qu'elles doivent faire pour intégrer à l'IA les valeurs des communautés autochtones.

Le but ultime du *tikanga* māori en tant que vision du monde est l'harmonie. C'est aussi ce que vise le projet de réseau énergétique : trouver un équilibre faisant intervenir protection de l'environnement et bien-être des collectivités autochtones. La mise en œuvre concrète de l'optique *tikanga* dans le domaine de l'IA contribuera à l'atteinte de ce but.

Si l'IA intègre des valeurs autochtones – que sa programmation relève d'un ou une Autochtone ou encore d'une personne qui applique des lignes directrices adaptées aux perspectives autochtones –, alors l'IA sera plus conforme à une conception autochtone.

En somme, il nous faut repenser l'outil qu'est l'IA. Pour mieux traiter les données associées aux droits ou aux principes des peuples autochtones, il importe nécessairement de concevoir les applications de l'IA en fonction de ces principes. Ce travail nécessite des discussions, une collaboration de diverses parties prenantes, et l'élaboration de modèles qui sauront inspirer les personnes chargées de la mise au point et de la programmation de l'IA de même que les Autochtones.

CONCLUSION

Ce chapitre a cherché à engager un dialogue critique au sujet de la valeur qu'une perspective autochtone apporterait à l'IA, en mettant l'accent sur la reconnaissance d'une vision autochtone du monde et sur les droits qui en découlent, et en présentant d'autres schèmes relationnels.

Intégrer une vision autochtone à un cadre qui en est habituellement exempt, comme c'est le cas pour l'IA, revient à imbriquer deux pièces incompatibles. Pour réussir, il faut adapter chacune d'elles. Tenir compte d'une vision autochtone du monde présente l'avantage d'offrir un cadre éthique et relationnel répandu. Néanmoins, la seule reconnaissance des droits ne suffit pas à défaut de garanties en matière de protection des renseignements personnels, de sécurité ou de propriété intellectuelle des données, ou encore d'une approche qui se penche sur le caractère procédural du droit à l'autodétermination et à l'essence de ce droit.

Si l'IA était envisagée comme un « outil » programmé par des Autochtones et s'il y avait une protection adéquate des droits fondamentaux, cela signifierait qu'il est possible d'intégrer une vision autochtone dans l'univers de l'IA. Nous ne devons toutefois pas oublier que tout système automatisé susceptible de recueillir des données de nature délicate menace le respect de la vie privée. À l'heure actuelle, les systèmes automatisés sont encore paramétrés par des humains, et il est nécessaire de tenir des débats et de mobiliser le public au sujet de ces technologies. L'implantation du réseau décrit dans l'étude de cas devra aussi tenir compte des risques, et prévoir dès l'étape de la conception des mesures relatives à la confidentialité, à la sécurisation et à la souveraineté sur les données.

RÉFÉRENCES

Anderson, B., Lin, S., Newing, A., Bahaj, A. et James, P. 2017. Electricity consumption and household characteristics : Implications for census-taking in a smart metered future. *Computers, Environment and Urban Systems*, vol. 63, pp. 58-67. https://doi.org/10.1016/j.compenvurbsys.2016.06.003

Aotea Great Barrier Island Local Board. 2020. Aotea Great Barrier Island Local Board Plan 2020. https://www.aucklandcouncil.govt.nz/about-auckland-council/how-auckland-council-works/local-boards/all-local-boards/great-barrier-local-board/Documents/aotea-great-barrier-local-board-plan-2020-english.pdf

Apperley, M. 2019. Modelling fractal-structured smart microgrids : Exploring signals and protocols. M. Negnevitsky et V. Sultan (dir.), *Proceedings of ENERGY 2019, The Ninth International Conference on Smart Grids, Green Communications and IT Energy-aware Technologies*, Athens, Greece : IARIA, pp. 13-17.

Benton, R., Frame, A., et Meredith, P. (dir.). 2013. *Te Mātāpunenga : A Compendium of References to the Concepts and Institutions of Māori Customary Law, compiled for Te Matahauariki Institute*. Wellington, Victoria University Press.

Bostrom, N. 2012. The Superintelligent Will : Motivation and Instrumental Rationality in Advanced Artificial Agents. *Minds and Machines*, vol. 22, nº 2, pp. 71-85. https://doi.org/10.1007/s11023-012-9281-3

Bostrom, N. 2014. *Superintelligence : Paths, Dangers and Strategies.* London, UK, Oxford University Press, pp. 1-21, 185-207.

Bradshaw-Martin, H. 2020. Could your self-driving car choose to kill you ? BBC Science Focus. 17 novembre 2020. https://www.sciencefocus.com/future-technology/could-your-self-driving-car-choose-to-kill-you/

Calo, R. 2017. *Artificial Intelligence Policy : A Roadmap.* UC Davis Law Review, vol. 51, pp. 399-435. https://doi.org/10.2139/ssrn.3015350

Cath, C. J. N. *et al.* 2016. Artificial Intelligence and the 'Good Society' : The US, EU, and UK Approach. *Oxford Internet Institute.* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2906249

Cavoukian, A. 2011. Privacy by Design – The 7 Foundational Principles. IAPP. https://iapp.org/media/pdf/resource_center/pbd_implement_7found_principles.pdf

Cleemput, S. 2018. *Secure and privacy-friendly smart electricity metering.* Dissertation, KU Leuven, Belgium. https://lirias.kuleuven.be/retrieve/509996

Commission européenne. 2021. Proposition de règlement du parlement européen et du conseil établissant des règles harmonisées concernant l'intelligence artificielle (législation sur l'intelligence artificielle) et modifiant certains actes législatifs de l'Union. COM(2021) 206. Bruxelles, Commission européenne. https://eur-lex.europa.eu/

d'Alessandro, B., O'Neil, C. et LaGatta, T. 2017. Conscientious classification : A data scientist's guide to discrimination-aware classification. *Big data*, vol. 5, nº 2, pp. 120-134.

Data Sovereignty Now. 2020. Data Sovereignty Now stimulating the data economy executive summary. https://datasovereigntynow.org ;voir aussi https://datasovereigntynow.org/wp-content/uploads/2020/09/stimulating-the-data-economy_executive-summary-1.pdf

Deloria, E. C. 1998. *Speaking of Indians.* Lincoln, University of Nebraska Press.

EDRi. 2021a. *An EU Artificial Intelligence Act for Fundamental Rights. Civil Society calls on the EU to put fundamental rights first in the AI Act*. 30 novembre 2021. https://edri.org/our-work/civil-society-calls-on-the-eu-to-put-fundamental-rights-first-in-the-ai-act/ (consulté le 13 décembre 2021).

EDRi. 2021b. *An EU Artificial Intelligence Act for Fundamental Rights : a civil society statement*. https://edri.org/wp-content/uploads/2021/12/Political-statement-on-AI-Act.pdf

Eliot, L. 2020. *Decisive Essays on AI and Law*. LBE Press Publishing, chap. 4.

European Parliament. (2015/2103(INL)) *EU Parliament Resolution with recommendations to the Commission on Civil Law Rules on Robotics*. https://www.europarl.europa.eu/doceo/document/JURI-PR-582443_EN.pdf?redirect

European Parliament Committee on Legal Affairs. 2016. *Civil Law Rules on Robotics* (2015/2103 (INL)). Brussels, Belgium : European Parliament. https://www.europarl.europa.eu/doceo/document/A-8-2017-0005_EN.html

Executive Office of the President National Science and Technology Council Committee on Technology. 2016. *Preparing for the Future of Artificial Intelligence*. Washington, D.C. http://www.eenews.net/assets/2016/10/12/document_gw_03.pdf

Fjelland, R. 2020. Why general artificial intelligence will not be realized. *Humanit Soc Sci Commun.,* vol. 7, nº 10, pp. 1-9. https://doi.org/10.1057/s41599-020-0494-4

Franklin, A. 2020. The Fourth Amendment in Your Shower : Naperville, Reasonable Expectations of Privacy, and the Intimate Nature of Electric Smart Meter Data. *NCL Rev.,* vol. 99, nº 4, pp. 1141-66. https://scholarship.law.unc.edu/cgi/viewcontent.cgi?article=6848&context=nclr

Geolitica. 2021. *Homepage*. https://geolitica.com

Greveler, U. *et al.* 2012. Multimedia content identification through smart meter power usage profiles. *Proceedings of the International Conference on Information and Knowledge Engineering (IKE),* p. 1. WorldComp.

Haskie, M. J. 2002. *Preserving Culture : Practicing the Navajo Principles of Hozho Doo K'e.* Dissertation, UMI No. 3077247, Ann Arbor, MI, Proquest. https://www.proquest.com/openview/34b9993f90d41bd6a482011691cb023a/1?pq origsite=gscholar&cbl=18750&diss=y

Heaven, W. D. 2020. *Predictive policing algorithms are racist. They need to be dismantled.* MIT Technology Review. https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/

House of Commons Science and Technology Committee. 2016. *Robotics and artificial intelligence.* Fifth Report of Session, London, UK, pp. 5-6, 16-24, 36-38. http://www.publications.parliament.uk/pa/cm201617/cmselect/cmsctech/145/145.pdf

Indigenous Peoples Council on Biocolonialism. http://www.ipcb.org

Indigenous Peoples Council on Biocolonialism et Harry, D. 1995. Patenting of Life and Its Implications For Indigenous Peoples. *Information About Intellectual Property Rights*, nº 7. Janvier 1995. http://www.ipcb.org/publications/briefing_papers/files/patents.html

Indigenous Peoples Council on Biocolonialism. 2006. *The Convention On Biological Diversity's International Regime On Access & Benefit Sharing : Background & Considerations For Indigenous Peoples*. Indigenous Peoples Council on Biocolonialism, Briefing Paper. http://www.ipcb.org/pdf_files/absbriefcop8.pdf

Internet Society. 2015. *The Internet of Things (IoT) : An Overview.* (White Paper Oct 2015), p. 5. https://www.internetsociety.org/wp-content/uploads/2017/08/ISOC-IoT-Overview-20151221-en.pdf

IWGIA and ILO. 2021. *Indigenous Peoples in a changing world of work – Exploring Indigenous Peoples economic and social rights through the Indigenous Navigator,* p. 20. https://indigenousnavigator.org/sites/indigenousnavigator.org/files/media/document/Indigenous%20peoples%20in%20a%20changing%20world%20of%20work%20-%20wcms_792208.pdf

Kahn-John, M. et Koithan, M. 2015. Living in Health, Harmony, and Beauty : The Diné (Navajo) Hózhó Wellness Philosophy. *Global Advances in Health and Medicine Journal*, vol. 4, nº 3, pp. 24-30. https://doi.org/10.7453/gahmj.2015.044

Kaplan, A. et Haenlein, M. 2019. Siri, Siri, in My Hand : Who's the Fairest in the Land ? On the Interpretations, Illustrations, and Implications of Artificial Intelligence. *Business Horizons, v*ol. 62, nº 1, pp. 15-25. https://doi.org/10.1016/j.bushor.2018.08.004

Karwe, M. et Müller, G. 2015. DPIP : A Demand Response Privacy Preserving Interaction Protocol. *International Conference on Business Information Systems* (pp. 224-234). Springer, Cham. https://link.springer.com/chapter/10.1007/978-3-319-26762-3_20

Keall, C. 2021. Why are our defences so shaky ? The Waikato DHB ransomware attack in 20 questions. *New Zealand Herald*, 29 mai. https://www.nzherald.co.nz/business/why-are-our-defences-so-shaky-the-waikato-dhb-ransomware-attack-in-20-questions/4NDSFQD6FST4LHH3UEIIRLABBY/

Kesserwan, K. 2018. *How Can Indigenous Knowledge Shape Our View Of AI Policy Options.* 16 février. https://policyoptions.irpp.org/magazines/february-2018/how-can indigenous-knowledge-shape-our-view-of-ai/

Kim, T. W. et Mejia, S. 2019. From Artificial Intelligence to Artificial Wisdom : What Socrates Teaches Us. *Computer*, vol. 52, nº 10, pp. 70-74. https://doi.org/10.1109/MC.2019.2929723

Kröger, J. 2019. Unexpected Inferences from Sensor Data : A Hidden Privacy Threat in the Internet of Things. L. Strous et V. Cerf (dir.), *Internet of Things. Information Processing in an Increasingly Connected World. IFIPIoT 2018. IFIP Advances in Information and Communication Technology,* vol. 548. Springer, Cham. https://doi.org/10.1007/978-3-030-15651-0_13

Kukutai, T. et Taylor, J. 2016. Indigenous Data Sovereignty Towards an Agenda. *Centre for Aboriginal Economic Policy Research (CAEPR),* vol. 38. ANU, Australia. http://doi.org/10.22459/CAEPR38.11.2016

Landi, H. 2019. *Lingering Impacts from Wannacry : 40 % of healthcare organizations hit by WannaCry in past 6 months*. Fierce Healthcare. 29 mai. https://www.fiercehealthcare.com/tech/lingering-impacts-from-wannacry-40-healthcare-organizations-suffered-from-attack-past-6-months

Lewis, J. 2020. *Creating ethical AI from Indigenous perspectives*. University of Alberta. Folio. https://www.ualberta.ca/folio/2020/10/creating-ethical-ai-from-indigenous-perspectives.html

Lewis, J. E., Arista, N., Pechawis, A. et Kite, S. 2019. Making Kin with the Machines. *Journal of Design and Science*. https://doi.org/10.21428/bfafd97b

Liu, H. Y. et Zawieska, K. 2017. From Responsible Robotics Towards a Human Rights Regime Oriented to the Challenges of Robotics and Artificial Intelligence. *Ethics and Information Technology*, 22, pp. 1-13. https://doi.org/10.1007/s10676-017-9443-3

Lovejoy, K. 2020. *How to manage cyber risk with a Security by Design approach*. EY. 7 février. https://www.ey.com/en_nz/consulting/how-to-manage-cyber-risk-with-a-security-by-design-approach

Maitra, S. 2020. Artificial Intelligence and Indigenous Perspectives : Protecting and Empowering Intelligent Human Beings. *AIES '20,* New York, NY. https://dl.acm.org/doi/10.1145/3375627.3375845

Marsden, M. 1992. *God, Man and Universe : A Māori View.* M. King (dir.) *Te Ao Hurihuri : Aspects of Māoritanga.* Auckland, Reed Books, p. 117.

Marsden, M. 2003. *The Natural World and Natural Resources.* C. Royal (dir.) *The Woven Universe Selected Writings of Rev Maori Marsden*. Masterton, Estate of Rev. Maori Marsden.

McCarthy, J. 2007. *What Is Artificial Intelligence ?* Computer Science Department, Stanford University. http://www-formal.stanford.edu/jmc/whatisai/node1.html

McCarthy, J., Minsky, M. L., Rochester, N. et Shannon, C. E. 2006. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazine*, vol. 27(4), p. 12. doi: 10.1609/aimag.v27i4.1904.

Mead, H. M. 2003. *Tikanga Māori : Living by Māori Values.* Wellington, Huia Publishers, p. 28.

Murrill, B. J., Liu, E. C. et Thompson, R. M. 2012. *Smart Meter Data : Privacy and Cybersecurity, report*. Washington, D.C. https://digital.library.unt.edu/ark:/67531/metadc87204/ (consulté le 10 septembre 2021) ; University of North Texas Libraries, UNT Digital Library, https://digital.library.unt.edu ; crediting UNT Libraries Government Documents Department.

Nagel, L. et Lycklama, D. 2021. Design Principles for Data Spaces. Position Paper. Version 1.0. Berlin. http://doi.org/10.5281/zenodo.5105744

Nations Unies. 2007. Déclaration des Nations Unies sur les droits des peuples autochtones. https://www.un.org/development/desa/indigenouspeoples/declaration-on-the-rights-of%20indigenous-peoples.html

New Zealand Government. 2020. *Privacy Act,* Wellington, New Zealand.

New Zealand Herald. 2021. *Waikato DHB cyber attack : 4200 people's personal details disclosed on dark web*. New Zealand Herald,10 septembre. https://www.nzherald.co.nz/nz/waikato-dhb-cyber-attack-4200-peoples-personal-details-disclosed-on-dark-web/LCSXDX4W3HTZ4FCISHAL4T32IM/

New Zealand Law Commission. 2001. *Māori Custom and Values in New Zealand Law.* Wellington, NZLC SP9. https://www.lawcom.govt.nz/sites/default/files/projectAvailableFormats/NZLC%20SP9.pdf

O'Neil, C. 2016. *Weapons of Math Destruction*. St Ives, Penguin, pp. 84-100.

PCI Security Standards Council. S. d. *Securing the Future of Payments Together.* https://www.pcisecuritystandards.org

Pham, C.T. et Månsson, D. 2019. A study on realistic energy storage systems for the privacy of smart meter readings of residential users. *IEEE Access*, 7, pp. 150262-70.

Phillips, A. M. et Mian, I. S. 2019. Governance and Assessment of Future Spaces : A Discussion of Some Issues Raised by the Possibilities of Human-Machine Mergers. *Development,* 62, pp. 66-80. https://doi.org/10.1057/s41301-019-00208-1

Posthumous, D. 2018. *All My Relatives : Exploring Lakota Ontology, Belief, and Ritual.* Lincoln, University of Nebraska Press.

Reimann, R. 2019. *TechDispatch #2 : Smart Meters in Smart Homes*. European Data Protection Supervisor,16 octobre. https://edps.europa.eu/data-protection/our-work/publications/techdispatch/techdispatch-2-smart-meters-smart-homes_fr

Russell, S. 2021. The history and future of AI. *Oxford Review of Economic Policy*, vol. 37, nº 3, pp. 509-520.

Supreme Court of New Zealand. 2021. *Trans-Tasman Resources Ltd v Taranaki-Whanganui Conservation Board* [2021] NZSC 127.

Technology Quarterly. 2020. *Driverless cars show the limits of today's AI*. The Economist. 13 juin. https://www.economist.com/technology-quarterly/2020/06/11/driverless-cars-show-the-limits-of-todays-ai

Te Mana Raraunga Māori Data Sovereignty Network. S. d. *Te Mana Raraunga – Māori Data Sovereignty Network Charter*. https://static1.squarespace.com/static/58e9b10f9de4bb8d1fb5ebbc/t/5913020d15cf7dde1df34482/1494417935052/Te+Mana+Raraunga+Charter+%28Final+%26+Approved%29.pdf

Toki, V. 2018. *Indigenous Courts, Self Determination and Criminal Justice*. Oxford, Routledge.

Union européenne. 2016. Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (règlement général sur la protection des données). *Journal officiel de l'Union européenne*, 4.5.2016, p 1-88. http://data.europa.eu/eli/reg/2016/679/oj

United States Court of Appeals. *Naperville Smart Meter Awareness v. City of Naperville*, nº 16-3766 (7th Cir. 2018).

Waitangi Tribunal. 2014. *He Whakaputanga me te Tiriti The Declaration and the Treaty : The Report on Stage 1 of the Te Paparahi o Te Raki Inquiry. Wai 1040.* https://forms.justice.govt.nz/search/Documents/WT/wt_DOC_85648980/Te%20Raki%20W.pdf

Walker, R. S. et Hamilton, M. J. 2018. Machine Learning with Remote Sensing Data to Locate Uncontacted Indigenous Villages in Amazonia. *PeerJ Preprints* 6 :e27307v1. https://doi.org/10.7287/peerj.preprints.27307v1

Weidenbener, L. 2019. Many Questions, Fewer Answers at Intersection of AI, Ethics. *Indianapolis Business Journal*, 40, pp. 20-21. https://www.ibj.com/articles/73969-innovation-issue-many-questions-fewer-answers-at-intersection-of-ai-ethics

Whaanga, H. 2020. AI : A New (R)evolution or the New Coloniser for Indigenous Peoples. J. E. Lewis (dir.), *Indigenous Protocol and Artificial Intelligence Position Paper.* Honolulu, Hawai'i : The Initiative for Indigenous Futures and the Canadian Institute for Advanced Research (CIFAR), p. 35. https://spectrum.library.concordia.ca/986506/7/Indigenous_Protocol_and_AI_2020.pdf

Whitt, L. A. 1998. Biocolonialism and the commodification of knowledge. *Science as Culture,* vol. 7, nº 1. https://doi.org/10.1080/09505439809526490

Williams, D. H. et Shipley, G. P. 2019. Limitations of the Western Scientific Worldview for the Study of Metaphysically Inclusive Peoples. *Open Journal of Philosophy*, vol. 9, pp. 295-317. doi:10.4236/ojpp.2019.93020

Williams, D. H. et Shipley, G. P. 2021. Enhancing Artificial Intelligence with Indigenous Wisdom. *Open Journal of Philosophy*, vol. 11, pp. 43-58. doi:10.4236/ojpp.2021.111005

WIPO. 2006. Bioethics and Patent Law : The Cases of Moore and the Hagahai People, septembre. http://www.wipo.int/wipo_magazine/en/2006/05/article_0008.html

Yang, L. *et al.* 2014. Inferring occupancy from opportunistically available sensor data. Pervasive Computing and Communications (PerCom), IEEE, pp. 60-68.

NOUVEL ÉCLAIRAGE PLUTÔT QUE RÉTROSPECTIVE : VOIR, RECONNAÎTRE, PRENDRE EN CONSIDÉRATION ET INSCRIRE LES PERSONNES LGBTI DANS LE CYCLE DE VIE DE L'INTELLIGENCE ARTIFICIELLE

JED HORNER

Praticien en sécurité et confiance digitales. Préalablement directeur de projet au Centre australien pour les droits de la personne, faculté de droit, UNSW Sydney.

ODD 3 - Bonne santé et bien-être
ODD 5 - Égalité entre les sexes
ODD 10 - Inégalités réduites
ODD 11 - Villes et communautés durables
ODD 16 - Paix, justice et institutions efficaces
ODD 17 - Partenariats pour la réalisation des objectifs

NOUVEL ÉCLAIRAGE PLUTÔT QUE RÉTROSPECTIVE : VOIR, RECONNAÎTRE, PRENDRE EN CONSIDÉRATION ET INSCRIRE LES PERSONNES LGBTI DANS LE CYCLE DE VIE DE L'INTELLIGENCE ARTIFICIELLE

Our future survival is predicated upon our ability to relate within equality.

– Audre Lorde (1980, p. 358)

RÉSUMÉ

Les préjugés et la discrimination, ainsi que les préjudices qu'ils causent, ne sont pas de nouvelles préoccupations. Le potentiel qu'a l'intelligence artificielle (IA) de perpétuer des préjugés et d'exacerber la discrimination est au premier plan des discussions mondiales sur les politiques grâce à l'intérêt renouvelé que suscite actuellement l'IA. Pour les groupes de population historiquement marginalisés, comme les personnes lesbiennes, gaies, bisexuelles, transgenres, intersexes, queers ou non binaires (LGBTI), ces préoccupations sont importantes et fondées dans des expériences d'exclusion juridique et sociale. En tant que décideurs et décideuses, membres des communautés touchées, développeurs et développeuses et représentants et représentantes du milieu des affaires, notre réponse à ces préoccupations devrait être de créer une IA responsable d'une manière qui est pratique, multidimensionnelle et à l'écoute des besoins des personnes LGBTI et autres. Une approche de cycle de vie, qui demeure ouverte et adaptable selon l'évolution des pratiques, est un prérequis. Pour contribuer à cette approche, je décris certaines pratiques précises qui, si elles sont adoptées à grande échelle, pourraient jouer un rôle central pour maximiser les retombées positives et réduire les préjudices découlant de l'IA pour les personnes LGBTI. Des audits sociaux élargis, l'adoption de normes internationales reconnues (et une mobilisation accrue dans le processus menant à leur établissement) et des évaluations régulières qui examinent l'adéquation structurale des lois et règlementations existantes afin de gérer les impacts

> de l'IA, notamment en ce qui a trait aux personnes LGBTI, ne sont que trois assises d'une telle approche, alors que nous cherchons à écrire un nouveau chapitre en matière de politiques technologiques plus responsables et plus inclusives.

INTRODUCTION

La triade équité, responsabilité et transparence (désignée sous l'acronyme FAT en anglais) figure en bonne place dans les discussions internationales sur l'intelligence artificielle (IA) responsable (Raji *et al.*, 2020 ; Selbst *et al.*, 2019). Alors que ces discussions mettaient auparavant l'accent sur des différences principalement liées au genre et à la « race », d'autres axes émergent de plus en plus, y compris l'orientation sexuelle, l'identité de genre et le statut intersexe. Les personnes lesbiennes, gaies, bisexuelles, transgenres, intersexes, queers ou non binaires (LGBTI) sont ainsi mises en évidence. Ces discussions, de même que les approches pratiques qu'elles engendrent à l'aide de principes, de boîtes à outils ou de cadres, parfois à un échelon supranational (Commission Européenne, 2021), mais souvent au sein d'équipes techniques, s'avèrent essentielles, bien qu'elles ne suffisent pas (Bowles, 2018 ; Nolan et Frishling, 2020). L'un des enjeux liés à l'adéquation des approches existantes a trait à l'héritage pernicieux et durable découlant de la discrimination *de jure* et *de facto*, qui a laissé une « marque indélébile dans la vie des personnes LGBTI, envisagées comme un groupe diversifié » (Horner, 2017, p. 99). Il nécessite une intervention élargie, plus englobante, qui prend au sérieux les facteurs structurels si enchevêtrés dans l'histoire de la discrimination.

Dans ce chapitre, je fais valoir que, pour concevoir une « IA responsable » d'une manière qui s'adapte aux besoins des personnes LGBTI, il faut adopter une « approche du cycle de vie » élargie multidimensionnelle qui continuera à se transformer à mesure que les bonnes pratiques évoluent, et que les manques et les angles morts se manifestent. Pour contribuer à l'énonciation de ce à quoi cette approche pourrait ressembler, je présente quelques pratiques à prendre en considération. Je reconnais le travail de ceux et celles qui ont façonné, créé et rehaussé ces approches pratiques, et j'en suis redevable. Ceux-là et celles-là comprennent, au fil de l'histoire, des entreprises américaines, le regretté révérend Leon Sullivan (Stewart, 2011), des groupes de réflexion (tels que Carnegie et la Fondation Ford) et des fantassins, dont mon propre oncle, qui, en cherchant à transformer les conditions de travail au sein d'entreprises multinationales actives en Afrique du Sud durant l'apartheid, a eu la témérité de demander un « changement évolutif » (Horner, 1971). En période d'intensification de la lutte politique, aucune de ces contributions ne devrait être simplement rejetée alors que nous nous précipitons pour créer de nouveaux cadres, sans toujours tenir compte des anciens. En effet, il ne faut pas abandonner les idées, la bonne volonté, les connaissances disciplinaires, les précédents réglementaires et les aspects des bonnes pratiques, comme l'ont déjà fait remarquer des praticiens de l'IA responsable (Raji *et al.*, 2020 ; Marchant, 2011 ; Mittelstadt *et al.*, 2016).

Je soutiens qu'en traçant cette approche élargie du cycle de vie qui tient compte des aspects politiques et historiques, nous devrions simultanément nous concentrer sur une approche multidimensionnelle. Une première dimension consiste à réaliser des *audits sociaux* afin de comprendre le contexte sociopolitique élargi dans lequel l'IA se développe, s'utilise, se jauge et s'évalue, et ce, avant d'examiner les incidences qu'elle a comme produit. Une autre dimension invite à l'adoption (et à la mise sur pied effective) de *normes internationales* reconnues, et une troisième concerne l'évaluation de l'*adéquation structurelle*

des cadres législatifs et réglementaires. Cette dernière préoccupation touche aux axes de différence précis particuliers que sont par exemple l'orientation sexuelle, l'identité de genre ou le statut intersexe dans les domaines de la vie publique et en lien avec l'IA. L'évaluation de tous ces aspects ne relève pas seulement d'une responsabilité des gouvernements, mais aussi des entreprises et de la société civile, comme l'histoire l'a compris (Gray et Karp, 1994 ; First, 1973). Compte tenu de l'évolution récente de la situation en Union européenne (Commission Européenne, 2021), aux États-Unis (National Security Commission on Artificial Intelligence, 2021) et dans d'autres pays, il s'avère urgent d'adopter une telle approche multidimensionnelle qui mobilise l'industrie, le gouvernement, la société civile, ainsi que les membres des communautés en cause. À défaut de le faire et en s'en remettant à des solutions techniques étroites, il est possible que les mauvaises expériences des personnes LGBTI en ce qui concerne le fonctionnement de l'IA deviennent des obstacles réels et concrets au plein respect de leurs droits fondamentaux. Puisque la production et la diffusion technologiques, et notamment leurs répercussions, ont historiquement été inégales et inéquitables, des divisions pourraient s'exacerber entre les populations des pays à revenu élevé et celles des économies émergentes, ainsi qu'entre différents groupes au sein des pays (Benjamin, 2019 ; Eubanks, 2018). Les personnes LGBTI font partie de ces gens vulnérables susceptibles de faire l'objet de violence et de discrimination. Voilà un élément sur lequel nous pouvons intervenir pour prévenir les écarts, au bénéfice de l'humanité, et plus particulièrement des personnes LGBTI. La question qui se pose est la suivante : Comment ?

VIDER LE PLACARD DE L'IA : ORIENTATION SEXUELLE, TRI SOCIAL ET ALGORITHMES INTELLIGENTS

Au cours de la première vague de COVID-19, un couple homosexuel établi aux États-Unis grâce à un visa de travail et touché par les effets de la pandémie sur le marché local, a choisi de générer son propre revenu en vendant son album en ligne (Golding-Young, 2020). Tous deux faisaient de la musique ensemble depuis huit ans. Ils ont publié une vidéo sur Facebook, sous forme de publicité payante, dans le but d'atteindre des admirateurs et admiratrices. Leur publication a été refusée, Facebook l'ayant qualifié de « contenu sexuellement explicite pour adultes », présumément en raison d'une photographie de leurs fronts se touchant, qu'ils utilisent depuis des années. Ils ont alors testé la plateforme, se disant que des règles similaires (vraisemblablement les « normes de la communauté ») s'appliqueraient aux images en apparence romantiques ou intimes d'un couple hétérosexuel. Ils rapportent qu'il en est allé autrement (Golding-Young, 2020). Il n'y aurait pas tout à fait deux ensembles de règles, mais sans doute deux interprétations des mêmes règles. Les membres du couple dénoncent la situation :

> We have been heartened recently by the improved representation of LGBTQ people on television, and we are grateful that most people we meet are accepting of our relationship. It's enough to make you think that maybe society has fully accepted that "love is love." Unfortunately, our recent experience with Facebook suggests otherwise. When Facebook's platform refused to allow us to fully express ourselves as both artists and a same-sex couple, it brought back painful memories of discrimination against the LGBTQ community (Golding-Young, 2020).

Il y a plusieurs façons d'aborder ce cas, que ce soit par exemple sous l'angle des ambiguïtés, des incohérences ou des problèmes éthiques. Une manière de l'envisager consiste à invoquer la notion de « boîte noire », selon laquelle on ne peut voir les décisions prises en coulisses, qui ne sont pas expliquées de manière adéquate aux utilisateurs et utilisatrices, aux consommateurs et consommatrices ou aux citoyens et citoyennes (Pascale, 2015 ; Rudin et Radin, 2019). Une autre option serait de faire référence au « placard de l'IA ». La métaphore du placard est largement utilisée et comporte de multiples significations. Une interprétation dominante en fait « un coin où se retirer dans l'intimité », alors qu'une autre renvoie à la notion de secrets ou de « squelettes dans le placard » (Sedgwick, 2008, p. 65).

Historiquement, le placard a rendu les lesbiennes, les gais, les bisexuels, les transgenres et parfois les intersexués à la fois visibles et invisibles, paradoxalement, selon la manière d'être ou le lieu. Les gens parlent de ce placard, sont conscients de son existence, mais ne s'entendent jamais tout à fait sur la signification à lui accorder. Il est omniprésent, tout comme l'IA l'est devenue. Les contours de cette notion sont façonnés par les lois, les médias populaires, les attitudes sociales et les dispositions individuelles (Horner, 2017). Il s'agit d'un espace de liminalité, quelque chose entre pleine citoyenneté sociale et marginalisation qu'une confluence de facteurs a imposée, quelque chose que le couple mentionné précédemment a sans doute vécu.

Certains pourraient considérer que la pertinence du placard s'érode, dans certains endroits, dans certaines conditions, pour certaines personnes. Dans un monde marqué par une connectivité numérique croissante, peut-être revêt-il simplement une nouvelle forme, restant témoin d'aspiration et de trépidation, de désir et de désespoir, de plaisir et de douleur, de danger et d'émancipation. Relativement à l'IA, il est utile de voir le placard sous cette optique, qui permet à la fois de prendre en compte les différentes significations rattachées à l'IA et d'invoquer ce que Bucher appelle l'« imaginaire algorithmique » de l'IA (2016, p. 31). Cet imaginaire englobe ce que les personnes LGBTI et le grand public considèrent comme les possibilités de l'IA et ses retombées, qu'elles soient réelles et actuelles ou imaginaires.

Comme l'exemple du couple homosexuel l'a fait ressortir, nous vivons déjà à l'ère des algorithmes intelligents. Ces algorithmes consomment, à une échelle sans précédent, des données servant à trier notre monde social, à façonner nos préférences et nos craintes pour ensuite y répondre, et ils permettent la monétisation d'aspects de nous-mêmes auparavant inaccessibles. Comme je l'ai souligné, ils surveillent et jugent également, de manière de plus en plus automatisée. Dans ce monde où l'information est devenue une monnaie en soi, une forme de capital qui remet en question la façon dont nous configurons et organisons la richesse et le pouvoir, une économie politique de l'information a émergé qui « instrumentalise la différence plutôt que la similitude » (Wark, 2019, p. 31). Cette économie politique de l'information, stimulée par le développement de produits dont elle dépend, a offert beaucoup de promesses aux personnes LGBTI, qui sont historiquement marginalisées, à la fois dans un sens *de jure* et *de facto* (Horner, 2017, p. 101). La résurgence de l'IA, comprise comme une partie essentielle de cette économie de l'information émergente, a promis la libre expression, la connectivité et, en fin de compte, une forme d'émancipation autrefois limitée par d'anciennes formes de technologie ou configurations du pouvoir politique.

Pensez au couple cherchant à gagner un revenu. Ou plutôt, pensez au sexe. Y succèdent de plus en plus des rencontres sexuelles de types totalement différents que des rencontres physiques grâce aux plateformes telles que Grindr et Scruff, qui s'adressent aux jeunes et aux moins jeunes générations (Albury *et al.*, 2017). Ces plateformes, des emblèmes de l'IA pour de nombreuses personnes LGBTI, organisent des contenus visuels (tels que des égoportraits), exploitent des données de localisation (gens à proximité, lieu épinglé sur une carte), et permettent d'exprimer des préférences (type de corps, âge, etc.). Ce faisant, elles facilitent les types de connexions numériques et physiques que les utilisateurs et utilisatrices pourraient souhaiter, produisant du plaisir, voire autre chose (Albury *et al.*, 2017 ; Race, 2009). Un nombre croissant de publications explorent les effets qu'ont les interactions sur ces plateformes pour les personnes LGBTI. Elles examinent par exemple comment la conception ou la configuration d'une plateforme, notamment l'ancienne fonctionnalité de Grindr permettant de classer les utilisateurs et utilisatrices en fonction de leur race, enracine les antagonismes sociaux existants et les expériences de discrimination raciale (Maslen, 2019).

Ces exemples, regroupés autour de ce qui est considéré comme de l'IA, laissent facilement entrevoir que les algorithmes intelligents jouent désormais un rôle pivot dans la manière dont les personnes LGBTI trouvent des partenaires sexuels, entrent en relation, s'engagent dans des activités commerciales, expriment des opinions politiques, et bien plus encore. Pour les personnes LGBTI, le rapport complexe aux promesses de l'IA, qu'il soit établi dans le (plus que) réseau social Facebook ou les applications

de rencontre, pourrait s'avérer « cruellement optimiste » (Berlant, 2011). Je fais en l'occurrence référence à Berlant (2011, p. 1), qui définit l'optimisme cruel comme « un état d'attachement à un objet (sentiment, relation, aspiration) avant la perte de celui-ci ; une vision très convaincante de la "belle vie" qui semble ultimement empêcher l'atteinte de telles aspirations en premier lieu ». Peut-être s'agit-il d'un avenir conjuré par l'IA (en matière de sexe, de revenu supplémentaire, d'une meilleure expérience de consommation ou d'une relation) ou d'une évasion de l'actuelle homophobie physique (seulement, elle sera remplacée en ligne par une forme d'homophobie infâme qui ne s'éteint pas, même quand on ferme sa porte d'entrée).

Ces considérations ne sont pas théoriques. Soulever ces questions permet de réorienter notre attention d'une technologie simplement instrumentale, vue comme un fait accompli, vers un engagement critique par lequel cerner et examiner notre rapport à l'IA en tant que personnes LGBTI. À partir de cette base, nous pouvons nous demander : Quelles cibles visons-nous quant à une forme particulière de technologie ? Que présumons-nous des capacités de celle-ci par rapport à nos désirs, à nos espoirs et à nos craintes ? Comment pouvons-nous gérer certains inconvénients auxquels nous faisons face, en tant que groupe diversifié, et que la technologie pourrait amplifier ? Soulever ces questions nous permet d'intervenir, avec d'autres et de manière réfléchie, en vue de redessiner la trajectoire qu'emprunte la technologie. Pour le faire efficacement, il faut toutefois énoncer ce que constitue l'IA, en passant par les domaines scientifique, populaire et politique.

DÉFINIR L'IA : DES CYCLES DE VIE, PAS *SEULEMENT* DES ALGORITHMES

Bien qu'il y ait une obsession quant aux algorithmes en tant que mandataires de l'IA et de ses répercussions sociales, politiques, environnementales et économiques, l'IA concerne finalement plus que les algorithmes (Metcalf et Crawford, 2016 ; Mittelstadt *et al.*, 2016 ; Raji *et al.*, 2020 ; Selbst *et al.*, 2019). Mittelstadt et ses collaborateurs et collaboratrices (2016, p. 2) ont fait valoir qu'il « est insensé d'envisager l'éthique des algorithmes sans tenir compte de la manière dont ceux-ci sont mis en œuvre et exécutés par des programmes informatiques, des logiciels et des systèmes d'information ». Il existe, au contraire, un cycle de vie de l'IA faisant intervenir des matières premières, de l'entrée de données, des algorithmes, de l'optimisation, des processus de vérification et une planification commerciale, autant d'éléments façonnés par la prise de décisions humaine. Ensemble, ces éléments forment l'*objet* que nous appelons IA. Pour décrire l'IA, certains ont adopté les termes *cyberphysique* ou *cybernétique*, qui comportent une façon d'en voir des aspects (Bell *et al.*, 2021). Afin d'expliquer une telle approche axée sur la chaîne de valeur et d'aider les gens à la visualiser, Bratton (2015) a introduit l'image puissante de la « pile », qui représente les composantes de l'informatique moderne et s'applique à l'IA. La pile de Bratton englobe la « couche terrestre » (par exemple, les composants de dispositifs techniques), l'interface et l'utilisateur ou utilisatrice. Elle s'apparente à une heuristique cherchant à comprendre ce qu'est l'IA, littéralement et matériellement, y compris dans une perspective d'économie politique. On pourrait envisager sous cet angle l'appareil d'une maison intelligente ou une application de rencontre, notamment, en pensant non seulement à l'IA, mais aussi à la puissance adjacente qui lui est nécessaire (par exemple, un système d'exploitation, des puces ou la batterie d'un téléphone intelligent). Cette conception de l'IA en tant que cycle de vie a des conséquences sur la manière de définir l'IA, l'attention se déplaçant de lignes de code isolées vers les processus d'entreprise, la mobilisation des parties prenantes et les décisions relatives à la manière dont la technologie est développée, adoptée, examinée. Elle comprend la façon d'en gérer l'évolution et les répercussions, y compris pour les personnes LGBTI. De telles décisions impliquent toujours des moments d'exclusion, lorsque nous choisissons de privilégier un facteur, un attribut ou un segment de marché ou de faire un compromis entre des formes de préjudice social et un impératif commercial (Mouffe, 2005). Par définition, une approche fondée sur le cycle de vie concerne donc la responsabilité, responsabilité d'entreprise ou autre, au moment de la prise de décisions.

Cette approche de l'IA encourage la responsabilité ou y donne lieu, justement parce que la discussion se déplace des notions d'inévitabilité ou de « technologie indisciplinée » vers l'idée que nous, en tant qu'humains et par « le politique », pouvons influer sur la trajectoire de la technologie en façonnant et en modelant la manière dont celle-ci se développe, devient nôtre et est mise à l'échelle (Mouffe, 2005). Elle aborde les effets potentiels de l'IA sur certains groupes particuliers. Alors la décision de recueillir à grande échelle des données sur l'orientation sexuelle, de manière à cibler la publicité politique ou de concevoir un produit tel qu'une application de drague ou de rencontre fondée sur une utilisation sélective (peut-être responsable) de données similaires, n'est pas qu'une réaction instinctive. Il s'agit plutôt d'une offre réfléchie du marché pour combler un besoin, satisfaire un segment ou profiter d'une occasion d'affaires, une offre étayée (ou non) par des considérations juridiques, éthiques et politiques. L'IA que nous voyons, que nous utilisons et à laquelle nous nous identifions n'est pas le fruit du hasard ni le produit d'un rêve technologique. Elle est construite par des humains, en fonction de leurs préjugés, de leur pouvoir et de leurs privilèges (Benjamin, 2019). En effet, même la capacité d'utiliser certains positionnements tels que gai ou bisexuel, ou encore la façon dont une équipe de produits définit des concepts comme « race » (noir, blanc, asiatique) ou le genre (homme, femme, non binaire, x, etc.) sont contingentes et évoluent dans le temps (Laclau, 2005).

En bref, il n'y a pas de produit, pas d'artéfact d'IA, sans un cycle de vie. Pourtant, je maintiens que le cycle lui-même est plus long et plus complexe que ce que des concepteurs et conceptrices ou des ingénieurs et ingénieures peuvent imaginer, puisqu'il englobe un large éventail de facteurs sociopolitiques. Dans une telle approche différenciée du cycle de vie, les décisions relatives à la conception de l'IA ou à son adoption ultérieure ne sont pas neutres et, selon ce point de vue, celles qui lui seraient défavorables n'appartiennent pas exclusivement aux équipes techniques (par exemple, l'ingénierie). Cette approche reconnaît l'importance du produit, de la marque et, plus largement, de l'appétit de l'organisation pour le risque et la responsbilité qui en découle, dans le cadre du cycle de vie.

Le projet de norme ISO/IEC DIS 22989 sur l'intelligence artificielle (relevant du comité technique JTC 1/SC 42) contient une définition semblable à celle à laquelle je fais référence. L'IA y est définie comme un « ensemble de méthodes ou d'entités automatisées qui créent, optimisent et appliquent un modèle pour que le système puisse, pour un ensemble donné de tâches prédéfinies, fournir des prédictions, des recommandations ou des décisions à l'issue de calculs. Les systèmes d'IA fonctionnent selon divers degrés d'automatisation » (ISO/IEC, 2020). Il est encourageant de constater que cette définition met l'accent sur un cycle de vie, sur une série de parties ou de fonctions interreliées (telles que l'optimisation algorithmique) faisant intervenir une automatisation à des degrés *divers*. Cette définition ne réduit pas l'IA à tous les cas d'automatisation, pas plus qu'elle ne met en cause des activités isolées telles que la collecte de données, même automatisée.

Bien que la définition que je préconise ne soit probablement pas contestée dans le milieu des affaires, étant donné les pratiques largement comprises de la stratégie d'entreprise, de la gestion des risques (PRISMA, 2020 ; Bemthuis *et al.*, 2020) et même, plus récemment, de la gestion des risques politiques (Rice et Zegart, 2018), il est important de la mettre à l'avant-plan. C'est qu'il y a des divergences, dans le cycle de vie de l'IA, dans la façon dont les universitaires, les praticiens et praticiennes, les institutions publiques et les entreprises la perçoivent, comprennent son impact potentiel et attribuent une responsabilité ou un blâme relativement à ses lacunes (Australian Human Rights Commission, 2021 ; Eubanks, 2018). L'approche de l'IA que j'ai décrite, celle du cycle de vie, embrasse une vision multipartite en ce qui a trait aux intervenants et intervenantes qui font l'IA, l'utilisent, en évaluent les effets, la réglementent et rendent compte de sa croissance, que ce soit à l'échelle nationale ou internationale.

ADOPTER UNE IA RESPONSABLE : ENTRER DANS LE CYCLE DE VIE

Marchant (2011, p. 200) affirme qu'en réponse aux développements technologiques susceptibles d'entraîner des préjudices, dont certains que j'ai décrits, nous pouvons : « 1) réduire la cadence de l'évolution ; ou 2) améliorer notre capacité d'adaptation ». La première option semble problématique, car elle nuit à la diffusion de la technologie et entrave le droit d'accéder aux avantages qu'offre la science relativement à la mise au point et à l'adoption de l'IA, ce qui va des réseaux neuronaux complexes à l'adaptation de l'apprentissage automatique. Elle pourrait aussi, par inadvertance, ancrer la domination de l'IA dans une zone géographique, ce qui poserait des problèmes socioéconomiques et des enjeux de sécurité (National Security Commission on Artificial Intelligence, 2021). La seconde option semble intuitive et nécessite la conception, l'examen et l'intégration d'approches qui sont techniquement fiables, adaptées aux contextes sociopolitiques et respectueuses des normes fondamentales en matière de droits humains. Il n'y a qu'à penser aux lois, aux principes, aux codes d'entreprise, aux audits et à un large éventail de pratiques existantes (Gray et Karp, 1994 ; Raji *et al.*, 2020 ; Whittlestone *et al.*, 2019). Parmi les exemples positifs récents et encourageants figure l'élaboration de ressources complètes par l'entremise de ABOUT ML (*Annotation and Benchmarking on Understanding and Transparency of Machine learning Lifecycles*), sous les auspices de l'institut de recherche Partnership on AI (Partnership on AI, 2021). Dans la foulée de ce projet et en élargissant son objectif, je décris ci-après des mesures précises à adopter, ensemble ou individuellement, pour favoriser une approche étendue du cycle de vie afin de gérer les retombées de l'IA sur les personnes LGBTI. Je décris la nature itérative et interreliée de cette approche dans la figure 1.

| FIGURE 1 |

Un modèle intégré et itératif d'une approche étendue du cycle de vie à l'IA.

Évaluations sociales

Facteurs à considérer et points de référence de base

- Instruments internationaux de droits humains
- Facteurs sociaux ou identitaires à considérer (par exemple l'orientation sexuelle, l'identité de genre, etc.)
- Rapports d'examen périodique universel du Conseil des droits de l'homme des Nations unies
- Objectifs de développement durable (ODD)
- Lois et politiques nationales
- Données sur les disparités sociales
- Voix des communautés affectées (par exemple littérature qualitative et contributions issues de processus participatifs)

Pratiques de l'entité

Adoption de normes internationales par des organismes d'élaboration de normes (OEN) reconnus

- Audits et certification
- Principes
- Règles de conception
- Fiches techniques et autres artefacts
- Recours, réparation et procédures de traitement des réclamations
- Gestion et révision de documentation du système

Suivi, examen et révision continus

Bilans de la législation, de la réglementation et des documents politiques contraignants

- Modification à la législation ou à la réglementation
- Changements dans les politiques stratégiques et opérationnelles
- Soumission de rapports aux processus multipartites (par exemple: institutions nationales des humains et le Comité des droits de l'homme des Nations unies)
- Influence sur les pratiques et processus (par exemple faire évoluer les bases de données)

Définir d'abord les enjeux : mener des audits sociaux étendus dans des domaines clés afin de cerner les enjeux et de mesurer leur ampleur

> Despite being the past participle of the Latin verb "dare," i.e., "to give" data instead are always produced by people, out of what they observe, fail to see, or suppress in the world in which they live. A corollary, in the case of people, is that a hallmark of privilege is who and what one can afford to ignore (Krieger, 2021, p. 2).

Il est souvent difficile d'agir sur ce qui nous reste invisible, ce sur quoi nous sommes incapables de nous mettre d'accord ou ce pour quoi nous n'avons pas déterminé un mode de mesure. Comme le note Krieger (2021, p. 2), c'est souvent la conséquence de décisions conscientes, des actes de commission et d'omission. Comprendre l'axe de l'IA responsable est difficile pour les personnes LGBTI en l'absence de consensus sur les aspects fondamentaux que sont les obligations, les besoins ou les préoccupations à une échelle nationale, sans parler d'une échelle régionale ou internationale. Nous pourrions, par exemple, énumérer certains droits fondamentaux, mais en omettre d'autres, ou nous appuyer sur les expériences limitées des membres d'une équipe pour cerner les risques connus et, vu nos angles morts, oublier ceux qui ont une forte probabilité de se matérialiser. Nous pourrions prêter attention à des disparités propres aux personnes LGBTI, en matière de santé ou d'emploi, mais pas à d'autres. De même, nous pourrions négliger des obstacles structurels persistants (par exemple, les lois qui désavantagent les personnes LGBTI) en privilégiant une analyse des pratiques sociales seulement. Il pourrait en être ainsi en partie à cause de la composition des équipes, qui pourraient privilégier des disciplines telles que l'anthropologie ou la sociologie, comme c'était auparavant le cas de l'informatique, et en partie à cause d'artéfacts qui se produisent dans le domaine, reflétant toujours une « façon de faire » acceptée.

Un modèle permettant une analyse et une prise de conscience plus approfondies de la situation de référence des personnes LGBTI, et fondant la *pratique* d'une IA responsable, repose sur un « audit social ». Bien qu'ils ne soient ni nouveaux ni nécessairement uniques (Nolan et Frishling, 2020), les audits offrent un processus complet pour analyser les données disponibles, repérer les lacunes quant à la disponibilité et à la qualité des données, et mettre au point de nouveaux modèles explicatifs. Ils sont déjà utilisés en développement de produits d'IA et peuvent s'ouvrir aux perspectives élargies des parties prenantes. Le modèle SMACTR (*Scoping, Mapping, Artifact Collection, Testing, and Reflection*) de Google en est d'ailleurs un exemple (Raji *et al.*, 2020). Cependant, l'audit social idéal que je décris est de nature structurelle et vise à fournir le matériel de base nécessaire à une évaluation des répercussions. Sa perspective est plus large, il se concentre sur des niveaux, et il examine le degré d'exposition aux préjudices et les voies qui continuent à désavantager les personnes LGBTI. L'objectif de ces audits sociaux est de vérifier les suppositions sur le risque, la vulnérabilité et la susceptibilité que pourraient avoir les décideurs et décideuses politiques, les gestionnaires de produits, les concepteurs et conceptrices et les ingénieurs et ingénieures. Pour ce faire, les audits s'appuient sur la combinaison de données, d'idées, de normes en matière de droits humains, ainsi que sur le matériel publié, et sont eux-mêmes façonnés par les voix des communautés en cause.

Il existe des modèles internationaux instructifs qui pourraient illustrer ce que constitue une approche idéale en matière d'audit social. Le Royaume-Uni, par exemple, a été le premier pays à mener un audit sur les disparités raciales, lequel a recueilli un ensemble de données accessibles à tous et à toutes dans des domaines d'analyse clés. Le processus d'audit a également fait ressortir les lacunes des normes de collecte de données dans de nombreux domaines critiques, ce qui est sans aucun doute utile pour fonder des interventions précises et, dans certains cas, adopter des mesures d'atténuation. Ce processus a été examiné par le Comité des femmes et de l'égalité de la Chambre des communes du Royaume-Uni (2018), qui l'a approuvé tout en demandant un plan d'action clair. La rapporteuse spéciale des Nations Unies sur les formes contemporaines de racisme, de discrimination raciale, de xénophobie et de l'intolérance qui y est associée a fait remarquer que « [l'audit des disparités raciales] et sa base de données sont dignes d'inspiration pour les gouvernements de partout dans le monde [et qu'elle] salue vivement cette initiative » (Haut-Commissariat des Nations Unies aux droits de l'homme, 2019).

Relativement aux personnes LGBTI, ces audits peuvent être réalisés par des entreprises, des gouvernements ou des organisations de la société civile, en tenant compte spécifiquement de la nature des préjudices connus auxquels ces personnes sont exposées. Sur le plan méthodologique, ils devraient toutefois s'améliorer et s'appuyer sur des points de vue élargis et des perspectives disciplinaires différentes, et prévoir des mesures proactives pour garantir que les communautés en cause sont représentées, non seulement dans la conception de ces processus, mais aussi au stade de l'analyse. Idéalement, ces audits devraient également se concentrer sur trois domaines de préoccupation clés servant de point de départ :

1. ***L'adéquation des données existantes :*** Examen des données, y compris celles sur les disparités sociales (degré de fiabilité des données, validité apparente, manière de les recueillir, détection ou non de l'ampleur des préjudices, par exemple, leur intensité et la durée de l'exposition à ceux-ci).

2. ***Les modèles explicatifs utilisés pour regrouper, analyser et décrire les données :*** Le cadre d'analyse dégage-t-il une orientation des résultats en fonction de suppositions au sujet des personnes ayant une identité de genre ou une orientation sexuelle en particulier ou de leurs caractéristiques sexuelles physiques ?

3. ***Des facteurs structurels (tels que les lois) dont les effets se répercutent dans des domaines clés :*** En quoi cela correspond-il aux données d'exposition ? Les voies empruntées pour observer ces effets législatifs et réglementaires sont-elles claires, précisées et expliquées ? Les lois ont-elles été abrogées récemment en raison d'un décalage par rapport aux attitudes sociales ou sont-elles enracinées depuis des décennies ?

Enfin, les audits sociaux devraient être entrepris par des équipes d'experts et d'expertes ayant des connaissances approfondies en gestion des risques, en recherche en sciences sociales, en ingénierie, en informatique et dans d'autres disciplines fondamentales. Étant donné la profondeur de l'analyse dans des domaines (tels que la santé publique) et des disciplines (telles que l'épidémiologie) qui mesurent les préjudices sociaux, l'utilisation de ces concepts, outils et méthodologies disciplinaires enrichirait le cadre d'analyse des audits sociaux (Krieger, 2021). Par exemple, dans l'évaluation des effets de préjudices proprement liés à l'IA ou encore de ceux découlant de son développement (Benjamin, 2019), les mesures ne devraient pas tenir compte que de la susceptibilité d'être exposé au préjudice et d'une éventuelle exposition. Elles pourraient également vérifier à quel degré l'exposition à des préjudices donnés contribue à maintenir ou à ancrer les disparités, « ce qui requiert l'élaboration d'hypothèses préliminaires sur les relations entre le moment ou l'intensité de l'exposition et les effets qui font l'objet de l'étude » (Krieger, 2000, p. 57). S'il y a lieu de détailler les résultats d'un tel audit social et de faire preuve de rigueur méthodologique, il importe également de présenter une analyse des données claire et d'illustrer les résultats pour un public de non-spécialistes. Après tout, l'objectif d'un tel audit social est de changer les pratiques, pas seulement les modes d'analyse. Si nous ne sommes pas en mesure de définir les disparités sociales qui perdurent et touchent les personnes LGBTI ou de comprendre les mécanismes qui les maintiennent en place, nous ne pouvons pas avoir une discussion éclairée sur ces enjeux, même lorsque ceux-ci sont remis en cause, figés ou amplifiés par l'IA. Les audits sociaux constituent une première étape nécessaire pour élargir et approfondir notre compréhension, notre engagement et, en fin de compte, nos analyses, afin d'en venir à une IA responsable.

Élaborer et adopter des normes internationales pour guider le développement, le déploiement et l'évaluation de l'IA, à grande échelle

Compte tenu du vécu de chaque personne, des contextes juridiques nationaux et de la diffusion géographique et socioéconomique de la technologie, les personnes LGBTI doivent réfléchir aux moyens les plus efficaces et les plus percutants d'influer sur une IA responsable à l'échelle mondiale. Les défenseurs et défenseuses des droits LGBTI l'ont fait avec succès par le passé, en s'appuyant sur

les droits civils et les mouvements de justice sociale naissants, pour mieux faire respecter leurs droits fondamentaux (Lixinski, 2020). Ils et elles y sont parvenus par des modifications de la législation nationale ou des initiatives de diversité et d'inclusion, par exemple.

Dans un monde de plus en plus multipolaire, un incitatif à poursuivre ce travail pourrait venir des mesures réglementaires prises en Europe (Commission Européenne, 2021), aux États-Unis ou à quelque endroit où il existe un marché concentré et un pouvoir réglementaire, où les institutions et les États-nations accepteraient de tourner le regard vers les risques associés à l'utilisation de l'IA et les préjudices sociaux qui en découlent pour les personnes LGBTI et les autres. Très souvent, ces changements nécessitent une interprétation et des mécanismes concrets pour favoriser non seulement la conformité, mais aussi un suivi et un examen continus. Ils créent des circonstances opportunes, dont beaucoup pourraient entraîner un échange de connaissances pratiques et une collaboration entre la société civile, les chercheurs et chercheuses, les organismes publics et les entreprises technologiques elles-mêmes. L'énergie et les efforts déployés pour comprendre cet enjeu doivent toutefois être mieux canalisés et dirigés, afin de produire des artéfacts utilisables à grande échelle et par des entités de toutes tailles. Une telle précaution permettra d'éviter la dispersion des bonnes pratiques, voire les pratiques propriétaires.

L'une des voies à suivre est la définition de normes internationales. Ayant connu des précédents, elle bénéficie également d'une infrastructure ouverte à une approche multipartite et elle embrasse le multilatéralisme. Par l'intermédiaire d'organismes de normalisation reconnus, tels que l'Organisation internationale de normalisation (ISO) ou la Commission électrotechnique internationale (CEI), de nombreuses entreprises technologiques responsables, comme Microsoft, Google et IBM, participent déjà à l'élaboration de normes relatives à la sécurité de l'information, à la gouvernance des technologies de l'information et, plus récemment, à l'intelligence artificielle. Grâce à leur modèle de gouvernance, ces partenariats peuvent fournir un processus complet et rigoureux qui soutient le langage, la structure et l'approche méthodologique essentiels à la création d'artéfacts (c'est-à-dire les normes, les spécifications techniques) dans le contexte des contrats commerciaux, de l'appel à la réglementation (le cas échéant) ainsi que de l'utilisation volontaire élargie au sein de l'industrie (Cihon, 2019).

L'élaboration de normes internationales n'entre pas nécessairement en contradiction avec les mesures réglementaires internationales relatives à l'IA qui concernent les personnes LGBTI. Au contraire, elle s'avère souvent complémentaire, voire nécessaire. Par exemple, le respect de la vie privée restera un droit essentiel à protéger pour les personnes LGBTI, pour tout être humain. Des normes internationales telles que l'ISO/CEI 27701, adaptées aux exigences du Règlement général sur la protection des données (RGPD) de l'Union Européenne et d'autres lois nationales sur le respect de la vie privée, fournissent déjà un cadre aux entreprises de toutes tailles qui souhaitent mettre en œuvre, de manière pratique, une approche complète de la gestion de l'information relative au respect de la vie privée (Standards Australia, 2020, p. 28).

Alors que l'élaboration des normes d'IA s'accélère, les organisations de la société civile LGBTI et les contributeurs et contributrices individuels ont des occasions manifestes de s'assurer que les normes émergentes en ce qui a trait à la partialité, à la gestion des risques ou à des domaines connexes tiennent dûment compte des préjudices sociaux particuliers que subissent les personnes LGBTI. Dans certains cas, on y parviendra en faisant en sorte que les normes reflètent les pratiques et les méthodes efficaces, et qu'elles y fassent référence. Dans d'autres, il faudra peut-être produire des rapports techniques précis, lorsque les pratiques sont récentes et qu'une étude approfondie s'impose avant de procéder à la normalisation. Une mesure intermédiaire pourrait être le travail de « prénormalisation », durant lequel les consortiums travaillent à cerner et à définir les enjeux afin d'ébaucher une proposition normative avant de se tourner vers un organisme national ou international de normalisation pour élaborer une norme qui se fonde sur un projet complet. Non seulement une telle approche augmente les chances qu'une proposition de nouvel élément de travail menant à une norme soit approuvée, mais elle améliore également la rigueur de la rédaction et les chances que plusieurs parties prenantes s'engagent dans le projet.

Comme des organisations de la société civile participent, à l'échelle nationale ou internationale, à ces démarches de définition de normes (qui visent par exemple plus largement les consommateurs et consommatrices), le défi particulier des personnes LGBTI est de stimuler un tel élan de mobilisation. Sur la scène nationale, nous pourrions poser les questions constructives suivantes : 1) Quelle organisation, du point de vue de la société civile, veille à la prise en compte des points de vue des personnes LGBTI dans le travail des organismes nationaux de normalisation concernant l'IA ? et 2) Utilise-t-on, dans le cadre du processus d'établissement des normes, des documents qui reflètent de manière adéquate les recherches menées sur les avantages et les préjudices associés à l'IA pour les personnes LGBTI ?

Il importe d'enrichir le processus d'élaboration des normes des points de vue de personnes LGBTI. Le défi est donc triple : d'abord, façonner l'élaboration des normes en s'appuyant sur les expériences vécues. Ainsi, codifier le tout d'une manière qui soit à la fois sensible et généralisable – nous avons tous et toutes des droits fondamentaux (voir la figure 2). Finalement, nous devons encourager l'adoption ultérieure de normes internationales reconnues afin de garantir un changement marqué des pratiques, des normes et du comportement du marché qui en résulte.

| FIGURE 2 |

Façonner les normes internationales : un modèle d'organisation et de participation pour les parties prenantes. LGBTI.

Évaluer l'adéquation structurelle des lois et des règlements

Enfin, les gouvernements devraient, de manière proactive et consolidée, dresser un bilan des lois, des règlements et même des directives politiques contraignantes qui, en matière d'IA, empêchent les personnes LGBTI de jouir pleinement de leurs droits fondamentaux. Les objectifs explicites de cette démarche seraient de cerner les développements empiriques de l'IA, dans la mesure où ils ont une incidence sur les personnes LGBTI, et d'évaluer l'adéquation des lois actuelles et leur cohérence avec les normes relatives aux droits humains. Ces évaluations ne devraient pas se limiter à la législation et à la réglementation propres à l'IA ou centrées sur la technologie, mais plutôt se concentrer sur les *domaines* ou les secteurs de la vie publique dans lesquels les répercussions se font sentir, et où il y aurait des écarts connus ou probables. Elles pourraient notamment se concentrer sur les écarts entre les objectifs et les résultats déclarés,par exemple quant à la pleine égalité en droit, et l'état de la législation en vigueur en matière de prévention de la discrimination. Dans ce cas, les exemptions religieuses qui s'appliquent actuellement pourraient influer sur la manière dont les entreprises ou les organisations embauchent et licencient du personnel en toute légalité (Horner, 2017). Ces mêmes lois pourraient ensuite façonner le fonctionnement des algorithmes de sélection de candidats à un poste, en dépit de l'inquiétude qu'ils suscitent et de la désapprobation d'un nombre croissant de consommateurs et consommatrices quant à pareilles utilisations de l'IA. Une fois de plus, il ne s'agit pas de faire ressortir que les « mauvais côtés » de l'IA, mais de se pencher sur les risques matériels et les préjudices subis par les personnes LGBTI ou les membres de toute autre communauté, dans tous les domaines de la vie publique. Agir ainsi permettrait nécessairement d'explorer des questions et, par exemple, de se demander si l'amplification numérique de la discrimination et d'autres préjudices sociaux de ce genre exacerbe les inégalités structurelles et les vulnérabilités observées au sein de la population (Horner, 2017 ; Bourgois *et al.*, 2017 ; Metzl et Hansel, 2014).

Si elle est menée avec intégrité, cette approche comporte l'avantage de se recentrer sur les personnes touchées par l'IA et de mettre en cause l'impact disproportionné et discriminatoire qu'a celle-ci sur les membres de certains groupes de la population (Raji *et al.*, 2020). Elle prévoit également une analyse sociale et juridique complète et réfléchie d'exemples précis, de défis et de solutions potentielles, y compris au moyen d'une réforme du droit. À l'ère du « capitalisme réglementaire », où « des interventions peuvent émerger de partout, au sein des réseaux, puis par des mécanismes de diffusion, rapidement se mondialiser » (Drahos, 2017, p. 776), cette approche invite au dialogue, à l'analyse et à la transparence. Elle procure aux gouvernements, à la société civile et aux entreprises technologiques une compréhension affinée et contextuelle qui fonde globalement leurs politiques.

En Australie, la Commission des droits de la personne a adopté une approche similaire de 2018 à 2021, puis a publié, au début de 2021, un rapport final qui décrit une série de mesures, certaines volontaires, d'autres structurelles, visant à faire face aux enjeux qui ont été cernés en matière d'IA (Australian Human Rights Commission, 2021). Cette approche élargie peut tenir compte des personnes LGBTI en fonction d'attributs protégés et en tant que groupe d'intérêt touché par l'IA. Ancrer plus fermement cette approche différenciée, en se concentrant sur divers groupes de la population tels que les personnes LGBTI, devrait fournir des indications instructives pour les initiatives à venir liées au bilan.

De tels bilans ne doivent par ailleurs pas être dressés uniquement par les gouvernements, ministères et organismes publics nationaux ou ceux relevant des États ou des provinces. Des associations professionnelles, des organismes de quasi-réglementation et d'autres parties prenantes sont appelés à jouer un rôle pivot et à faire preuve de leadership. À titre d'exemple, la New Zealand Law Foundation a mené, en collaboration avec l'Université d'Otago, un large processus consultatif qui a abouti à une analyse détaillée des répercussions de l'IA sur les droits fondamentaux en Nouvelle-Zélande en se référant à la structure des droits humains adoptée par le pays (Gavaghan *et al.*, 2019). Cette démarche influe sur la façon dont la Nouvelle-Zélande gérera désormais les répercussions de l'IA, y compris pour les personnes LGBTI, dans des domaines comme la justice pénale (prévention de la criminalité) et le droit du travail.

CONCLUSION

Pour les personnes LGBTI, réduire les angles morts de l'IA, qui reflètent les préjugés, les aspirations, les craintes et les oublis des êtres humains qui conçoivent et « entraînent » les systèmes d'IA, nécessite d'envisager l'IA comme un cycle de vie. Repérer, remettre en question et pallier efficacement ces angles morts, qui se manifestent par des préjugés et une discrimination matérielle, exige une grande créativité et une volonté de tirer parti d'observations concrètes. Il s'agit notamment de tenir compte des diverses expériences vécues par les membres des communautés LGBTI touchées, expériences qui restent bien trop réelles pour nombre d'entre eux et elles, ainsi que de concepts et de méthodologies issus de disciplines qui s'intéressent à l'étude des préjudices tels que la discrimination, par exemple l'épidémiologie (Horner, 2017). À ce jour, il semble qu'on n'ait pas suffisamment tiré profit de ces manières de faire au sein du mouvement naissant de promotion d'une IA responsable.

Par le biais d'une activité structurée, réfléchie et concrète s'appuyant sur un travail qui est déjà en cours (Raji *et al.*, 2020 ; Australian Human Rights Commission, 2021 ; National Security Commission on Artificial Intelligence, 2021), il est possible de corriger la trajectoire de l'IA et ses répercussions sur les personnes LGBTI en tenant mieux compte des droits fondamentaux. Comme l'affirme Bowles (2018, p. 197), « [p]lutôt que de s'attaquer tête première au problème de la culture, mieux vaut se concentrer sur le changement concret. Le fait d'aller au-delà des besoins de l'utilisateur ou de l'utilisatrice et de l'entreprise et d'envisager la société en tant que partie prenante amène ceux et celles qui travaillent dans les technologies à apprécier la place qu'ils et elles occupent dans une large collectivité régie par un contrat social ».

Il faut pour ce faire recourir à ce que j'appelle une approche du « cycle de vie élargi », approche faisant intervenir, dans le contexte d'un ensemble plus large de mesures, les éléments suivants : 1) des audits sociaux, 2) l'adoption de normes internationales reconnues et 3) des bilans consolidés réguliers de la législation et de la réglementation. Chacune de ces dimensions devrait expressément intégrer les perspectives des personnes LGBTI, en tant que membres d'une communauté qui a historiquement été marginalisée. Leur mise en œuvre nécessitera des efforts continus, collectifs et coordonnés qui se matérialiseront par des mesures et des pratiques spécifiques. Ce sont des efforts que les membres des communautés LGBTI ont l'habitude de fournir et de soutenir. Pour reprendre des paroles inspirantes de Treichler (1999, p. 1) écrites en plein cœur d'une autre crise de santé publique qui a touché les personnes LGBTI et, plus particulièrement, les hommes gais :

> it is the careful examination of language and culture that enables us, as members of intersecting social constellations, to think carefully about ideas in the midst of a crisis : to use our intelligence and critical faculties to consider theoretical problems, develop policy, and articulate long term social needs, even as we acknowledge the urgency of the [...] crisis and try to satisfy its relentless demands for immediate action.

Voilà des aptitudes que nous possédons collectivement et que nous devons exploiter pour faire en sorte que l'IA responsable, en tant qu'ensemble évolutif de pratiques concrètes et réelles, efface les angles morts de l'IA tels qu'ils se manifestent pour les personnes LGBTI. Si nous voulons vraiment établir des relations égalitaires, voilà l'approche de référence qu'attendent les prochaines générations et même la nôtre.

RÉFÉRENCES

Albury, K., Burgess, J., Light, B., Race, K. et Wilken, R. 2017. Data cultures of mobile dating and hook-up apps : Emerging issues for critical social science research. *Big Data & Society,* vol. 4, nº 2, pp. 1-11.

Australian Human Rights Commission. 2021. *Human Rights and Technology : Final Report*. Sydney, Australian Human Rights Commission.

Bell, G., Gould, M., Martin, B., McLennan, A. et O'Brien, E. 2021. Do more data equal more truth ? Toward a cybernetic approach to data. *Australian Journal of Social Issues,* vol. 56, nº 2, pp. 213-222.

Bemthuis, R., Iacob, M.-E. et Havinga, P. 2020. A design of the resilient enterprise : A reference architecture for emergent behaviors control. *Sensors*, vol. 20, nº 22, p. 667.

Benjamin, R. 2019. *Race After Technology*. London, Polity Press.

Berlant, L. 2011. *Cruel Optimism.* Durham, N.C., Duke University Press.

Bourgois, P., Holmes, S., Sue, K. et Quesada, J. 2017. Structural vulnerability : Operationalizing the concept to address health disparities in clinical care. *Academic Medicine,* vol. 92, nº 3, pp. 299-307.

Bowles, C. 2018. *Future Ethics.* East Sussex : New Next Press.

Bratton, B. 2015. *The Stack : On Software and Sovereignty.* Cambridge, M.A. : MIT Press.

Bucher, T. 2016. The algorithmic imaginary : Exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society,* vol. 20, nº 1, pp. 30-44.

Cihon, P. 2019. *Standards for AI Governance : International Standards to Enable Global Co-ordination in AI Research & Development*. Oxford, Future of Humanity Institute (University of Oxford).

Commission Européenne. 2021. *Impact Assessment of the Regulation on Artificial Intelligence*. Bruxelles : Commission Européenne. https://digitalstrategy.ec.europa.eu/en/library/impact-assessment-regulation-artificial–intelligence

Drahos, P. 2017. Regulating capitalism's processes of destruction. P. Drahos (éd.), *Regulatory Theory : Foundations and Applications.* Canberra, The Australian National University, pp. 761-783.

Eubanks, V. 2018. *Automating Inequality : How High-Tech Tools Profile, Police, and Punish the Poor.* New York, St. Martin's Press.

First, R. 1973. The South African Connection : From Polaroid to Oppenheimer. *Issues : A Journal of Opinion,* vol. 3, nº 2, pp. 2-6.

Gavaghan, C., Knott, A., Maclaurin, J., Zerilli, J. et Liddicoat, J. 2019. *Government Use of Artificial Intelligence in New Zealand*. Wellington, New Zealand Law Foundation et University of Otago.

Golding-Young, S. 2020. *Facebook's Discrimination Against the LGBT Community*. ACLU. https://www.aclu.org/news/lgbtq-rights/facebooks-discrimination-against-the-lgbt-community/

Gray, K. R. et Karp, R., E. 1994. Corporate social responsibility : The Sullivan Principles and South Africa. *Visions in Leisure and Business,* vol. 12, nº 4, article 2.

Haut-Commissariat des Nations Unies aux droits de l'homme. 2019. *End of Mission Statement of the Special Rapporteur on Contemporary Forms of Racism, Racial Discrimination, Xenophobia and Related Intolerance at the Conclusion of Her Mission to the United Kingdom of Great Britain and Northern Ireland.* https://www.ohchr.org/en/NewsEvents/Pages/DisplayNews.aspx?NewsID=23073&LangID=E

Horner, D. B. 1971. *United States corporate investment and social change in South Africa*. Johannesburg, South African Institute of Race Relations.

Horner, J. 2017. Expanding the gaze : LGBTI people, discrimination and disadvantage in Australia. A. Durbach, B. Edgeworth, et V. Sentas (éd.), *Law and Poverty in Australia : 40 Years After the Poverty Commission.* Sydney, Federation Press, pp. 92-102.

House of Commons Women and Equalities Committee. 2018. *Race Disparity Audit : Third Report of Session 2017–19*. London, House of Commons.

ISO/IEC. 2020. *Draft International Standard 22989*. Geneva, ISO/IEC.

Kreiger, N. 2000. Discrimination and health. Berkman, L. F. et Kawachi, I. (éd.) *Social Epidemiology.* Oxford, Oxford University Press, pp. 36-75.

Kreiger, N. 2021. Structural racism, health inequities, and the two-edged sword of data : Structural problems require structural solutions. *Frontiers in Public Health,* vol. 9.

Laclau, E. 2005. Populism : What's in a name ? Panizza, F. (éd.) *Populism and the Mirror of Democracy.* London, Verso, pp. 32-49.

Lixinski, L. 2020. Rights litigation piggybacking : Legal mobilization strategies in LGBTIQ international human rights jurisprudence. *Florida Journal of International Law,* vol. 31, nº 3, pp. 273-314.

Lorde, A. 1980. Age, race, class, and sex : Women redefining difference. Rothenberg, P. S. (éd.) *Racism and Sexism : An Integrated Study.* New York, St. Martin's Press, pp. 352-359.

Marchant, G. E. 2011. Addressing the pacing problem. Marchant, G. E. A., Braden, R. et Herkert, J. R. (éd.) *The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight.* Dordrecht, Springer, pp. 199-205.

Maslen, A. T. 2019. *White for White : An Exploration of Gay Racism on the World's Most Popular Platform for Gay and Bisexual Men.* London, London School of Economics and Political Science.

Metcalf, J. et Crawford, K. 2016. Where are human subjects in Big Data research ? The emerging ethics divide. *Big Data & Society,* vol. 3., nº 1.

Metzl, J. M. and Hansel, H. 2014. Structural competency : Theorizing a new medical engagement with stigma and inequality. *Social Science & Medicine,* vol. 103, pp. 126-133.

Mittelstadt, B., Allo, P., Taddeo, M., Wachter, S. et Floridi, L. 2016. The Ethics of Algorithms : Mapping the Debate. *Big Data & Society*, vol. 3, nº 2, pp. 1-26.

Mouffe, C. 2005. *On The Political.* London, Routledge.

National Security Commission on Artificial Intelligence (NSCAI). 2021. *Final Report*. Washington, D.C., NSCAI.

Nolan, J. et Frishling, N. 2020. Human rights due diligence and the (over) reliance on social auditing in supply chains. Deva, S. et Birchall, D. (éd.), *Research Handbook on Human Rights and Business.* Cheltenham, Edward Elgar Publishing, pp. 108-129.

Pasquale, F. 2015. *The Black Box Society : The Secret Algorithms that Control Money and Information.* Cambridge, M.A. : Harvard University Press.

Partnership on AI. 2021. *ABOUT ML Resources Library*. https://partnershiponai.org/about-ml-resources-library/

PRISMA. 2020. *Guidelines to Innovate Responsibly : The PRISMA Roadmap to Integrate Responsible Research and Innovation (RRI) in Industrial Strategies.* Rome : Italian Association for Industrial Research.

Race, K. 2009. *Pleasure Consuming Medicine : The Queer Politics of Drugs.* Durham, N.C., Duke University Press.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson B., Smith-Loud, J., Theron, D. et Barnes, P. 2020. Closing the AI accountability gap : Defining an end-to-end framework for internal algorithmic auditing. Conference on Fairness, Accountability, and Transparency (FAT* 20'), 27-30 janvier Barcelona. New York, ACM.

Rice, C. et Zegart, A. 2018. *Political Risk : How Businesses and Organisations Can Anticipate Global Insecurity.* London, Weidenfeld & Nicolson.

Rudin, C. et Radin, J. 2019. Why are we using black box models in AI when we don't need to ? A lesson from an explainable AI competition, *Harvard Data Science Review,* vol. 1, nº 2.

Sedgwick, E. K. 2008. *Epistemology of the Closet.* Berkeley, University of California Press.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S. et Vertesi, J. Fairness and abstraction in sociotechnical systems. FAT* '19 : Conference on Fairness, Accountability, and Transparency (FAT* '19), 29-31 janvier Atlanta. New York, ACM.

Standards Australia. 2020. *An Artificial Intelligence Standards Roadmap : Making Australia's Voice Heard*. Sydney, Standards Australia.

Stewart, J. B. 2011. Amandla ! The Sullivan Principles and the battle to end apartheid in South Africa, 1975-87. *Journal of African American History,* vol. 96, nº 1, pp. 62-89.

Treichler, P. A. 1999. *How to Have Theory in an Epidemic : Cultural Chronicles of AIDS.* Durham, N.C., Duke University Press.

Wark, M. 2019. *Capital is Dead : Is This Something Worse ?* London, Verso.

Whittlestone, J., Nyrup, R., Alexandrova, A., Dihal, K. et Cave, S. 2019. *Ethical and societal implications of algorithms, data, and artificial intelligence : A roadmap for research*. London, Nuffield Foundation.

INNOVATION INCLUSIVE EN MATIÈRE D'INTELLIGENCE ARTIFICIELLE : DE LA FRAGMENTATION À L'UNITÉ

ÉLIANE UBALIJORO

Directrice générale, Sustainability in the Digital Age, et directrice du pôle canadien de Future Earth, Montréal, Canada.

GUYLAINE POISSON

Professeure agrégée, Département des sciences de l'information et de l'informatique, Université d'Hawaï à Mānoa, Honolulu, Hawaï, États-Unis.

NAHLA CURRAN

Étudiante de premier cycle, Département d'économie, de philosophie et de science politique, Université de la Colombie-Britannique – Okanagan, Kelowna, Canada.

KYUNGIM BAEK

Professeure agrégée, Département des sciences de l'information et de l'informatique, Université d'Hawaï à Mānoa, Honolulu, Hawaï, États-Unis.

NILUFAR SABET-KASSOUF

Responsable des programmes stratégiques, Sustainability in the Digital Age et Future Earth, Montréal, Canada.

MÉLISANDE TENG

Étudiante au doctorat, Mila et Université de Montréal, et stagiaire LEADS, Future Earth, Montréal, Canada.

ODD 5 - Égalité entre les sexes

ODD 9 - Industrie, innovation et infrastructure

ODD 10 - Inégalités réduites

ODD 11 - Villes et communautés durables

ODD 16 - Paix, justice et institutions efficaces

ODD 17 - Partenariats pour la réalisation des objectifs

INNOVATION INCLUSIVE EN MATIÈRE D'INTELLIGENCE ARTIFICIELLE : DE LA FRAGMENTATION À L'UNITÉ

RÉSUMÉ

L'intelligence artificielle (IA) façonne l'avenir de l'humanité. Mais qu'arrive-t-il quand seulement une partie de la société est présente à la table autour de laquelle cet avenir se définit ? L'innovation a tendance à se centrer sur le profit et la croissance, et à tout simplement reléguer l'inclusivité en marge. Il en résulte une ère numérique fragmentée où se multiplient les préjugés, les disparités et les inégalités, et qui commande des compromis sociaux relativement au bien-être général ainsi que des compromis technologiques quant à la performance et à la fiabilité de l'IA. Dans ce chapitre, nous examinons le rôle que jouent, au moment de définir des politiques en matière d'IA et de susciter la mobilisation de parties prenantes, le fossé numérique, le manque de diversité et de représentation dans le domaine de l'IA et celui des sciences, de la technologie, de l'ingénierie et des mathématiques (STIM) ainsi que l'incidence de l'innovation sur le financement de la recherche et ce qui la motive au sein du milieu universitaire, du gouvernement ou de l'industrie. Nous envisageons les conséquences qu'il y a à travailler en vase clos en vue d'améliorer la diversité et l'inclusion et examinons l'insuffisance d'une telle approche à susciter un changement systémique. Nous soutenons qu'il faut changer notre conception de l'innovation et envisager une innovation inclusive, et expliquons comment entraîner un tel changement. Inscrire l'inclusivité au cœur de l'avenir que nous façonnons au moyen de la technologie nous permettra de passer d'une ère numérique fragmentée à une unité qui profitera à tous et à toutes.

INTRODUCTION

L'intelligence artificielle (IA) est un moteur d'innovation dans une large gamme de secteurs et d'industries qui présentent différents besoins et doivent résoudre divers problèmes. En tant que telle, elle ne devrait pas évoluer en vase clos dans le seul domaine de la technologie ni d'aucune autre discipline. La conception, la mise au point, le déploiement et l'évaluation de technologies telles que l'intelligence artificielle sont complexes et nécessitent une interdisciplinarité. Mais qu'est-ce que l'innovation ? Il importe de répondre à cette question pour comprendre ce qui motive ou oriente actuellement l'innovation technologique.

Les innovations découlant de technologies et d'applications basées sur l'IA transforment rapidement de nombreux aspects de nos vies. Malheureusement, ce n'est pas toujours pour le bien de l'humanité. Dans ce chapitre, nous soutenons que les conséquences involontaires d'une IA inadaptée (empreinte de disparités et d'inégalités, et touchée par un nombre croissant de préjugés) peuvent être abordées en passant de l'idée d'« innovation » à celle d'« innovation inclusive ». L'innovation inclusive fait référence aux « moyens par lesquels de nouveaux biens et services sont offerts pour et/ou par ceux et celles qui ont été exclus du courant dominant du développement, en particulier les milliards d'individus vivant avec les revenus les plus faibles », ce qui élargit et diversifie finalement l'éventail de parties prenantes (Heeks *et al.*, 2013, p. 1). Ce changement de mentalité devrait s'appliquer à toutes les étapes de la mise au point et du déploiement de technologies basées sur l'IA, ainsi que dans l'élaboration des politiques et l'amélioration des systèmes. Les innovations en matière d'IA devraient être intrinsèquement inclusives et interdisciplinaires. Selon D^re^ Katia Walsh (citée dans Larsen, 2021), « l'intelligence artificielle résulte de l'intelligence humaine, a été rendue possible par ses vastes talents et est également sujette à ses limites. Par conséquent, il est impératif que toutes les équipes qui travaillent dans la technologie et l'IA soient aussi diversifiées que possible ». La mesure dans laquelle l'IA peut vraiment profiter à la planète entière et même au-delà est liée au degré de diversité dont elle tient compte.

Dans ce chapitre, nous examinons certains des principaux obstacles à une innovation véritablement inclusive et les mesures nécessaires pour garantir que personne ne soit laissé pour compte. Ces obstacles comprennent les écarts et les préjugés inhérents qui touchent actuellement l'IA, la manière d'en financer les projets, ainsi que les motifs qui sous-tendent les investissements et orientent l'IA. L'incidence d'une technologie qui ne serait pas suffisamment inclusive ainsi que le manque d'efforts pour en pallier les causes sont largement sous-estimés et constituent un obstacle à la mise au point d'une IA qui profite à tous et à toutes. Les technologies de l'IA façonnent l'avenir de l'humanité, et une réflexion approfondie s'impose à ce sujet si nous souhaitons progresser convenablement.

QUÊTE D'INNOVATION

Depuis qu'Alan Turing (1950) a expliqué comment construire et tester des machines intelligentes et que l'expression « intelligence artificielle » a été adoptée en 1956 (McCarthy *et al.*, 1955), des succès et des revers ont parsemé les soixante-dix ans d'histoire moderne de l'IA. L'engouement dont celle-ci fait l'objet depuis une dizaine d'années a été stimulé par l'accès à de grandes quantités de données, des ordinateurs abordables et rapides, et le développement de techniques d'apprentissage automatique, en particulier l'apprentissage profond. De nos jours, l'IA a imprégné de nombreux aspects de nos vies quotidiennes, des fils d'actualité des réseaux sociaux aux achats en ligne, en passant par la découverte de médicaments (Fleming, 2018 ; Jiménez-Luna *et al.*, 2021 ; Lada *et al.*, 2021) et la lutte contre les épidémies (Cho, 2020 ; Zeng *et al.*, 2021). L'IA est l'une des forces majeures qui révolutionnent la société humaine et elle entraîne dans son sillage une nouvelle « ère numérique ».

Malheureusement, ces avancées ont créé une fracture mondiale d'un nouveau type, entre les riches et les pauvres en technologie. Les progrès rapides de l'IA ont creusé et élargi le fossé numérique et amplifié les *a priori* existants, par exemple, quant au niveau d'instruction, au genre, à la race, et à la distinction entre pays ou populations riches et pauvres (Carter *et al.*, 2020). Il est intéressant de noter que les préjugés et les écarts observés actuellement dans l'IA reflètent en partie ceux qui ont tourmenté nos sociétés pendant des siècles. Alors que les avancées technologiques ont souvent amplifié le colonialisme au fil de l'histoire, il existe un réel danger d'aller vers de nouvelles formes de colonialisme renforcées par les technologies numériques et par ce qui propulse actuellement l'innovation dans ce domaine. Le colonialisme numérique se produit lorsque « des entreprises technologiques d'envergure extraient et analysent les données des utilisateurs et utilisatrices et qu'elles s'en servent à des fins de profit et d'influence sur le marché, obtenant un gain non négligeable pour avoir été la source des données » (Coleman, 2019, p. 417). C'est « l'exercice d'une autorité impériale dans la structure de l'écosystème numérique, logiciels, matériel électronique et connectivité de réseau, qui donne alors lieu à des formes de domination connexes » (Kwet, 2019, p. 1). Prenons à titre d'exemple la division du travail (celui de « travailleurs et travailleuses invisibles de l'IA », ce personnel chargé d'annoter des données, souvent issu de communautés moins privilégiées, et endurant l'isolement et des conditions de travail souvent difficiles) (Gray et Suri, 2019) et les préjugés dans les données utilisées pour entraîner les systèmes d'IA. Au fur et à mesure que nous concevons, multiplions et rassemblons les données qui alimentent les machines afin qu'elles apprennent, nous transférons inévitablement nos préjugés à l'IA. S'attaquer aux enjeux jusque-là négligés du développement et des politiques en matière d'IA servira non seulement à améliorer la technologie elle-même, mais également à lutter contre les préjugés et les écarts systémiques actuels et futurs (Li, 2020).

Il nous semble manifeste que, selon la manière dont nous envisageons les technologies basées sur l'IA, deux avenues s'ouvrent à nous : soit, par inadvertance, de nouvelles formes de colonialisme se perpétueront dans l'ère numérique (Voskoboynik, 2018), soit l'humanité pourra aller de l'avant, d'une manière inclusive, dans la poursuite d'un objectif commun consistant à régler les problèmes mondiaux actuels et à proposer de manière concertée des innovations ayant des retombées bénéfiques. Comment pouvons-nous nous assurer de suivre la bonne voie ?

REDÉFINITION DES PARTIES PRENANTES DE L'IA

L'une des clés du succès des chercheurs et chercheuses du monde universitaire est d'obtenir du financement pour leurs recherches, et de publier dans des revues de premier plan et des actes de conférences. Pour y parvenir, il est conseillé aux personnes dont la carrière débute de donner la priorité à la recherche innovante dans leurs travaux. Mais qu'est-ce que l'innovation ? Lors de l'évaluation des propositions de recherche, les agences de financement envisagent l'innovation en tant qu'« activités et concepts créatifs, originaux et transformateurs » ou que « méthodes, approches, technologies de pointe ou concepts uniques et innovants » (National Science Foundation [NSF] and National Aeronautics and Space Administration [NASA] comme cité dans Falk-Krzesinski et Tobin, 2015, p. 15). Vu les occasions de financement limitées, un projet doit souvent montrer des avancées notables dans le domaine pour être considéré comme innovant. Cela signifie souvent qu'il faut s'efforcer d'employer la méthode la plus récente et la plus rapide qui soit et nécessitant de nombreuses ressources : technologies, algorithmes et systèmes dotés d'une puissance de calcul élevée, importante capacité de stockage, Internet rapide et fiable ou accès cellulaire (Thompson *et al.*, 2020, p. 2).

Ces contraintes réduisent la portée de l'IA et sa capacité de répondre aux besoins mondiaux qui, à l'heure actuelle, se limitent principalement à ce qui engendre des profits dans les pays du Nord. En plus des problèmes inhérents qu'elles posent pour une société juste et équitable, elles représentent également

des défis techniques dans la mise au point d'une « IA digne de confiance et vérifiable » (Dengel *et al.*, 2021, p. 91) qui est adaptable en contexte de ressources limitées. Comme l'indiquent Dengel et autres (2021, p. 93),

> current research evaluation methods and academic criteria tend to favor vertical, short-term, narrow, highly focused, community – and discipline-dependent research. It is the responsibility of all scientists in the academic world to foster a methodological shift that facilitates (or at least does not penalize) long-term, horizontal, interdisciplinary, and very ambitious research.

Cette optique s'applique également à l'industrie. Tel qu'il a été mentionné lors de la Conférence des Nations Unies sur le commerce et le développement à propos de la technologie et de l'innovation : « Comme pour toute nouvelle technologie, de nombreuses entreprises, lorsqu'elles innovent et produisent de nouveaux biens et services, ont tendance à se concentrer sur les consommateurs et consommatrices à revenu élevé qui peuvent supporter le coût initial élevé de ces produits » (UNCTAD, 2021, p. 125). Malheureusement, on néglige souvent la pertinence qu'auraient ces nouvelles technologies pour les pays en développement (’Utoikamanu, 2018).

Pour apporter les changements nécessaires dans la recherche et le développement, nous devons inclure toutes les parties prenantes qui sont touchées par les innovations en IA, pas seulement celles qui en bénéficient actuellement. De cette façon, nous pouvons percevoir la créativité d'une manière élargie qui comprend l'adaptabilité des techniques et leurs nouvelles applications.

Afin d'élargir les capacités de l'IA et de passer à des innovations plus inclusives, nous devons d'abord saisir les préjugés que les normes actuelles en matière d'innovation entraînent et perpétuent à la fois.

Fossé numérique et innovation inclusive

L'Organisation de coopération et de développement économiques définit le fossé numérique comme « l'écart qui sépare des individus, des ménages, des entreprises et des zones géographiques de différents niveaux socioéconomiques en ce qui a trait tant à l'accès aux technologies de l'information et de la communication (TIC) qu'à l'utilisation d'Internet dans une grande variété d'activités » (OCDE, 2001, p. 5).

Le fossé se creuse davantage entre le Nord et le Sud. Par exemple, en 2018, 80 % de la population en Europe utilisait Internet, contre seulement 25 % de la population en Afrique subsaharienne (UNCTAD, 2021, p. 78). Les ressources financières et technologiques se concentrent principalement dans les pays du Nord ou y sont acheminées, ce qui exclut souvent les parties prenantes du Sud de la scène mondiale de la recherche scientifique et de l'innovation (Chan *et al.*, 2021 ; Garcia, 2021 ; Mishra, 2021 ; Reidpath et Allotey, 2019 ; Skupien et Rüffin, 2020). De plus, des écarts importants s'observent au sein des pays. La pandémie de COVID-19 a d'ailleurs attiré l'attention sur ces écarts alors que le monde passait en mode virtuel pour le travail, les achats, les services de santé et l'éducation, ce qui nécessite un ordinateur et une connexion Internet (United Nations, 2020). Bien qu'il s'agisse là d'un problème général et que cet exemple particulier soit récent, dans le domaine de la recherche et de l'éducation en matière d'IA, le manque de ressources causé par l'écart technologique est un problème à régler de toute urgence.

Le fossé technologique contribue de manière notable au manque de diversité des innovations en IA ou, pourrions-nous dire, il néglige ou sous-estime certaines grandes innovations. Comme ce domaine et son financement dépendent de la technologie de pointe, l'écart favorise l'étude et la recherche menées par des personnes issues de milieux socioéconomiques privilégiés (American University, 2020). Celles qui n'ont pas les moyens d'acheter un ordinateur ou dont l'accès Internet est lent ou inexistant sont mises de côté, ce qui a un effet sur le domaine dans son ensemble. Lorsque la grande majorité de la recherche en technologies basées sur l'IA est le fait de personnes provenant de milieux semblables, le reste du

monde en est évincé. La technologie est souvent conçue et améliorée par des scientifiques et pour une partie de la population de leur pays, d'une région du monde. Il en résulte une inégalité que viendra alimenter la prochaine innovation technologique.

Pour réduire le fossé technologique, nous devons reconsidérer les critères d'évaluation de la qualité de l'innovation en matière d'IA, et l'innovation inclusive s'avère primordiale. Ainsi, on favoriserait de meilleures technologies basées sur l'IA adaptées à différentes communautés et on élargirait le bassin des parties prenantes. Alors que nous examinons comment les grands défis de ce siècle touchent de manière disproportionnée les personnes marginalisées, il est essentiel d'amener ces dernières à la pointe de l'innovation pour mettre l'innovation technologique au service de l'humanité et de la planète.

Diversité et représentation en IA et dans les STIM

Tant que l'inclusion sera perçue comme un « acte de charité » ou comme le fait d'accepter de revoir à la baisse nos ambitions de faire progresser la technologie, nous nous retrouverons avec des technologies tendancieuses. L'innovation inclusive doit être vue pour ce qu'elle est : un moyen d'aspirer à une technologie inclusive et équilibrée qui bénéficiera à tous et à toutes. Si nous voulons des machines capables de résoudre des problèmes complexes, nous devons les exposer à une grande variété de données. Cela signifie que des personnes possédant une expertise et des expériences diverses doivent participer à toutes les opérations du processus de développement de l'IA et des technologies basées sur l'IA : acquisition de données pour entraîner les systèmes d'IA, conception, mise au point, déploiement, exploitation, surveillance et maintenance. Une telle diversité devrait également caractériser tout ce qui concerne les politiques ou la prise de décisions liées à l'IA. Un manque de diversité et de représentation de toutes les parties prenantes pose le risque d'une omission par ignorance (pas nécessairement intentionnelle), ce qui rend le problème difficile à régler (*Coded Bias*, 2020).

Les enjeux de la diversité et de la représentation ne sont bien sûr pas inhérents à l'IA. Historiquement, le domaine des sciences, de la technologie, de l'ingénierie et des mathématiques (STIM) a eu une base à prédominance masculine blanche (Dancy *et al.*, 2020, p. 1). La marginalisation dans les domaines des STIM touche indéniablement de nombreuses communautés, y compris les peuples autochtones, les personnes handicapées et les communautés LGBTI (Miller et Downey, 2020 ; Schneiderwind et Johnson, 2020). Dans ce chapitre, nous observerons plus particulièrement les préjugés sexistes, racistes et socioéconomiques.

| TABLEAU 1 |

Pourcentage de personnes employées aux États-Unis dans des professions informatiques et mathématiques (Bureau of Labor Statistics, 2010 ; 2020).

	2010	2020
Femmes	25,8	25,2
Hommes	74,2*	74,8*
Personnes blanches	77,2*	65,4
Personnes noires ou afro-américaines	6,7	9,1
Asiatiques	16,1	23,0

* Pourcentage estimé

Préjugés sexistes

Le tableau 1 montre une répartition professionnelle inégale dans le domaine de l'informatique et des mathématiques. Bien qu'il y ait généralement eu de légères améliorations en une décennie, la tendance n'est guère encourageante. Sans surprise, les femmes ne représentent, au sein des métiers de l'informatique et des mathématiques, que le quart du personnel et, fait alarmant, leur représentation a légèrement diminué au cours des dix dernières années. Bien que nous puissions attribuer ces mauvais résultats aux organisations axées sur la science qui n'embauchent pas autant de femmes que d'hommes (*Picture a Scientist*, 2020), l'écart de genre dans ce domaine commence bien plus tôt. Dans l'enfance, les filles font face à des stéréotypes relatifs aux STIM provenant de leurs parents, des normes sociales et des enseignants et enseignantes, ce qui en décourage beaucoup de s'intéresser à ce domaine (Hill, 2020). Les filles qui le font ou qui réussissent bien en mathématiques ou en sciences ne poursuivent pas toujours une carrière dans les STIM parce qu'elles pensent que ces professions sont « inappropriées à leur genre » (Hill *et al.*, 2010, p. 22). Les facteurs observés dès un jeune âge démotivent de nombreuses femmes et filles très tôt. Les graves préjugés sexistes qui circulent par ailleurs dans les milieux de travail peuvent amener les femmes à délaisser une carrière dans les STIM. Ces préjugés ont trait, sans s'y limiter, à l'environnement de travail, aux responsabilités familiales et aux préjugés implicites (Hill *et al.*, 2010, pp. 24-25).

Les préjugés inconscients envers les femmes peuvent représenter un obstacle majeur à la réussite et à l'avancement professionnel, et même motiver la décision de quitter le métier. Les lettres de recommandation rédigées pour des femmes illustrent par exemple les conséquences de tels préjugés. Les traits de personnalité y sont souvent relevés plutôt que l'expertise technique (Trix et Psenka, 2003, p. 215). De telles manifestations des préjugés inconscients réduisent la participation des femmes dans la conception de technologies basées sur l'IA et leur présence dans le monde scientifique en tant que décideuses politiques engagées. De plus, les femmes qui connaissent du succès dans leur domaine sont plus méprisées et moins appréciées que les hommes qui réussissent, ce qui rend l'environnement de travail malsain et difficile à changer. Dans les STIM relevant du secteur privé, les femmes quittent leur emploi en raison de l'incertitude quant à leurs chances d'avancement, du sentiment d'isolement, d'un environnement peu favorable et d'un horaire de travail intense (Hill *et al.*, 2010, p. 24). N'ayant pas d'occasions d'avancement et étant victimes de microagressions constantes, les femmes n'ont aucune raison de rester dans un environnement où on essaie de les pousser vers la sortie.

En ce qui concerne l'état civil et les responsabilités familiales, il existe également des différences évidentes entre les hommes et les femmes. Dans les facultés universitaires de STIM, les femmes célibataires sont plus susceptibles d'occuper un poste menant à la permanence que celles qui sont mariées. En ce qui a trait aux enfants, les femmes s'abstiendraient d'en avoir ou retarderaient la maternité en raison d'exigences que le domaine leur impose et de l'approche traditionnelle selon laquelle il revient principalement à la femme de s'occuper des enfants (Hill *et al.*, 2010, p. 26). De plus, une étude sur le maintien en fonction du personnel en génie a révélé que les femmes étaient plus susceptibles de délaisser le travail que les hommes en raison de problèmes de conciliation travail-famille (Frehill *et al.*, 2008). Ces facteurs de discrimination fondés sur le genre contribuent tous au bas nombre de candidatures féminines et au faible maintien en fonction des femmes dans les STIM.

Préjugés racistes

De 2010 à 2020, il y avait une tendance à la hausse du nombre de personnes non blanches employées dans des professions du domaine de l'informatique et des mathématiques. Les facteurs qui contribuent à une faible présence de ces personnes dans le domaine sont semblables à ceux qui expliquent l'écart de genre. Dans cette section, nous nous concentrerons sur la sous-représentation des personnes noires, asiatiques, et hispaniques ou latino-américaines dans les STIM. Dès un jeune âge, les préjugés inconscients influent sur la décision d'un étudiant ou d'une étudiante de poursuivre ses études ou de les abandonner.

Une étude a révélé que les élèves noirs et noires à faible revenu auxquels au moins une personne noire a enseigné en 3e, 4e ou 5e année sont 29 % moins susceptibles d'abandonner l'école secondaire (Dodge, 2018). Au niveau secondaire, années où les STIM sont généralement présentées, les jeunes peuvent commencer à s'y intéresser avant d'aller au collège ; ces chances ne sont pas égales pour tous et toutes. Une étude de Teach for America a révélé qu'« une école sur quatre [aux États-Unis] propose des cours d'informatique » (Dodge, 2018). En règle générale, les écoles des quartiers aisés, qui ont une population étudiante majoritairement blanche, offrent d'aborder l'informatique, mais les jeunes issus de minorités ou à faible revenu n'ont pas cette chance. Sans y avoir été familiarisés au préalable à l'école, les jeunes cultivent difficilement de l'intérêt pour cette matière et croient peu en leurs chances d'étudier au collège dans un domaine perçu comme exigeant un talent inné (Leslie *et al.*, 2015 ; Miller, 2017 ; Riegle-Crumb *et al.*, 2019). Ils ont aussi souvent un faible sentiment d'identité et d'appartenance à la culture de l'« informaticien typique » (Metcalf *et al.*, 2018, p. 613). Cet état de fait perpétue la croyance selon laquelle les personnes issues de minorités (tout comme les femmes) ne sont pas destinées à faire carrière dans les STIM, malgré l'intérêt qu'elles y portent (Dodge, 2018). Nous pouvons observer les conséquences de tels préjugés dans le système éducatif (aux États-Unis, par exemple) en examinant le faible pourcentage de personnes noires, asiatiques et hispaniques se joignant au marché du travail dans les STIM (Barber *et al.*, 2020 ; Clark et Hurd, 2020). Le milieu de travail lui-même peut devenir un autre champ de bataille quand les défis présents dans le système éducatif ont été surmontés. Le racisme et les préjugés touchant les STIM nuisent considérablement à la diversité dans le domaine (McGee et Bentley, 2017 ; McGee, 2020). À San Francisco, par exemple, 60 % des personnes noires et 42 % des Asiatiques et des Hispaniques travaillant dans les STIM subissent de la discrimination raciale (Dodge, 2018). Celle-ci ne prend pas toujours la forme d'un discours haineux. Comme les femmes, les personnes de minorités ethniques subissent des écarts salariaux et des microagressions, ne se voient pas offrir des promotions ou des projets importants, et moins de valeur est accordée à leur travail (Dodge, 2018). Ces facteurs contribuent tous à créer un milieu de travail néfaste qui non seulement nuit aux groupes minoritaires, mais diminue l'intérêt que ces derniers portent au domaine et conduit des personnes à le quitter complètement (Dodge, 2018). Par conséquent, attirer et retenir davantage de populations minoritaires dans l'enseignement des STIM est une première étape nécessaire pour atténuer les *a priori* dans l'IA.

Nous avons précédemment fait ressortir deux écarts majeurs qui surviennent tôt dans le système éducatif, à savoir les préjugés inconscients et le manque d'accès aux cours d'informatique pour les enfants issus de minorités. Le personnel éducatif est susceptible d'avoir des préjugés inconscients, et il ne faudrait pas sous-estimer son rôle dans le fait que les enfants s'intéressent ou non aux STIM (Bushweller, 2021). Il est donc important de prévoir une formation appropriée sur les préjugés tôt dans le parcours éducatif, car les écarts qui y naissent creusent ceux qu'on observe plus tard dans les STIM (Warikoo *et al.*, 2016).

En ce qui concerne le manque de cours d'informatique, une solution possible à ce problème consiste à soutenir les organisations à but non lucratif offrant des programmes d'informatique, idéalement celles qui sont dirigées par des personnes de minorités ethniques. L'embauche d'enseignants et enseignantes, le don de technologies à jour et la recherche d'un environnement approprié sont des composantes essentielles de la réussite. Le fait qu'il y ait des responsables issus de minorités est un avantage ; les jeunes réussissent mieux lorsqu'ils et elles apprennent de personnes ayant un parcours semblable au leur (Rosen, 2018). L'intégration de programmes parascolaires au sein de la communauté, des programmes menés par la communauté, représente également un moyen d'encourager les cours d'informatique dans le cheminement scolaire en faisant augmenter la demande générale à cet égard. Il s'agit d'une piste de solution à ne pas négliger pouvant mener à l'innovation inclusive en IA.

Préjugés socioéconomiques

Enfin, un autre préjugé important dans le domaine a trait au statut socioéconomique. Une étude de Yale a révélé que la façon dont un individu prononce certains mots est révélatrice de son statut social (Cummings, 2019). Bien qu'une prononciation distincte ne soit pas un problème majeur en soi, le statut socioéconomique d'une personne peut influer sur une décision d'embauche. La même étude portant sur 274 « individus ayant une expérience de l'embauche » a révélé que, sans aucune information sur les compétences, ces responsables estimeraient que les candidats et candidates de statut socioéconomique élevé sont plus aptes à occuper le poste que ceux et celles de statut modeste (Cummings, 2019), et qu'ils et elles bénéficieraient d'un meilleur salaire et de chances accrues d'obtenir des primes.

Cette idée préconçue est généralisable à l'ensemble de la main-d'œuvre. Cependant, si nous revenons aux préjugés racistes dans les STIM, nous observons qu'il y a une intersection entre race et revenu, et que le genre recoupe également ces deux facteurs. Aux États-Unis, la population de nombreux quartiers défavorisés est fortement composée de groupes ethniques minoritaires, plus particulièrement de personnes noires ou latino-américaines. Cela découle d'une longue histoire de discrimination qui a ségrégué et ghettoïsé ces groupes ethniques, ainsi qu'à un financement réduit au strict minimum (Firebaugh et Acciai, 2016, p. 13372). Il en résulte des écoles mal financées et un accès limité aux emplois. Que des employeurs privilégient les candidats et candidates à revenu élevé s'ajoute au fait que le statut socioéconomique d'une personne influe dès son plus jeune âge sur son avenir à long terme. Il n'y a aucune obligation pour les entreprises d'embaucher un certain pourcentage de leurs recrues dans les quartiers défavorisés. Sans diversité socioéconomique dans les STIM, il y a peu de représentation d'une partie importante de la population et la technologie mise au point pour aider les populations de ces quartiers ne sera teintée que de l'optique de personnes à revenu élevé.

Il est particulièrement important de tenir compte de la diversité dans les STIM pour résoudre certains des principaux défis actuels de l'IA. L'un de ces défis concerne la « pénurie de talents », c'est-à-dire « le manque d'experts et d'expertes hautement qualifiés dans la construction de systèmes d'IA » (Dengel *et al.*, 2021, p. 93). Comme nous l'avons vu, une partie importante de la population est actuellement exclue du développement et ne peut offrir son talent et son expertise en IA en raison de préjugés systémiques (quant au sexe, à la race, au statut socioéconomique) même dans les pays qui sont sur le côté technologiquement riche du fossé numérique. Un autre défi majeur concerne l'efficacité des systèmes d'IA au regard de la représentativité insuffisante des données qui les alimentent (Kuhlman *et al.*, 2020).

Les humains nourrissent un algorithme « vierge » à partir de leurs expériences limitées et de leurs préjugés et, petit à petit, cet algorithme apprend à reproduire des comportements en conséquence. En fin de compte, on obtient une technologie peu fiable sans qu'il y ait faute de sa part. Elle n'a fait que son devoir : apprendre et reproduire ce qu'on lui a appris.

Biais algorithmique appliqué

Le manque de ressources et le fossé technologique qui sont à l'origine du manque de diversité dans la recherche posent particulièrement problème quand il est question des préjugés acquis par les technologies basées sur l'IA. De tels biais algorithmiques se manifestent dans des applications diverses, notamment dans des technologies de reconnaissance faciale ou des outils servant à embaucher du personnel. Dans le documentaire *Coded Bias*, Joy Buolamwini, chercheuse au MIT Media Lab, a découvert que l'IA de son projet Aspire Mirror – un « dispositif qui vous permet de vous regarder et de voir une réflexion sur votre visage en fonction de ce qui vous inspire ou de ce avec quoi vous espérez sympathiser » basé sur un logiciel de détection de visage[60] – ne reconnaissait pas son visage de femme noire (*Coded Bias*, 2020).

60. Pour plus d'information sur le projet *Aspire Mirror*, voir : : http://www.aspiremirror.com/

Elle a dû mettre un masque blanc pour que la machine détecte son visage. Cela pourrait sembler une simple erreur ou un bogue du logiciel, mais cette technologie a déjà de réelles applications, et l'expérience de Buolamwini a été reproduite mille fois.

L'un des usages les plus courants des technologies basées sur l'IA vise la surveillance et la sécurité, généralement la reconnaissance faciale. *Coded Bias* explore cette question en détail. Buolamwini y explique que, puisque l'algorithme de reconnaissance faciale est programmé par des hommes blancs, il est alimenté de visages blancs et masculins. Après avoir mentionné ce problème à des entreprises comme Microsoft et IBM, Buolamwini a constaté qu'IBM avait amélioré la précision de son algorithme pour qu'il reconnaisse non seulement la couleur de la peau, mais aussi le genre, comme le montre le tableau 2.

| TABLEAU 2 |

Degré de précision de l'algorithme d'IBM, en 2017 et en 2018 (Buolamwini, 2019).

	2017	2018
Couleur de peau et genre		
Foncée/masculin	88,0 %	99,4 %
Claire/masculin	99,7 %	99,7 %
Foncée/féminin	65,3 %	83,5 %
Claire/féminin	92,9 %	97,6 %

En juin 2020, M. Williams, un Noir du Michigan, a été arrêté pour vol après une reconnaissance faciale du voleur (Hill, 2020). Les policiers ayant eu confiance dans l'algorithme, ils l'ont arrêté sans effectuer de vérifications au préalable (par exemple, vérifier son alibi, interroger des témoins, etc.). Ce dernier a ensuite été libéré et il y a eu abandon des accusations, mais l'erreur commise par l'algorithme et le mauvais travail des policiers auraient pu coûter la vie à M. Williams[61]. Étant donné la surabondance de caméras en place, le recours à la reconnaissance faciale comme outil de surveillance devient lentement une réalité, et l'identification erronée et la poursuite de personnes innocentes peuvent monter en flèche (Raji *et al.*, 2020). Dans le même ordre d'idée que les préjugés présents dans les activités de maintien de l'ordre et de sécurité, la technologie basée sur l'IA affecte par ailleurs inégalement des policiers et policières à certaines communautés (Heaven, 2020). Il y a toujours eu une surveillance excessive des communautés non blanches, des secteurs dits « ghettoïsés ». Un algorithme, par exemple celui que décrit l'article d'Osoba et Welser (2017), apprendra en se fondant sur ces données historiques sur

61. Un incident similaire s'est produit en 2017. Un travailleur palestinien a été arrêté à tort à cause de la traduction automatique de Facebook, qui a rendu son « bonjour », écrit en arabe, par « attaquez-les » en hébreu et « faites-leur du mal » en anglais. Voir Berger (2017).

l'affectation des ressources policières. Il apprendra à accroître la vigilance dans les secteurs où la tendance en matière de criminalité semble plus élevée, et conduira à une répartition inéquitable de l'effectif policier, qui mènera à son tour à une « criminalisation » inéquitable (Osoba et Welser IV, 2017, pp. 14-15). Cela entraînera une augmentation du nombre de personnes issues de minorités emprisonnées en raison de délits mineurs, comme la possession de marijuana, les excès de vitesse ou la situation d'itinérance, ce qui amplifie les préjugés inhérents au système (Heaven, 2020 ; O'Donnell, 2019). Ne pas corriger ces préjugés les renforcera au sein des systèmes d'IA, qui continueront alors à affecter l'effectif policier de manière inégale dans les communautés marginalisées.

Une quantité substantielle de biais algorithmiques marque également le processus de recrutement en IA et en informatique. L'excès de confiance dans les algorithmes creusera les écarts créés par les préjugés à l'embauche, ce qui se produit souvent à l'insu des employeurs (Hickok, 2020). Comme Bogen (2019) l'explique dans son article, l'IA intervient à plusieurs égards dans l'embauche, avant même la réception de candidatures. Les offres d'emploi ciblées faites dans Facebook, LinkedIn et Indeed contribuent à renforcer les préjugés racistes et sexistes en déterminant « qui est le plus susceptible de cliquer sur l'annonce » (Bogen, 2019). Une étude conjointe de l'Université Northeastern et de l'Université de Californie du Sud s'est penchée sur la diffusion asymétrique des offres d'emploi dans Facebook. Par exemple, dans les cas extrêmes, les emplois de caissier ou caissière « atteindraient un public à 85 % féminin », tandis que les postes dans les entreprises de taxi « atteindraient un public composé de 75 % de personnes noires », même si les employeurs offrent les postes à tous les groupes démographiques (Ali *et al.*, 2019, p. 4). C'est que l'algorithme a retenu la préférence des responsables du recrutement à l'égard des candidats et candidates, et il cible les personnes qui correspondent à cette préférence. Encore une fois, l'algorithme a pour tâche d'adapter et de répliquer les données qu'il reçoit, et d'apprendre de ces données.

Tout au long du processus d'embauche, l'algorithme peut éliminer un nombre important de candidats et candidates ayant de l'expérience, mais n'étant pas associés aux expressions ou aux mots-clés employés pour entraîner l'algorithme (Bogen, 2019). Certains algorithmes se fondent également sur des décisions d'embauche antérieures pour déterminer les candidatures à rejeter, ce qui peut perpétuer la discrimination (Dastin, 2018). D'autres outils d'embauche détermineront qui aura du succès dans un poste en fonction de l'expérience passée, d'évaluations de rendement, de la durée d'emploi et parfois de l'absence de renseignements négatifs comme des mesures disciplinaires (Bogen, 2019). Ces algorithmes d'embauche se trouvent bien sûr dans d'autres domaines que celui de l'IA. Les algorithmes d'embauche reproduisent et amplifient les préjugés humains dont nous discutons dans ce chapitre (quant au genre, à la race ou au statut socioéconomique), perpétuant le cercle vicieux qui alimente le manque de diversité en informatique et en programmation de l'IA.

Les problèmes liés à l'IA abordés jusqu'à présent (manque de diversité, biais algorithmique appliqué, recherche en vase clos au sein d'une seule discipline et innovation non inclusive) restent souvent sous-évalués en ce qui concerne les effets qu'ils ont tant sur la qualité de la technologie que sur la main-d'œuvre, et plus largement sur l'avenir de l'humanité. En renonçant à une innovation véritablement inclusive, nous compromettons essentiellement le bien-être et la prospérité dans le monde ainsi que des normes élevées en matière de performance et de fiabilité de l'IA (Dengel *et al.*, 2021) au nom du profit à court terme. L'actuelle structure de financement de la recherche en IA est l'un des principaux facteurs qui alimentent cette compromission.

STRUCTURES DE FINANCEMENT ET MOTIVATIONS

Les obstacles à une IA véritablement inclusive se fondent sur la manière de financer la recherche dans ce domaine et les raisons de le faire. Dans l'état actuel des choses, les projets financés par l'industrie ou des organismes du secteur public ne mettent malheureusement pas l'accent sur l'inclusion et la diversité. Souvent, ce ne sont pas des projets interdisciplinaires ou collaboratifs, et ils ne tiennent pas compte de la croissance du capital humain, social et naturel autant que du rendement de l'investissement. Les innovations qui en découlent influent à leur tour sur l'orientation que les décideurs et décideuses font prendre aux nouvelles technologies, ce qui se répercute ensuite sur la façon de répartir le financement. Ainsi, un cercle vicieux complexe se perpétue.

Bien qu'il ne soit pas intentionnel, un cercle vicieux se crée vu le caractère indissociable des projets de recherche en IA, des sources de financement et des politiques, et il se renforce en raison d'une diversité limitée des parties prenantes bénéficiant des innovations en matière d'IA ou ayant de l'influence à cet égard.

IA et monde universitaire

Que ce soit directement ou indirectement, des entreprises mènent la plupart des recherches sur l'IA ou soutiennent les nouvelles technologies. Comme l'indique le rapport fédéral sur le financement de la recherche et du développement de 2021 du Congressional Research Service (2020), 54 % de la recherche appliquée et 85 % du développement aux États-Unis ont été financés par des entreprises (consulter la figure 1). Une évaluation récente de la politique et du financement en matière d'IA au Canada montre que même le financement public est principalement destiné à « l'industrie et [au] milieu universitaire ayant des liens avec l'industrie. Le milieu universitaire sert souvent d'intermédiaire entre l'industrie et le gouvernement. Indirectement, ces fonds peuvent toujours profiter à des organisations à but lucratif » (Brandusescu, 2021, p. 37). Cette réalité peut fortement entrer en jeu dans la recherche universitaire et l'élaboration de politiques ainsi que dans le degré d'influence qu'ont les entreprises sur l'innovation en IA.

En matière d'IA, le secteur privé est inextricablement présent dans le milieu universitaire. Selon le *Artificial Intelligence Index Report* produit par la Stanford Institute for Human-Centered Artificial Intelligence (Zhang *et al.*, 2021, p. 21), plus de 15 % des publications évaluées par des pairs en 2019 provenaient d'entreprises présentes dans tous les grands pays et les grandes régions du monde. L'industrie absorbe également la majorité de l'expertise en IA issue du milieu universitaire (en 2019, 65 % des doctorants et doctorantes en IA en Amérique du Nord ont intégré l'industrie après avoir obtenu leur diplôme) (Zhang *et al.*, 2021, p. 4). Les entreprises parrainent également de nombreuses conférences et plusieurs ateliers dans le domaine, ou y sont très présentes (Alford, 2021). Par exemple, lors de l'International Conference on Learning Representation (ICLR) tenue en 2021, près de 30 % des communications étaient le fait d'entreprises telles que Google, Amazon, IBM et Facebook. De plus, quatre communications de Google et une de Facebook faisaient partie des huit ayant obtenu un prix d'excellence (ICLR, 2021).

La plupart des entreprises en IA sont guidées par des programmes de recherche et développement fortement influencés par la demande du marché et le rendement de l'investissement. Les innovations qui révolutionnent le domaine, qui apportent de nouveaux éléments d'actif ou qui élargissent les horizons sont au cœur de ces programmes, et l'expertise et les compétences acquises dans le milieu universitaire constituent une ressource à part entière. C'est l'une des raisons pour lesquelles l'industrie finance la recherche universitaire. Il s'ensuit que de nombreux projets de recherche portent l'empreinte des besoins de l'industrie. Cette dynamique entre l'industrie et le monde universitaire crée une tension double. Premièrement, puisque la plupart des protagonistes de l'industrie de l'IA et de la technologie se concentrent dans les pays du Nord (Chan *et al.*, 2021), les écarts qui empêchent l'innovation

de vraiment profiter à tous et à toutes se creusent davantage. Deuxièmement, le domaine s'oriente de manière disproportionnée vers les intérêts du secteur privé plutôt que le bien public.

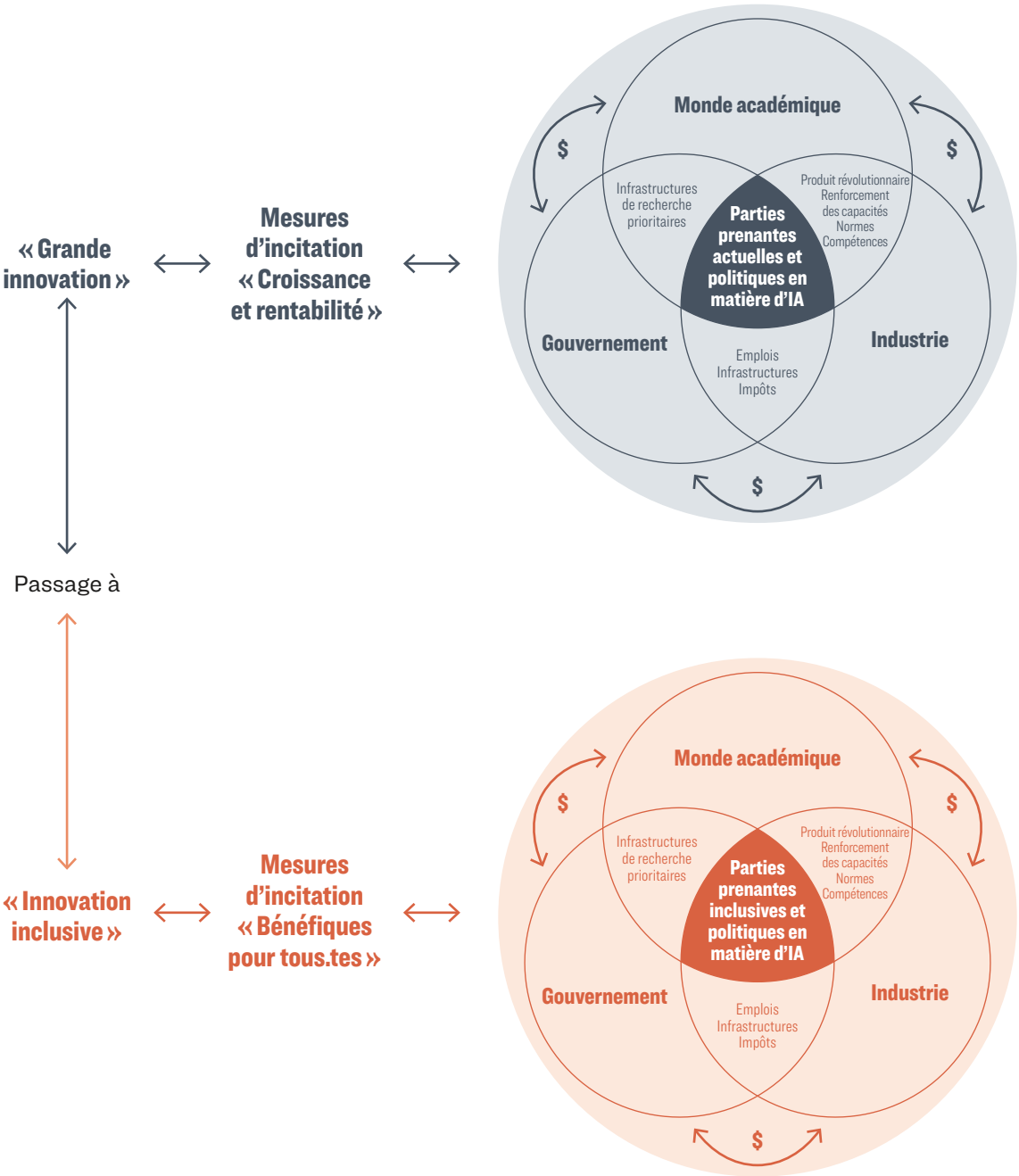
Des collaborations étroites entre l'industrie et le milieu universitaire ne posent pas problème en soi ; elles peuvent profiter à la recherche et à l'enseignement dans les établissements universitaires (Etzioni, 2019). Le capitalisme des parties prenantes, « une forme de capitalisme dans laquelle les entreprises recherchent la création de valeur à long terme en tenant compte des besoins de toutes leurs parties prenantes et de la société dans son ensemble », peut être envisagé comme une solution viable pour les personnes et la planète (Schwab et Vanham, 2021). Il oblige cependant les industries à placer l'interdisciplinarité et l'innovation inclusive au cœur de leur stratégie en matière d'IA, en opérant finalement un virage de la demande du marché vers les biens publics et en intégrant des personnes et des communautés marginalisées parmi les principales parties prenantes.

Boucles de rétroaction : financement public, motivations du secteur privé, politiques

L'enthousiasme suscité par l'IA réduit le financement de la recherche fondamentale au profit de la recherche appliquée et de la « grande innovation » commercialisable à court ou moyen terme. Ainsi, l'IA change rapidement les règles du jeu quant au financement des secteurs public et privé. La recherche et le développement appliqués sont souvent motivés par le potentiel de rendement sur l'investissement (en ce qui a trait à la fois au profit et à la croissance). Actuellement, la recherche appliquée est essentiellement financée par le secteur privé (Congressional Research Service, 2020, figure 1). Étant donné que l'industrie finance également indirectement la recherche universitaire (par exemple en soutenant des programmes de financement public) (Brandusescu, 2021), il est difficile de ne pas percevoir la forte influence du secteur privé sur l'orientation que prend l'IA. En outre, cette influence se répercute sur les stratégies de croissance économique qui, à leur tour, influent également sur les programmes de financement public (voir la figure 1).

| FIGURE 1 |

Structure de financement de l'IA et motivations. Adaptée des modèles de Kimatu (2016) et d'Ondimu (2012).



Ainsi, « il vaut la peine de se demander comment l'économie de l'innovation subit l'influence d'intérêts privés et du pouvoir privé et, par extension, comment se rédigent les politiques publiques en matière d'IA » (Brandusescu, 2021, p. 38). Compte tenu du mécanisme de rétroaction, en IA, entre le financement public et les innovations au sein du secteur privé, la nécessité de placer l'inclusivité au cœur de ces innovations n'a jamais été aussi pressante si nous voulons des technologies basées sur l'IA qui profitent à tout le monde et jouissent de la confiance de tous et de toutes. L'innovation est un moteur important de la recherche et du financement, et elle influe directement sur le milieu universitaire, le gouvernement et l'industrie. Passer de la « grande innovation » à l'« innovation inclusive » peut modifier les dynamiques de la recherche et du financement, les politiques en matière d'IA et la mobilisation des parties prenantes de telle sorte que personne ne soit laissé pour compte.

Il y a des avantages à faire sortir l'IA de l'actuel vase clos de la science et de la technologie, et d'élargir ses horizons à des domaines tels que les neurosciences, la linguistique informatique, l'éthique, la sociologie et l'anthropologie (Rahwan *et al.*, 2019, p. 477). Il s'agit notamment d'accroître l'interdisciplinarité et, au-delà de la technique, d'intégrer des compétences qui font cruellement défaut à l'IA en général et qui, de ce fait, entravent ses progrès (Dengel *et al.*, 2021).

La loi nationale sur l'intelligence artificielle adoptée aux États-Unis en 2020 (United States Congress, 2020) vise à répartir le financement de la recherche sur l'IA et ses applications au sein d'un plus large éventail d'organismes gouvernementaux, c'est-à-dire au-delà de la défense nationale, qui, auparavant, dirigeait principalement les politiques américaines en la matière (Delgado et Levy, 2021). Cette loi ainsi que d'autres politiques et initiatives amorcent un changement dans le fonctionnement des agences de financement. Elles reconnaissent la nécessité de revoir les priorités de financement : « L'intelligence artificielle devient de plus en plus un domaine hautement interdisciplinaire nécessitant l'expertise d'un large éventail de disciplines scientifiques et d'autres matières universitaires qui ont traditionnellement évolué de manière indépendante et qui continuent d'affronter des obstacles culturels et institutionnels nuisant à une collaboration à grande échelle » (United States Congress, 2020).

Comme l'a cependant indiqué le Congrès américain dans l'une de ses conclusions : « Les investissements fédéraux actuels et les mécanismes de financement sont largement insuffisants pour susciter et soutenir la collaboration interdisciplinaire et celle des secteurs public et privé à grande échelle, lesquelles seront nécessaires au perfectionnement de systèmes d'intelligence artificielle fiables aux États-Unis » (United States Congress, 2020, pp. 3-4). Cela n'est pas surprenant si l'on considère la forte influence du secteur privé sur les critères de financement de la recherche et de l'innovation décrits précédemment. De plus, compte tenu de ces critères, il n'est pas rare que des chercheurs ou chercheuses adaptent leurs travaux pour qu'ils correspondent aux occasions de financement. Par conséquent, en plus du fait que les politiques publiques en matière d'IA sont prises dans un cercle vicieux, la qualité de l'IA en elle-même n'a qu'à satisfaire aux exigences du marché. Malheureusement, la ligne d'action actuelle en ce qui a trait à l'IA est principalement tracée par un nombre limité de parties prenantes sans grande diversité (Delgado et Levy, 2021). Afin de revoir les motivations et de briser le cycle, les critères de financement devraient prioritairement se focaliser sur l'innovation inclusive. Il faut mettre en balance, d'une part, le profit et la croissance et, d'autre part, les occasions d'apprécier un capital humain, naturel et social florissant.

Un soutien accru du gouvernement et de l'industrie à des projets communautaires, collaboratifs et interdisciplinaires constitue une façon d'accorder la priorité à l'innovation inclusive. En ce moment, la recherche de projets innovants et prestigieux signifie malheureusement trop souvent qu'on ne choisit pas de travailler sur une IA inclusive dans un contexte de ressources limitées et en vue d'avoir un effet localement, car cela ne révolutionne pas expressément le domaine à court terme et n'attire pas de financement. À titre d'exemple, la National Science Foundation (NSF, 2021) a pris l'engagement d'augmenter le financement de l'IA appliquée. Bien que cet engagement vise à diversifier la recherche, ne pas ancrer de telles initiatives dans l'innovation inclusive risque d'orienter les fonds vers l'innovation technologique et de les éloigner de la recherche fondamentale ne présentant pas une viabilité commerciale

à court terme, ce qui défavorise les étudiants et étudiantes qui s'intéressent à la recherche axée sur le bien public (Viglione, 2020). Changer les motivations qui alimentent actuellement le cercle vicieux du financement permettrait à un nombre accru de chercheurs et chercheuses d'entreprendre avant la mi-carrière des projets centrés sur l'innovation inclusive et nourrirait l'expertise en matière d'IA inclusive. Nous soutenons qu'une innovation inclusive est la seule véritable innovation qui devrait être envisagée en IA si nous voulons que celle-ci profite à tous et à toutes. Cet objectif est irréalisable lorsque l'innovation en IA se produit en vase clos et qu'elle est principalement parrainée par l'industrie, sous la pression des actionnaires et au nom du profit.

RUISSELLEMENT

Pour justifier l'actuel manque d'inclusion dans l'innovation, le concept d'économie de ruissellement a parfois été étendu à une « science du ruissellement ». Une forte concentration de ressources et d'universitaires dans les pays du Nord devrait « produire la meilleure science », dont « des méthodes, des théories et des idées » s'écouleront vers ceux du Sud (Reidpath et Allotey, 2019, p. 1). Tout comme l'économie de ruissellement, ce modèle n'est pas viable et constitue en fait le contraire de la réalité (Reidpath et Allotey, 2019, p. 1), en partie en raison de ce qui motive le financement et oriente la demande du marché, comme il en a été question précédemment. Par exemple, la combinaison de la forte demande en ressources et d'un assouplissement de la réglementation et des mesures de protection de la vie privée dans les pays du Sud entraîne une exploitation accrue des ressources, tant humaines (forage de données) que naturelles (matériaux) (Arezki, 2021 ; Arun, 2020, p. 594 ; Mishra, 2021).

Il est également important de tenir compte de l'environnement politique de régions où les applications basées sur l'IA sont déployées. Dans bien des cas, les technologies mises au point pour quelques groupes privilégiés s'avèrent nuisibles dans des régions n'ayant pas autant de ressources. Le rapport d'enquête des Nations Unies sur le génocide des Rohingyas a par exemple fait remarquer que « Facebook [avait] été un instrument utile à qui cherchait à répandre la haine » (Human Rights Council, 2018, p. 34). Cela démontre l'effet puissant que les médias sociaux peuvent avoir sur les droits humains lorsqu'on les utilise là où l'environnement politique et médiatique est malsain.

L'innovation rapide se produit souvent aux dépens des personnes qui devraient profiter du ruissellement, alors que des répercussions néfastes l'emportent de manière disproportionnée sur tout avantage potentiel (Schia, 2018, p. 827). Ce déséquilibre n'est que renforcé par la continuelle exclusion de personnes et de communautés marginalisées en tant que parties prenantes clés. Comme l'a déclaré Shirley Chisholm : « S'ils ne vous accordent pas une place à table, apportez une chaise pliante ». Il ne faut pas sous-estimer l'importance de la représentation, aussi difficile soit-il d'assurer celle-ci. Et même quand on y parvient, le travail est loin d'être terminé.

RÉELLE SIGNIFICATION DE « AVOIR UNE PLACE À TABLE »

Les enjeux décrits dans ce chapitre ne sont bien sûr pas uniquement liés à l'IA ou à la façon de percevoir l'innovation. Ils sont représentatifs de problèmes systémiques plus larges qui évoluent quotidiennement. L'un des principaux obstacles au moment de les régler est le fait que les efforts actuels sont fournis en vase clos plutôt qu'abordés selon une perspective systémique.

De nombreuses initiatives relatives à l'IA sont déjà en place, telles que AI4ALL[62], l'African Master's in Machine Intelligence (AMMI)[63], Quantum Leap Africa (QLA)[64], le Centro de Excelencia en Inteligencia Artificial en Medellín[65] et l'African Supercomputing Center (ASCC) de l'Université Mohammed VI Polytechnique (UM6P)[66] au Maroc, et font beaucoup pour que l'IA profite à tous et à toutes et qu'elle comprenne une pluralité de points de vue.

Des entreprises telles que Google et Microsoft, des fondations, des responsables politiques et des organismes de financement publics investissent d'une certaine manière dans des projets d'IA pour le bien social[67] et, ce faisant, financent des projets qui ne seraient probablement pas sélectionnés en fonction des directives actuelles qui régissent le financement. Si de telles initiatives ne s'ancrent pas dans l'innovation inclusive, elles peuvent avoir un effet pervers : accroître la marginalisation de groupes minoritaires (Latonero, 2019). Ainsi, le fait que ces fonds s'inscrivent à l'extérieur du cycle de financement habituel accentue la mise en marge de projets plutôt que de s'y attaquer.

Une mise en marge se produit également sur le plan individuel. Lorsque les groupes marginalisés ont pour unique porte d'entrée dans la recherche innovante des programmes spécialisés, il y a accentuation du « syndrome de l'imposteur » (Tulshyan et Burey, 2021) si commun aux personnes issues de minorités travaillant en sciences ou occupant un poste de haut niveau. Ces programmes visent à réduire l'écart en augmentant la présence de ces personnes dans les coulisses et dans la prise de décisions. Ils se penchent toutefois rarement sur des problèmes systémiques plus larges étant à la source de préjudices, d'environnements toxiques, ou de collègues toxiques qui perpétuent l'idée que les personnes de groupes minoritaires doivent être « invitées » à se joindre au cercle des scientifiques ou des responsables. En plus de subir un manque de confiance en soi qui, nous l'avons vu, commence dès l'enfance, ces personnes s'imposent une panoplie d'obstacles et d'attentes après avoir été réduites au silence « à la table » par les autres et trop souvent par elles-mêmes.

Malheureusement, ces programmes d'inclusion sont souvent perçus comme des efforts « suffisants » pour réduire les écarts (Puritty *et al.*, 2017). Bien sûr, ils ne le sont pas concrètement. Nous le constatons dans le pourcentage de femmes occupant des emplois en mathématiques et en informatique aux États-Unis, qui est passé de 25,8 % en 2010 à 25,2 % en 2020 (voir le tableau 1). Ce sont de bonnes initiatives, alors pourquoi ne fonctionnent-elles pas comme prévu ? Comme le mentionnent Dengel et autres (2021, p. 90), « … nous avons encore besoin de beaucoup de travail de recherche et d'un changement de paradigme en IA afin de mettre au point une véritable IA pour l'humanité, une IA centrée sur l'humain ». Nous soutenons qu'une condition importante à l'instauration de ce paradigme consiste à inscrire l'inclusivité au centre de l'innovation plutôt que de le faire dans la périphérie ou après-coup.

Nous ne sommes pas les premières à exposer tous les préjugés et les problèmes que soulèvent les technologies de l'IA. Nous ne sommes pas non plus les premières à mentionner l'ampleur des progrès accomplis. Mais il est important de continuer à rehausser les normes en ce qui a trait à l'inclusivité et à l'innovation. Ces mesures ne peuvent que contribuer à améliorer les structures au sein desquelles

62. Consulter le site Web de l'initiative AI4ALL (2021) pour de l'information.

63. Voir AIMS (2021) pour des renseignements sur l'African Institute for Mathematical Sciences et l'African Master's in Machine Intelligence.

64. Consulter le site Web de Quantum Leap Africa (2021) pour des détails supplémentaires.

65. Ce centre a été mis en place grâce à un partenariat entre l'organisme colombien Ruta N et l'Institute for Robotic Process Automation & Artificial Intelligence. Consulter le site Web de Ruta N (2018) pour de l'information.

66. L'ASCC relève de l'UM6P. Des renseignements figurent sur le site Web (ASCC 2020).

67. Parmi ces initiatives figure le Google.com Impact Challenge pour les femmes et les filles (2021), AI for Good Research Lab (Microsoft, 2021) et Creating Sustained Social Impact (Microsoft Corporate Citizenship, 2021).

nous travaillons et les innovations auxquelles nous aspirons (comme nous l'avons vu avec les systèmes de reconnaissance faciale, par exemple). Selon Giridharadas (2021), un auteur connu pour sa critique de la prise de décision exclusive des élites concernant la manière de résoudre les problèmes mondiaux, tous les grands défis exigent des solutions publiques, institutionnelles, démocratiques et universelles. L'action et la prise de décision collectives sont plus bénéfiques pour tout le monde que l'action individuelle :

> all grand challenges[...] require public, institutional, democratic and universal solutions. They need to solve the problem at the root and for everyone. What we do together is more interesting, compelling, more powerful, more valuable than what we do alone. Current neo-liberal myth is that what we do alone is better and more beautiful than what we do together. We need to bring back the notion that we live in society within which we have interdependence. Valuing what we do together needs to be reclaimed. Only this collective intelligence will allow us to solve the grand challenges we face.

CONCLUSION

La mise au point de nouveaux algorithmes, le perfectionnement des ressources informatiques et la disponibilité de données abondantes ont entraîné la récente vague d'innovation en IA. Ces changements entraînent des transformations dans un large éventail d'industries et de secteurs qui vont probablement révolutionner la société, comme l'ont fait les révolutions industrielles passées. Ainsi, l'humanité risque une fois de plus de perpétuer un changement de système aux répercussions inéquitables, poussé notamment par des mentalités coloniales et des clivages socioéconomiques. En particulier, le fossé numérique creuse les inégalités en ce qui a trait à l'accès à l'IA et aux conséquences néfastes des préjugés humains transmis aux technologies basées sur l'IA.

Aborder les enjeux de l'IA jusque-là négligés commence par s'attaquer à nos propres préjugés. Nous avons discuté dans ce chapitre de la manière dont les préjugés des humains influent fortement sur l'IA et le domaine des STIM en général. Premièrement, la qualité et la précision des systèmes basés sur l'IA sont compromises par le manque de diversité des données et des ressources humaines à toutes les étapes de la mise au point de l'IA. Deuxièmement, ce problème est amplifié par la marginalisation de groupes entiers au sein des STIM en fonction du genre, de la race et du statut socioéconomique, ce qui accentue également la « pénurie des talents » qui préoccupe actuellement le domaine de l'IA. Cela se produit alors que la demande en IA monte en flèche (même si la technologie est inadaptée).

Il en découle que le financement, privé et public, est pris dans un cercle vicieux que renforcent les motivations que sont le profit à court terme et la croissance économique, ce qui oriente la trajectoire de l'IA et l'ère numérique. Ce cercle vicieux est commun là où règne la croyance en une croissance rapide axée sur la « grande innovation ». Comme nous le soutenons dans ce chapitre, l'attention doit se tourner vers l'« innovation inclusive », ce qui favorisera la diversité des points de vue et un renforcement des capacités, en particulier au sein de communautés marginalisées et pauvres en ressources. Pour que l'intelligence artificielle reflète vraiment le pouvoir de la conscience humaine, elle doit représenter la beauté et le pouvoir de la diversité.

L'interdépendance croissante des systèmes et des enjeux mondiaux décentre l'attention, l'amenant du pur profit vers une valorisation du capital naturel, humain et social. Il n'y a aucun moyen pour les gens et la planète de prospérer sans ce changement, et il est essentiel d'accorder la priorité aux solutions locales qui incarnent les principes éthiques universels de confiance, de responsabilité et d'empathie. Bien qu'il soit tentant de soutenir d'abord la croissance rapide et le profit à court terme au nom de l'« innovation », agir ainsi limiterait inévitablement le potentiel de nos systèmes d'IA au profit de quelques groupes privilégiés plutôt que de l'humanité dans son ensemble. Une fois que la recherche et le développement en matière d'IA seront guidés par l'innovation inclusive, nous serons en mesure de passer d'une IA et d'une ère numérique fragmentées à une unité qui profitera à tous et à toutes, y compris aux générations à venir.

RÉFÉRENCES

AI4ALL. 2021. Page d'accueil. https://ai-4-all.org/

AIMS (African Masters in Machine Intelligence). 2021. Page d'accueil. https://aimsammi.org

Alford, A. 2021. AI conference recap : Google, Microsoft, Facebook, and others at ICLR 2021. *InfoQ*. https://www.infoq.com/news/2021/06/conference-recap-iclr-2021/

Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A. et Rieke, A. 2019. Discrimination through optimization : How Facebook's ad delivery can lead to skewed outcomes. *Proceedings of the ACM on Human-Computer Interaction,* vol. 3, nº 199, pp. 1-30. https://dl.acm.org/doi/pdf/10.1145/3359301

American University. 2020. *Understanding the digital divide in education*. School of Education Online, blogue, 15 décembre. https://soeonline.american.edu/blog/digital-divide-in-education

Arezki, R. 2021. Transnational governance of natural resources for the 21st century. Brookings Institution, blogue, 7 juillet. https://www.brookings.edu/blog/future-development/2021/07/07/transnational-governance-of-natural-resources-for-the-21st-century/

Arun, C. 2020. AI and the Global South : Designing for other worlds. M. D. Dubber, F. Pasquale et S. Das (dir.), *The Oxford Handbook of Ethics of AI*. Oxford, Oxford University Press, pp. 58-606.

ASCC (African SuperComputing Center). 2020. Page d'accueil. https://ascc.um6p.ma

Barber, P.H., Hayes, T. B., Johnson, T. L. et Márquez-Magaña, L. 2020. Systemic racism in higher education. *Science*, vol. 369, nº 6510, pp. 1440-1441. https://www.science.org/doi/pdf/10.1126/science.abd7140

Berger, Y. 2017. Israel arrests Palestinian because Facebook translated "good morning" to "attack them". Haaretz. https://www.haaretz.com/israel-news/palestinian-arrested-over-mistranslated-good-morning-facebook-post-1.5459427

Bogen, M. 2019. All the ways hiring algorithms can introduce bias. *Harvard Business Review*. https://hbr.org/2019/05/all-the-ways-hiring-algorithms-can-introduce-bias

Brandusescu, A. 2021. *Artificial intelligence policy and funding in Canada : Public investments, private interests*. Centre for Interdisciplinary Research on Montréal, pp. 11-51. https://www.mcgill.ca/centre-montreal/files/centre-montreal/aipolicyandfunding_report_updated_mar5.pdf

Buolamwini, J. 2019. Compassion through computation : Fighting algorithmic bias. Vidéo, World Economic Forum, YouTube. https://youtu.be/5PGYOYZKsdY

Bureau of Labor Statistics (United States). 2010. Employed persons by detailed occupation, sex, race, and Hispanic or Latino ethnicity, labor force statistics from the current Population Survey. https://www.bls.gov/cps/aa2010/cpsaat11.pdf

––––. 2020. Employed persons by detailed occupation, sex, race, and Hispanic or Latino Ethnicity, labor force statistics from the current Population Survey. https://www.bls.gov/cps/cpsaat11.pdf

Bushweller, K. 2021. How to get more students of color into STEM : Tackle bias, expand resources. Education Week, article en ligne, 2 mars. https://www.edweek.org/technology/how-to-get-more-students-of-color-into-stem-tackle-bias-expand-resources/2021/03

Carter, L., Liu, D., et Cantrell, C. 2020. Exploring the intersection of the digital divide and artificial intelligence : A hermeneutic literature review. *AIS Transactions on Human-Computer Interaction,* vol. 12, nº 4, pp. 253-275. https://aisel.aisnet.org/thci/vol12/iss4/5/

Chan, A., Okolo, C. T., Terner, Z. et Wang, A. 2021. *The limits of global inclusion in AI development*. Association for the Advancement of Artificial Intelligence. https://arxiv.org/pdf/2102.01265.pdf

Cho, A. 2020. Artificial intelligence systems aim to sniff out signs of COVID-19 outbreaks. *Science*, 12 mai. https://www.sciencemag.org/news/2020/05/artificial-intelligence-systems-aim-sniff-out-signs-covid-19-outbreaks <

Clark, U.S. et Hurd, Y.L. 2020. Addressing racism and disparities in the biomedical sciences. *Nature Human Behaviour,* vol. 4, nº 8, pp. 774-777. https://www.nature.com/articles/s41562-020-0917-7

Coded Bias. 2020. Film cinématographique, 7th Empire Media, Brooklyn, réalisation de Shalini Kantayya.

Coleman, D. 2019. Digital colonialism : The 21st century scramble for Africa through the extraction and control of user data and the limitations of data protection laws. *Michigan Journal of Race and Law,* vol. 24, nº 2, pp. 417-439. https://repository.law.umich.edu/mjrl/vol24/iss2/6

Congressional Research Service (United States). 2020. *Federal Research and Development (R&D) Funding : FY2021*. https://fas.org/sgp/crs/misc/R46341.pdf

Cummings, M. 2019. Yale study shows class bias in hiring based on few seconds of speech. *YaleNews*, 21 octobre. https://news.yale.edu/2019/10/21/yale-study-shows-class-bias-hiring-based-few-seconds-speech

Dancy, M., Rainey, K., Stearns, E., Mickelson, R. et Moller, S. 2020. Undergraduates' awareness of white and male privilege in STEM. *International Journal of STEM Education,* vol. 7, nº 52, pp. 1-17. https://stemeducationjournal.springeropen.com/articles/10.1186/s40594-020-00250-3

Dastin, J. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. Londres, *Reuters*, 10 octobre. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Delgado, F. A. et Levy, K. 2021. A community-centered research agenda for AI innovation policy. *Cornell Policy Review*, 4 mai. https://www.cornellpolicyreview.com/a-community-centered-research-agenda-for-ai-innovation-policy/

Dengel, A., Etzioni, O., DeCario, N., Hoos, H., Li, F., Tsujii, J. et Traverso, P. 2021. Next big challenges in core AI technology. B. Braunschweig et M. Ghallab (dir.), *Reflections on Artificial Intelligence for Humanity* Lecture Notes in Computer Science, vol. 12600, Springer, Cham, pp. 90-115. https://doi.org/10.1007/978-3-030-69128-8_7

Dodge, A. 2018. What you need to know about the stem race gap. Redondo Ozobot, blogue, 20 février. https://ozobot.com/blog/need-know-stem-race-gap

Etzioni, O. 2019. AI academy under siege. *Inside Higher Ed*, 20 novembre. https://www.insidehighered.com/views/2019/11/20/how-stop-brain-drain-artificial-intelligence-experts-out-academia-opinion

Falk-Krzesinski, H. J. et Tobin, S. C. 2015. How do I review thee ? Let me count the ways : A comparison of research grant proposal review criteria across US federal funding agencies. *The Journal of Research Administration,* vol. 46, nº 2, pp. 79-94. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4892374/

Firebaugh, G. et Acciai, F. 2016. For blacks in America, the gap in neighborhood poverty has declined faster than segregation. *Proceedings of the National Academy of Sciences,* vol. 113, nº 47, pp. 13372-13377. https://www.pnas.org/content/pnas/113/47/13372.full.pdf

Fleming, N. 2018. How artificial intelligence is changing drug discovery. *Nature,* vol. 557, nº 7706, pp. 55-57. link.gale.com/apps/doc/A572639347/AONE

Frehill, L. M., Di Fabio, N., Hill, S., Traeger, K., et Buono, J. 2008. Women in engineering : A review of the 2007 literature. *SWE Magazine,* vol. 54, pp. 6-30.

Garcia, E. 2021. The international governance of AI : Where is the Global South ? The Good AI, blogue, 28 janvier. https://thegoodai.co/2021/01/28/the-international-governance-of-ai-where-is-the-global-south/

Giridharadas, A. 2021. Philanthropy and the state : Who is funding what and why ? Vidéo, UCL Institute for Innovation and Public Purpose. https://www.youtube.com/watch?v=fOAkNu7Y6f4

Google. 2021. Google.org Impact Challenge pour les femmes et les filles 2021. https://impactchallenge.withgoogle.com/womenandgirls2021

Gray, M. L. et Suri, S. 2019. *Ghost Work : How to Stop Silicon Valley from Building a New Global Underclass*. Boston, Houghton Mifflin Harcourt.

Heaven, W.D. 2020. Predictive policing algorithms are racist. They need to be dismantled. *MIT Technology Review*, 17 juillet. https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice

Heeks, R., Amalia, M., Kintu, R. et Shah, N. 2013. Inclusive innovation : Definition, conceptualisation and future research priorities. *Manchester Center for Development Informatics,* nº 53, pp. 1-28. https://www.researchgate.net/publication/334613068_Inclusive_Innovation_Definition_Conceptualisation_and_Future_Research_Priorities

Hickok, M. 2020. Why was your job application rejected ? Bias in recruitment algorithms. Institut d'éthique en IA de Montréal. https://montrealethics.ai/why-was-your-job-application-rejected-bias-in-recruitment-algorithms-part-1/

Hill, C. 2020. The STEM gap : Women and girls in science, technology, engineering and math. AAUW, section sur les ressources. https://www.aauw.org/resources/research/the-stem-gap/

Hill, C., Corbett, C. et St. Rose, A. 2010. *Why so Few ? Women in Science, Technology, Engineering, and Mathematics*. Washington DC, AAUW. https://www.aauw.org/app/uploads/2020/03/why-so-few-research.pdf

Hill, K. 2020. Wrongfully accused by an algorithm. *The New York Times*, 24 juin. https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html

Human Rights Council. 2018. *Report of the Independent International Fact-Finding Mission on Myanmar*. Genève. https://www.ohchr.org/Documents/HRBodies/HRCouncil/FFM-Myanmar/A_HRC_39_64.pdf

ICLR (International Conference on Learning Representations). 2021. Announcing ICLR 2021 Outstanding Paper Awards. https://iclr-conf.medium.com/announcing-iclr-2021-outstanding-paper-awards-9ae0514734ab

Jiménez-Luna, J., Grisoni, F., Weskamp, N. et Schneider, G. 2021. Artificial intelligence in drug discovery : Recent advances and future perspectives. *Expert Opinion on Drug Discovery*, vol. 16, nº 9, pp. 1-11. https://www.tandfonline.com/doi/pdf/10.1080/17460441.2021.1909567

Kimatu, J.N. 2016. Evolution of strategic interactions from the triple to quad helix innovation models for sustainable development in the era of globalization. *Journal of Innovation and Entrepreneurship,* vol. 5, nº 16, pp. 1-7. https://innovation-entrepreneurship.springeropen.com/articles/10.1186/s13731-016-0044-x

Kuhlman, C., Jackson, L. et Chunara, R. 2020. No computation without representation : Avoiding data and algorithm biases through diversity. *arXiv Preprint.* https://arxiv.org/pdf/2002.11836.pdf

Kwet, M. 2019. Digital colonialism : US empire and the new imperialism in the Global South. *Race & Class,* vol. 60, nº 4, pp. 3-26. https://journals.sagepub.com/doi/pdf/10.1177/0306396818823172

Lada, A., Wang, M. et Yan, T. 2021. How machine learning powers Facebook's news feed ranking algorithm. engineering at Meta, blogue, 26 janvier. https://engineering.fb.com/2021/01/26/ml-applications/news-feed-ranking/

Larsen, J. 2021. Levi-Strauss' Dr. Katia Walsh on why diversity in AI and ML is non-negotiable. San Francisco, *VentureBeat*. https://venturebeat.com/2021/08/02/levi-strauss-dr-katia-walsh-on-why-diversity-is-non-negotiable-in-ai-and-machine-learning/

Latonero, M. 2019. Opinion : AI for good is often bad. *Wired*, 18 novembre. https://www.wired.com/story/opinion-ai-for-good-is-often-bad/

Leslie, S., Cimpian, A., Meyer, M. et Freeland, E. 2015. Expectations of brilliance underlie gender distributions across academic disciplines. *Science,* vol. 347, nº 6219, pp. 262-265. https://www.science.org/doi/full/10.1126/science.1261375

May, A.. 2020. Dr. Fei-Fei Li*:* "We can make humanity better in so many ways." *Artificial Intelligence in Medicine*, 12 décembre https://ai-med.io/ai-champions/dr-fei-fei-li-we-can-make-humanity-better-in-so-many-ways/

McCarthy, J., Minsky, M. L., Rochester, N. et Shannon, C. E. 1955. A proposal for the Dartmouth Summer Research Project on Artificial Intelligence. http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf

McGee, E. O. 2020. Interrogating structural racism in STEM higher education. *Educational Researcher,* vol. 49, nº 9, pp. 633-644. https://journals.sagepub.com/doi/full/10.3102/0013189X20972718

McGee, E. et Bentley, L. 2017. The troubled success of Black women in STEM. *Cognition and Instruction,* vol. 35, nº 4, pp. 265-289. https://www.tandfonline.com/doi/pdf/10.1080/07370008.2017.1355211

Metcalf, H.E., Crenshaw, T.L., Chambers, E.W. et Heeren, C. 2018. Diversity across a decade : A case study on undergraduate computing culture at the University of Illinois. *Proceedings of the 49th ACM Technical Symposium on Computer Science Education. Association of Computing Machinery*, pp. 610-615. https://dl.acm.org/doi/abs/10.1145/3159450.3159497

Microsoft Corporate Citizenship. 2021. Creating sustained societal impact. https://www.microsoft.com/en-hk/sparkhk/creating-sustained-societal-impact

Microsoft Research. 2021. AI for Good Research Lab*:* Overview. https://www.microsoft.com/en-us/research/group/ai-for-good-research-lab/

Miller, O. 2017. The myth of innate ability in tech. Blogue personnel, 9 janvier. http://omojumiller.com/articles/The-Myth-Of-Innate-Ability-In-Tech

Miller, R. A. et Downey, M. 2020. Examining the STEM climate for queer students with disabilities. *Journal of Postsecondary Education and Disability,* vol. 33, nº 2, pp. 169-181. https://www.researchgate.net/publication/334654579_Examining_the_STEM_Climate_for_Queer_Students_with_Disabilities

Mishra, S. 2021. Opinion : Is AI deepening the divide between the Global North and South ? *Newsweek*, 9 mars. https://www.newsweek.com/ai-deepening-divide-between-global-north-south-opinion-1574141

NSF. 2021. National Science Foundation Graduate Research Fellowship Program. https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=6201

OCDE [Organisation de coopération et de développement économiques]. 2001. *Understanding the Digital Divide*. https://www.oecd.org/digital/ieconomy/1888451.pdf.

O'Donnell, R. M. 2019. Challenging racist predictive policing algorithms under the equal protection clause. *New York University Law Review,* vol. 94, nº 544, pp. 544-580. https://www.nyulawreview.org/wp-content/uploads/2019/06/NYULawReview-94-3-ODonnell.pdf

Ondimu, S. 2012. Possible approaches to commercialisable university research in Kenya. *The 7th KUAT Scientific, Technological and Industrialization Conference*, pp. 1-16. https://www.researchgate.net/publication/328095915_Possible_Approaches_to_Commercialisable_University_Research_in_Kenya

Osoba, O. et Welser IV, W. 2017. *An Intelligence in our Image : The Risks of Bias and Errors in Artificial Intelligence*. Santa Monica, RAND Corporation. https://www.rand.org/pubs/research_reports/RR1744.html

Picture a Scientist. 2020. Film cinématographique. Uprising Production, Antarctica, réalisation de Ian Cheney et Sharon Shattuck.

Puritty, C., Strickland, L. R., Alia, E., Blonder, B., Klein, E., Kohl, M. T., McGee, E., Quintana, M., Ridley, R. E., Tellman, B. et Gerber, L. R. 2017. Without inclusion, diversity initiatives may not be enough. *Science,* vol. 357, nº 6356, pp. 1101-1102. https://www.science.org/doi/full/10.1126/science.aai9054

Quantum Leap Africa. 2021. Preparing Africa for the Coming Quantum Revolution. https://quantumleapafrica.org

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., Mcelreath, R., Mislove, A., Parkes, D. C., Pentland, A. S, Roberts, M. E., Shariff, A., Tenenbaum, J. B. et Wellman, M. 2019. Machine behaviour. *Nature,* vol. 568, nº 7753, pp. 477-486. http://doi:10.1038/s41586-019-1138-y

Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J. et Denton, E. 2020. Saving face : Investigating the ethical concerns of facial recognition auditing. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society,* pp. 145-151. https://dl.acm.org/doi/pdf/10.1145/3375627.3375820

Reidpath, D. et Allotey, P. 2019. The problem of "trickle-down science" from the Global North to the Global South. *BMJ Global Health,* vol. 4, nº 4, pp. 1-3. https://gh.bmj.com/content/bmjgh/4/4/e001719.full.pdf

Riegle-Crumb, C., King, B. et Irizarry, Y. 2019. Does STEM stand out ? Examining racial/ethnic gaps in persistence across postsecondary fields. *Educational Researcher,* vol. 48, nº 3, pp. 133-144. https://journals.sagepub.com/doi/pdf/10.3102/0013189X19831006

Rosen, J. 2018. Black students who have one black teacher are more likely to go to college. Johns Hopkins University Hub, 12 novembre. https://hub.jhu.edu/2018/11/12/black-students-black-teachers-college-gap/

Ruta N. 2018. *Ruta N Medellín : Centro de Innovación y Negocios Inicio.* https://www.rutanmedellin.org/es/

Schia, N. N. 2018. The cyber frontier and digital pitfalls in the Global South. *Third World Quarterly,* vol. 39, nº 5, pp. 821-837. https://www.tandfonline.com/doi/pdf/10.1080/01436597.2017.1408403

Schneiderwind, J. et Johnson, J. M. 2020. Why are students with disabilities so invisible in STEM education ? *Education Week*, 27 juillet. https://www.edweek.org/education/opinion-why-are-students-with-disabilities-so-invisible-in-stem-education/2020/07

Schwab, K. et Vanham, P. 2021. What is stakeholder capitalism ? *European Business Review*. https://www.europeanbusinessreview.eu/page.asp?pid=4603

Skupien, S. et Rüffin, N. 2019. The geography of research funding : Semantics and beyond. *Journal of Studies in International Education,* vol. 24, nº 1, pp. 24-38. https://journals.sagepub.com/doi/pdf/10.1177/1028315319889896

Thompson, N. C., Greenewald, K., Lee, K. et Manso, G. F. 2020. The computational limits of deep learning. *MIT Initiative on the Digital Economy Research Brief,* vol. 4, pp. 1-16. https://arxiv.org/pdf/2007.05558.pdf

Trix, F. et Psenka, C. 2003. Exploring the color of glass : Letters of recommendation for female and male medical faculty. *Discourse & Society,* vol. 14, nº 2, pp. 191-220. https://journals.sagepub.com/doi/pdf/10.1177/0957926503014002277?casa_token=HxYIongTjWcAAAAA:47ON3AJXqdCxq3j__UTwOJvuuaJvaITBSeOFzG0I7MMO3GrcRvrTWTaXOUWjdvBPd6sUh0X8Veat

Tulshyan, R. et Burey, J. A. 2021. Stop telling women they have imposter syndrome. *Harvard Business Review*, 11 février. https://hbr.org/2021/02/stop-telling-women-they-have-imposter-syndrome

Turing, A. M. 1950. Computing machinery and intelligence. *Mind,* vol. 54, nº 236, pp. 433-460.

UNCTAD. 2021. *Technology and Innovation Report : Catching Technological Waves – Innovation with Equity*. New York, United Nations Publications. https://unctad.org/webflyer/technology-and-innovation-report-2021

United Nations. 2020. *Digital divide " a matter of life and death" amid COVID-19 crisis, Secretary-General warns virtual meeting, stressing universal connectivity key for health, development*. Communiqué de presse, 11 juin. https://www.un.org/press/en/2020/sgsm20118.doc.htm

United States Congress. 2020. *H.R.6216* – National Artificial Intelligence Initiative Act of 2020, pp. 1-56. https://www.congress.gov/bill/116th-congress/house-bill/6216/text#toc-H7A238FDF26594A338CB94267854F51D4

'Utoikamanu, F. 2021. Closing the technology gap in least developed countries. *UN Chronicle*. https://www.un.org/en/chronicle/article/closing-technology-gap-least-developed-countries

Viglione, G. 2020. NSF grant changes raise alarm about commitment to basic research. *Nature,* vol. 584, nº 7820, pp. 177-178. . https://www.nature.com/articles/d41586-020-02272-x

Voskoboynik, D. M. 2018. To fix the climate crisis, we must face up to our imperial past. *OpenDemocracy*, 8 octobre. https://www.opendemocracy.net/en/opendemocracyuk/to-fix-climate-crisis-we-must-acknowledge-our-imperial-past/

Warikoo, N., Sinclair, S., Fei, J. et Jacoby-Senghor, D. 2016. Examining racial bias in education : A new approach. *Educational Researcher,* vol. 45, nº 9, pp. 508-514. https://journals.sagepub.com/doi/full/10.3102/0013189X16683408

Zeng, D., Cao, Z. et Neill, D. B. 2021. Artificial intelligence-enabled public health surveillance – from local detection to global epidemic monitoring and control. L. Xing, M. L. Giger et J. K. Min [dir.], *Artificial Intelligence in Medicine,* pp. 437-453. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7484813/

Zhang, D., Mishra, S., Brynjolfsson, E., Etchemendy, J., Ganguli, D., Grosz, B., Lyons, T., Manyika, J., Niebles, J. C., Sellitto, M., Shoham, Y., Clark, J. et Perrault, R. 2021. *Artificial intelligence index report*. Stanford Institute for Human-Centered Artificial Intelligence. https://aiindex.stanford.edu/report/

PARADOXES DE LA PARTICIPATION DANS LA GOUVERNANCE INCLUSIVE DE L'IA : QUATRE APPROCHES CLÉS QUANT AU DISCOURS DU SUD ET DE LA SOCIÉTÉ CIVILE

MARIE-THERESE PNG

Candidate au doctorat au sein de l'Oxford Internet Institute et boursière de Google DeepMind, Marie-Therese Png a auparavant été conseillère en technologie au sein du Groupe de haut niveau sur la coopération numérique, des Nations Unies. Elle conseille actuellement des organisations sur l'éthique liée aux grands modèles linguistiques et les répercussions environnementales des infrastructures d'information.

ODD 8 - Travail décent et croissance économique

ODD 9 - Industrie, innovation et infrastructure

ODD 10 - Inégalités réduites

ODD 11 - Villes et communautés durables

ODD 15 - Vie terrestre

ODD 16 - Paix, justice et institutions efficaces

ODD 17 - Partenariats pour la réalisation des objectifs

PARADOXES DE LA PARTICIPATION DANS LA GOUVERNANCE INCLUSIVE DE L'IA : QUATRE APPROCHES CLÉS QUANT AU DISCOURS DU SUD ET DE LA SOCIÉTÉ CIVILE

RÉSUMÉ

Il a été estimé que l'intelligence artificielle (IA) pouvait engendrer des retombées économiques supplémentaires d'environ 13 billions de dollars américains d'ici 2030. Les pays et les entités les mieux placés pour en profiter sont toutefois ceux qui, au Nord, possèdent le plus de pouvoir économique, tandis que ceux qui sont déjà désavantagés – et qui proviennent démesurément du Sud – en paient le prix. Des initiatives de gouvernance inclusive de l'IA visent à remédier à cette répartition inégale, mais doivent d'abord s'attaquer à des problèmes structurels qui renforcent les inégalités.

De plus, les initiatives de gouvernance inclusive n'abordent pas en priorité des enjeux d'une importance particulière pour les pays du Sud : domination occidentale par rapport aux infrastructures et à la réglementation, droit de propriété exclusif, incompatibilités culturelles ou contextuelles, extraction de données et de matériaux, tenue de bêta-tests, ou encore droits des travailleurs et travailleuses.

Ce chapitre présente une approche méthodique en vue d'aborder ces enjeux fondamentaux, approche que peuvent adopter ceux et celles qui s'emploient à rendre réellement inclusive la gouvernance de l'IA ainsi que des domaines connexes tels que la législation du commerce mondial, la propriété intellectuelle, les normes techniques et la certification, et les droits humains. Il contient quatre recommandations pour parvenir à une gouvernance mondiale de l'IA qui soit inclusive et efficace : comprendre le discours que tient le Sud sur l'IA, définir, en collaboration, le rôle formel que jouent la société civile ainsi que les représentants ou représentantes des États et de l'industrie dans les processus de gouvernance mondiale de l'IA, cerner et lever les obstacles qui freinent la participation des parties prenantes du Sud et aborder le contexte historique des inégalités géopolitiques touchant la gouvernance de l'IA.

INTRODUCTION

Alors que la gouvernance de l'intelligence artificielle (IA) traverse des phases déterminantes de son évolution et que le « triopole » de l'Amérique du Nord, de la Chine et de l'Europe en mène les activités, ceux et celles qui sont à la tête d'initiatives en matière de gouvernance reconnaissent qu'ils et qu'elles ont la responsabilité de s'assurer que le déploiement de l'IA et la réglementation connexe n'intègrent pas d'inégalités à l'échelle nationale ou internationale. Nous constatons que les efforts se multiplient pour mettre l'IA « au service du bien » et pour y faire participer la société civile et les parties prenantes des pays du Sud. Cet élan se fonde sur la logique selon laquelle les paramètres, les garde-fous et les mécanismes de protection doivent être définis par ceux et celles qui connaissent les travers de l'IA et font face à leurs conséquences, ce que ne peuvent faire adéquatement les personnes qui sont à l'abri des risques que pose l'IA en raison du pouvoir et de la sécurité dont elles jouissent au sein d'institutions (Ulnicane *et al.*, 2020 ; Milan et Gutiérrez, 20158 ; Schiff *et al.*, 2021).

Par ailleurs, « les pratiques de gouvernance mondiale engendrent souvent des effets sociaux concurrents, par lesquels des tendances inclusives se combinent avec des tendances plutôt exclusives » (Pouliot et Thérien, 2017). Cela mène aux « paradoxes de la participation », dans lesquels l'inclusion peut exister alors que les dommages structurels persistent, et par lesquels les méthodes qui visent à accroître la participation citoyenne entraînent néanmoins une domination encore plus grande de *l'establishment* (Cleaver, 1999 ; Bliss and Neumann, 2008 ; Ahmed, 2012). Ce chapitre cherche à vérifier si les initiatives liées à la gouvernance inclusive de l'IA bénéficient concrètement à ceux et celles qui sont davantage exposés aux risques que posent les systèmes d'IA. Il propose que l'inclusion dans la gouvernance de l'IA passe par une réforme structurelle – redistribution des ressources, établissement d'un plan d'action et partage du pouvoir décisionnel (Fraser, 2005) – et que, au-delà de l'inclusion, les parties prenantes du Sud et la société civile contribuent à instaurer d'autres mécanismes de gouvernance.

Ce chapitre propose donc une approche méthodique résumée en quatre recommandations que peuvent adopter, pour satisfaire à des exigences de base, ceux et celles qui cherchent à ce que la gouvernance de l'IA soit véritablement inclusive, que ce soit au sein des organisations élaborant des normes techniques, des gouvernements, des organisations internationales ou de l'industrie. Des sections examinent les préoccupations concrètes du Sud, notamment l'exploitation des ressources naturelles, la main-d'œuvre bon marché du secteur numérique, des régimes de financement, et la domination de l'Occident dans la réglementation.

Le chapitre suggère quatre conditions préalables à une gouvernance mondiale efficace de l'IA.

Recommandation 1 : Comprendre le discours que tient le Sud sur l'IA (société civile mondiale, représentants ou représentantes des États ou de l'industrie, discours public) afin de cibler et d'intégrer de manière valable les demandes et les objectifs du Sud, et comprendre en quoi ils s'harmonisent ou non avec le processus de gouvernance en place.

Recommandation 2 : Définir, en collaboration, le rôle formel que jouent la société civile ainsi que les représentants ou représentantes des États et de l'industrie dans les processus de gouvernance mondiale de l'IA. Cette mesure est nécessaire pour que l'intégration des parties prenantes du Sud soit productive plutôt que performative, et elle vise une restructuration qui rendra les processus de gouvernance fiables et complets.

Recommandation 3 : Cerner et lever les obstacles qui empêchent les parties prenantes du Sud d'accéder au pouvoir décisionnel en matière de structures et d'infrastructures, et éviter de tomber dans les « paradoxes de la participation ». Il importe pour ce faire d'examiner tant le potentiel et les limites des parties prenantes du Sud que les processus historiques ayant fondé les inégalités.

Recommandation 4 : Aborder le contexte historique des inégalités géopolitiques touchant la gouvernance de l'IA en analysant le pouvoir et les dynamiques historico-politiques, par exemple des précédents quant à une asymétrie du pouvoir et l'exclusion transnationale observée dans la gouvernance mondiale d'autres technologies émergentes.

PRÉSENTATION DES CONCEPTS CLÉS

Intelligence artificielle

L'intelligence artificielle (IA) est une vaste branche de l'informatique qui vise, au moyen d'un ensemble de techniques, à concevoir des machines capables d'effectuer des tâches qui, traditionnellement, nécessitaient des fonctions cognitives humaines, par exemple la perception visuelle, la reconnaissance vocale, la prise de décisions et la traduction.

Dans ce chapitre, nous abordons l'IA de manière globale à titre de domaine de recherche, de technologie sous-jacente aux produits numériques et d'industrie. Nous envisageons également l'IA sous l'angle de son utilité politique, de sa matérialité, soit le matériel et les infrastructures qu'on y trouve (centres de données, unités de traitement graphique, etc.), des chaînes d'approvisionnement et de la dépendance à l'égard de la main-d'œuvre humaine (par exemple, le personnel qui gère ces données ou fait de l'annotation) (Crawford, 2021).

Gouvernance mondiale

La gouvernance mondiale constitue une évolution collaborative de l'éthique, de la politique et de la réglementation quant à des « questions qui sont devenues trop complexes pour qu'un seul État parvienne à les aborder ». Elle est le « produit d'un changement du paradigme néolibéral en matière de relations politiques et économiques internationales » (Jang *et al.*, 2016). Comme celle qui se rapporte à d'autres technologies émergentes (Ulnicane *et al.*, 2021), la gouvernance de l'IA fait interagir des structures multipartites afin de combler les lacunes d'une gouvernance dans laquelle des parties prenantes « du secteur privé et de la société civile assument des rôles d'autorité qui étaient auparavant considérés comme du ressort de l'État » (Jang *et al.*, 2016).

Sud et Nord

Il existe des distinctions entre la façon dont le discours politique dominant au Nord et le discours émergeant du Sud conceptualisent l'IA.

Depuis la fin de la Guerre froide, on associe le Nord à des États et à des économies stables, tandis que le Sud fait référence à des États-nations économiquement défavorisés. Mieux comprendre ce qu'est le Sud passe par le fait de le situer dans le concept de la géographie déterritorialisée des effets externes du capitalisme – où le Sud global n'est pas seulement défini comme les pays situés dans l'hémisphère Sud géographique, mais représente également « les peuples assujettis à l'intérieur des frontières des pays plus riches, de sorte qu'il y a des Suds dans le Nord géographique et Nords dans le Sud géographique » (Mahler, 2017). Il est également essentiel de tenir en compte que ces personnes supportent de manière disproportionnée les coûts de l'extractivisme et de l'exploitation par les économies capitalistes. Les perspectives du Sud se centrent sur les priorités, les préoccupations et le point de vue d'une majorité à l'échelle mondiale. Comme l'expriment Singh et Guzmán (2021), « nous considérons le Sud comme un impératif pour nous concentrer sur l'expérience que vivent les populations exclues, réduites au silence et marginalisées lorsqu'elles font face aux données et à l'IA au quotidien ».

Le discours dominant sur l'IA est mené par l'« Ouest » et le Nord, c'est-à-dire par l'Europe occidentale et l'Amérique du Nord, et il fait référence aux figures traditionnelles de pouvoir d'États puissants, industrialisés et riches, et, dans le cadre du modèle de gouvernance multipartite qui voit le jour, de l'industrie, des organisations de normalisation ainsi que des institutions militaires liées à la recherche et au financement. Il reproduit aussi la hiérarchie politique, épistémique, économique et morale qui s'est instaurée pendant la colonisation européenne. Comme l'ont dit Glissant et Dash (1999) : « L'Occident n'est pas à l'Ouest. Ce n'est pas un lieu, c'est un projet. »

Tant le Sud que le Nord sont hétérogènes. Compte tenu de la complexe pluralité des parties prenantes du Sud, il est crucial d'examiner les limites et l'utilité du « Sud » comme cadre d'analyse des asymétries de pouvoir actuelles et de la répartition inégale des risques liés à l'IA. D'une part, le Sud présente l'utilité d'unifier l'établissement d'une solidarité et, d'autre part, il omet l'hétérogénéité et l'incongruité interne du discours du Sud sur l'IA. Les « Suds » (Connell, 2007 ; Comaroff et Comaroff, 2016) représentent divers « régimes politiques, niveaux de développement, idéologies et intérêts géopolitiques » (Weiss, 2016) qui engendrent une contestation régionale et posent de réelles limites à la coordination et à la mobilisation collective. Le discours sur l'IA en provenance des Suds intervient « sur un large spectre allant de l'optimisme relatif au bond en avant et à la transformation numérique des sociétés, d'un côté, au pessimisme lié à la souffrance humaine résultant de nouvelles formes de capitalisme des données et de colonialisme, de l'autre » (Singh, 2021).

Les plans d'action du Sud et du Nord ne doivent pas être considérés comme intrinsèquement dichotomiques ou antagonistes. De nombreux pays « occupent une position qui se situe entre celle du Nord et celle du Sud », par exemple au sein de l'« Est » (Müller, 2018). En outre, la binarité Nord-Sud ne tient pas compte des peuples soumis à l'intérieur des frontières des pays riches ni des situations opposées « des Suds économiques dans le Nord géographique et des Nords dans le Sud géographique » (Mahler, 2017).

La binarité Nord-Sud se brouille devant le leadership qu'exercent le gouvernement chinois et l'industrie de ce pays dans la gouvernance de l'IA et en recherche et développement appliquée. Les puissants noms de l'industrie technologique sont les GAFAM (Google, Amazon, Facebook, Apple, Microsoft) aux États-Unis et les BATX (Baidu, Alibaba, Tencent, Xiaomi) en Chine. Le pouvoir qu'a la Chine en matière géopolitique, dans la recherche et la production ou dans l'établissement de normes façonne inévitablement le discours dominant et s'avère d'une grande importance pour le Sud. Comme l'indique Lee (2019) au sujet des pays en développement : « À moins qu'ils ne souhaitent plonger leur population dans la pauvreté, ils seront obligés de négocier avec le pays leur fournissant la plupart de leurs logiciels d'IA – que ce soit la Chine ou les États-Unis – pour ainsi essentiellement devenir économiquement dépendants de ces pays. »

RECOMMANDATION 1 – COMPRENDRE LE DISCOURS QUE TIENT LE SUD SUR L'IA

Une véritable participation des parties prenantes du Sud aux processus de gouvernance de l'IA, ainsi que l'intégration de leurs demandes et de leurs objectifs, nécessite de comprendre adéquatement le discours que tiennent différents pays du Sud sur l'IA ainsi que les activités de la société civile, des représentants et représentantes des États et de l'industrie, des organismes de recherche et, plus généralement, du public. Dans cette partie, nous résumons un sous-ensemble de tendances qui caractérisent le Sud et faisons ressortir des contrastes entre le discours qui porte sur l'IA « pour et par » le Sud et le discours dominant sur la gouvernance de l'IA.

Aperçu du discours que tient le Sud sur l'IA

Le discours que tient le Sud sur l'IA est de nature plurielle. Ainsi, nombre de discours s'appuient sur des travaux de longue date visant à instaurer des infrastructures numériques qui répondent aux besoins et aux préoccupations des pays à faible revenu ou à revenu intermédiaire, des populations habituellement marginalisées et des écosystèmes. Ces considérations sont couramment négligées par les décideurs et décideuses en place, qui s'accommodent d'un certain *statu quo* dominant. Les discours que tient le Sud sur l'IA ont tendance à s'inspirer de pratiques antihégémoniques ; ils se penchent sur les effets de l'impérialisme ainsi que sur des critiques constructives des structures capitalistes menant à des pratiques d'exploitation inadmissibles, inéquitables et nuisibles. Ces discours sur la gouvernance de l'IA ne proviennent pas exclusivement du Sud, mais aussi d'institutions et de communautés du Nord qui ne participent pas encore à grande échelle aux processus internationaux de gouvernance de l'IA.

Le Nord et le Sud abordent de manière contrastée les cadres normatifs de l'IA, la manière de cerner les problèmes et l'évaluation des risques. Compte tenu des politiques liées à la technologie (Winner, 1980), les études qui se centrent sur les Suds ou les communautés marginalisées subissant les préjudices de l'IA relèvent notamment de l'informatique postcoloniale (Irani *et al.*, 2010), de l'informatique décoloniale (Ali, 2016), de l'extractivisme de données (Couldry et Mejias, 2019 ; Ricaurte, 2019 ; Crawford, 2021), de l'IA culturellement adaptée et des droits humains (Mhlambi, 2020 ; Kak, 2020), du colonialisme des données (Birhane, 2020), de la souveraineté sur les données autochtones (Rainie *et al.*, 2019), des pratiques féministes (D'Ignazio et Klein, 2020), du design équitable (Costanza-Chock, 2020) et de la justice relative aux données (Milan et Terré, 2019 ; Taylor, 2017). Des communautés de solidarité transnationale et des mesures collectives voient également le jour, que ce soit par exemple la conférence de 2017 sur l'IA et l'inclusion, l'association Article 19, la Web Foundation, la coopérative Tierra Común, le Non Aligned Technologies Movement, le Global Data Justice Project, le Technology Justice Lab, l'initiative Big Data Sur, l'organisation Black in AI ou le Digital Asia Hub.

Fait important, les discours sur l'IA centrés sur le Sud ont mis en évidence la réalité physique et le travail humain derrière l'IA ou, pour ainsi dire, le « poids du nuage ». Pollicy, un organisme ougandais, présente succinctement des domaines d'action et de défense des intérêts relativement à ce qu'il appelle l'« extractivisme numérique ». Outre l'exploitation des ressources naturelles et de la main-d'œuvre bon marché du secteur numérique, ses membres attirent l'attention sur « les flux financiers illicites, l'extraction de données, les monopoles quant aux infrastructures, les prêts numériques, les structures de financement, la tenue de bêta-tests et la gouvernance des plateformes » (Iyer *et al.*, 2021). Puisque le discours dominant sur la gouvernance de l'IA néglige plusieurs de ces enjeux, la participation de parties prenantes du Sud s'avère très pertinente pour améliorer l'évaluation des risques et la gouvernance en recherche et développement, les infrastructures, les chaînes d'approvisionnement, ainsi que le déploiement et la réglementation de l'IA.

Préoccupations concrètes du Sud

Cette section présente les préoccupations concrètes du Sud, notamment quant aux différences culturelles, à la domination occidentale par rapport aux infrastructures et à la réglementation, au droit de propriété exclusif, aux incompatibilités contextuelles, à la durabilité et à l'extraction, à la tenue de bêta-tests, et aux droits des travailleurs et travailleuses.

Infrastructure occidentale, réglementation et droit de propriété exclusif

Il est essentiel que les pays du Sud conçoivent des infrastructures technologiques dont ils sont propriétaires pour que les Suds tirent profit du développement de l'IA (Rayment, 1983 ; Mbembe et Nuttall, 2004). Que le Sud adopte les « paysages et [l']histoire de l'Euro-Amérique quant aux infrastructures et à la réglementation » (Raval *et al.*, 2021), de même que ceux de la Chine, soulève des inquiétudes sur

la consolidation du pouvoir des gouvernements et des industries de l'Europe occidentale, de l'Amérique du Nord et de la Chine. Sampath décrit les enjeux critiques abordés dans le discours dominant sur la gouvernance de l'IA qu'il faut s'employer à résoudre en vue d'atténuer la dynamique de dépendance-extraction établie entre le Sud et le Nord. Ces enjeux comprennent des « modèles liés à la fabrication, à l'approvisionnement, au développement et à la tarification des technologies » (Sampath, 2021), les avantages du « premier arrivé » sur le marché et les partenariats public-privé inégaux.

La législation portant sur la propriété intellectuelle s'avère également déterminante pour le Sud. L'organisation intergouvernementale South Centre (2020), qui regroupe 54 pays en développement, dénonce le fait que les entreprises et les pays du monde développé possèdent le monopole des droits de propriété intellectuelle, ce qui brime l'autonomie des pays du Sud. Or, renverser un tel monopole est très complexe. Au départ, il faut notamment que les offices régionaux de propriété intellectuelle renforcent leurs capacités et élaborent des lignes directrices, en misant sur une coopération Sud-Sud, puis il faut prévoir des mesures incitatives pour que le secteur privé s'attarde aux besoins des pays en développement.

En ce qui concerne l'acquisition de technologies, le transfert de celles-ci (Gopalakrishnan et Santoro, 2004) et l'assistance technique représentent des solutions efficaces à court terme, mais ne permettent pas aux gouvernements et aux populations du Sud de réellement bénéficier des gains économiques engendrés par les technologies. En particulier, les droits attachés aux brevets peuvent « considérablement réduire le transfert de technologie, car ils entraînent des droits de permis élevés et peuvent donc freiner l'adaptation des connaissances aux réalités locales » (Kane, 2010).

Un monopole semblable touche les brevets associés aux vaccins contre la COVID-19. De riches pays industrialisés tels que les États-Unis, le Royaume-Uni et certains États de l'Union européenne ont empêché une augmentation de la production de vaccins à l'échelle mondiale. En refusant une proposition de dérogation à l'Accord sur les ADPIC[68] de l'Organisation mondiale du commerce, dérogation qui allait suspendre les brevets sur le matériel médical lié à la COVID-19 nécessaire à une immunité collective, ils ont favorisé les profits des sociétés pharmaceutiques (Vawda, 2021).

Différences culturelles et incompatibilités contextuelles

Importer au Sud et employer dans un nouveau contexte des systèmes d'IA dont la conception et l'entraînement reposent sur des ensembles de données recueillies dans des contextes occidentaux, et reflétant donc des caractéristiques particulières en matière de démographie, d'identité, de parenté, de religion, de culture, de sociopolitique, de réglementation et d'infrastructure, entraîne sans surprise des « conséquences involontaires ». Un algorithme d'apprentissage entraîné au moyen d'ensembles de données nord-américaines ne peut être directement mis en œuvre en Amérique centrale, en Afrique ou en Asie sans risque de causer des incompatibilités contextuelles (Neupane et Smith, 2017). Il y a moyen d'éviter ces conséquences « involontaires », ou plutôt imprévues, en s'associant à des spécialistes locaux et aux communautés en cause pour élaborer des technologies d'IA et des politiques à cet égard. La collecte participative de données (Graham *et al.*, 2015) pourrait accroître le besoin d'ensembles de données pertinentes sur le plan local dans les économies en développement (Quinn *et al.*, 2014) et permettre l'entraînement de modèles d'IA reprenant des « caractéristiques uniques qui ne sont pas présentes dans d'autres environnements » (Lee *et al.*, 2020).

Les sujets de préoccupation qui émergent au Nord ont généralement trait aux préjugés et à l'équité, à la responsabilité, à la transparence, à l'IA explicable et à l'IA responsable (Singh, 2021). La réelle pertinence de traduire ces concepts et de les importer dans les contextes sociopolitiques, économiques, culturels, linguistiques et relatifs aux infrastructures du Sud a fait l'objet d'études et elle est contestée

68. Accord sur les aspects des droits de propriété intellectuelle qui touchent au commerce.

(Raval *et al.*, 2021). Dans une allocution qu'il a faite lors d'une conférence sur l'éthique de l'information qui s'est tenue à Pretoria en 2007, le professeur Rafael Capurro a décrit une « éthique de l'information pour et par l'Afrique » qui fait ressortir le monopole des traditions éthiques occidentales (hétérogènes) dans l'éthique, la gouvernance et la réglementation liées aux technologies de l'information et de la communication ainsi qu'aux systèmes d'information automatisés.

Le comité d'éthique traditionnelle de l'Institute of Electrical and Electronics Engineers (IEEE, 2019) postule que les monopoles occidentaux nuisent à l'élaboration de normes relatives à l'IA qui soient pertinentes à l'échelle mondiale, ce qui est un « projet intrinsèquement chargé de valeurs, car il statue sur les critères normatifs à intégrer au réseau mondial » (Wong, 2016). L'application généralisée des valeurs culturelles et politiques occidentales peut délégitimer « la plausibilité de [l'innovation responsable] fondée sur des valeurs locales, en particulier lorsque ces valeurs entrent en conflit avec les valeurs démocratiques libérales » qui « ne permettent pas aux scientifiques et aux personnes qui conçoivent des technologies d'être reconnus en tant que membres du réseau mondial de la recherche et de l'innovation » (Wong, 2016). À titre d'exemple, la confidentialité des données continue d'être envisagée dans une optique occidentale (Arora, 2018).

Durabilité et extraction

Comme l'explique Crawford (2021), « l'économie des données se fonde sur le maintien d'une ignorance de l'environnement ». À l'heure où l'IA parvient à optimiser la consommation énergétique et à soutenir la mise au point de technologies vertes, les cadres de gouvernance ne peuvent omettre les coûts environnementaux de l'infrastructure liée à l'IA et à l'information (Parikka, 2015). Les préoccupations environnementales gravitent toujours en périphérie du discours dominant sur la gouvernance de l'IA, et des initiatives émergentes, par exemple le projet de stratégie d'IA responsable à l'égard de l'environnement du Partenariat mondial sur l'intelligence artificielle (Clutton-Brock *et al.*, 2021), passent à côté de nombreuses préoccupations soulevées par le Sud.

Les modèles d'entraînement de l'apprentissage automatique requièrent beaucoup d'énergie, et les systèmes d'IA reposent sur des infrastructures matérielles (centres de données, unités de traitement graphique, semi-conducteurs), ce qui stimule la demande de minéraux de terres rares. En 2020, le Parlement européen a rapporté que « l'extraction du nickel, du cobalt et du graphite pour en faire usage dans les batteries lithium-ion, que l'on trouve couramment dans les voitures électriques et les téléphones intelligents, avait déjà endommagé l'environnement, et [que l'IA allait] vraisemblablement accroître cette demande » (Bird *et al.*, 2020 ; Khakurel *et al.*, 2018). Une demande accrue limite l'approvisionnement en métaux de terres rares, d'autant plus qu'il faut accéder à des environnements complexes pour les extraire, ce qui intensifie l'automatisation de l'exploitation minière et de l'extraction des métaux (Khakurel *et al.*, 2018). De plus, Microsoft, Google et Amazon ont signé des contrats prévoyant la fourniture d'outils à l'industrie pétrolière et gazière afin d'optimiser l'extraction (Greenpeace, 2020 ; Conger *et al.*, 2020), ce qui contribue à la dégradation de notre environnement et des services écologiques.

En 2021, l'Observer Research Foundation a organisé un atelier sur les risques environnementaux que présente l'IA, particulièrement pour les communautés marginalisées, en utilisant des cadres relatifs au racisme environnemental. Il est reconnu que l'industrie de l'extraction nuit d'abord aux groupes racisés, vulnérables et négligés, étant donné l'exploitation des travailleurs et travailleuses, le travail des enfants, la violence de l'État envers les communautés autochtones et l'augmentation de la violence sexiste (Legassik, 1974).

Ainsi, le caractère « éthique » et « responsable » du déploiement et de la gouvernance des systèmes d'IA nécessite une évaluation des risques et des coûts associés à l'ensemble du système, une évaluation à laquelle procèdent actuellement des praticiens et praticiennes, chercheurs et chercheuses, et

défenseurs et défenseuses des intérêts du Sud. L'évaluation touche l'automatisation des inégalités (Eubanks, 2018 ; Noble, 2018), l'infrastructure matérielle des systèmes d'IA et les chaînes d'approvisionnement en matériaux (Crawford, 2021).

Droits des travailleurs et travailleuses, et travail fantôme

Une évaluation des risques et des coûts associés à l'ensemble du système prend également en considération les personnes qui effectuent du travail ou du « travail fantôme » (Gray et Suri, 2019) consistant à annoter « les grands volumes de données nécessaires afin de faire ressortir les éléments de sens commun qui rendent les données utiles pour une tâche déterminée » (Mohamed *et al.*, 2020). Des travailleurs ou travailleuses fantômes, souvent établis au Sud, sont engagés par des entreprises ou plateformes d'annotation spécialisées. Celles-ci fournissent effectivement des emplois, mais manquent souvent de structures ou de politiques de responsabilisation pour protéger les gens de l'exploitation, par exemple de pratiques comme une retenue de la rémunération ou le refus de reconnaître « les droits des travailleurs et travailleuses à des conditions sûres et dignes » (Irani et Silberman, 2013), ce qui touche les personnes économiquement vulnérables, notamment dans les administrations où le droit du travail est peu établi (Yuan, 2018).

Les syndicats mondialisés et les coalitions de travailleurs et travailleuses de l'industrie de la technologie, tels Turkoptikon et UNI Global Union, ou des projets de recherche comme Fairwork représentent des sources d'expertise pour comprendre les préjudices et les risques auxquels font face les travailleurs et travailleuses faisant rouler l'économie de l'IA. Ils sont efficaces parce qu'ils centralisent les connaissances de ces derniers, ce qui s'avère nécessaire à la mise en place de solides garde-fous et d'une réglementation qui les protègent.

Tenue de bêta-tests

Tant le travail fantôme que la tenue de bêta-tests constituent des pratiques industrielles potentiellement exploitantes. Les bêta-tests contribuent au « réglage précis des premières versions de systèmes logiciels et aident à cerner les problèmes liés à leur utilisation dans des conditions réelles, avec de vrais utilisateurs ou utilisatrices » (Mohamed *et al.*, 2020). Or, il a été démontré que, à l'occasion de bêta-tests, des entreprises exposent des groupes déjà vulnérables aux risques d'un produit. Mentionnons par exemple le déploiement d'un système de prévision des actes criminels par Palantir à La Nouvelle-Orléans, et l'essai d'une technologie d'analyse des données électorales par Cambridge Analytica tenu au Kenya et au Nigeria plutôt que dans des démocraties occidentales (Mohamed *et al.*, 2020). Il existe une tendance à sélectionner des communautés systématiquement moins protégées ou plus exposées aux risques que d'autres, ou à effectuer des bêta-tests dans des États n'offrant ni garanties ni réglementation quant à l'utilisation des données, ce qui profite aux entreprises, car les modalités des tests enfreindraient les lois de leur pays d'origine (UNCTAD, 2013).

RECOMMANDATION 2 – COMPRENDRE ET FORMALISER LES RÔLES DU SUD ET DE LA SOCIÉTÉ CIVILE

Pour qu'il y ait une réelle intégration des parties prenantes du Sud aux processus de gouvernance de l'IA, c'est-à-dire une cogouvernance, nous devons comprendre et formaliser les rôles que jouent ces parties prenantes. Ces rôles doivent être définis en collaboration avec des représentants et représentantes de la société civile, de l'industrie et des États du Sud, et établis dans des structures réformées garantissant que l'intégration aux processus de gouvernance est productive, et non performative.

Les rôles proposés pour les parties prenantes du Sud et présentés dans ce chapitre sont de quatre ordres : (i) contester les mécanismes d'une gouvernance excluante, (ii) fournir une expertise légitime quant à l'interprétation et à la contextualisation des risques, des préoccupations, des exigences et des enjeux formulés. (iii) établir des structures de responsabilisation en démocratie au sein des États et des processus de gouvernance internationale et (iv) proposer de nouveaux mécanismes de gouvernance.

Interprétation légitime et contextualisation des risques

Peu importe le contexte géographique ou social, il revient à ceux et celles qui subissent les contrecoups des systèmes d'IA de cerner les préjudices évitables et d'encadrer la mise en œuvre de ces systèmes. Les risques que pose l'IA et les garde-fous visant à les éviter ne peuvent être déterminés de manière adéquate par les personnes qui se trouvent à l'abri de ces risques en raison du pouvoir et de la sécurité dont elles bénéficient au sein des institutions (Ulnicane *et al.*, 2021 ; Milan et Treré, 2019 ; Schiff *et al.*, 2021). Les parties prenantes du Sud ont la capacité légitime d'interpréter les enjeux auxquels elles font face et de braquer les projecteurs sur des risques largement mis de côté dans les discussions générales sur la gouvernance de l'IA. Étant donné l'importante crédibilité accordée aux États et à des institutions d'élite pour la plupart établis en Europe occidentale, en Amérique du Nord et en Chine, la société civile et les parties prenantes du Sud bénéficient actuellement d'une légitimité, d'une visibilité et d'une influence moindres.

L'idée d'orienter le savoir sur celui des personnes les plus vulnérables aux risques est employée depuis longtemps en recherche-action participative et dans une approche critique des études du développement, et c'est le cas dans l'évaluation des risques liés aux produits réalisée avec les communautés touchées. Une telle pratique se résume bien par le slogan « rien sur nous sans nous », issu des traditions politiques de l'Europe centrale (Smogorzewski, 1938) et repris plus tard par le mouvement des droits des personnes handicapées relativement à la mise au point de technologies innovantes (Werner et PROJIMO, 1998).

Comprendre que les groupes touchés ont une expertise légitime est également essentiel pour contrebalancer la tendance qu'a le discours dominant sur la gouvernance en IA à universaliser les notions associées aux dommages d'une manière qui pourrait ne pas s'appliquer à différentes cultures, régions ou administrations. Il en découle une proposition selon laquelle il faudrait adopter, pour coordonner les interventions à l'échelle mondiale, non pas des normes universelles, mais bien des normes comparables. Les régions, les États et les villes « doivent être en mesure de répondre aux exigences particulières qu'ont leurs citoyens et citoyennes sur le plan social, économique et culturel » (Abdala *et al.*, 2020). L'« universalisation » impose souvent la perspective hégémonique, incompatible avec le point de vue local, des économies de l'information matures de l'Amérique du Nord et de l'Europe occidentale (Mignolo, 2012). Arora (2018) fait ressortir ces limites dans le contexte de la confidentialité à l'ère numérique : « alors que les entreprises technologiques étendent leur portée dans le monde entier, on continue d'envisager la notion de confidentialité à travers un prisme ethnocentrique. Cette notion se fonde de manière disproportionnée sur des données empiriques provenant de groupes démographiques occidentaux et blancs de la classe moyenne ». Arora fait valoir, entre autres choses, une perspective du Sud dans laquelle la réglementation en matière de confidentialité « rend leur dignité à ceux et celles qui se trouvent en marge, en assurant le respect de la confidentialité en fonction du contexte ».

Contestation et responsabilisation en démocratie

Le rôle du Sud relativement au discours sur la gouvernance liée à l'IA en est à tout le moins un de contestation d'une gouvernance excluante et des processus institutionnalisés qui négligent de tenir compte des communautés marginalisées ou qui leur causent du tort (Marchetti, 2016). Les défis et les interventions sont déterminés par des acteurs étatiques et une « économie politique de la résistance » en expansion menée par l'activisme de la société civile (Taylor, 2017 ; Torres, 2017 ; Milan et Treré,

2019). L'« activisme des données » décrit « de nouvelles formes de participation politique et d'engagement citoyen à l'ère de la mise en données » et vise à atténuer réellement des préjudices pourtant évitables que pourraient causer des systèmes d'IA (Milan et Velden, 2016).

En tant qu'« espace non gouvernemental et non commercial d'association et de communication » (Jaeger, 2007), la société civile constitue une bonne structure de responsabilisation pour les organes de gouvernance étatiques et mondiaux. Ces organes souffrent souvent du « manque de mécanismes formels de responsabilisation en démocratie au sein des États » et, au lieu de cela, « les conseils exécutifs des organismes mondiaux de réglementation se composent principalement de bureaucrates très peu au fait des contextes dans lesquels leurs décisions se répercutent », ce qui illustre la nature inaccessible et opaque des principaux processus de gouvernance mondiale de l'IA (McGlinchey *et al.*, 2017). Sur la scène internationale, la société civile « se concentre principalement sur l'élaboration de cadres politiques qui intègrent la responsabilisation en démocratie » en soutenant, en renforçant et en appliquant une pression pour réformer la législation, les garde-fous de la réglementation et les cadres fondés sur les droits. La société civile joue également un rôle de diffusion et renforce la coopération quant à des préoccupations concrètes du Sud (Marchetti, 2016) qui sont exclues des plans d'action transnationaux sur la gouvernance en IA.

La société civile mondiale a soutenu la responsabilisation en veillant à la transparence des activités de gouvernance mondiale, en surveillant et en examinant les politiques, en demandant réparation pour les erreurs et les préjudices imputables aux organismes de réglementation, et en proposant la création de mécanismes formels de responsabilisation en ce qui a trait à la gouvernance mondiale (Scholte, 2004). Étant donné que tous les secteurs de la société en bénéficient, ces efforts devraient profiter des ressources qui leur font souvent défaut quant à la gouvernance en IA, et il faudrait compenser la participation volontaire à des consultations (McGlinchey *et al.*, 2021).

La Tech Equity Coalition, fondée par le bureau de Washington de l'Union américaine pour la société civile (ACLU Washington), illustre la manière dont la société civile a amélioré la responsabilisation en démocratie en protégeant les droits et les libertés civiles des communautés marginalisées devant des technologies de plus en plus puissantes. La coalition a recours à des politiques, à la recherche, au règlement et à des procès. Quand la législation en vigueur semble injuste ou qu'une loi est enfreinte, elle s'adresse de manière stratégique aux tribunaux, aux assemblées législatives et aux communautés pour s'assurer que des organisations privées ou des organismes financés par le gouvernement respectent la loi. Ses membres élaborent également des propositions de politiques en partenariat avec les communautés directement touchées, préconisant des politiques adaptées aux communautés de même que des lois qui prévoient des mesures de protection à l'égard des technologies d'IA et de leurs données. Lorsqu'il est nécessaire de le faire, la coalition demande que cesse l'utilisation de technologies manifestement nuisibles, par exemple les outils permettant une surveillance policière excessive.

Source de nouveaux mécanismes de gouvernance

Il existe une distinction entre l'appel du Sud et celui de la société civile pour établir de nouvelles formes d'organisations politiques ainsi que d'autres objectifs politiques et types d'actions en vue de régir les technologies de l'IA de manière sûre et équitable (Milan et Velden, 2016). Avant d'insister sur la représentation des parties prenantes du Sud dans les processus de gouvernance de l'IA, nous devons nous interroger sur la faisabilité de cet objectif, car les processus dominants en matière de gouvernance de l'IA peuvent tout simplement ne pas s'y prêter structurellement.

> Nous devons changer les institutions qui, historiquement, ont été mises en place comme des outils d'avancement et de contrôle pour certaines personnes, mais d'exclusion de nombreuses autres, et pas seulement les fignoler ou chercher des moyens de créer un espace. Sans cela, l'avancement de

> la science et de la technologie continuera à profiter à ceux et celles qui ont gouverné au fil de l'histoire, au détriment de ceux et celles qui ont été exclus[69] (Sampath, 2021).

Des groupes dirigés par des Autochtones, comme la Global Indigenous Data Alliance, et d'autres groupes partageant le point de vue du Sud réfléchissent à la conception des prochaines technologies de l'IA et à la gouvernance à cet égard. Les appels à une réforme des mécanismes de gouvernance actuels et à la création de nouveaux mécanismes tiennent compte du fait que des facteurs administratifs, culturels, économiques et épistémologiques du colonialisme européen teintent la gouvernance mondiale (Quijano, 2000 ; Sampath, 2021).

Les forums de coopération Sud-Sud tels que l'initiative des Nations Unies pour la coopération Sud-Sud, le Groupe des 77 et le mouvement des non-alignés, qui ont joué un rôle clé dans les mouvements de décolonisation et d'indépendance, représentent de nouveaux mécanismes de gouvernance. Bien qu'ils soient dirigés par des États et aux prises avec des tensions entre l'État et la société civile, ces organismes servent de plateforme aux pays du Sud afin de consolider et de protéger des intérêts collectifs, de promouvoir la coopération géopolitique Sud-Sud, d'affirmer l'autodétermination par une action multilatérale et de prendre part au renforcement des capacités par des moyens qui érodent la dynamique de dépendance envers les États et l'industrie du Nord (Weiss, 2016).

RECOMMANDATION 3 – CERNER ET LEVER LES OBSTACLES À LA PARTICIPATION DU SUD

Taxonomie des pièges : obstacles à la gouvernance inclusive de l'IA

Lors du séminaire de Chatham House au cours duquel cet institut britannique a publié son rapport de 2021 sur la gouvernance inclusive, les personnes présentes se sont fait demander quels étaient les principaux obstacles à la participation inclusive en matière de gouvernance mondiale. Les réponses mentionnaient la géopolitique, le financement, la capacité, les déséquilibres de pouvoir, les élites, le déficit démocratique, la langue et le manque de mécanismes d'inclusion. D'autres problèmes ont également été soulevés, tels que la faible gouvernance nationale, l'esprit néocolonial, les jeux d'influence cachés, la fragmentation des efforts, la méfiance, la titrisation des créances, les obstacles juridiques, le racisme, les mécanismes rigides et la fracture numérique (Chatham House, 2021). Un grand nombre de ces problèmes observés par le public ne sont pas encore abordés dans les initiatives de gouvernance inclusive de l'IA.

Les efforts visant à supprimer l'exclusion transnationale et intersectorielle systématique en ce qui a trait à la gouvernance mondiale des technologies émergentes ne sont pas nouveaux (Ulnicane *et al.*, 2021). Pour atténuer de manière responsable les dynamiques d'exclusion et leurs effets néfastes en aval, et pour assurer une répartition équitable des retombées de l'IA, il importe absolument de comprendre les obstacles systémiques, ou les « pièges », qui empêchent les parties prenantes issues de la société civile du Sud d'accéder à différentes formes de pouvoir. Ces pièges comprennent, entre autres choses, la culture organisationnelle, les logiques structurelles d'exclusion, l'utilisation de termes généraux, un savoir-faire technologique insuffisant, l'appropriation, et des mesures incitatives financières défavorables. La métaphore des pièges provient de Selbst et autres (2020), qui définissent

69. *We need to change the institutions that have historically been set up as tools of advancement and control for some, to the exclusion of many, and not just tweak it or look for ways to create space. Without this, advancement of science and technology will continue to benefit those who governed historically at the expense of those who were excluded.*

ceux-ci comme des « modes de défaillance résultant de l'incapacité à prendre en considération ou à comprendre adéquatement les interactions qui existent entre les systèmes techniques et les mondes sociaux ».

Culture organisationnelle

Schiff et autres (2021) notent que le monde de la recherche « ferait bien de se focaliser sur les contextes organisationnels ou sectoriels qui façonnent l'éthique de l'IA et sur la manière dont ces contextes guideraient les priorités et les actions liées à l'éthique ». Les processus de gouvernance de l'IA comprennent des structures organisationnelles et des normes d'exclusion qui reflètent plus largement des inégalités de pouvoir interpersonnelles, géopolitiques et historiques (Wilson, 2000). Les motivations organisationnelles qui modèlent la gouvernance de l'IA peuvent nuire à l'inclusion, qu'elles touchent « un avantage compétitif, la planification ou l'intervention stratégiques, ou le fait d'aborder la responsabilité sociale ou le leadership » (Schiff *et al.*, 2020). Des structures de financement, et les paramètres qui servent à les évaluer, sous-tendent également des motivations individuelles qui laissent place au carriérisme et à l'opportunisme politique.

Les initiatives de gouvernance de l'IA sont souvent opaques, et on y contribue sur invitation en fonction de la crédibilité institutionnelle, de la compatibilité linguistique (le travail s'y fait généralement en anglais) et de plans d'action établis par de puissantes figures des États ou de l'industrie. Même quand elles sont invitées à y prendre part, les parties prenantes du Sud ou de la société civile font face à des « filtres institutionnels » (McGlinchey, 2021) et sont tenues en périphérie de la prise de décision finale. De nouvelles structures institutionnelles sont également « continuellement en train d'émerger, et le défi sur le plan de l'intégration est donc sans cesse renouvelé » (Marchetti, 2016), ce qui requiert des efforts et des ressources supplémentaires pour s'adapter.

Logiques structurelles

Étroitement liées à la culture organisationnelle, les logiques structurelles sont ancrées dans les institutions et les bureaucraties qui façonnent les processus de gouvernance. Elles normalisent la manière dont une personne doit ou ne doit pas raisonner, et indiquent quelles sont les prises de position politiques appropriées dans une certaine sphère sociale. Nous constatons qu'il y a, dans la gouvernance de l'IA, une reproduction d'élans interventionnistes et de modèles qui tiennent peu compte du Sud (Latonero, 2019 ; Vinuesa *et al.*, 2020), et même un amalgame entre la capacité des ressources et la capacité inhérente des parties prenantes marginalisées. Taylor (2019) explique que, bien que les parties prenantes du Sud subissent effectivement des échecs systémiques, « une cause fondamentale de l'échec des projets de développement réside dans les attitudes, adoptées par défaut, de paternalisme, de recherche d'une solution technologique et d'inclusion prédatrice ».

Les normes diplomatiques en matière de gouvernance mondiale de l'IA servent à réduire les frictions qui existent nécessairement entre les parties prenantes et favorisent un comportement performatif et hiérarchique. Ce qui est considéré comme diplomatique et poli « traduit souvent les structures de pouvoir en place et renforce les modèles d'interaction existants [...] La politesse reflète et favorise généralement les structures de pouvoir dominantes » (Roberts, 2018). Si les parties prenantes du Sud et de la société civile jouent un rôle de contestation quant aux processus de gouvernance de l'IA, il n'est pas surprenant que les contestations soient qualifiées de non diplomatiques, voire d'impolies, car elles enfreignent les normes dominantes. Les normes diplomatiques obscurcissent parfois les dynamiques de pouvoir et justifient l'exclusion de ceux et celles qui, en défendant les personnes marginalisées, causent des remous (Kazmi, 2012 ; McConnell *et al.*, 2012). Afin de survivre, les organisations défendant l'intérêt public doivent souvent opérer « comme un sous-système de la politique mondiale plutôt que de s'opposer au système de l'extérieur » (Jaeger, 2007).

Utilisation de termes généraux

Le langage employé dans la gouvernance de l'IA fait de plus en plus l'objet d'un examen minutieux. En se penchant sur les initiatives de l'IA « au service du bien », Green (2019) affirme que « bien » ne suffit pas, faisant référence à la définition limitée et vague de l'expression « bien social ». Tant la connaissance socioculturelle que le savoir-faire technique sont essentiels pour établir des définitions utiles.

En comparant l'engagement du secteur privé et celui des organisations non gouvernementales et du secteur public relativement aux questions éthiques liées à l'IA, Schiff et autres (2021) remarquent que les préoccupations de ce second groupe ont « une plus large portée éthique quant au nombre de sujets couverts, [qu'elles] ont plus souvent trait à la loi et à la réglementation, et [qu'elles] résultent davantage de processus participatifs ». Elles fondent ainsi des stratégies précises pour aborder des enjeux sociaux, éthiques et politiques liés à l'IA. Le caractère général des termes adoptés par les entités privées ou les organisations internationales penche par ailleurs vers une interprétation qui protège de manière opportune les intérêts des riches gouvernements ou représentants et représentantes de l'industrie. Comme pour les normes diplomatiques, l'emploi de termes généraux favorise la structure de pouvoir dominante, contrairement à ce que fait la précision sociopolitique des approches plus critiques ou radicales.

Par exemple, le Groupe d'experts indépendants de haut niveau sur l'intelligence artificielle, constitué par la Commission européenne, a récemment publié les projets de loi parmi les plus importants pour régir l'IA. Des membres de la société civile ont toutefois fait remarquer que ceux-ci ne respectent pas les normes de base en matière de protection des droits numériques et qu'ils comportent des lacunes qui exposent les citoyens et citoyennes à des utilisations abusives ou malveillantes de l'IA. La nature générale des projets de loi laisse à l'industrie technologique une grande marge de manœuvre pour se réglementer elle-même ; « de nombreux groupes industriels se sont dits soulagés que les règlements ne soient pas plus stricts, tandis que des groupes de la société civile ont déclaré que ceux-ci auraient dû aller plus loin » (Satariano, 2021).

Appropriation et récupération

L'utilisation de termes et d'expressions apparemment largement acceptés tels que *droits humains*, *durabilité*, *interdépendance*, *souveraineté* ou même *inclusion* et *participation* varie radicalement au sein des groupes de parties prenantes. Ces termes et expressions risquent également d'être repris par des intérêts particuliers (Ulnicane *et al.*, 2020) et redéfinis pour servir les priorités du secteur privé ou des pays riches, qui se voient conférer « le pouvoir de le faire en raison de leur statut d'élite, de leurs connaissances spécialisées ou de leur éventuelle capacité à entraver des engagements ou des objectifs essentiels » (Selznick, 2015).

Nous faisons référence à l'appropriation en utilisant en parallèle le terme *récupération*, pour décrire « le processus par lequel des idées et des images politiquement radicales sont déformées, reprises, absorbées, désamorcées, incorporées, annexées ou transformées en marchandises [...] et interprétées selon une perspective neutralisée, inoffensive ou socialement conventionnelle » (Downing *et al.*, 2001). Cette *récupération* s'oppose au révolutionnaire *détournement*, « un plagiat subversif qui détourne le langage et l'imagerie du spectacle de l'utilisation qu'on prévoyait en faire » (Downing *et al.*, 2001). Le *détournement* consiste à reprendre des expressions des systèmes dominants et à les retourner contre elles-mêmes et à l'écart de leur définition habituelle, au service de la contestation ou de la mobilisation.

Participation

L'inclusion se résume souvent à des procédures ou à une statistique, sert des fins de marketing, met de l'avant la vertu ou constitue une « inclusion d'optique ». L'inclusion et la participation ont certainement lieu d'être même si une structure tolère encore les préjudices, et il y a « peu de preuves

selon lesquelles la participation est efficace à long terme pour améliorer concrètement les conditions des personnes les plus vulnérables ou [qu'elle constitue] une stratégie de changement social » (Cleaver, 1999). Cette apparente contradiction illustre les « paradoxes de la participation » (Cleaver, 1999 ; Bliss et Neumann, 2008 ; Williams, 2004 ; Ahmed, 2012).

Il est donc nécessaire de comprendre que l'objectif de l'inclusion est de réformer les structures par une redistribution des ressources allouées, une redéfinition du plan d'action et une autre répartition du pouvoir de décision (Fraser, 2005). Selon le spectre des dynamiques d'inclusion et d'exclusion fourni par Marchetti (2016), qui couvre l'ostracisme, l'exclusion, l'appropriation, l'inclusion et l'intégration, des éléments auxquels pourraient s'ajouter la réforme structurelle et l'altérité, l'inclusion n'est que le premier pas dans la bonne direction pour s'éloigner d'une exclusion dommageable.

Interdépendance

Au cours des dernières décennies, le terme *interdépendance* s'est fait populaire pour décrire la « coopération internationale devant un ordre mondial de plus en plus complexe et globalisant » (Keohane et Nye, 1977 ; Coate *et al.*, 2015). Le plan d'action du Secrétaire général des Nations Unies pour la coopération numérique (paru en 2020) nous situe dans « l'ère de l'interdépendance numérique ». Mais à qui ces relations d'interdépendance profitent-elles ? Manifestement, le discours dominant n'aborde pas de manière adéquate les asymétries de pouvoir au sein des actuelles relations d'interdépendance. Les économies du Nord dépendent de l'extraction continue des ressources naturelles et de la main-d'œuvre du Sud, et le Sud systématise sa dépendance envers le Nord en ce qui a trait aux biens de consommation, à l'infrastructure numérique, à la réglementation commerciale et financière, etc.

Droits humains

Selon Latonero (2018), « pour que l'IA profite au bien commun, sa conception et son déploiement, à tout le moins, devraient éviter de heurter les valeurs humaines fondamentales. Les droits humains reconnus internationalement énoncent ces valeurs de manière affirmée et globale ». Ils comprennent le droit à la sécurité sociale, le droit au travail, la liberté d'expression, le respect de la vie privée, la sécurité sociale et le droit de ne pas subir de discrimination (Arun, 2020). Des politiques et même des logiciels et du matériel contribuent à mettre en œuvre ces droits.

Des organisations de la société civile soulignent que, dans la pratique, une appropriation de longue date des cadres relatifs aux droits humains (Peck, 2011) et la question de savoir qui bénéficie effectivement des droits humains se sont reportées dans le discours portant sur l'IA et les droits humains. L'intégration de ces droits dans la mise au point et le déploiement des systèmes d'IA, tant par les gouvernements que par les entreprises privées, ne repose souvent que sur des paroles. La recherche du profit et la concurrence qui s'exerce sur le marché international se font au détriment des droits individuels et collectifs.

Dans le contexte où c'est largement le Nord qui conçoit les technologies de l'IA qu'il exporte vers le Sud et qui veille à la réglementation, il importe d'accorder une attention particulière aux « pièges associés aux droits humains, particulièrement ceux qui se rapportent à la critique selon laquelle ces droits seraient trop occidentaux, trop individualistes, trop abstraits ou de portée trop étroite pour former la base d'une bonne gouvernance de l'IA » (Smuha, 2020). Certains documents reprennent de telles critiques, par exemple *Ubuntu as an Ethical and Human Rights Framework for AI Governance*, de Mhlambi (2021).

Incitatifs financiers

Les tensions entre les objectifs prosociaux (orientés vers les personnes) et les objectifs économiques (axés vers le profit) (Schiff *et al.*, 2021) entraînent des dynamiques asymétriques entre le Nord et le Sud. Ainsi, les décideurs et décideuses politiques devraient faire preuve d'astuce quant à la façon dont le financement influe sur les programmes de gouvernance et les dynamiques de participation ou d'exclusion.

Les initiatives de gouvernance de l'IA semblent actuellement récompenser la centralisation du pouvoir, ce que Gurumurthy (2021) décrit comme « des discours hégémoniques sur l'IA qui servent le capitalisme néolibéral ».

La capacité d'une organisation de la société civile à participer de manière significative aux processus de gouvernance de l'IA dépend de la disponibilité des ressources (financement, personnel, avantages politiques, temps, expertise, réseaux), de l'accès aux réseaux de gouvernance ou de la visibilité par rapport à ceux-ci, et de la capacité de négocier quant aux normes, à la langue, aux protocoles et aux priorités dans le cadre d'une dynamique de pouvoir déséquilibrée (Marchetti, 2016 ; Milan et Gutiérrez, 2015).

Dans les sphères de gouvernance, notamment à l'Organisation des Nations Unies (ONU), des formulations telles que « l'avenir du multilatéralisme est le multipartisme » encouragent l'inclusion de parties prenantes non étatiques dans les processus de gouvernance mondiale. Des représentants et représentantes d'entreprises ont assurément un rôle légitime à jouer dans les processus de gouvernance de l'IA. Une bonne partie de la gouvernance de l'IA s'effectue dans les entreprises, et les gouvernements du Nord et du Sud se procurent les produits d'IA de celles-ci. Le multipartisme a toutefois fait en sorte que les entreprises ont gagné en influence politique et en attrait dans les processus de gouvernance mondiale. Microsoft, par exemple, a un bureau au sein de l'ONU. Comme l'indique un rapport du Département des affaires économiques et sociales de l'ONU, « une force importante qui teinte la gouvernance à l'échelle nationale et internationale est celle des grandes entreprises, qui exercent des pressions en faveur de lois et de politiques servant leurs intérêts » (UN DESA, 2014).

Par ailleurs, les géants de la technologie jouent un rôle dans le financement de la recherche et exercent une influence sur des groupes d'universitaires que l'on consulte dans la gouvernance de l'IA. Comme le soulignent Abdala et autres (2020), « les géants de la technologie peuvent activement déformer le paysage universitaire pour qu'il réponde à ses besoins ». Les stratégies consistent à influer sur « les décisions prises par les universités financées » ou sur « les questions et les plans de recherche des scientifiques pris isolément » (Abdala *et al.*, 2020), et à manipuler le milieu universitaire pour éviter la réglementation (Ochigame, 2019).

La réglementation de l'IA s'externalise. Elle est confiée à des organismes de normalisation privés, par exemple le Comité européen de normalisation en électronique et en électrotechnique, l'Institut des ingénieurs électriciens et électroniciens, la Verband Deutscher Elektrotechniker, l'Organisation internationale de normalisation, etc., qui subissent également un fort lobbyisme de la part de l'industrie, ce qui peut les faire dévier de manière importante d'enjeux essentiels (Veale et Borgesius, 2021) tels que les droits humains. Les organisations de la société civile – qui comptent d'une compréhension empirique des questions de droits humains sur le terrain – jouent un rôle primordial dans l'établissement de normes là où l'expertise en matière de droits humains manque sérieusement (Cath, 2020 ; ten Oever et Cath, 2017). Le financement s'avère également un obstacle de taille à la participation de la société civile dans la réglementation de l'IA. Non seulement les géants de la technologie poursuivent des objectifs souvent incommensurables, mais leurs ressources dépassent de loin celles de la société civile et des parties prenantes du Sud.

Limites auxquelles se heurtent le Sud et la société civile

Il ne suffit pas de repérer les obstacles à la gouvernance inclusive au sein des sphères dominantes en matière de gouvernance de l'IA. Bien que les parties prenantes du Sud et de la société civile aient un rôle fondamental à jouer dans la protection des droits civils et humains et dans la responsabilisation de la gouvernance de l'IA, elles se heurtent à des limites en ce qui a trait à l'atténuation efficace des préjudices. Elles font notamment face à des contraintes liées à l'infrastructure et à la connectivité, au savoir-faire technologique, à la dépolitisation de la société civile, et à des tensions entre les gouvernements et la société civile.

En tant que « mécanisme » de responsabilisation, la société civile parvient à des résultats positifs qui restent hautement contextuels (Grimes, 2008). Ceux-ci ne doivent pas être considérés comme des acquis ; ils doivent plutôt servir à consolider une gouvernance protectrice et efficace. Citant Gramsci, Sassoon (2014) nous rappelle que « par-dessus tout [...] nous ne devons pas idéaliser la société civile » ; nous devons reconnaître l'hétérogénéité des objectifs, des priorités et des motivations qui sont siennes. Au-delà de la société civile institutionnalisée, il existe notamment des mouvements sociaux, des groupes d'activistes, et des organisations populaires non officiellement constituées, qui observent de plus près les effets de l'automatisation sur les droits et l'égalité.

L'atténuation des objectifs politiques pour survivre dans les sphères de gouvernance élitistes est considérée comme une « dépolitisation » (Jaeger, 2007). Elle serait un « double mouvement » dans lequel les organisations de la société civile jouent un rôle tant de dépolitisation que de politisation. Les organisations exercent leurs activités « aussi bien à l'intérieur qu'à l'extérieur du système politique de la société mondiale » (Jaeger, 2007), au sein de « l'ordre établi » et parfois en tant que « bloc contre-hégémonique » (Katz, 2006).

En outre, les récits selon lesquels les processus dominants de l'IA s'approprient les efforts de la société civile sont incomplets. Premièrement, les « modèles risquent d'adopter des hypothèses indûment simplistes quant à une victimisation passive des institutions au sein des sociétés démocratiques libérales » (Pils, 2019). Deuxièmement, la société civile fait preuve d'agentivité de diverses façons, soit en contestant (elle résiste, démantèle des structures, bâtit la solidarité), en collaborant, en se conformant aux attentes, ou en étant complice du pouvoir financier ou politique dominant.

Il est donc important de recueillir « d'habiles comptes rendus au sujet de la représentation et de faire ressortir les pratiques contestataires » (Dryzek, 2012) relatives à la gouvernance de l'IA afin de mener une refonte fondée sur des données probantes pour protéger efficacement les plus vulnérables. Il faut se rappeler que l'inclusion des gouvernements du Sud dans les processus de gouvernance de l'IA ne profitera pas toujours concrètement à l'ensemble de la population, en particulier aux groupes déjà vulnérables aux inégalités croissantes observables au sein des pays. On saisit bien les tensions entre la société civile et les États ou les gouvernements. Il est réducteur de supposer qu'il y a alignement des parties prenantes du Sud, à savoir des États et des organisations de la société civile, que leurs objectifs se confondent ou qu'elles adoptent toutes une perspective postcoloniale. La répression et la violence d'État existent dans les pays du Nord et du Sud. Bien que centrales, l'atteinte de l'autodétermination et la restructuration en vue de redistribuer équitablement certains avantages à l'échelle internationale (commerce, propriété, influence géopolitique, etc.) ne résolvent pas les problèmes nationaux que soulève la société civile. Au Kenya, par exemple, le Nubian Rights Forum et la Kenya Human Rights Commission ont intenté avec succès une action en justice pour contester le système gouvernemental national intégré de gestion de l'identité, alléguant qu'il violait « les droits à la vie privée, à l'égalité et à la non-discrimination inscrits dans la constitution du Kenya » (Mahmoud, 2019).

RECOMMANDATION 4 – ABORDER LE CONTEXTE HISTORIQUE DES DÉSÉQUILIBRES DE POUVOIR DANS LA GOUVERNANCE DE L'IA

Déséquilibres de pouvoir historiques

Les pièges de la gouvernance inclusive décrits à la section 3.1, par exemple, la culture organisationnelle, les logiques structurelles, l'utilisation de termes généraux, l'appropriation, et les incitatifs financiers, se trouvent plus largement dans des structures institutionnelles et géopolitiques empreintes de pouvoir et d'inégalités. Il y a des inégalités géopolitiques dans la gouvernance de l'IA ; le poids du discours formulé par le Nord (au sein de l'Union européenne et dans des entreprises technologiques d'Europe et d'Amérique du Nord) et par la Chine en est une. Ainsi, une représentation accrue des parties prenantes du Sud doit également s'inscrire dans une « analyse élargie du pouvoir et des dynamiques ou tensions politiques » (ÓhÉigeartaigh *et al.*, 2020). En examinant de plus près la dialectique répétitive de l'inclusion et de l'exclusion, nous pouvons « mieux comprendre la stratégie d'élaboration des politiques publiques mondiales, y compris sa dynamique de pouvoir » (Pouliot et Thérien, 2017). Pour ce faire, nous devons nous demander comment ces dynamiques se sont systématisées au fil du temps.

Colonialité du pouvoir dans la gouvernance de l'IA

Nous ne pouvons comprendre les inégalités actuelles relatives à l'IA ni anticiper ce qu'elles deviendront sans tenir compte de leur trajectoire historique. Les avantages du « premier arrivé » et la voie de l'exclusion qui a entraîné une dépendance, soit des éléments observables aujourd'hui, sont en quelque sorte des vestiges du colonialisme. Le manque d'engagement sincère à comprendre les racines historiques de l'exclusion dans les sphères de gouvernance de l'IA s'explique en partie, selon Sampath (2021), par « les explications technocentriques du progrès et de l'industrialisation » qui « sont profondément ancrées dans un contexte social plus large qui nous encourage à faire fi des racines historiques des inégalités, inégalités qui, en fait, ne se prêtent pas à une solution purement technologique ».

Les vestiges contemporains du colonialisme européen qui se traduisent en des inégalités de pouvoir géopolitique ou interpersonnel sont repris dans le concept de « colonialité » (Quijano, 2000). La colonialité du pouvoir offre un cadre nécessaire pour comprendre comment se répartissent, au Sud et au Nord, les préjudices et les avantages liés aux systèmes d'IA. Elle se prête aux études émergentes sur le colonialisme et le capitalisme des données, qui reconnaissent qu'il y a continuité dans l'exploitation, l'extraction et la dépossession coloniales en ce qui concerne l'utilisation de la main-d'œuvre, des ressources matérielles et des données dans l'industrie de l'IA (Thatcher *et al.*, 2016 ; Ricaurte, 2019 ; Couldry et Mejias, 2019 ; Birhane, 2020 ; Zuboff, 2019 ; Irani *et al.*, 2010 ; Ali, 2016).

Quatrième révolution industrielle

McKinsey a estimé que l'IA pouvait alimenter « une production économique supplémentaire d'environ 13 billions de dollars américains d'ici 2030, [et ainsi rehausser] le PIB mondial d'environ 1,2 % par an » (Bughin *et al.*, 2018). Bien que l'économie numérique et celle de l'IA aient des retombées au Sud, les initiatives et les partenariats liés au commerce et aux infrastructures et voyant le jour dans le contexte de la quatrième révolution industrielle ne reconnaissent pas comme ils le devraient que les avantages du « premier arrivé » et la voie de l'exclusion (et la dépendance) persistent, et que se perpétuent des déséquilibres de pouvoir historiques. « Les mieux placés pour profiter de la prolifération des systèmes d'IA sont ceux et celles qui ont le plus de pouvoir économique » (Chan *et al.*, 2021). Les avantages économiques découlant de l'IA correspondent parfois aux « hyperboles selon lesquelles les mégadonnées (*big data*) et l'économie des données sont la nouvelle "frontière de l'innovation", et qu'elles sont "rentables"

et "sources de profits" pour tous et toutes » (Sampath, 2021). Une telle interprétation ne reconnaîtrait pas l'enrichissement sélectif ni l'« effet Matthieu » (Fernández-Villaverde *et al.*, 2021) qu'entretient le système économique capitaliste aux frontières du Nord et du Sud.

La première révolution industrielle a été stimulée par l'extraction et l'exploitation de la main-d'œuvre, du savoir et des ressources naturelles des colonies européennes, ce qui a été rendu possible par les efforts militaires des gouvernements coloniaux de l'Europe occidentale, puis de l'Amérique du Nord. Les régimes coloniaux étaient structurés autour de systèmes juridiques, politiques, commerciaux et raciaux inégaux, qui persistent aujourd'hui. Il est particulièrement intéressant de noter que l'emploi d'une terminologie associée à la « révolution industrielle » et au « Sud » renvoie directement à l'histoire coloniale, qui n'est pourtant pas reconnue dans les discours dominants au sujet de la gouvernance de l'IA, en partie parce que ses répercussions ne touchent pas particulièrement les riches pays industrialisés.

Le travail fantôme et la tenue de bêta-tests sont des pratiques dans lesquelles s'inscrit une continuité historique se rapportant à l'extraction et à l'exploitation des ex-colonies par les anciens États colonisateurs (Mosco et Wasko, 1988 ; Agrawal *et al.*, 2019 ; Keskin et Kiggins, 2021). En plus de façonner le traitement des travailleurs et travailleuses et d'exposer les populations marginalisées aux risques que posent les bêta-tests, le mode de production capitaliste maintient la fracture économique Nord-Sud (Arrighi, 2001) d'une manière qui doit être reconnue et abordée de façon dialectique quand il est question des inégalités Nord-Sud. Nous pouvons estimer que la quatrième révolution industrielle apporte à la fois des avantages tangibles (en ce qui a trait aux soins de santé, aux communications, à l'agriculture, au marché du travail et à l'éducation), mais qu'elle entraîne aussi le techno-impérialisme (Sampath, 2021), le capitalisme racialisé et le capitalisme de la surveillance.

Souveraineté dans l'histoire

« Il est temps pour l'Europe d'être numériquement souveraine », est-il déclaré dans une lettre conjointe adressée à la Commission européenne par des dirigeants politiques du continent (ERR News, 2021). En l'occurrence, la souveraineté numérique fait référence à la propriété des données, à l'utilisation et au stockage de celles-ci, ainsi qu'à « l'augmentation du potentiel technologique de l'Europe et de sa capacité à établir des valeurs et des règles dans un monde centré sur la technologie, que dominent d'autres pays » (European Union, 2021).

Le Sud énonce également ses exigences en matière de souveraineté numérique, les économies émergentes souhaitant tirer profit de leurs propres données. En Afrique, par exemple, des infrastructures capitales (câbles sous-marins, réseaux terrestres de fibre optique et centres de données) appartiennent dans une large mesure à des entreprises de télécommunications qui ne sont pas africaines. Les données personnelles de la population sont en grande partie conservées dans des serveurs hébergés à l'étranger, comme en Irlande, étant donné que de nombreux pays africains ne disposent pas de centres de données nationaux. Des initiatives telles que Smart Africa et le Forum sur l'administration fiscale africaine élaborent des politiques relatives à la confidentialité des données et à la fiscalité pour atténuer l'accès illimité aux données nationales et la monétisation qu'en font les géants de la technologie, au détriment des économies locales en croissance (Velluet, 2021 ; Elmi, 2020).

Les mouvements qui s'intéressent à la collecte, à la propriété et à l'utilisation des données et, dans l'ensemble, à la souveraineté des données autochtones, défendent le « droit des peuples autochtones de contrôler les données provenant de leurs communautés et de leurs terres et les concernant, en énonçant les droits tant individuels que collectifs liés à l'accès aux données et à la vie privée » (Rainie *et al.*, 2019).

Étant donné les diverses notions liées à la souveraineté, il est crucial de comprendre quels éléments le discours sur la gouvernance de l'IA légitime et quels éléments il met de côté (et pourquoi). Les sphères de gouvernance européenne, africaine et autochtone défendent différemment l'autodétermination territoriale et numérique, ce qui découle du colonialisme européen. Kovacs et Ranganathan (2019)

formulent un avertissement contre toute mise en œuvre aveugle de la souveraineté, et nous rappellent qu'« il est important de se demander dans quelles conditions il devient possible de réclamer la souveraineté malgré [un passé de violence] ».

La conception moderne de la « souveraineté » se fonde sur les traités de Westphalie, signés au XVII[e] siècle « lorsqu'un nouvel ordre politique a été reconnu » et que la paix a été rétablie dans le Saint-Empire romain germanique après une longue période de violence (de Graaf et Kampmann, 2018). Ces traités de paix auraient favorisé le développement économique et technologique, et ensemencé la « culture de la sécurité » en Europe. Le voisinage ne constituant plus une menace, l'expansion impériale européenne a pu aller de l'avant, sur la base d'un système international hiérarchisé et racisé. Nous voyons cette hiérarchie à l'œuvre à différents égards dans l'exploitation et la dépossession continues du Sud dans l'économie moderne des données : main-d'œuvre bon marché, flux financiers illicites, extraction de données, exploitation des ressources naturelles, monopoles quant aux infrastructures, structures de financement, tenue de bêta-tests, etc. (Iyer *et al.*, 2021 ; Birhane, 2020).

CONCLUSION

> Les initiatives de gouvernance de l'IA qui cherchent à intégrer la société civile et les parties prenantes du Sud afin de réévaluer la répartition des risques doivent sérieusement examiner les tendances à l'exclusion dans l'histoire et la politique, mieux comprendre le travail que mènent le Sud et la société civile, envisager les obstacles et les pièges à une réelle inclusion, et aller au-delà des paradoxes de la participation. Une réelle inclusion requiert une redistribution structurelle du pouvoir, ce vers quoi les instances de gouvernance en place ne tendent pas. La juste distribution des retombées d'une IA transformationnelle n'est possible que par le recours à d'autres modèles épistémologiques et stratégies de développement et de gouvernance en provenance du Sud. Enfin, il importe d'effectuer un virage vers une cogouvernance participative pour que s'amenuisent « les nouvelles asymétries de pouvoir [...] et une exclusion draconienne » (Sampath, 2021).

RÉFÉRENCES

Abdala, M.B., Ortega, A. et Pomares, J. 2020. *Managing the transition to a multi-stakeholder artificial intelligence governance*. G20 Insights. https://www.g20-insights.org/policy_briefs/managing-the-transition-to-a-multi-stakeholder-artificial-intelligence-governance/

Agrawal, A., Gans, J. et Goldfarb, A. 2019. Artificial Intelligence : The ambiguous labor market impact of automating prediction. *National Bureau of Economic Research,* vol. 3, nº 2, pp. 31-50.

Ahmed, S. 2012. *On Being Included : Racism and Diversity in Institutional Life*. Durham, NC, Duke University Press.

Ali, S.M. 2016. A brief introduction to decolonial computing. *XRDS : Crossroads, The ACM Magazine for Students,* vol. 22, nº 4, pp. 16-21.

Arora, P. 2018. Decolonizing privacy studies. *Television & New Media,* vol. 20, nº 4, pp. 366-378.

Arrighi, G. 2008. Historical perspectives on states, markets and capitalism, East and West. *The Asia-Pacific Journal,* vol. 6, nº 1.

Arun, C. 2020. AI et the Global South. Dub, Pasquale and Das (éditeurs.) . *The Oxford Handbook of Ethics of AI*, New York : Oxford University Press, pp. 587-606.

Bird, E., Fox-Skelly, J., Jenner, N., Larbey, R., Weitkamp, E. et Weitkamp, A. 2020. *The Ethics of Artificial Intelligence : issues and initiatives*. European Parliament Scientific Foresight Unit. https://www.europarl.europa.eu/RegData/Etudes/STUD/2020/634452/EPRS_STU(2020)%20634452_EN.pdf

Birhane, A. 2020. Algorithmic Colonization of Africa. *SCRIPT-ed*, vol. 17, nº 2, pp. 389-409.

Bliss, F. et Neumann, S. 2008. Participation in international development discourse and practice : "State of the art" and challenges. *INEF-Report,* vol. 94. https://duepublico2.uni-due.de/receive/duepublico_mods_00027005

Bughin, J., Seong, J., Manyika, J., Chui, M. et Joshi, R. 2018. *Notes from the AI Frontier : Modeling the Impact of AI on the World Economy.* McKinsey Global Institute.

Capurro, R. 2009. Intercultural information ethics : Foundations and applications. *Signo y Pensamiento,* vol. 28, nº 55, pp. 66-79. http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0120-48232009000200004

Cath, C. 2020. What's wrong with loud men talking loudly ? The IETF's culture wars. https://hackcur.io/whats-wrong-with-loud-men-talking-loudly-the-ietfs-culture-wars/

Chan, A., Okolo, C.T., Terner, Z. et Wang, A. 2021. The limits of global inclusion in AI development. *arXiv.org*. https://arxiv.org/abs/2102.01265

Chatham House 2021. *Reflections on building more inclusive global governance*. https://www.chathamhouse.org/sites/default/files/2021-04/2021-04-15-reflections-building-inclusive-global-governance.pdf

Cleaver, F. 1999. Paradoxes of Participation : Questioning participatory approaches to development. *Journal of International Development,* vol. 11, pp. 597-612. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.461.2819

Clutton-Brock, P., Rolnick, D., Donti, P.L. et Kaack, L.H. 2021. *Climate Change and AI.* Global Partnership on AI Report.

Coate, R.A., Griffin, J.A. et Elliott-Gower, S. 2015. Interdependence in international organization and global governance. *Oxford Research Encyclopedia of International Studies*.

Comaroff, J. et Comaroff, J.L. 2016. *Theory from the south, or, How Euro-America is evolving toward Africa*. Londres, New York, Routledge, Taylor & Francis Group.

Conger, R., Robinson, H. et Sellschop, R. 2020. Inside a mining company's AI transformation. https://www.mckinsey.com/industries/metals-and-mining/how-we-help-clients/inside-a-mining-companys-ai-transformation

Connell, R. 2007. The Northern theory of globalization. *Sociological Theory,* vol. 25, nº 4, pp. 368-385.

Costanza-Chock, S. 2020. *Design Justice : Community-Led Practices to Build the Worlds We Need*. Cambridge, MA. Cambridge, The MIT Press.

Couldry, N. et Mejias, U. A. 2019. *The costs of connection : How data is colonizing human life and appropriating it for capitalism*. Stanford, CA, Stanford University Press.

Crawford, K. 2021. *Atlas of AI*. New Haven, Yale University Press.

de Graaf, B. et Kampmann, C. 2018. The Peace of Westphalia also had its dark side. Communiqué de presse du Religion and Politics Cluster of Excellence du 19 septembre 2018. Université de Münster https://www.uni-muenster.de/Religion-und-Politik/en/aktuelles/2018/sep/PM_Westfaelischer_Frieden_hatte_auch_Schattenseiten.html

d'Ignazio, C. et Klein, L.F. 2020. *Data feminism*. Cambridge, MIT Press, pp. 97-123.

Downing, J. D. H., Downing, J., Ford, T.V., Gil, G. and Stein, L. 2001. *Radical Media : Rebellious Communication and Social Movements*. London : SAGE Publications.

Dryzek, J. S. 2012. Global civil society : The progress of post-Westphalian politics. *Annual Review of Political Science*, vol. 15, nº 1, pp. 101–19. https://doi.org/10.1146/annurev-polisci-042010-164946

Elmi, N. 2020. Is Big Tech setting Africa back ? *Foreign Policy*. https://foreignpolicy.com/2020/11/11/is-big-tech-setting-africa-back/

ERR News 2021. Estonia, EU countries propose faster "European digital sovereignty." *ERR News*. https://news.err.ee/1608127618/estonia-eu-countries-propose-faster-european-digital-sovereignty

Eubanks, V. 2019. *Automating Inequality : How high-tech tools profile, police, and punish the poor.* New York, St. Martin's Press.

European Union. 2021. European Digital Sovereignty Conference. https://eu-digitalsovereignty.com/.

Fernández-Villaverde, J., Mandelman, F., Yu, Y. et Zanetti, F. 2021. « The 'Matthew effect' and market concentration : Search complementarities and monopsony Power. » CAMA Working Paper No. 22/2021. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3787787

Fraser, N. 2005. *Reframing Justice*. Uitgeverij Van Gorcum.

Glissant, É, et Dash, J.M. 1999. *Caribbean discourse : selected essays*. Charlottesville : University Press of Virginia.

Gopalakrishnan, S. et Santoro, M.D. 2004. Distinguishing between knowledge transfer and technology transfer activities : The role of key organizational factors. . University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1495508

Graham, M., De Sabbata, S. et Zook, M.A. 2015. Towards a study of information geographies : (im) mutable augmentations and a mapping of the geographies of information. *Geo : Geography and Environment,* vol. 2, nº 1, pp. 88-105.

Gray, M. et Suri, S. 2019. *Ghost Work : How to Stop Silicon Valley from Building a New Global Underclass.* New York, Houghton Mifflin Harcourt.

Green, B. 2019. "Good" isn't good enough. NeurIPS Joint Workshop on AI for Social Good, Vancouver, Canada.

Greenpeace. 2020. *Oil in the Cloud : How Tech Companies are Helping Big Oil Profit from Climate Destruction.* https://www.greenpeace.org/usa/reports/oil-in-the-cloud/

Grimes, M. 2008. Contestation or Complicity : Civil Society as Antidote or Accessory to Political Corruption. https://www.semanticscholar.org/paper/Contestation-or-Complicity%3A-Civil-Society-as-or-to-Grimes/45e98ef85f3abad31de4f473f522de7154cf849d

Gurumurthy, A. 2020. How to make AI work for people and planet. *openDemocracy.* https://www.opendemocracy.net/en/oureconomy/how-make-ai-work-people-and-planet/

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. 2019. *Classical Ethics in A/IS.* https://Standards.ieee.org/Wp-Content/Uploads/Import/Documents/Other/Ead_classical_ethics_ais_v2.Pdf

Irani, L. et Silberman, S. 2013. Turkopticon : Interrupting worker invisibility in Amazon mechanical turk. CHI 2013 : Changing Perspectives, Paris, France. http://crowdsourcing-class.org/readings/downloads/ethics/turkopticon.pdf

Irani, L., Vertesi, J., Dourish, P., Philip, K. et Grinter, R.E. 2010. Postcolonial computing. Proceedings of the 28th international conference on Human factors in computing systems – CHI 2010.

Iyer, N., Achieng, G., Borokini, F. et Ludger, U. 2021. *Automated Imperialism, Expansionist Dreams : Exploring Digital Extractivism in Africa.* Pollicy. https://archive.pollicy.org/wp-content/uploads/2021/06/Automated-Imperialism-Expansionist-Dreams-Exploring-Digital-Extractivism-in-Africa.pdf

Jaeger, H.-M. 2007. "Global civil society" and the political depoliticization of global governance. *International Political Sociology,* vol. 1, nº 3, pp. 257-277.

Jang, J., McSparren, J. et Rashchupkina, Y. 2016. Global governance : Present and future. Palgrave Communications, vol. 2, article 15045. https://www.nature.com/articles/palcomms201545

Kak, A. 2020. The Global South is everywhere, but also always somewhere. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.*

Kane, C. 2010. The relationship between IP, iechnology iransfer, and development. *Intellectual Property Watch.* 30 août. https://www.ip-watch.org/2010/08/30/the-relationship-between-ip-technology-transfer-and-development/

Katz, H. 2006. Gramsci, hegemony, and global civil society networks. *Voluntas : International Journal of Voluntary and Nonprofit Organizations,* vol. 17, nº 4, pp. 333-348. https://www.jstor.org/stable/27928042 (consulté le 18 mai 2021).

Kazmi, Z. 2012. Polite Anarchy and Diplomacy. *Palgrave Macmillan History of International Thought*. New York : Palgrave Macmillan. https://link.springer.com/chapter/10.1057%2F9781137028136_7

Keohane, R.O. et Nye, J.S. 1977. *Power and Interdependence.* Boston, :Little Brown Cop.

Keskin, T. et Kiggins, R., D. 2021. *Towards an International Political Economy of Artificial Intelligence.* Cham, Suisse, Palgrave Macmillan.

Khakurel, J., Penzenstadler, B., Porras, J., Knutas, A. et Zhang, W. 2018. The Rise of Artificial Intelligence under the Lens of Sustainability. *Technologies,* vol. 6, nº 4, article 100. https://doi.org/10.3390/technologies6040100

Kovacs, A. et Ranganathan, N. 2019. *Data sovereignty, of whom ? Limits and suitability of sovereignty frameworks for data in India*. Data Governance Network Working Paper 03.

Latonero, M. 2019. *Stop surveillance humanitarianism*. *New York Times*, 11 juin. https://www.nytimes.com/2019/07/11/opinion/data-humanitarian-aid.html

Lee, J., Gamundani, A. et Stinckwich, S. 2020. Blogue : The Inaugural AI Expert Consultation Meeting Recap : What's next for AI in Africa ? United Nations University Institute in Macau, Chine. https://cs.unu.edu/news/news/ai-expert-consultation-meeting.html

Lee, K.-F. 2019. *AI Superpowers : China, Silicon Valley, and the New World Order.* New York, HarperCollins.

Legassick, M. 1974. South Africa : Capital accumulation and violence. *Economy and Society,* vol. 3, nº 3, pp. 253-291.

Mahler, A.G. 2017. Global South. Oxford Bibliographies Online Datasets.

Mahmoud, M. 2019. Stopping the digital ID register in Kenya – A stand against discrimination. *Namati* (blogue). 25 avril. https://namati.org/news-stories/stopping-the-digital-id-register-in-kenya-a-stand-against-discrimination/

Marchetti, R. 2016. Global Civil Society. Extrait de *International Relations*. Bristol, E-International Relations. https://www.e-ir.info/2016/12/28/global-civil-society/

Mbembe, A. et Nuttall, S. 2004. Writing the World from an African metropolis. *Public Culture,* vol. 16, nº 3, pp.347-372.

McConnell, F., Moreau, T. and Dittmer, J. 2012. Mimicking state diplomacy : The legitimizing strategies of unofficial diplomacies. *Geoforum*, vol. 43, No. 4, pp.804–814.

McGlinchey, S., Walters, R. et Scheinpflug, C. 2017. *International relations theory*. Bristol : E-International Relations. https://www.e-ir.info/publication/international-relations-theory/

McGlinchey, S., Waters, R. et Scheinpflug, C. 2021. Global civil society as a response to transnational exclusion. Bristol : E-International Relations. https://socialsci.libretexts.org/@go/page/11133

Mhlambi, S. 2020. From rationality to relationality : Ubuntu as an ethical and human rights framework for artificial intelligence governance. Carr Center Discussion Paper Series (2020-009). https://carrcenter.hks.harvard.edu/publications/rationality-relationality-ubuntu-ethical-and-human-rights-framework-artificial

Mignolo, W. 2012. *Local histories/global designs : Coloniality, subaltern knowledges, and border thinking*. Princeton, NJ. Oxford, Princeton University Press.

Milan, S. et Gutiérrez, M. 2015. "Citizens" media meets big data : the emergence of data activism. *Mediaciones,* vol. 11, nº 14, pp. 120-133.

Milan, S. et Treré, E. 2019. Big Data from the South(s) : Beyond data universalism. *Television & New Media,* vol. 20, nº 4, pp. 319-335.

Milan, S. et van der Velden, L. 2016. The alternative epistemologies of data activism. *Digital Culture & Society,* vol. 2, nº 2.

Mohamed, S., Png, M.-T. et Isaac, W. 2020. Decolonial AI : Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology,* vol. 33.

Mosco, V. et Wasko, J. 1988. *The political economy of information*. Madison, Wisconsin, The University of Wisconsin Press.

Müller, M. 2018. In Search of the Global East : Thinking between North and South. *Geopolitics,* vol. 25, nº 3, pp. 1-22.

Neupane, S. et Smith, M.L. 2017. *Artificial Intelligence and Human Development.* Ottawa, Canada, International Development Research Centre. https://idl-bnc-idrc.dspacedirect.org/handle/10625/56949

Noble, S. U. 2018. *Algorithms of Oppression : How Search Engines Reinforce Racism.* New York, New York University Press.

Ochigame, R. 2019. How Big Tech manipulates academia to avoid regulation. *The Intercept*. 20 décembre. https://theintercept.com/2019/12/20/mit-ethical-ai-artificial-intelligence/

ÓhÉigeartaigh, S. S., Whittlestone, J., Liu, Y., Zeng, Y. et Liu, Z. 2020. Overcoming barriers to cross-cultural cooperation in AI ethics and governance. *Philosophy & Technology,* vol. 33, nº 4, pp. 571-593.

Parikka, J. 2015. *A Geology of Media*. Minneapolis, London : University of Minnesota Press.

Peck, J. 2011. *Ideal illusions : How the U.S. government co-opted human rights*. New York, Henry Holt, Godalming.

Pils, E. 2019. *The Risks of complicity in transnational civil society repression : An argument for institutional responses*. https://ecpr.eu/Events/Event/PaperDetails/44842

Pouliot, V. et Thérien, J.-P. 2017. Global governance in practice. *Global Policy,* vol. 9, nº 2, pp. 163-172.

Quijano, A. 2000. Coloniality of power and eurocentrism in latin america. *International Sociology,* vol. 15, nº 2, pp. 215-232.

Quinn, J., Frias-Martinez, V. et Subramanian, L. 2014. Computational sustainability and artificial intelligence in the developing world. *AI Magazine,* vol. 35, nº 3, p. 36.

Rainie, S.C., Kukutai, T., Walter, M., Figueroa-Rodríguez, O.L., Walker, J. et Axelsson, P. 2019. Indigenous data sovereignty. African Minds and the International Development Research Centre (IDRC). https://researchcommons.waikato.ac.nz/handle/10289/12918

Raval, N., Kak, A. et Calcaño, A. 2021. *A new AI lexicon : Responses and challenges to the critical AI discourse*. AI Now Institute. https://medium.com/a-new-ai-lexicon/a-new-ai-lexicon-responses-and-challenges-to-the-critical-ai-discourse-f2275989fa62

Rayment, P.B.W. 1983. Intra-"industry" specialisation and the foreign trade of industrial countries. *Controlling Industrial Economies.* The Vienna Institute for Comparative Economic Studies. London, Palgrave Macmillan pp. 1-28.

Ricaurte, P. 2019. Data epistemologies, the coloniality of power, and resistance. *Television & New Media,* vol. 20, no 4, pp. 350-365.

Roberts, A. 2018. *Politeness as process, manners as method*. OECD Observatory of Public Sector Innovation. 26 juillet. https://oecd-opsi.org/politeness-as-process-manners-as-method/

Sampath, P. G. 2021. Technology and inequality : Can we decolonise the digital world ? *South Views,* vol. 215. https://www.southcentre.int/wp-content/uploads/2021/04/SouthViews-Sampath.pdf

Sassoon, A. S. 2014. *Gramsci and Contemporary Politics : Beyond Pessimism of the Intellect.* London : Routledge.

Satariano, A. 2021. *"Europe proposes strict rules for artificial intelligence."* The New York Times, 21 avril. https://www.nytimes.com/2021/04/16/business/artificial-intelligence-regulation.html

Schiff, D., Borenstein, J., Biddle, J. et Laas, K. 2021. AI ethics in the public, private, and NGO sectors : A review of a global document collection. *IEEE Transactions on Technology and Society,* vol. 2, nº 1, pp. 31-42.

Scholte, J.A. 2004. Civil society and democratically accountable global governance. *Government and Opposition,* vol. 39, nº 2, pp. 211-233.

Selznick, P. 2015. *TVA and the grass roots : A study of politics and organization*. New Orleans, LA, Quid Pro Books.

Singh, R. 2021. Mapping AI in the Global South. A new project to identify sites and vocabularies of digital IDs and AI. *Data & Society*, 26 janvier 26. Blogue.. https://points.datasociety.net/ai-in-the-global-south-sites-and-vocabularies-e3b67d631508

Singh, R. et Lara Guzmán, R. 2021. Parables of AI in/from the Global South. *Data & Society*. https://datasociety.net/announcements/2021/07/13/call-for-participants-parables-of-ai-in-from-the-global-south/

Smogorzewski, K. 1938. Poland's foreign relations. *The Slavonic and East European Review,* vol. 16, nº 48, pp. 558-571.

Smuha, N.A. 2020. Beyond a human rights-based approach to AI governance : Promise, pitfalls, plea. *Philosophy & Technology,* vol. 34, pp. 91-104.

South Centre 2020. Submission by the South Centre to the Draft Issues Paper on Intellectual Property Policy and Artificial Intelligence (WIPO/IP/AI/2/GE(20/1). https://www.southcentre.int/wp-content/uploads/2020/02/Submission-by-SC-to-the-Draft-Issues-Paper-on-IP-Policy-and-AI.pdf

Taylor, K.-Y. 2019. *Race for Profit : How Banks and the Real Estate Industry Undermined Black Homeownership.* Chapel Hill, University of North Carolina Press.

Taylor, L. 2017. What is data justice ? The case for connecting digital rights and freedoms globally. *Big Data & Society,* vol. 4, nº 2, doi :10.1177/2053951717736335

ten Oever, N. et Cath, C. 2017. Research into human rights protocol considerations. Internet Research Task Force, article 19. https://datatracker.ietf.org/doc/html/rfc8280

Thatcher, J., O'Sullivan, D. et Mahmoudi, D. 2016. Data colonialism through accumulation by dispossession : New metaphors for daily data. *Environment and Planning D : Society and Space,* vol. 34, nº 6, pp. 990-1006. https://doi.org/10.1177/0263775816633195

Torres, G. 2017. Taking a look at institutional resistance to citizen empowerment. *DATACTIVE*. Blogue. https://data-activism.net/2017/02/blog-taking-a-look-at-institutional-resistance-to-citizen-empowerment-through-data/

Ulnicane, I., Eke, D.O., Knight, W., Ogoh, G. et Stahl, B.C. 2021. Good governance as a response to discontents ? Déjà vu, or lessons for AI from other emerging technologies. *Interdisciplinary Science Reviews,* vol. 46, nº 1-2, pp. 71-93.

Ulnicane, I., Knight, W., Leach, T., Stahl, B.C. et Wanjiku, W.-G. 2020. Framing governance for a contested emerging technology : Insights from AI policy. *Policy and Society,* vol. 40, nº 2, pp. 1-20.

UNCTAD. 2013. *Information economy report 2013 : The cloud economy and developing countries*. Genève, United Nations Conference on Trade and Development.

UN DESA. 2014. *Global Governance and Global Rules for Development in the Post-2015 Era*. Department of Economic and Social Affairs, Committee for Development Policy. https://www.un.org/en/development/desa/policy/cdp/cdp_publications/2014cdppolicynote.pdf

Vawda, Y. 2021. The TRIPS COVID-19 waiver, challenges for Africa and decolonizing intellectual property. *South Centre Policy Brief,* vol. 99. https://www.southcentre.int/policy-brief-99-august-2021/

Veale, M. et Borgesius, F. Z. 2021. *Demystifying the Draft EU Artificial Intelligence Act. Computer Law Review International*. Vol. 22, No. 4, pp. 97–112. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3896852

Velluet, Q. 2021. Can Africa salvage its digital sovereignty ? The Africa Report. 16 avril. https://www.theafricareport.com/80606/can-africa-salvage-its-digital-sovereignty/

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S.D., Tegmark, M. et Nerini, F. 2020. The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications,* vol. 11, nº 1.

Weiss, T. G. 2016. Rising powers, global governance, and the United Nations. *Rising Powers Quarterly,* vol. 1, nº 2, pp.7-19.

Werner, D. et PROJIMO. 1998. *Nothing about us without us : Developing innovative technologies for, by and with disabled persons*. Palo Alto : Healthwrights.

Williams, G. 2004. Evaluating participatory development : Tyranny, power and (re)politicisation. *Third World Quarterly*, Vol. 25, No. 3, pp.557–578. https://www.jstor.org/stable/3993825?seq=1

Winner, L. 1980. Do artifacts have politics ? *Daedalus*, Vol. 109, No. 1, pp. 121-136. http://www.jstor.org/stable/20024652

Wong, P.-H. 2016. Responsible innovation for decent nonliberal peoples : A dilemma ? *Journal of Responsible Innovation,* vol. 3, nº 2, pp. 154-168.

Yuan, L. 2018. How cheap labor drives china's A.I. ambitions. *The New York Times*. 25 novembre. https://www.nytimes.com/2018/11/25/business/china-artificial-intelligence-labeling.html

Zuboff, S. 2019. *The Age of Surveillance Capitalism : The Fight for a Human Future at the New Frontier of Power.* New York, Publicaffairs.

DÉMOCRATISER L'ÉLABORATION DE POLITIQUES EN MATIÈRE D'IA

STEFAN RIEZEBOS

Directeur de recherche, Innovation for Policy Foundation. Titulaire d'une maîtrise en politique et en économie de l'Université Erasmus de Rotterdam, il compte plus de sept ans d'expérience à titre de conseiller principal pour le gouvernement néerlandais, se spécialisant dans la réglementation de l'économie des plateformes (GAFA).

TIM GELISSEN

Directeur des services-conseils, Innovation for Policy Foundation. Titulaire d'une maîtrise en économie internationale de l'Université Tilburg, il a travaillé 10 ans avec le gouvernement néerlandais. Avec i4Policy, il a soutenu plusieurs processus de co-création dans différents pays, plus récemment au Rwanda et au Nigeria.

RAASHI SAXENA

Spécialiste des programmes d'IA, Innovation for Policy Foundation. Titulaire d'un baccalauréat en ingénierie des télécommunications de l'Université technique Visvesvaraya, en Inde, elle possède une vaste expérience dans les domaines de technologie et innovation et dans le rôle de consultante en matière d'impact social.

au nom de l'Innovation for Policy Foundation.

ODD 9 - Industrie, innovation et infrastructure

ODD 10 - Inégalités réduites

ODD 11 - Villes et communautés durables

ODD 16 - Paix, justice et institutions efficaces

ODD 17 - Partenariats pour la réalisation des objectifs

DÉMOCRATISER L'ÉLABORATION DE POLITIQUES EN MATIÈRE D'IA

RÉSUMÉ

De nos jours, l'intelligence artificielle (IA) se dissimule de plus en plus dans la prise de décisions. Les gouvernements constatent que l'IA se répand et entrevoient les conséquences qu'elle pourrait avoir d'un point de vue économique et social. Plus de 50 pays ont publié des politiques visant à tirer profit des avantages de l'IA, tout en protégeant l'intérêt public. La mise en œuvre et la croissance des systèmes d'IA s'accompagnent de menaces aux droits humains, et soulèvent des questions et des dilemmes fondés sur des valeurs. Nous insistons sur le fait qu'il ne doit pas revenir à une seule partie prenante de réduire l'incidence de l'IA. Dans l'élaboration de politiques en matière d'IA, il importe d'adopter une approche multipartite participative faisant intervenir un large éventail de parties prenantes telles que des responsables des politiques, des figures de la société civile, des citoyens et citoyennes, des universitaires, ainsi que la communauté technique et le secteur privé. Une telle mesure améliorera la qualité de la prise de décisions, créera un environnement propice à l'apprentissage et à la délibération, et nous aidera à nous défaire d'une perspective monolithique sur l'IA. À partir d'études de cas de certains pays, d'ateliers, d'entretiens avec des spécialistes et d'une expérience acquise dans plus d'une douzaine de pays, nous faisons ressortir cinq leçons concrètes quant à l'élaboration de politiques inclusives en matière d'IA.

INTRODUCTION

De nos jours, l'intelligence artificielle (IA) se dissimule de plus en plus dans la prise de décisions. Elle a des répercussions dans divers domaines ayant une portée économique ou sociale, et touche notre quotidien de façons insoupçonnées par la plupart d'entre nous. Les systèmes d'IA n'intéressent plus seulement les chercheurs et chercheuses ou les adeptes de science-fiction ; leur adoption s'accélère

dans toute la société. Par exemple, la Rétrospective 2022 de Spotify, compilée au moyen de l'IA, suscitera des débats animés au sujet des préférences musicales[70], des médecins s'aident de l'IA dans leurs diagnostics, d'autres systèmes d'IA déterminent de manière autonome l'admissibilité à un prêt personnel.

Même si l'IA constitue un domaine de recherche depuis les années 1950, ce n'est que récemment qu'elle a quitté les laboratoires pour se répandre dans la société (The Netherlands Scientific Council for Government Policy, 2021). Trois principaux facteurs ont conduit à cette évolution. Premièrement, la puissance informatique s'est accrue ces dernières années, ce qui permet d'effectuer des calculs de plus en plus complexes (selon la loi de Moore). Deuxièmement, la quantité de données disponibles a explosé alors que, parallèlement, les coûts de stockage diminuaient. Troisièmement, plusieurs percées scientifiques permettent maintenant à l'IA de dégager des tendances à partir de multiples couches de données. Cette capacité à scruter davantage les données – qu'on appelle « apprentissage profond » – a ouvert la voie aux applications de l'IA en usage aujourd'hui.

Comparer l'IA à d'autres technologies polyvalentes telles que le moteur à vapeur, l'électricité ou l'ordinateur met en lumière son potentiel considérable. Selon des estimations du McKinsey Global Institute (2018), l'IA devrait engendrer d'ici 2030 des retombées économiques de l'ordre de 13 billions de dollars américains à l'échelle mondiale, ce qui correspond à une hausse d'environ 1,2 % de la croissance annuelle. L'Union internationale des télécommunications fait observer que si une telle incidence s'avérait, elle serait tout à fait comparable à celle qu'ont eue d'autres technologies polyvalentes au cours de l'histoire (ITU, 2018). Ces chiffres traduisent le potentiel qu'a l'IA de transformer l'industrie et de multiplier les retombées sociétales, par exemple en contribuant à pourvoir des besoins liés à l'alimentation, à la santé ainsi qu'à l'approvisionnement en eau ou en énergie, de même qu'en soutenant la transition vers la carboneutralité (Vinuesa *et al.*, 2020).

Il ne faut pas en conclure que l'IA n'aura qu'un effet positif dans le monde. Vinuesa *et al.* (2020) affirment que le recours à l'IA pourrait empêcher l'atteinte de 59 des cibles liées aux objectifs de développement durable établis par les Nations Unies. Les inconvénients seraient même variés et comprendraient un accroissement des inégalités de revenu à l'échelle nationale et internationale, une concentration exacerbée du marché ainsi qu'une polarisation des emplois. Le recours à des systèmes d'IA – notamment pour la reconnaissance faciale, la reconnaissance de formes ou les hypertrucages – peut soulever des préoccupations en matière de confidentialité, contribuer à propager de la désinformation, ou encore entraîner la violation de droits humains et de libertés individuelles.

Les gouvernements constatent que l'IA se répand et entrevoient les conséquences qu'elle pourrait avoir. Le Canada a été l'un des premiers pays à publier une stratégie nationale en matière d'IA, en 2017. Plusieurs autres l'ont imité par la suite en ébauchant des stratégies, des politiques ou des lois relatives à l'IA. À ce jour, plus de 50 documents de la sorte ont été publiés[71]. Bien que chaque pays définisse ses propres priorités et les enjeux qu'il privilégie, la plupart des politiques en matière d'IA visent un même objectif : tirer profit des avantages de l'IA tout en protégeant l'intérêt public[72].

On affirme bien souvent que les politiques en matière d'IA et les lignes directrices relatives à l'éthique sont issues d'un « large consensus ». Elles sont toutefois massivement conçues par des pays économiquement développés et de nature faiblement inclusive (Crawford, 2021). C'est ce que révèle en fait

70. D'aucuns pourraient accuser l'algorithme employé par Spotify d'avoir retenu des plaisirs coupables dans la liste des chansons les plus écoutées de l'année.

71. Voir des exemples de stratégies pour l'IA, voir the Future of Life Institute (https://futureoflife.org/ai-policy/), OECD.AI Policy Observatory (https://oecd.ai/en/dashboards) et l'initiative Globalpolicy.ai.

72. Nous employons « politique en matière d'IA » en tant que terme générique qui désigne tant les stratégies relatives à l'IA que les plans d'action ou les propositions de politiques prévoyant officiellement des mesures concrètes (notamment des mesures législatives).

le clivage qui s'observe entre les pays du Nord et ceux du Sud en ce qui a trait à leur intervention ; le continent africain n'en est toujours qu'aux premiers stades de l'élaboration de politiques relatives à l'IA (Smart Africa, 2021). En outre, Gwagwa *et al.* font remarquer qu'« une caractéristique de certaines discussions sur les politiques en matière d'IA [...] a été la marginalisation ou l'exclusion du point de vue des pays du Sud » (2020, p. 16), tandis que l'organisation AlgorithmWatch (2020) constate que l'écrasante majorité des lignes directrices concernant l'éthique proviennent de l'Europe et des États-Unis.

Seule une poignée de pays a eu recours à des processus participatifs pour parvenir à un consensus, informer le public ou délibérer sur d'éventuelles solutions stratégiques. Voilà qui contraste nettement avec les principaux défis qu'engendre l'avènement des systèmes d'IA, un phénomène qui soulève des dilemmes fondés sur des valeurs ainsi que des problèmes complexes nécessitant des compromis. De tels dilemmes et problèmes sont d'une importance et d'une complexité telles qu'il ne revient pas à un seul groupe de parties prenantes de les aborder.

Ce chapitre s'articule autour d'une question : Comment pouvons-nous démocratiser l'élaboration de politiques en matière d'IA ? Nous estimons que de véritables approches multipartites font partie de la solution. Responsables des politiques, figures de la société civile, utilisateurs et utilisatrices, citoyens et citoyennes, universitaires, communauté technique et secteur privé doivent tous participer à la réflexion sur les dilemmes et les problèmes que posent les systèmes d'IA afin de parvenir à des résultats satisfaisants pour tous. Il est d'ailleurs temps que nous prêtions attention à tout le processus et non seulement aux résultats.

Nous entamons notre raisonnement dans la partie 1 en nous penchant sur le besoin de réglementer les technologies et applications d'IA. Nous abordons dans la partie 2 les concepts de multipartisme et de délibération, avant d'expliquer dans la partie 3 comment ils s'appliquent à l'élaboration d'une politique en matière d'IA. La partie 4 expose les principes généraux soutenant les processus multipartites efficaces, ce que viennent illustrer les deux études de cas présentées dans la partie 5. La partie 6 fait ressortir des leçons concrètes quant à la conception d'approches multipartites, puis on conclut notre propos.

Le travail dont il est question dans ce chapitre est le fruit d'une collaboration entre l'Innovation for Policy Foundation (i4Policy) et l'UNESCO, et tire parti des conclusions de cinq ateliers, d'entretiens avec des spécialistes, de l'examen de plus de 20 études de cas, et d'une expérience concrète des approches multipartites acquise dans plus d'une douzaine de pays.

---

1. RISQUES ENTOURANT LA MISE EN ŒUVRE DE SYSTÈMES D'IA

Dans cette partie, nous décortiquons les risques que posent les systèmes d'IA. Nous reprenons les constats de spécialistes ayant participé à deux ateliers, tenus en septembre et en octobre 2021, au sujet des répercussions de l'IA sur les droits humains et les libertés fondamentales. Nous recensons par ailleurs des écrits supplémentaires. Nous montrons l'ampleur que pourraient avoir d'éventuelles répercussions néfastes. À défaut de limiter les effets des systèmes d'IA au moyen de politiques ayant précisément cette visée, il faudra corriger d'autres politiques pour atténuer les contrecoups d'une diffusion débridée de l'IA dans nos sociétés. Nous examinons ci-après en détail trois risques que posent l'IA : i) la disparité entre pays du Nord et pays du Sud, ii) l'effet particulièrement néfaste de l'IA sur des groupes ayant été marginalisés au cours de l'histoire, iii) les enjeux du respect de la vie privée et de la surveillance. Ces risques justifient qu'on investisse temps et efforts dans la conception de politiques participatives inclusives.

Disparité entre pays du Nord et pays du Sud

Tout d'abord, il existe une grande disparité entre les pays Nord et ceux du Sud en ce qui concerne la mise au point, le déploiement et l'usage de l'IA. Un large fossé numérique subsiste entre pays développés et pays en développement, et ce fossé constitue un obstacle récurrent au développement (UNCTAD, 2021). Son existence déteint sur le rôle que jouent les pays du Sud dans les discussions sur l'IA et son évolution. Une méta-analyse de Jobin *et al.* (2019) montre que les valeurs des États-Unis et de l'Occident pèsent dans 67 % des principes éthiques liés à l'IA. Les pays d'Afrique, d'Amérique du Sud et d'Amérique centrale ainsi que d'Asie centrale (exception faite de l'Inde) ne sont pas représentés dans les données. Dans le même ordre d'idées, la nette majorité des applications d'IA en usage en Afrique subsaharienne ne sont pas produites en Afrique ou conçues pour ce continent (Oxford Insights, 2020). Birhane (2020) explique qu'il s'agit là d'un problème puisque les systèmes de valeurs diffèrent d'une culture à l'autre, et que ce qui est considéré comme un problème ou une solution diffère donc également. Pour ce faire, elle reprend un exemple frappant présenté par Black et Richmond (2019), selon lequel des mesures de dépistage précoce du cancer du sein avérées efficaces en Occident ne le sont pas en Afrique subsaharienne en raison de l'âge moyen des patientes (qui y est inférieur), d'un stade plus avancé de la maladie et d'un accès limité à certains traitements.

Récemment, des publications ont exposé des théories liées au colonialisme et au capitalisme de données, lesquelles considèrent les données comme une ressource qu'on exploite (Mohamed *et al.*, 2020 ; Kwet, 2019). De plus, des spécialistes du domaine de la militarisation de l'IA ont lancé un avertissement quant à la sous-représentation des pays du Sud dans les discussions. Or, les menaces à la paix et à la sécurité se font d'abord sentir dans les zones de conflit des pays en développement. Une publication du Conseil de sécurité des Nations Unies permet d'illustrer ce propos. Elle indique qu'en Libye, des systèmes d'armes létales autonomes ont servi, en mars 2020, à attaquer des troupes et un convoi transportant du matériel logistique qui battaient en retraite. Le rapport fait ressortir que des systèmes d'armes ont été programmés pour attaquer des cibles « sans qu'il soit besoin d'établir une connexion des données entre l'opérateur et la munition » (Garcia, 2019 ; Nations Unies. Conseil de sécurité, 2021, p. 20).

L'UNESCO (2021) et Gwagwa *et al.* (2021) soulignent que même si le déploiement de l'IA en Afrique relève de divers pays, touche différentes populations et soulève des problèmes très variés, la grande majorité des pays du continent n'ont adopté aucun cadre stratégique à cet égard. Cette absence de cadre risque de tenir les pays africains en marge des discussions sur l'éthique et les normes relatives à l'IA, en plus de freiner la mise sur pied d'environnements de travail propices au développement et à l'usage de l'IA.

Effet particulièrement néfaste sur des groupes marginalisés

Par ailleurs, l'IA peut avoir un effet particulièrement néfaste sur des groupes ayant été marginalisés au fil de l'histoire. L'IA qui se fonde sur des données non représentatives contribue à exacerber les iniquités sociales et économiques. C'est qu'elle reproduit les disparités ou les préjugés que comportent les données servant à l'entraîner. Pensons par exemple à une entreprise dont le personnel est surtout composé d'hommes. Si les données historiques relatives à l'embauche au sein de cette entreprise servent à entraîner un outil de recrutement intelligent, celui-ci montrera une préférence pour les candidats, au détriment des candidates. Voilà exactement ce qui est arrivé chez Amazon (Polli, 2019). Les préjugés des systèmes d'IA peuvent être induits par les données, ou encore par les programmeurs et programmeuses. Puisqu'il incombe à des êtres humains de cerner des problèmes et de déterminer la validité de résultats, il est possible qu'ils intègrent leurs préjugés personnels dans un système (UNESCO, 2019 ; Barocas et Selbst, 2016).

Prince et Schwarcz (2020) illustrent ces deux types de préjugés. Ils allèguent que l'IA engendre parfois une « discrimination par procuration », ce qui survient quand une variable apparemment neutre se substitue, dans un modèle, à une variable dont l'usage est interdit en raison de son caractère discriminatoire reconnu. Les préjugés induits par les programmeurs et programmeuses sont ceux qui découlent de l'emploi intentionnel d'une telle variable substitut (ce qui s'appelle un « masquage »). Par ailleurs subsiste aussi un risque de discrimination par procuration non intentionnelle, c'est-à-dire de préjugés induits par les données. En effet, l'IA reçoit un entraînement fondé sur de vastes ensembles de données, mais reste aussi libre de déduire des relations entre des variables en fonction de résultats attendus. Cette caractéristique augmente la probabilité qu'elle observe des corrélations puis qu'elle applique de nouvelles variables substituts à des groupes marginalisés, ce qui risque d'avoir un effet discriminatoire. Le code postal constitue un exemple à cet égard. À première vue, il semble s'agir d'une variable neutre qui n'indique qu'un emplacement. Cette donnée est toutefois intimement corrélée au statut socioéconomique et à l'origine ethnique. Murray (2013) a même forgé l'expression *super zip*[73] en référence aux codes postaux des secteurs les plus aisés et influents des États-Unis. L'existence de cette corrélation signifie que l'inclusion du code postal dans un ensemble de données peut involontairement entraîner de la discrimination envers des groupes minoritaires.

Enjeux du respect de la vie privée et de la surveillance

Enfin, notre droit au respect de la vie privée – soit un droit fondamental – est menacé par l'IA. Protéger ce droit est une vaste tâche qui ne vise pas seulement l'information substantielle contenue dans des documents, mais également les métadonnées, étant donné que l'analyse de celles-ci donne aussi des indications sur la conduite d'une personne (HCDH, 2018). De plus, il n'y a pas que l'examen ou l'usage de données par les algorithmes ou les programmeurs et programmeuses qui nuisent au respect de la vie privée. Comme le soutient Bernal, « plusieurs des risques majeurs interviennent quand les données sont recueillies ; l'existence de données crée le risque » (2016, p. 249). En fait, notre droit au respect de la vie privée est menacé à différentes étapes du processus : cueillette des données, analyse, évaluation des résultats.

Les technologies de surveillance basées sur l'IA La technologie de surveillance de l'IA est un domaine dans lequel il convient de faire très attention à la protection du droit à la vie privée. Ce type de système d'IA se répand à un rythme incroyable . Feldstein (2019) le démontre à l'aide du AI Global Surveillance Index, soit un outil qui compile des données sur la surveillance par des systèmes d'IA dans 176 pays. Des spécialistes, des activistes et des organisations non gouvernementales ont soulevé des questions légitimes quant aux conséquences du recours aux technologies de reconnaissance faciale sur le respect de la vie privée et la liberté d'expression (consulter par exemple Moraes *et al.*, 2021 et Mudongo, 2021). En 2018, le Haut-Commissariat des Nations Unies aux droits de l'homme a conclu que de nombreux États continuaient d'exercer une surveillance de masse et d'intercepter des communications (HCDH, 2018). Bien que plusieurs États prétendent que la surveillance de masse soit nécessaire pour assurer la sécurité nationale, un tel recours aux technologies de reconnaissance faciale n'est pas acceptable selon la législation internationale relative aux droits humains (Privacy International, 2019).

En résumé, les systèmes d'IA constituent une menace sérieuse à l'égard des droits humains et des libertés fondamentales quand on en permet la diffusion débridée. Il est essentiel de protéger l'intérêt public en instaurant des politiques visant précisément l'IA. Nous démontrons dans les deux parties qui suivent qu'une approche délibérative multipartite convient pour élaborer de telles politiques.

73. Aux États-Unis, le code postal est appelé *zip code* (N. d. T.).

2. MULTIPARTISME : ORIGINE ET CONTEXTE

La participation de parties prenantes représente depuis longtemps l'une des stratégies de gouvernance publique. Hofmann (2016) montre que cette participation a traditionnellement été associée à des enjeux internationaux tels que les conditions de travail ou les normes environnementales. Elle invoque à titre d'exemple la structure tripartite de l'Organisation internationale du travail (OIT) qui réunit des représentants et représentantes des gouvernements, employeurs et travailleurs et travailleuses. Fondée en 1919, l'OIT reste à ce jour le seul organisme des Nations Unies à avoir une structure tripartite.

Le terme anglais *multi-stakeholder*[74] a été forgé dans les années 1990. Il a nettement gagné du terrain depuis, en particulier en ce qui a trait à la gouvernance mondiale. Scholte (2020) indique que les approches multipartites mondiales sont devenues une solution de rechange au multilatéralisme international, puisqu'elles étaient de plus en plus perçues comme une manière de composer avec des problèmes complexes et épineux touchant des personnes et des organisations à l'échelle de la planète[75]. Il ajoute que l'intérêt du multipartisme réside dans le fait que le mariage de diverses sources d'information et perspectives peut améliorer l'efficacité de la résolution de problèmes mondiaux et mener à une mise en commun des ressources permettant d'aborder ces problèmes.

En parallèle, une « vague délibérative » touche l'élaboration de politiques nationales, tel que le fait remarquer l'Organisation de coopération et de développement économiques (OCDE, 2020). Il est souvent fait référence à de telles initiatives nationales en employant les expressions « démocratie délibérative » ou « démocratie participative ». L'OCDE fait valoir, d'une manière semblable à Scholte, que le caractère de plus en plus complexe de l'élaboration de politiques ainsi que l'échec à trouver des solutions à certains des problèmes les plus criants ont incité des politiciens et politiciennes, des responsables des politiques, des organisations de la société civile ainsi que des citoyens et citoyennes à réfléchir à la manière souhaitable de prendre collectivement des décisions d'intérêt public au 21e siècle.

Derrière le concept de gouvernance multipartite se trouve l'idée de créer un forum de dialogue pouvant mener à un consensus sur un ensemble de valeurs et d'objectifs communs. Cette idée s'ancre dans l'éthique de la discussion d'Habermas, théorie selon laquelle la morale et les normes découlent d'un processus dans lequel des personnes opposent leurs points de vue. Si toutes les parties examinent rationnellement les arguments des unes et des autres, elles devraient parvenir ensemble à brosser un meilleur portrait de l'enjeu, ce qui devrait les amener à revoir leur position. Il s'agit de répéter ce processus jusqu'à ce que les parties en cause parviennent à une décision acceptable pour toutes (Habermas, 1989 ; Martens *et al.*, 2019).

Cette approche normative comporte de grands avantages. Premièrement, les approches multipartites renforcent l'inclusivité, car elles facilitent la participation de représentants et représentantes de divers groupes pour la plupart minoritaires, par exemple des jeunes, des personnes défavorisées ou des femmes (Adam *et al.*, 2007). Deuxièmement, les processus participatifs créent un environnement propice à l'apprentissage, à la délibération, et à l'élaboration de recommandations éclairées. Ils permettent de mieux saisir les sujets de préoccupation et d'intérêt des autres parties prenantes (Faysse, 2006). Troisièmement, le fait d'intégrer plus d'expertise et de diversité dans les processus de prise de décisions et d'encourager l'obtention d'un consensus devrait améliorer la qualité de la prise de décisions (Souter, 2017).

74. *Multi-stakeholder* signifie littéralement « plusieurs parties prenantes ». Son emploi adjectival se traduit donc par « multipartite ». De ce terme découle aussi *multistakeholderism,* traduit en l'occurrence par « multipartisme » (N. d. T.).

75. Pour en apprendre davantage sur l'évolution de la gouvernance multipartite, consulter aussi Dingwerth (2008), Brockmyer et Fox (2015), et Gleckman (2018).

Il ne faut pas pour autant en conclure que le multipartisme est une panacée. Plusieurs spécialistes montrent qu'en pratique, les initiatives multipartites ne sont pas toujours à la hauteur des attentes en ce qui a trait à la confiance (Sloan et Oliver, 2013), à des enjeux de légitimité (Bäckstrand, 2006), ou à la quantité de ressources et au temps qu'elles demandent (Moog *et al.*, 2014). En outre, des asymétries de pouvoir surgissent parfois quand des parties ne contribuent pas de manière équitable, qu'il soit question de savoir, de financement ou d'accès à l'information (Fransen et Kolk, 2007). Les approches multipartites posent parfois problème dans certains processus à court terme, car les mécanismes formels de prise de décisions y sont parfois nébuleux (Faysse, 2006).

Avant d'adapter une approche multipartite, il importe de déterminer, d'une part, si le type d'approche retenu est applicable et, d'autre part, s'il y a moyen d'en pallier les faiblesses. Dans la partie qui suit, nous faisons valoir qu'une telle approche convient tout à fait à l'élaboration de politiques en matière d'IA. Dans le reste du chapitre, nous nous intéressons à la conception d'une approche multipartite optimale.

---

3. NÉCESSITÉ D'UNE APPROCHE MULTIPARTITE DANS L'ÉLABORATION DE POLITIQUES EN MATIÈRE D'IA

Buhmann et Fieseler (2021) formulent deux raisons pour lesquelles les approches communicatives et délibératives constituent des solutions appropriées relativement aux politiques en matière d'IA. Premièrement, les vastes ramifications des systèmes d'IA dans la société et leur rapide prolifération tant dans la sphère publique que dans la sphère privée de nos vies devraient selon eux faire de ces systèmes un sujet de préoccupation politique majeur dans son ensemble. Deuxièmement, le manque de transparence et d'imputabilité autour des systèmes d'IA requiert de recourir au « pouvoir épistémique » de la délibération en vue d'améliorer la compréhension qu'on a des enjeux et de faciliter la rétroaction au moyen de processus itératifs faisant intervenir des parties prenantes habilitées à le faire. Nous considérons par ailleurs une troisième raison. Tel que l'a démontré l'OCDE, les processus délibératifs conviennent davantage quand le sujet à l'étude soulève des dilemmes fondés sur des valeurs, des problèmes complexes nécessitant des compromis, ou des enjeux de longue haleine qui vont au-delà de mesures à court terme et des cycles électoraux (OCDE, 2020). Nous discutons de chacune de ces raisons ci-après.

Vastes ramifications de l'IA dans la société

L'éventuelle incidence des systèmes d'IA s'explique plus facilement quand on considère que ceux-ci relèvent d'une technologie polyvalente. L'organisation The Netherlands Scientific Council for Government Policy (2021) estime que l'IA s'apparente à une telle technologie et que son incidence potentielle a une envergure comparable à celle des moteurs à vapeur ou à combustion, de l'électricité, ou encore de l'ordinateur.

Trois facteurs caractérisent les technologies polyvalentes : i) leur omniprésence, ii) leur constant perfectionnement et iii) le fait qu'elles soient source d'innovations complémentaires. Premièrement, ces technologies se répandent de manière généralisée dans différents domaines et processus de production en plus de se trouver dans des produits et services, d'où leur omniprésence. Selon ce qu'on entend sur le terrain, il devient de plus en plus difficile pour les citoyens et citoyennes d'éviter d'avoir affaire à des systèmes d'IA alors que les entreprises et les organismes publics intègrent celle-ci dans leurs produits et services courants. Deuxièmement, le perfectionnement fait référence à la vitesse d'évolution technologique. Force est de constater que l'IA n'est pas statique, mais qu'elle continue plutôt d'évoluer et de s'améliorer grâce à des avancées de l'informatique, à la baisse des coûts associés à la collecte des données ou à leur stockage, ainsi qu'à la recherche scientifique en cours. Troisièmement, en ce qui a trait au caractère

innovant, les technologies et procédés connectés deviendront plus efficaces et mèneront à des gains de productivité à mesure que les entreprises et les gouvernements intégreront l'IA dans leurs produits et services. Pour reprendre les mots de Trajtenberg, l'IA a le potentiel d'entraîner « une vague d'innovations complémentaires dans un large et sans cesse croissant éventail de domaines d'application » (2018, p. 176).

L'incidence des systèmes d'IA est donc vaste et devrait être un sujet de préoccupation politique général, ce qui requiert la participation d'un grand ensemble de parties prenantes touchées par la présence croissante des systèmes d'IA.

Besoin de transparence et d'imputabilité

La transparence et l'explicabilité (caractère explicable et interprétable) font souvent défaut en ce qui a trait aux systèmes d'IA. Elles sont toutefois des conditions préalables pour assurer la confiance, le fonctionnement adéquat de régimes légaux, et la capacité d'évaluer ou éventuellement de remettre en question certains résultats. Pour étayer notre propos, nous considérons ci-après deux points de vue : l'un de nature plutôt technique et l'autre se rapportant à la communication.

D'un point de vue technique, les systèmes d'IA peuvent s'avérer complexes ; en expliquer le fonctionnement est donc parfois long et difficile à faire. Ce constat s'applique particulièrement aux modèles d'apprentissage profond, qui scrutent des couches de données à un niveau plus poussé que sont capables de le faire des humains. Le fait qu'il y ait nécessairement un compromis à faire entre la précision de ces modèles et leur interprétabilité relève d'un mythe, tel que le mentionne Rudin (2019), qui soutient de plus que ce mythe a amené des chercheurs et chercheuses à renoncer à concevoir des modèles interprétables. Elle ajoute qu'il y a, de la part du secteur privé, un fort intérêt commercial à garder les modèles secrets. Il importe toutefois de souligner qu'il existe des solutions techniques pour favoriser la transparence, bien souvent sans difficulté (consulter aussi Hall et Gill, 2019 ; Guidotti *et al.*, 2018).

Un second point de vue relève de la communication. Il y a dans la société une compréhension limitée de l'IA. Il s'avère souvent difficile de saisir le vocabulaire qu'emploient ceux et celles qui s'y connaissent, qui font par exemple référence à l'IA en tant que « boîte noire » ou parlent d'« apprentissage automatique » et de « mégadonnées ». Voilà qui peut faire dérailler les discussions sur l'IA. Or, il est important que les membres de la société civile ainsi que les utilisateurs et utilisatrices comprennent bien le fonctionnement général des systèmes auxquels ils ont affaire afin de saisir les conséquences qu'il y a à les utiliser, d'en débattre, de déterminer les améliorations possibles et, éventuellement, de s'insurger contre des répercussions fâcheuses. Les processus délibératifs contribuent en parallèle à favoriser les mécanismes d'apprentissage et fournissent une rétroaction sur le fonctionnement des systèmes d'IA.

Dilemmes fondés sur des valeurs, et compromis

Les processus délibératifs sont d'autant plus appropriés quand le sujet à l'étude soulève des dilemmes fondés sur des valeurs, des problèmes complexes nécessitant des compromis, ou des enjeux de longue haleine qui vont au-delà de mesures à court terme et des cycles électoraux. Des systèmes d'IA qu'on ne cesse de sophistiquer entraînent de tels dilemmes et compromis (voir par exemple Winfield, 2019).

Utilisons l'enjeu des discours haineux en ligne pour illustrer un tel dilemme moral. Les systèmes d'IA sont la solution privilégiée par des entreprises technologiques pour repérer, catégoriser et supprimer les discours haineux largement diffusés en ligne (voir par exemple Gorwa *et al.*, 2020). En pratique, ces dernières font toutefois face à des défis méthodologiques, techniques et éthiques. Elles sont notamment chargées de préserver un équilibre entre la liberté d'expression et la protection contre des préjudices, et doivent également veiller au respect du droit à la vie privée des utilisateurs et utilisatrices. De plus,

les entreprises technologiques doivent pouvoir expliquer le fondement des décisions prises par leurs systèmes d'IA et elles sont responsables d'atténuer les dommages découlant des préjugés sociaux qui y sont encodés (Llansó *et al.*, 2020).

Dans cette partie, nous avons montré l'existence de nombreux éléments justifiant la nécessité d'une approche multipartite. La conception et l'usage de systèmes d'IA comportent des dilemmes moraux et entraînent des répercussions variées qui auront un effet à long terme sur la société. Il importe de favoriser la poursuite des apprentissages et des délibérations en vue d'accroître la transparence et l'imputabilité en ce qui a trait à l'IA.

---

4. PRINCIPES GÉNÉRAUX DES PROCESSUS MULTIPARTITES EFFICACES

Concevoir un cadre stratégique relatif à l'IA requiert la recherche d'un équilibre entre le soutien à l'innovation et l'atténuation des risques qu'elle pose. L'un des plus importants défis au moment de réglementer les technologies polyvalentes consiste à déterminer le moment où il importe de fixer des règles et le degré de sévérité de celles-ci. Le dilemme de Collingridge résume bien la difficulté de la tâche. Il pose qu'au cours des phases préliminaires de mise au point, la nature et les répercussions d'une nouvelle technologie restent difficiles à déterminer et qu'il est par conséquent difficile de réglementer celle-ci. Plus tard, quand cette technologie entraîne des conséquences néfastes, elle est souvent si imbriquée dans l'économie et la société qu'il est encore là très difficile de la réglementer (Collingridge, 1981). Cette difficulté commande l'élaboration de politiques agiles et souples afin de composer avec l'évolution constante de l'IA. Elle sous-tend notamment la pertinence de choisir des types de politiques pour lesquelles les coûts liés aux erreurs sont faibles ainsi que l'importance accrue d'un mécanisme de surveillance et d'évaluation efficace. Faire intervenir des parties prenantes représente sans doute une manière judicieuse d'élaborer de telles politiques.

Un large pan de la documentation couvre la mobilisation des parties prenantes. Au cours des dernières années, des leçons ont été tirées dans des régions, des cultures et des domaines divers. À titre d'exemple, Renn *et al.* (2020) ainsi que Ambole *et al.* (2021) ont rendu compte de la participation de parties prenantes dans le domaine de l'énergie. Mustalahti et Rakotonarivo (2014) ont analysé la participation communautaire dans la réduction des émissions découlant de la déforestation et de la dégradation des forêts en Tanzanie. García-López et Arizpe (2010) ont étudié le recours à des processus participatifs pour régler des conflits au sujet de la production de soja au Paraguay et en Argentine. Hoogesteger (2012) s'est, quant à lui, intéressé à la participation en contexte de démocratisation de la gestion des eaux dans les Andes équatoriennes. Leurs analyses présentent des éléments communs : i) les approches multipartites permettent aux décideurs et décideuses de cibler des enjeux de manière prioritaire et de prendre des décisions soutenues par des données, ii) elles favorisent la croissance à long terme et la pérennité en améliorant la représentation et l'influence des groupes marginalisés et iii) à l'échelle locale, il arrive souvent que les communautés ne disposent pas des ressources adéquates pour ficeler et gérer leurs propres projets. Fait important à retenir : les communautés locales bénéficient de la collaboration de spécialistes de l'extérieur pouvant les aider à concevoir leurs projets ainsi qu'à obtenir un soutien financier et logistique.

Il est possible de tirer des leçons à partir de l'expérience vécue avec d'autres innovations technologiques. Van der Spuy (2017) brosse une vue d'ensemble de l'évolution de la participation multipartite en matière de gouvernance d'Internet. Son analyse montre que les processus de participation doivent intégrer un certain nombre de valeurs pour s'avérer efficaces dans l'atteinte d'un consensus et l'amélioration de la prise de décisions. La chercheuse considère que les approches multipartites doivent être i) inclusives, ii) diverses, iii) collaboratives, iv) transparentes, v) équitables, vi) modulables et pertinentes, vii) sûres

et discrètes, viii) imputables et légitimes, et enfin ix) réactives. Nous soutenons que ces principes généraux peuvent constituer un référentiel de toute approche multipartite visant à concevoir une politique en matière d'IA.

Abordant précisément le sujet qui nous intéresse, Buhmann et Fieseler (2021) suggèrent que quatre principes de communication orientent la délibération sur l'IA. Ces principes sont de nature légèrement plus pratique que ce qui précède et sont inspirés d'une recherche théorique menée par Nanz et Steffek (2005). Le premier principe pointe le besoin d'un accès à des instances délibératives formelles. Toutes les personnes aptes à prendre la parole et à agir, et particulièrement celles qui pourraient souffrir des conséquences néfastes des algorithmes et de leurs décisions, devraient bénéficier d'un accès équitable à un forum ouvert aspirant à mettre des enjeux en lumière et à encourager la discussion. Bondi *et al.* (2021) rappellent ce principe et soulignent la nécessité d'adopter une approche communautaire pour évaluer la réussite des projets d'IA. Il importe de mentionner un prérequis important : susciter une prise de conscience et améliorer les compétences en matière d'IA constituent des gages d'efficacité des processus délibératifs, particulièrement dans les pays en développement. Ces mesures font augmenter le nombre de personnes pouvant participer au débat et les encouragent à le faire. Elles favorisent l'inclusion et la diversité au sein d'un processus délibératif.

Le deuxième principe énonce que ceux et celles qui participent à un processus délibératif devraient avoir accès à un maximum d'information sur les enjeux en cause, les possibles solutions et leurs conséquences. À cet égard, rendre l'information librement accessible à toutes les parties prenantes constituerait une pratique exemplaire, mais ne suffirait pas à satisfaire ce principe. Les parties prenantes ne sont pas également informées ou ne bénéficient pas d'un même accès à l'information. L'atteinte d'un niveau équivalent de savoir relève bien sûr de l'utopie, mais ce principe fait que ceux et celles qui possèdent de vastes connaissances et l'information (c'est-à-dire la communauté technique, le secteur privé et les responsables des politiques) ont la responsabilité de s'assurer que les autres parties prenantes (habituellement issues de la société civile) ont l'occasion d'affiner leur propre savoir concernant les enjeux et qu'elles disposent de ressources pour le faire.

Le troisième principe invite à prendre en considération tous les arguments possibles au cours du processus. Cette mesure est une condition préalable à la sauvegarde du caractère raisonnable du discours et de la délibération, et elle garantit des résultats équilibrés. Enfin, Buhmann et Fieseler signalent qu'un processus délibératif se doit d'être réactif au regard des préoccupations ou des suggestions des parties prenantes. Même si toutes les conditions mentionnées précédemment sont réunies, la participation ne joue sur les résultats que si les décideurs et décideuses se montrent ouverts aux avis et acceptent qu'une certaine influence soit exercée sur le processus décisionnel.

Pour aider les responsables des politiques à mettre ces principes en œuvre, l'Université de Washington et l'Université de Montréal ont conçu deux guides pratiques bien utiles. Le premier, intitulé *Diverse Voices*, aide les responsables à accroître la diversité dans l'élaboration d'une politique et contient une méthode visant à intégrer dans la démarche des groupes sous-représentés (Magassa *et al.*, 2017 ; consulter aussi Young *et al.*, 2019). Le second guide explique en des termes simples mais précis ce qu'est l'IA, présente les liens qu'elle entretient avec d'autres technologies et justifie la nécessité de délibérer sur l'IA (Dilhac *et al.*, 2020). En outre, la Westminster Foundation for Democracy (WFD) et la newDemocracy Foundation ont étudié la nouvelle vague délibérative qui touche les processus démocratiques en Afrique, en Asie et en Amérique latine. Les cas de démocratie délibérative et les leçons qui en ont été tirées s'avèrent utiles et représentent une source d'inspiration (WFD, 2021 et Kimaili 2021).

5. ÉTUDES DE CAS : CHILI ET INDE

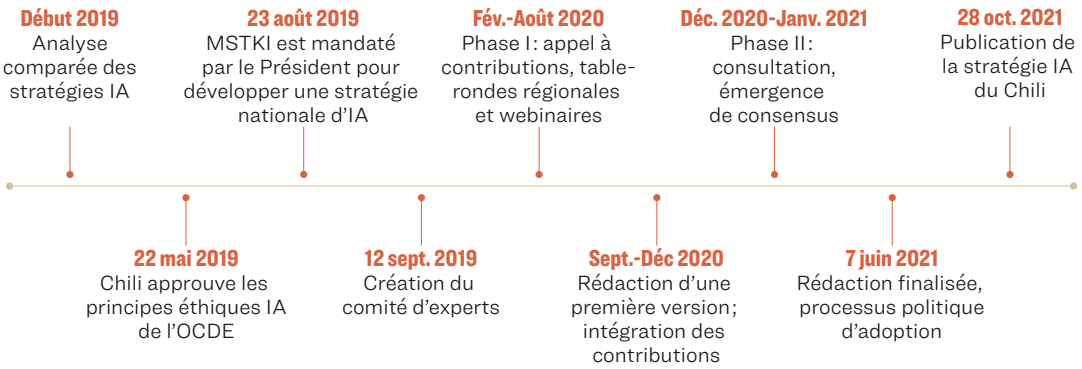
Depuis la publication de la première stratégie nationale en matière d'IA par le Canada, en 2017, plusieurs pays ont entrepris d'élaborer leurs propres stratégies, plans d'action ou politiques. À ce jour, la plupart des pays de l'OCDE ont adopté ou mis en œuvre des initiatives, mais de nombreux pays du Sud doivent encore s'y mettre. En Afrique, par exemple, l'Égypte et la Mauritanie sont actuellement les seuls pays à avoir une stratégie visant précisément l'IA, mais du travail est déjà en cours à cet égard au Rwanda, au Kenya, au Ghana, au Nigéria, en Afrique du Sud, en Tunisie et en Ouganda (Effoduh, 2020). Le Future of Life Institute[76], l'Observatoire OCDE des politiques relatives à l'IA (OCDE.AI, 2021) et la plateforme Globalpolicy.AI[77] publient tous un portrait à jour de l'évolution de la situation à l'échelle internationale.

Dans cette partie du chapitre, nous scrutons l'approche adoptée par le Chili et l'Inde, respectivement, en distinguant trois phases du processus d'élaboration d'une politique : i) la planification, ii) l'élaboration et iii) la mise en œuvre et l'évaluation. Nous discutons des étapes suivies durant chacune de ces phases en faisant ressortir des décisions clés et en observant particulièrement la nature de la participation.

Étude de cas : Chili

| FIGURE 1 |

Étapes clés de l'élaboration de la stratégie en matière d'IA du Chili.



Le Chili a commencé à élaborer sa stratégie en matière d'IA en 2019 (voir la figure 1). Le gouvernement chilien mentionnait alors que l'essor de l'IA fondait le besoin d'agir de manière préventive à l'égard des changements que l'IA pouvait engendrer dans la société. La stratégie a été publiée en octobre 2021 à la suite d'un processus participatif qui s'était déroulé en deux rondes (MCTCI, 2021b). La stratégie s'articule autour de trois axes (facteurs favorables ; élaboration et adoption ; éthique, et aspects réglementaires et socioéconomiques) et réclame la mise au point et l'usage d'une IA centrée sur l'humain

76. Pour plus d'information, voir le site de l'institut : https://futureoflife.org/ai-policy/.

77. Pour plus d'information sur GlobalPolicy.AI : https://globalpolicy.ai/fr/key-focus-areas/.

qui soit sûre, inclusive, mondialisée et profitable à la société. La version définitive du document prévoit 70 actions prioritaires à mener à court terme (plan d'action) et 180 initiatives à mettre en place au cours de la période 2021-2030 (stratégie en matière d'IA).

L'approche adoptée au Chili constitue un excellent exemple d'un processus participatif multipartite. Au cours de la phase de la planification, les responsables ont d'abord mené une analyse comparative des stratégies et des politiques d'autres pays en ce qui a trait à l'IA. Les résultats de cette analyse ont été présentés au cabinet présidentiel chilien en août 2019. Ce dernier a mandaté le ministère de la Science, de la Technologie, du Savoir et de l'Innovation (le MCTCI) pour qu'il élabore une stratégie nationale en matière d'IA, une initiative qu'allaient mener un comité de spécialistes ainsi qu'un comité interministériel. Leur tâche consistait à rédiger une ébauche de la stratégie, laquelle serait ensuite publiée pour recueillir des commentaires du grand public.

Après une période d'agitation sociale ayant secoué le Chili en 2019, les spécialistes et les responsables du projet ont délaissé l'approche descendante et linéaire au profit d'une approche ascendante participative multipartite. Au lieu de rédiger une ébauche de la stratégie, les spécialistes ont compilé une liste de thèmes pertinents par rapport à l'IA, laquelle a servi de guide au cours de la première phase du processus participatif multipartite. Cette phase, engagée en février 2020, comprenait trois volets : l'appel à organiser sa propre table ronde (un formulaire de commentaires modèle était accessible en ligne), l'organisation de tables rondes régionales par le ministère, et la tenue de webinaires en ligne dans lesquels des spécialistes effectueraient de la sensibilisation et renforceraient les capacités. Pour faciliter ces démarches, on a fourni un guide sur la participation publique, des fonctionnaires ont pris part à des tables rondes quand c'était nécessaire, et le gouvernement a accordé du financement public.

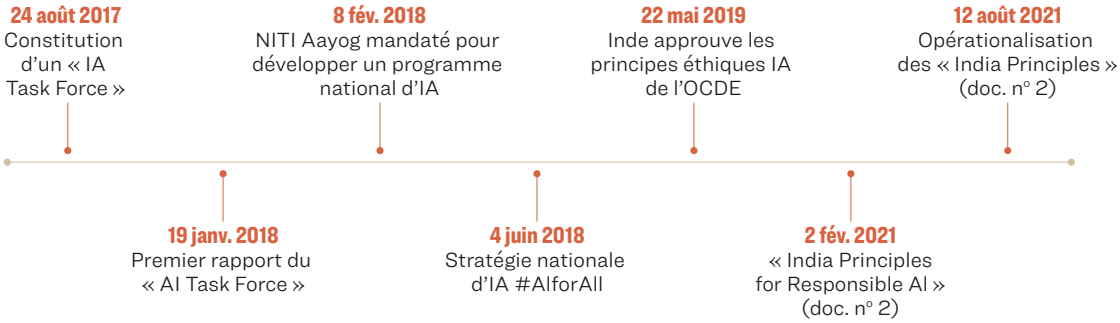
Le caractère unique de ce processus participatif relève du nombre de parties prenantes mobilisées. En six mois, plus de 1 300 personnes et organisations ont tenu leur propre table ronde ou fourni leurs commentaires en ligne, 69 tables rondes régionales ont été organisées, auxquelles 400 personnes ont participé et les webinaires (dont la moitié ont été présentés par des femmes) ont rejoint un auditoire de 6 600 personnes. Il y a également eu une variété de participants et participantes dans ce processus. Des commentaires reçus en ligne, 36 % provenaient de la société civile, et nombreux sont ceux et celles qui ont indiqué n'avoir jamais participé à l'élaboration de politiques auparavant.

À partir de l'information recueillie, les spécialistes et autres responsables ont rédigé une première ébauche de la stratégie. En décembre 2020, cette ébauche a été publiée en ligne et a été l'objet de la seconde ronde de consultation publique. Cette fois, les participants et participantes ont formulé de nouvelles questions ou d'autres commentaires, et ils et elles ont évalué leur degré d'adhésion aux objectifs de la politique en matière d'IA et à certains de ses éléments. Le processus consultatif a mené à l'approbation, à plus de 80 %, des objectifs et des principes contenus dans l'ébauche. Des données qualitatives ont indiqué que les participants et participantes avaient apprécié l'approche ascendante ainsi que son aspect éducatif (MCTCI, 2021a). La phase de l'élaboration de la stratégie s'est conclue en juin 2021 par le traitement de l'information recueillie, ce qui a ensuite fait place à la démarche d'adoption par les instances politiques. Cinq mois plus tard, soit le 28 octobre, la stratégie et le plan d'action du Chili en matière d'IA étaient publiés.

Étude de cas : Inde

| FIGURE 2 |

Étapes clés de l'élaboration de la stratégie en matière d'IA de l'Inde.



En Inde, le processus d'élaboration d'une politique s'est amorcé par la mise sur pied d'un groupe de travail sur l'IA (voir la figure 2). Après que ce dernier eut présenté un rapport, le groupe de réflexion NITI Aayog a reçu le mandat d'élaborer une stratégie nationale relative à l'IA. Un tel document – qui prône la technologie inclusive et dont le titre se traduirait par *Stratégie de l'Inde #IAPourTous* – a été publié au cours de l'été 2018 (NITI Aayog, 2018). Des discussions sont en cours depuis pour faire de cette stratégie une politique publique. Après une vaste démarche consultative auprès de spécialistes ainsi que de représentants et représentantes de la société civile et du secteur privé, NITI Aayog a récemment publié deux documents sur l'approche à adopter. Ils servent de feuille de route pour établir un écosystème de l'IA en Inde et présentent une mise à jour quant au processus d'élaboration d'une politique (NITI Aayog, 2021a, 2021b).

Puisque ce processus est toujours en cours en Inde, nous n'en présentons pas tous les détails, mais faisons plutôt ressortir, dans cette section, un certain nombre d'éléments dignes d'intérêt. En premier lieu, observons le groupe de travail qui a reçu le mandat de scruter le panorama de l'IA en Inde et de fournir des recommandations sur le rôle du gouvernement. Il a présenté ses conclusions en janvier 2018 (Kamakoti, 2018). Deux points saillants méritent tout particulièrement notre attention. D'une part, le groupe de travail sur l'IA se compose de 18 membres aux parcours divers ; ils et elles viennent du domaine de l'IA, de la fonction publique ou encore des milieux de la santé, du droit ou des finances. La diversité est déterminante au sein de cette équipe, qui exécute le travail préparatoire quant à l'approche et à la perspective à adopter par rapport l'IA. D'autre part, le groupe de travail a mis en ligne un site Web afin de solliciter les commentaires du public sur des enjeux liés à l'IA. Il s'agit d'une bonne pratique, particulièrement à l'étape de la planification. Faire intervenir le public tôt dans la démarche permet non seulement d'esquisser le panorama de l'IA, mais également d'obtenir de précieux renseignements sur la manière dont la société civile perçoit l'IA, ce qui est crucial pour en évaluer le potentiel (degré de compréhension et d'adoption).

Un deuxième élément digne de mention est le portail national INDIAai, lancé en 2020[78]. Financé tant par le gouvernement que par le secteur privé, le portail vise à rassembler au même endroit toute l'information relative à l'IA et à renforcer l'écosystème de l'IA en Inde. Il est déjà à la source de plusieurs initiatives remarquables, par exemple la conception de programmes éducatifs pour les jeunes, la mise en ligne d'un robot de clavardage cherchant à combattre la désinformation sur la COVID-19 et le lancement d'une mission nationale qui se penchera sur la traduction. Ce dernier projet a pour objectif de lever les barrières linguistiques dans un pays où il est fait bon usage de l'anglais, ce qui s'avère d'autant plus pertinent que 22 langues officielles sont employées en Inde, en plus d'un millier d'autres langues ou dialectes (Census of India, 2011).

Nous tenons dans un troisième temps à attirer l'attention sur la publication de deux manuels issus d'une collaboration entre des organismes publics et des parties prenantes d'Inde. Plus précisément, l'un cible les développeurs et développeuses de l'IA et traite de la protection et de la confidentialité des données, tandis que l'autre explique aux entreprises émergentes comment atténuer les préjugés de l'IA. Ils ont été rédigés et publiés, respectivement, par GIZ India (2021) en étroite collaboration avec le Data Security Council of India, et par INDIAai (2021). Ces manuels présentent des astuces pratiques et des conseils fondés sur la recherche universitaire, des principes éthiques largement reconnus et le contexte réglementaire de l'Inde. Ils constituent une manière efficace de transmettre des connaissances à jour à un lectorat ciblé, qu'il s'agisse de développeurs et développeuses de l'IA ou d'entrepreneurs et entrepreneuses. Ils pourraient également s'avérer utiles aux responsables des orientations politiques d'autres pays.

Le dernier élément digne de mention est le sommet mondial RAISE (acronyme anglais signifiant « IA responsable pour un renforcement social »), qui s'est tenu en ligne en 2020. Organisé par le ministère de l'Électronique et des Technologies de l'information, il a rassemblé des responsables de politiques, des spécialistes de l'IA, des penseurs et penseuses, des influenceurs et influenceuses, des praticiens et praticiennes, ainsi que des jeunes venant d'Inde et de l'étranger. Il s'est déroulé en 48 séances, a duré 85 heures, a accueilli plusieurs centaines de conférenciers et conférencières, et a atteint une participation de 79 000 personnes de 147 pays. Avec RAISE, l'Inde a réitéré son engagement envers la responsabilisation de l'IA à grande échelle et a offert à la communauté internationale de l'IA un forum où partager des idées, ce qui est une facette importante de la mobilisation multipartite.

6. LEÇONS TIRÉES DE L'EXPÉRIENCE

Dans les parties précédentes, nous avons abordé un certain nombre de critères caractérisant les processus participatifs efficaces, partagé des preuves de la mise en pratique des approches multipartites, puis examiné les approches adoptées par le Chili et l'Inde. Ces éléments combinés permettent de tirer plusieurs leçons.

Premièrement, la clarté s'avère essentielle durant la phase de la planification et devrait être un élément constitutif d'un processus multipartite inclusif pour la politique en matière d'IA. Les gouvernements ont la responsabilité de conduire l'élaboration de politiques. Ils doivent se montrer clairs devant le secteur privé, les spécialistes et la société civile quant au processus à suivre. Si la prévisibilité revêt toujours de l'importance en matière de réglementation, le contenu des politiques relatives à l'IA reste, quant à lui, difficile à prévoir, car il est bien souvent en pleine préparation. Il devient alors d'autant plus important de clarifier le processus à suivre afin d'offrir aux parties prenantes une certaine garantie qu'elles pourront faire entendre leur point de vue et défendre leurs intérêts. Le Chili représente un bon exemple à cet égard ;

78. Le portail est accessible à cette adresse : https://indiaai.gov.in/.

les responsables des politiques ont rapidement annoncé que le processus participatif s'effectuerait en deux rondes. Cette annonce a permis aux parties prenantes de bien savoir quand et comment elles seraient amenées à y participer.

Deuxièmement, chacune des études de cas montre que de recourir à une équipe de spécialistes ou à un groupe de travail tôt dans le processus d'élaboration d'une politique comporte des avantages. La tâche des spécialistes consiste habituellement à scruter le panorama de l'IA et à déterminer les priorités stratégiques ainsi que l'éventuel avantage concurrentiel lié à la recherche et au développement sur l'IA de même qu'au déploiement de celle-ci. Dans une telle approche, la démarche de sélection est également capitale et doit faire preuve d'inclusivité si l'on souhaite obtenir un portrait équilibré de l'IA (possibilités, faiblesses, priorités). Les spécialistes devraient défendre des points de vue et des intérêts divers. Le processus et les résultats ne seront toutefois inclusifs que si l'on s'emploie activement à recruter des parties prenantes étant habituellement tenues en marge, par exemple des représentants et représentantes de groupes désavantagés ou des organisations de jeunes. En Inde par exemple, le groupe de travail montrait de la diversité sur le plan de la formation scolaire.

Troisièmement, élaborer une stratégie ou des politiques en matière d'IA nécessite un mandat bien défini. Un seul ministère ou organisme public devrait ultimement se charger de coordonner l'élaboration de la stratégie d'ensemble ou du cadre stratégique, puisque l'IA influe sur le secteur politique et donc sur les responsabilités de presque tous les ministères et organismes publics. Les études de cas montrent que tant le ministère chilien de la Science, de la Technologie, du Savoir et de l'Innovation (le MCTCI) que le groupe de réflexion indien NITI Aayog ont obtenu un mandat du président de leur pays. Toutefois, le fait d'avoir un mandat n'implique pas une prise de décision unilatérale : les approches multipartites sont inefficaces lorsque les décisions sont prises sans tenir compte de l'apport des parties prenantes. Le mandat peut plutôt être utilisé pour inciter les agences et institutions gouvernementales concernées à se joindre à la délibération.

Quatrièmement, en ce qui concerne la phase de l'élaboration, il faut s'efforcer de recourir à un processus ouvert et inclusif. Dans bien des pays, il en va désormais de l'intérêt général de mener des consultations, mais celles-ci durent souvent trop peu, pâtissent d'une structure rigide ou présentent une efficacité discutable. Le processus employé par Innovation for Policy[79] et l'approche chilienne montrent qu'il est possible d'innover. On peut par exemple offrir aux parties prenantes le moyen d'annoter les ébauches, de s'exprimer sur des éléments clés ou de répondre à des questions précises. Il est également possible de sonder leur degré d'adhésion à une version préliminaire des documents. Agir de la sorte garantit que les responsables des politiques obtiennent une rétroaction pertinente et que les parties prenantes se sentent engagées dans le processus permettra aux décideurs d'être attentifs aux commentaires et suggestions des participants.

Cinquièmement, considérons la phase de la mise en œuvre et de l'évaluation. Les leçons à tirer à cet égard ont trait à la nécessité et à l'importance de suivre un plan de mise en œuvre de court terme. Une stratégie relative à l'IA présente un cadre général d'orientation qui doit néanmoins s'assortir d'actions concrètes. Une pratique exemplaire consiste donc à élaborer tant une stratégie en matière d'IA qu'un plan d'action à court terme. Agir ainsi entraîne une prise en charge de l'enjeu et crée un sentiment d'urgence alors que toutes les parties prenantes doivent définir des actions concrètes, gérer le budget et s'entendre sur une répartition des responsabilités. En outre, la tenue d'actions à court terme continuera à mobiliser les parties prenantes et offrira bientôt des résultats tangibles aux décideurs et décideuses, au milieu politique et à la société civile.

79. Voir https://participedia.net/method/642.

CONCLUSION

Dans ce chapitre, nous avons montré que le principe de la gouvernance multipartite s'applique à la conception de politiques en matière d'IA et nous avons présenté quelques éléments de base pour la conception de politiques inclusives. L'IA peut être considérée comme une technologie polyvalente qui transforme notre façon de travailler et de vivre. Il est de plus en plus difficile d'y renoncer, et nous commençons à peine à comprendre l'influence qu'elle aura sur nos vies. Il ne faut surtout pas oublier que l'IA se transforme, apprend et évolue constamment. Sa mise en œuvre et sa prolifération s'accompagnent de menaces aux droits humains ; elles soulèvent des questions et des dilemmes fondés sur des valeurs. Nous avons également établi que les approches multipartites parviennent à faire avancer les discussions. La société doit s'assurer d'exploiter les avantages de l'IA de manière responsable et durable, ainsi que de neutraliser les risques qu'elle pose en matière de droits et libertés. Nous avons la conviction qu'une approche délibérative multipartite s'avère fondamentale dans la conception d'une politique relative à l'IA parce que : i) la délibération mène à un cadre stratégique fondé sur un ensemble de valeurs et d'objectifs largement partagés, ii) le cadre stratégique proposé permet de s'adapter rapidement à la rétroaction des participants et participantes, et il crée un sentiment de responsabilité chez les parties prenantes, iii) le processus participatif accroît la sensibilisation et renforce les capacités par rapport à l'IA et iv) de nouvelles relations lient les parties prenantes, ce qui favorise la poursuite des discussions sur la politique en matière d'IA et son élaboration.

Finalement, même si la participation importe à toutes les étapes de l'élaboration d'une politique, nous exhortons ceux et celles qui en sont responsables, les organisations de la société civile, les chercheurs et chercheuses, la communauté technique, le secteur privé, ainsi que les citoyens et citoyennes en cause, à prêter une attention particulière à la mobilisation des parties prenantes dès le début du processus. Puisque l'IA ne semble pas bien comprise, qu'elle recourt à un vocabulaire vague et qu'elle entraîne toute une gamme de répercussions, accroître la sensibilisation, renforcer les capacités et corriger les perspectives trop optimistes ou sombres à son égard devraient être des visées fondamentales de toute politique en la matière de même que de tout processus participatif.

Ce chapitre a mis en évidence cinq leçons pour le développement multipartite de l'IA. Il s'agit d'une version condensée de la publication de l'UNESCO et de i4Policy (2022), qui présente dix éléments constitutifs d'une conception politique inclusive. Nous espérons que ces publications donneront au lecteur les moyens de pousser à la démocratisation de l'élaboration des politiques d'IA.

RÉFÉRENCES

Adam, L., James, T. et Wanjira, A. M. 2007. *Frequently Asked Questions about Multi-Stakeholder Partnerships in ICTs for Development*. Association for Progressive Communications (APC). https://www.apc.org/sites/default/files/catia_ms_guide_EN-1.pdf

AlgorithmWatch. 2020. *In the Realm of Paper Tigers : Exploring the Failings of AI Ethics Guidelines*. https://algorithmwatch.org/en/ai-ethics-guidelines-inventory-upgrade-2020/

Ambole, A., Koranteng, K., Njoroge, P. et Luhangala, D. L. 2021. A review of energy communities in sub-Saharan Africa as a transition pathway to energy democracy. *Sustainability,* vol. 13, nº 4, p. 2128. https://doi.org/10.3390/su13042128

Bäckstrand, K. 2006. Multi-stakeholder partnerships for sustainable development : Rethinking legitimacy, accountability and effectiveness. *European Environment,* vol. 16, nº 5, pp. 290-306. https://doi.org/10.1002/eet.425

Barocas, S. et Selbst, A. D. 2016. *Big Data's disparate impact*. *104 California Law Review* No. 671. https://doi.org/10.2139/ssrn.2477899

Bernal, P. 2016. Data gathering, surveillance and human rights : Recasting the debate. *Journal of Cyber Policy,* vol. 1 (September), pp. 1-22. https://doi.org/10.1080/23738871.2016.1228990

Birhane, A. 2020. Algorithmic colonization of Africa. *SCRIPTed,* vol. 17, nº 2, pp. 389-409. https://doi.org/10.2966/scrip.170220.389

Black, E. et Richmond, R. 2019. Improving early detection of breast cancer in sub-Saharan Africa : Why mammography may not be the way forward. *Globalization and Health,* nº 15, article 3. https://doi.org/10.1186/s12992-018-0446-6

Bondi, E., Xu, L., Acosta-Navas, D. et Killian, J. A. 2021. Envisioning Communities : A Participatory Approach Towards AI for Social Good. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 425-36. Virtual Event USA : ACM. https://doi.org/10.1145/3461702.3462612

Brockmyer, B. et Fox, J. A. 2015. *Assessing the evidence : The effectiveness and impact of governance-oriented multi-stakeholder initiatives*. Transparency and Accountability Initiative, September 20. https://papers.ssrn.com/abstract=2693379

Buhmann, A. et Fieseler, C. 2021. Towards a deliberative framework for responsible innovation in artificial intelligence. *Technology in Society,* vol. 64 (February) : 101475. https://doi.org/10.1016/j.techsoc.2020.101475

Census of India. 2018. *Census of India 2011 : Language*. New Delhi, Office of the Registrar General. https://censusindia.gov.in/

Collingridge, D. 1981. *The Social Control of Technology*. New York, Palgrave Macmillan.

Crawford, K. 2021. *Atlas of AI : Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven, Yale University Press.

Dilhac, M.-A., Mai, V., Mörch, C.-M., Noiseau, P. et Voarino, N. 2020. *Penser l'intelligence artificielle responsable : un guide de délibération.* Montréal, Algora Lab-Mila, https://observatoire-ia.ulaval.ca/penser-lintelligence-artificielle-responsable-un-guide-de-deliberation/

Dingwerth, K. 2008. Private transnational governance and the developing world : A comparative perspective. *International Studies Quarterly,* vol. 52, nº 3, pp. 607-34. https://doi.org/10.1111/j.1468-2478.2008.00517.x

Faysse, N. 2006. Troubles on the way : An analysis of the challenges faced by multi-stakeholder platforms. *Natural Resources Forum,* vol. 30, nº 3 (August), pp. 219-29. https://onlinelibrary.wiley.com/doi/10.1111/j.1477-8947.2006.00112.x

Feldstein, S. 2019. *The Global Expansion of AI Surveillance : Working Paper*. Carnegie Endowment for International Peace. https://carnegieendowment.org/files/WP-Feldstein-AISurveillance_final1.pdf

Fransen, L. W. et Kolk, A. 2007. Global rule-Setting for business : A critical analysis of multi-stakeholder standards. *Organization,* vol. 14, nº 5, pp. 667-84. https://doi.org/10.1177/1350508407080305

Garcia, E. 2019. The militarization of artificial intelligence : A wake-up call for the global South. https://doi.org/10.2139/ssrn.3452323

García-López, G. A., et Arizpe, N. 2010. Participatory processes in the soy conflicts in Paraguay and Argentina. *Ecological Economics,* vol. 70, nº 2, pp. 196-206. https://doi.org/10.1016/j.ecolecon.2010.06.013

GIZ India. 2021. *Handbook on Data Protection and Privacy for Developers of Artificial Intelligence (AI) in India.* New Delhi. https://www.dsci.in/content/privacy-handbook-for-ai-developers

Gleckman, H. 2018. *Multistakeholder Governance and Democracy : A Global Challenge*. London, Routledge

Gorwa, R., Binns, R. et Katzenbach, C. 2020. Algorithmic content moderation : Technical and political challenges in the automation of platform governance. *Big Data & Society,* vol. 7, nº 1. https://doi.org/10.1177/2053951719897945

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D. et Giannotti, F. 2018. A survey of methods for explaining black box models. *ArXiv :1802.01933 [Cs]*, juin. http://arxiv.org/abs/1802.01933

Gwagwa, A., Kachidza, P., Siminyu, K. et Smith, M. 2021. Responsible artificial intelligence in sub-Saharan Africa : Landscape and general state of play. AI4D. https://idl-bnc-idrc.dspacedirect.org/handle/10625/59997

Gwagwa, A., Kraemer-Mbula, E., Rizk, N., Rutenberg, I. et De Beer, J. 2020. Artificial Intelligence (AI) Deployments in Africa : Benefits, challenges and policy dimensions. *The African Journal of Information and Communication*, nº 26 (December), pp. 1-28. https://doi.org/10.23962/10539/30361

Habermas, J. 1989. *The Structural Transformation of the Public Sphere : An Inquiry into a Category of Bourgeois Society*. Translated by Thomas Burger. Studies in Contemporary German Social Thought. Cambridge, MA, USA : MIT Press.

Hall, P. et Gill, N. 2019. *An Introduction to Machine Learning Interpretability*. 2nd ed. O'Reilly Media. https://www.oreilly.com/library/view/an-introduction-to/9781098115487/

HCDH. 2018. *Le droit à la vie privée à l'ère du numérique : rapport du Haut-Commissaire des Nations Unies aux droits de l'homme*. A/HRC/39/29. Haut-Commissariat des Nations Unies aux droits de l'homme. https://undocs.org/fr/A/HRC/39/29

Hofmann, J. 2016. Multi-stakeholderism in internet governance : Putting a fiction into practice. *Journal of Cyber Policy,* vol. 1, nº 1, pp. 29-49. https://doi.org/10.1080/23738871.2016.1158303

Hoogesteger, J. 2012. Democratizing water governance from the grassroots : The development of Interjuntas-Chimborazo in the Ecuadorian Andes. *Human Organization,* vol. 71, nº 1, pp. 76-86. https://doi.org/10.17730/humo.71.1.b8v77j0321u28863

INDIAai. 2021. *Mitigating Bias in AI : A Handbook for Startups.* https://indiaai.s3.ap-south-1.amazonaws.com/docs/AI+Handbook_27-09-2021.pdf

ITU. 2018. *Assessing the Economic Impact of Artificial Intelligence*. 1st Issue Paper on Emerging Trends. http://handle.itu.int/11.1002/pub/81202956-en

Jobin, A., Ienca, M. et Vayena, E. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence,* vol. 1, nº 9, pp. 389-99. https://doi.org/10.1038/s42256-019-0088-2

Kamakoti, V. 2018. *Report of the Artificial Intelligence Task Force*. Government of India. https://dipp.gov.in/sites/default/files/Report_of_Task_Force_on_ArtificialIntelligence_20March2018_2.pdf.

Kimaili, K. 2021. Lessons from deliberative democracy in Africa. Westminster Foundation for Democracy. August 25. https://www.wfd.org/commentary/lessons-deliberative-democracy-africa
Kwet, M. 2019. Digital colonialism : US empire and the new imperialism in the global South. *Race & Class,* vol. 60, nº 4, pp. 3-26. https://doi.org/10.1177/0306396818823172

Llansó, E., Van Hoboken, J., Leerssen, P. et Harambam, J. 2020. Artificial intelligence, content moderation, and freedom of expression. *The Transatlantic Working Group Papers Series*. Annenberg Public Policy Center of the University of Pennsylvania. https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/05/Artificial_Intelligence_TWG_Llanso_Feb_2020.pdf

Magassa, L., Young, M. et Friedman, B. 2017. *Diverse Voices : A How-to Guide for Facilitating Inclusiveness in Tech Policy*. Tech Policy Lab, University of Washington. https://techpolicylab.uw.edu/wp-content/uploads/2017/10/TPL_Diverse_Voices_How-To_Guide_2017.pdf

Martens, W., van der Linden, B. et Wörsdörfer, M. 2019. How to assess the democratic qualities of a multi-stakeholder initiative from a Habermasian perspective ? Deliberative Democracy and the Equator Principles Framework. *Journal of Business Ethics,* vol. 155, nº 4, pp. 1115-33. https://doi.org/10.1007/s10551-017-3532-4

McKinsey Global Institute. 2018. *Notes from the AI Frontier : Modeling the Impact of AI on the World Economy*. Discussion Paper. https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-modeling-the-impact-of-ai-on-the-world-economy

MCTCI, Ministerio de Ciencia, Tecnología, Conocimiento e Innovacíon. 2021a. *Consulta Pública de Inteligencia Artificial : Informe de Resultados*. Santiago de Chile, Gobierno de Chile. https://minciencia.gob.cl/areas-de-trabajo/inteligencia-artificial/politica-nacional-de-inteligencia-artificial/proceso-de-elaboracion/

———. 2021b. *Política Nacional de Inteligencia Artificial*. Santiago de Chile, Gobierno de Chile. https://minciencia.gob.cl/areas-de-trabajo/inteligencia-artificial/politica-nacional-de-inteligencia-artificial/

Mohamed, S., Png, M. et Isaac, W. 2020. Decolonial AI : Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology,* vol. 33, nº 4, pp. 659-84. https://doi.org/10.1007/s13347-020-00405-8

Moog, S., Spicer, A. et Böhm, S. 2014. The politics of multi-stakeholder initiatives : The crisis of the Forest Stewardship Council. *Journal of Business Ethics*, May. https://doi.org/10.1007/s10551-013-2033-3

Moraes, T. G., Almeida, E. C. et Pereira, J. R. L. 2021. Smile, you are being identified ! Risks and measures for the use of facial recognition in (semi-)public spaces. *AI and Ethics,* vol. 1, nº 2, pp. 159-72. https://doi.org/10.1007/s43681-020-00014-3

Mudongo, O. 2021. *Africa's Expansion of AI Surveillance : Regional Gaps and Key Trends*. Policy Brief. Cape Town, Research ICT Africa. https://researchictafrica.net/publication/africas-expansion-of-ai-surveillance-regional-gaps-and-key-trends/

Murray, C. 2013. *Coming Apart : The State of White America, 1960-2010*. Illustrated edition. New York, N.Y, Crown Forum.

Mustalahti, I. et Rakotonarivo, O. S. 2014. REDD+ and empowered deliberative democracy : Learning from Tanzania. *World Development,* vol. 59 (July), pp. 199-211. https://doi.org/10.1016/j.worlddev.2014.01.022

Nanz, P. et Steffek, J. 2005. Assessing the democratic quality of deliberation in international governance : Criteria and research strategies. *Acta Politica,* vol. 40, nº 3, pp. 368-83. https://doi.org/10.1057/palgrave.ap.5500118

Nations Unies. Conseil de sécurité. 2021. *Lettre datée du 8 mars 2021, adressée à la présidente du Conseil de sécurité par le Groupe d'experts sur la Libye créé par la résolution 1973 (2011) du Conseil de sécurité.* https://digitallibrary.un.org/record/3905159

NITI Aayog. 2018. *National Strategy for Artificial Intelligence #AIForAll.* https://www.niti.gov.in/national-strategy-artificial-intelligence

———. 2021a. *Approach Document for India. Part 1: Principles for Responsible AI. Responsible AI #AIForAll.* https://indiaai.gov.in/research-reports/responsible-ai-part-1-principles-for-responsible-ai

———. 2021b. *Approach Document for India. Part 2: Operationalizing Principles for Responsible AI. Responsible AI #AIForAll.* https://indiaai.gov.in/research-reports/responsible-ai-part-2-operationalizing-principles-for-responsible-ai

OCDE. 2020. *Innovative Citizen Participation and New Democratic Institutions: Catching the Deliberative Wave*. OECD. https://doi.org/10.1787/339306da-en

OCDE.AI (2021). *Politiques et stratégies nationales en matière d'IA.* D'après des données de la Commission européenne et de l'OCDE. https://oecd.ai/fr/dashboards

Effoduh, J. O. 2020. *7 ways that African states are legitimizing artificial intelligence. Open AIR.* 20 octobre. https://openair.africa/7-ways-that-african-states-are-legitimizing-artificial-intelligence/

Oxford Insights. 2020. *Government AI Readiness Index 2020*. Ottawa, IDRC. https://www.oxfordinsights.com/government-ai-readiness-index-2020

Polli, F. 2019. Using AI to Eliminate Bias from Hiring. *Harvard Business Review*, October 29, 2019. https://hbr.org/2019/10/using-ai-to-eliminate-bias-from-hiring

Prince, A. et Schwarcz, D. 2020. Proxy discrimination in the age of artificial intelligence and Big Data. *Iowa Law Review,* vol. 105, nº 1257, p. 63.

Privacy International. 2019. *Guide to International Law and Surveillance 2.0.* https://privacyinternational.org/sites/default/files/2019-04/Guide to International Law and Surveillance 2.0.pdf

Renn, O., Ulmer, F. et Deckert, A. 2020. *The Role of Public Participation in Energy Transitions.* Cambridge, MA, Academic Press.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *ArXiv:1811.10154 [Cs, Stat]*, septembre. http://arxiv.org/abs/1811.10154

Scholte, J. A. 2020. *Multistakeholderism Filling the Global Governance Gap?* The Global Challenges Foundation. https://globalchallenges.org/multistakeholderism-filling-the-global-governance-gap/

Sloan, P. et Oliver, D. 2013. Building trust in multi-stakeholder partnerships: Critical emotional incidents and practices of engagement. *Organization Studies,* vol. 34, nº 12, pp. 1835-68. https://doi.org/10.1177/0170840613495018

Smart Africa. 2021. *Blueprint: Artificial Intelligence for Africa*. Kigali, Smart Africa, GIZ et GFA Consulting. https://smartafrica.org/knowledge/artificial-intelligence-for-africa/

Souter, D. 2017. *Inside the Information Society: Multistakeholder Participation, a Work in Progress.* Association for Progressive Communications. https://www.apc.org/en/blog/inside-information-society-multistakeholder-participation-work-progress

The Netherlands Scientific Council for Government Policy. 2021. *Mission AI*. The New System Technology (English Summary). The Hague, WRR. https://www.wrr.nl/publicaties/rapporten/2021/11/11/opgave-ai-de-nieuwe-systeemtechnologie

Trajtenberg, M. 2018. Artificial Intelligence as the Next GPT: A Political-Economy Perspective. *NBER Chapters*. National Bureau of Economic Research. https://econpapers.repec.org/bookchap/nbrnberch/14025.htm

UNCTAD. 2021. *Digital Economy Report 2021*. New York, United Nations Publications.

UNESCO. 2019. *Steering AI and Advanced ICTs for Knowledge Societies : A Rights, Openness, Access, and Multi-Stakeholder Perspective,* vol. 14. UNESCO Series on Internet Freedom. Paris, UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000372132

———. 2021. *Évaluation des besoins en intelligence artificielle en Afrique.* Paris, Unesco. https://unesdoc.unesco.org/ark:/48223/pf0000375410_fre

UNESCO et i4 Policy. 2022. Multistakeholder AI development : 10 building blocks for inclusive policy design. United Nations Educational, Scientific and Cultural Organization (UNESCO) et Innovation for Policy Foundation (i4Policy). https://unesdoc.unesco.org/ark:/48223/pf0000382570

United Nations General Assembly. 2015. *Resolution Adopted by the General Assembly on 25 September 2015*. A/RES/70/1. United Nations. https://documents-dds-ny.un.org/doc/UNDOC/GEN/N15/291/89/PDF/N1529189.pdf

Van der Spuy, A. 2017. *What If We All Governed the Internet ? Advancing Multistakeholder Participation in Internet Governance*. UNESCO Series on Internet Freedom 11. Paris, UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000259717

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M. et Nerini, F. F. 2020. The Role of Artificial Intelligence in Achieving the Sustainable Development Goals. *Nature Communications,* vol. 11, nº 1,article nº 233. https://doi.org/10.1038/s41467-019-14108-y

WFD (Westminster Foundation for Democracy). 2021. *An Introduction to Deliberative Democracy for Members of Parliament.* https://www.wfd.org/wp-content/uploads/2021/09/WFD_newDemocracy_An-introduction-to-deliberative-democracy-for-members-of-parliament_2021.pdf

Winfield, A. 2019. Ethical standards in robotics and AI. *Nature Electronics,* vol. 2, nº 2, pp. 46-48. https://doi.org/10.1038/s41928-019-0213-6

Young, M., Magassa, L. et Friedman, B. 2019. Toward inclusive tech policy design : A method for underrepresented voices to strengthen tech policy documents. *Ethics and Information Technology,* vol. 21, nº 2, pp. 89-103. https://doi.org/10.1007/s10676-019-09497-z

PROPRIÉTÉ ET GESTION DE L'INFORMATION SUR LE COMPORTEMENT D'APPRENTISSAGE EN IAED

SHITANSHU MISHRA

Ph. D., responsable des technologies de l'information à l'Institut Mahatma Gandhi d'éducation pour la paix et le développement durable de l'UNESCO.

DAN SHEFET

Avocat à la Cour d'appel de Paris, France.

ANANTHA KUMAR DURAIAPPAH

Ph. D., directeur à l'Institut Mahatma Gandhi d'éducation pour la paix et le développement durable de l'UNESCO.

- **ODD 4** - Éducation de qualité
- **ODD 5** - Égalité entre les sexes
- **ODD 9** - Industrie, innovation et infrastructure
- **ODD 10** - Inégalités réduites
- **ODD 11** - Villes et communautés durables
- **ODD 16** - Paix, justice et institutions efficaces
- **ODD 17** - Partenariats pour la réalisation des objectifs

PROPRIÉTÉ ET GESTION DE L'INFORMATION SUR LE COMPORTEMENT D'APPRENTISSAGE EN IAED

RÉSUMÉ

L'intelligence artificielle en éducation (IAED) concerne les outils, les techniques et les méthodologies fondés sur l'intelligence artificielle (IA) que nous utilisons pour automatiser des processus que doivent mener des ordinateurs dans un environnement d'apprentissage amélioré par la technologie. Les « données sur les apprenants et apprenantes » sont devenues une commodité indispensable dans le développement de systèmes d'IAED. Contrôler la chaîne d'approvisionnement de cette commodité consiste en fait à contrôler l'IAED. Or, il se peut très bien que la vraie question soit de savoir si l'IA pour l'éducation est un patrimoine commun de l'humanité. Si c'est le cas, la commodité indispensable que sont les données sur les apprenants et apprenantes sera donc invariablement un patrimoine commun de l'humanité. Il est nécessaire de constituer un commun, soit le cadre ou le mécanisme de gouvernance, qui gère cet héritage commun. Ce chapitre présente des informations sur la manière dont le développement des systèmes IAED dépend des données, explique pourquoi les données sur l'éducation devraient être considérées comme un patrimoine commun de l'humanité et propose la nécessité et la structure d'un commun pour gérer ce patrimoine.

INTRODUCTION

Le 10 novembre 2021, l'UNESCO a publié son rapport intitulé « Les futurs de l'éducation » dans le cadre de la 41e session de la Conférence générale de l'UNESCO. Ce rapport plaide en faveur d'un nouveau contrat social pour l'éducation fondé sur les principes de la non-discrimination, de la justice sociale, du respect de toutes les vies, de la dignité humaine et de la diversité culturelle. L'appel à un contrat social sous-entend que les citoyens et citoyennes abandonnent certaines de leurs libertés naturelles à l'État ou même à une entité internationale ou intergouvernementale en échange de services ou de biens précis convenus dans

le contrat (Castiglen, 2015). Dans ce cas, le service rendu consisterait à fournir, de manière équitable, une éducation de grande qualité à tous les citoyens et à toutes les citoyennes sous forme d'un bien de « communs ».

Or, l'éducation sera fondamentalement différente à ce qu'elle a été au cours des 300 dernières années, lorsque les systèmes d'éducation actuels ont émergé pour répondre aux besoins de la révolution industrielle et de la croissance économique. Alors que nous passons à un monde numérique, que nous appelons maintenant plus couramment le métavers, les systèmes d'éducation futurs seront principalement numériques. L'intelligence artificielle (IA) a le potentiel d'être « l'agent » principal pour offrir des conseils aux apprenants et apprenantes dans le but d'améliorer leurs apprentissages et, par conséquent, leur potentialité. Fait intéressant, la discussion actuelle sur l'IA s'accompagne de discussions sur l'éthique en IA. « L'éthique en IAED » (Holmes *et al*., 2021) traite généralement des risques liés au suivi des étudiants et étudiantes, de la confidentialité des données, du consentement éclairé, de l'interprétation des données, de la propriété des données, de l'accès aux données, de la responsabilité, etc. Cependant, dans ce chapitre, nous nous concentrerons sur la propriété et la gestion des données relatives à l'éducation.

Si ce nouveau contrat social en matière d'éducation dans le métavers est pour assurer la prestation équitable d'une éducation de qualité, des paramètres clés doivent être établis. Le processus d'acquisition des informations sur le comportement d'apprentissage des apprenants et apprenantes, la gestion équitable, efficace et efficiente de ces données et le maintien des protocoles de confidentialité les plus élevés pour ces données sont parmi les paramètres clés que nous discuterons en détail dans ce chapitre.

Cette information sur le comportement d'apprentissage se distingue des ressources de connaissances qu'utilisent les apprenants et apprenantes. Par exemple, un module de mathématiques (une ressource de connaissances) utilisé dans des classes fait partie de cette dernière catégorie de ressources qui peuvent être privées ou faire partie d'une base de données de ressources en libre accès. Toutefois, l'expérience d'apprentissage de l'apprenante alors qu'elle progresse dans ce module est ce que nous appelons l'information sur le comportement d'apprentissage (ICA). Dans ce chapitre, l'ICA signifie plusieurs types de données sur les apprenants et les apprenantes recueillies de différentes modalités, comme le journal du système, les données d'oculométrie, les données physiologiques, les textes, les images, les vidéos, etc. Ce type d'information peut être recueilli en suivant les performances des étudiants et étudiantes, leurs stratégies, leurs séquences d'attention, leurs idées fausses, le temps qu'ils et elles consacrent à une question, le nombre de fois qu'ils et elles retournent à la section précédente du module, le nombre de fois qu'ils et elles tentent de terminer une section ou de répondre à un test et, plus récemment, les émotions qu'ils et elles éprouvent.

Actuellement, l'entité propriétaire de la plateforme d'apprentissage sur laquelle est offert le module est aussi propriétaire de l'ICA. Cette ICA est le bassin « génétique » dans lequel les algorithmes d'IA associés à la plateforme d'apprentissage puisent pour aider les apprenants et apprenantes à avoir une expérience d'apprentissage plus efficace et efficiente. Ce bassin génétique comprend les données d'ICA d'autres apprenants et apprenantes qui ont vécu des expériences d'apprentissage semblables. Comme c'est le cas de tous les algorithmes d'IA, plus le nombre de données utilisées pour son entraînement est grand, plus les interventions d'apprentissage proposées par l'IA seront efficaces et efficientes.

La question de l'équité se pose lorsque la qualité du bassin de données qu'utilisent les entités qui appartiennent et qui gèrent les données diffère. Ceci sous-entend que la qualité de l'apprentissage des apprenants et apprenantes variera en fonction de l'entité qui possède le meilleur bassin de données, ce qui, à son tour, suggère que les entités possédant de meilleurs bassins de données peuvent offrir une meilleure expérience d'apprentissage, mais à un coût qui pourrait exclure certains groupes d'apprenants et d'apprenantes, particulièrement ceux et celles issus de groupes marginalisés ou défavorisés. Le prix détermine alors qui peut ou ne peut pas avoir accès aux avantages des données regroupées qui, selon notre définition susmentionnée, font partie d'un patrimoine commun.

Le défi émerge donc de la question du partage de l'ICA entre les diverses plateformes d'apprentissage et de la recherche d'équité. Comme l'a souligné Borgman, en plus des quatre arguments courants en faveur d'un partage des données, à savoir la reproduction de la recherche, la mise des biens publics à la disposition du public, l'utilisation des investissements dans la recherche et la facilitation de la recherche et l'innovation, l'utilisation de l'IA en éducation en suggère un cinquième : l'amélioration de l'apprentissage par l'offre d'une expérience d'apprentissage personnalisée (Borgman, 2015 ; Margolis *et al.*, 2014 ; Wilkinson *et al.*, 2016). L'option d'offrir plusieurs trajectoires d'apprentissage plutôt qu'une seule est au cœur du partage des données et de l'utilisation de l'IA en éducation.

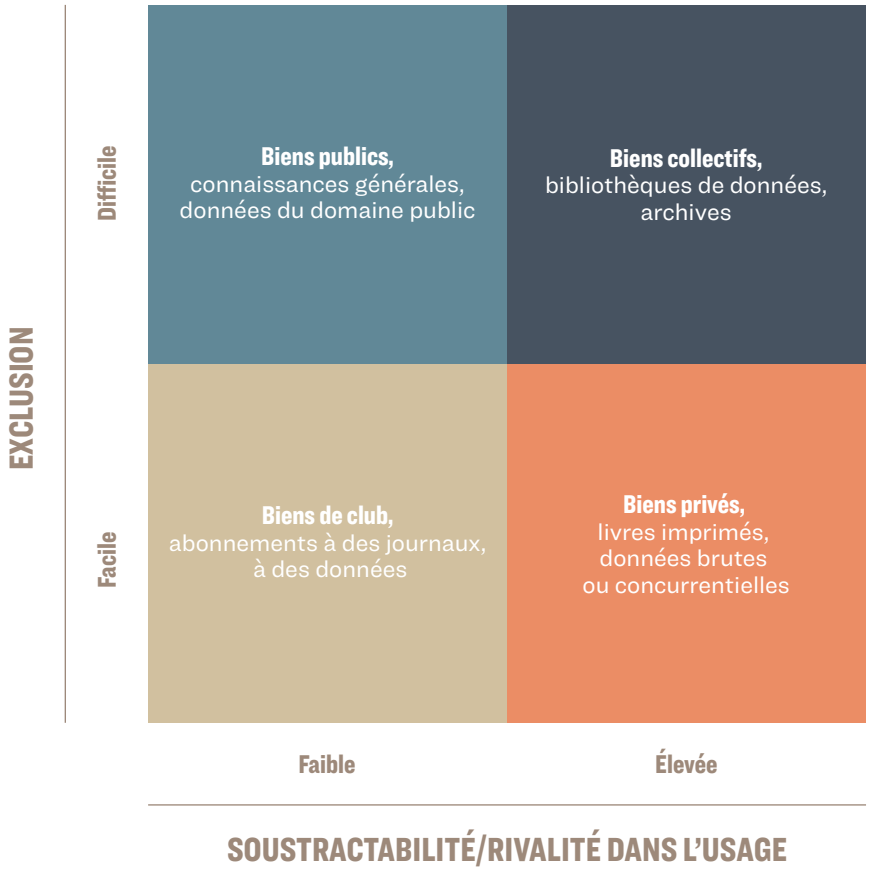
En principe, chaque personne appartient son ICA. Ces données deviennent utiles seulement lorsqu'elles sont utilisées d'une manière collective et regroupées avec des données semblables recueillies au fil du temps auprès du plus grand nombre de personnes possible. Un grand bassin de données est un prérequis pour la création de modèles d'IA précis et, par conséquent, de systèmes d'IAED efficaces. À elle seule, chaque donnée individuelle a peu de valeur. La personne propriétaire des données ne retire donc aucun avantage à moins que son information soit regroupée à celle de plusieurs autres personnes. Ce n'est qu'à ce moment que des gains en matière d'apprentissage sont réalisés.

Par conséquent, ceci nous mène à la question primordiale à savoir de quelle manière réunir toutes les informations générées par les apprenants et apprenantes distribués dans l'espace et dans le temps afin que les algorithmes d'IA puissent les utiliser au profit des apprenants et apprenantes. Comme susmentionné, l'ICA générée par chaque apprenant et chaque apprenante est, en principe, la propriété de cet apprenant ou de cette apprenante. Cependant, l'ICA collective de tous les apprenants et de toutes les apprenantes sur une plateforme d'apprentissage en particulier appartient actuellement à l'entité propriétaire de ladite plateforme. Cette même entité pourrait être propriétaire de l'algorithme d'IA utilisé pour soutenir les apprenants et apprenantes ou elle pourrait avoir acquis des licences d'entreprises d'IA pour utiliser des algorithmes d'IA particuliers pour bâtir une plateforme d'apprentissage IAED rentable.

En plus des facteurs comme la qualité, la fiabilité et l'asymétrie (qui entraîne des préjugés) des données dans la base de données, le volume de la base de données détermine, comme susmentionné, l'efficacité des plateformes d'apprentissage fondées sur l'IAED. Plus la base de données d'ICA est grande, meilleure sera la qualité des interventions d'apprentissage qu'offriront les algorithmes d'IA respectifs. En plus des données d'ICA, la qualité des interventions d'apprentissage dépendra aussi de la qualité des algorithmes d'IA. Nous posons maintenant les deux questions de ce chapitre. D'abord, le bassin collectif d'ICA devrait-il appartenir aux entités offrant les plateformes d'apprentissage en tant que bien privé, bien de club offert par quelques entités, bien public ou bien collectif ? Ensuite, les algorithmes utilisés pour fournir les interventions d'apprentissage devraient-ils eux aussi être un patrimoine commun de l'humanité ou une ressource privée, publique, de club, ou collective ?

| FIGURE 1 |

Types de biens (figure adaptée de Ostrom et Ostrom, 1977 ; Hess et Ostrom, 2007).



Selon la classification Hess-Ostrom illustrée dans la figure 1, le système actuel d'attribution de la propriété des plateformes d'apprentissage classe essentiellement l'ICA en tant que bien de club offert aux utilisateurs et utilisatrices qui s'abonnent à la plateforme d'apprentissage et qui y sont membres. Cette option produit des résultats sous-optimaux, puisque le bassin de données se limite uniquement aux personnes abonnées à la plateforme, ce qui mène habituellement à des monopoles dans le secteur et produit des résultats non équitables, comme cela est souligné plus haut dans cet article. Les monopoles s'entendent essentiellement pour fixer les prix afin de s'emparer du marché et, par le fait même, empêchent d'autres d'offrir des options plus concurrentielles aux consommateurs et consommatrices qui, dans ce cas, sont les apprenants et apprenantes. Si l'éducation doit être un contrat social, le fait de la reléguer aux monopoles peut ne pas servir la société comme prévu. Le degré d'efficacité devient une fois de plus un enjeu si l'ICA est considérée comme étant un bien privé, puisque l'établissement de la propriété de l'information générée entre les apprenants et apprenantes et les plateformes d'apprentissage devient

une tâche non triviale. Pour résoudre le problème d'efficacité, nous avons donc l'option de traiter l'ICA comme une ressource publique ou collective. Dans le premier cas, bien que les gouvernements puissent offrir le service, la question associée à leur capacité de fournir seuls d'une manière efficace un bien dans le cadre de règlements bien intentionnés se pose, particulièrement en raison des coûts élevés de transaction et de la probabilité élevée de capture réglementaire, et ce, surtout lorsque les principaux acteurs et actrices du secteur sont des entités privées à but lucratif (Beales *et al.,* 2017).

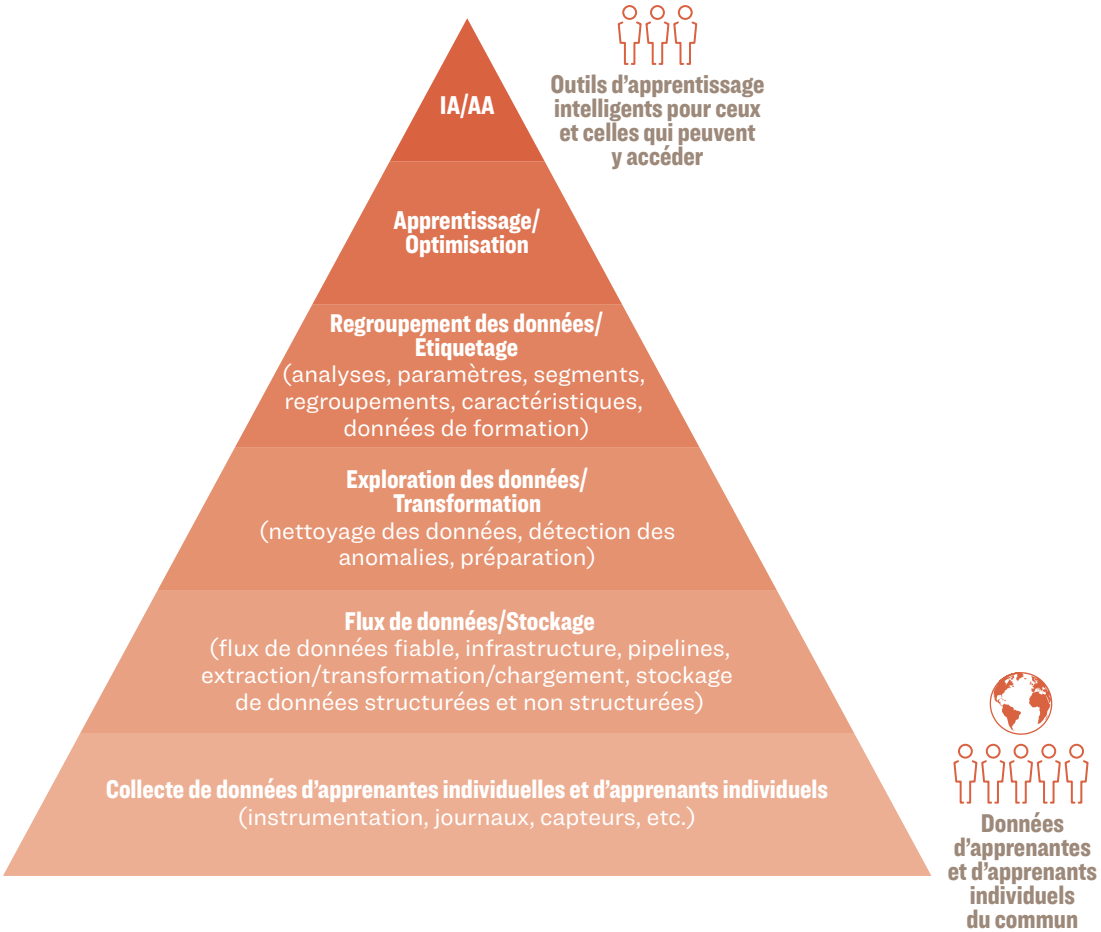
Il est donc nécessaire d'établir une structure de gouvernance à laquelle participent les apprenants et apprenantes, les gouvernements ainsi que les entités derrière les plateformes d'apprentissage qui peut dépasser les frontières nationales et faire en sorte que l'ICA soit regroupée à l'échelle mondiale au profit de chacun des apprenants et de chacune des apprenantes partout dans le monde. Une structure de gouvernance fondée sur un commun qui permet à chaque partie qui est une propriétaire et utilisatrice potentielle ou un propriétaire et utilisateur potentiel du service se prête bien à cette fin, considérant la nature regroupée et collective de l'ICA. La section suivante fournit une analyse approfondie de la nature de l'ICA, de la manière dont elle est consommée dans une plateforme d'apprentissage fondée sur l'IA et de la raison pour laquelle il est nécessaire de la traiter comme une ressource collective.

DONNÉES D'ICA ET IA EN ÉDUCATION

L'intelligence artificielle en éducation (IAED) concerne les outils, les techniques et les méthodologies fondés sur l'IA utilisés pour automatiser les processus que doivent mener des ordinateurs dans un environnement d'apprentissage amélioré par la technologie. La figure 2 illustre la dépendance de tout système d'IAED sur les données recueillies des apprenants individuels et des apprenantes individuelles (données d'ICA).

| FIGURE 2 |

La hiérarchie des besoins de la science des données (figure adaptée de Rogati, 2017).



Dans la figure 2, nous constatons que les données sont au bas de la hiérarchie des besoins de la science des données (Rogati, 2017). Pour discuter de la manière dont les données d'ICA de l'IAED sont utilisées et devraient être utilisées, il devient impératif de décomposer les systèmes d'IA et de comprendre comment ils sont construits. Cela fournirait un portrait plus clair de la contribution des données d'ICA au processus et des défis que doivent relever les développeurs et développeuses de systèmes d'IAED, tout en utilisant les données pour bâtir des plateformes d'apprentissage fondées sur l'IAED. Les trois principaux défis pour le développement et la mise en œuvre d'un système d'IA sont le manque de données, le manque d'infrastructure et le manque de talents possédant les compétences requises pour assurer un développement de l'IA efficace et réussi (Ernst *et al.*, 2019). Il en va de même pour les systèmes d'IAED. De ces derniers, la disponibilité des données semble être un défi perpétuel. La disponibilité d'un grand volume de données diversifiées est nécessaire pour veiller à ce que le système d'IA (ou d'IAED) soit efficace et produise des résultats justes et équitables.

Tout processus de développement de système d'IA commence généralement par une compréhension et la définition du problème que doit résoudre l'IA. Ensuite, il est nécessaire d'assurer la disponibilité des données, ce qui requiert l'administration de processus de collecte de données approfondies. Après s'être assurés que la disponibilité des données est adéquate, les développeurs et développeuses d'IA doivent veiller à ce que les données soient bien stockées et organisées afin qu'elles puissent être facilement accessibles dans d'autres processus. Cette étape est suivie de l'exploration et du prétraitement des données. L'exploration des données est nécessaire pour vérifier si les données représentent correctement les événements concernés, ce qui permet d'évaluer les hypothèses et la compréhension des données d'IA. Le processus de nettoyage des données est essentiel, puisque des données non nettoyées peuvent mener à un entraînement inexact, duquel peuvent découler de mauvaises décisions et conclusions et de piètres analyses, particulièrement si une grande quantité de mégadonnées est considérée. Le nettoyage des données consiste à retirer ou à mettre à jour l'information incomplète, incorrecte, redondante ou sans importance. Cela consiste aussi à éliminer toute asymétrie dans les données et à considérer la normalisation des données (Jeni *et al.*, 2013) afin qu'elles apparaissent de manière semblable dans tous les registres et champs. Le nettoyage des données maximise l'exactitude de l'ensemble de données, sans nécessairement manipuler les données disponibles.

Le nettoyage des données est suivi par le regroupement des données et l'ingénierie des caractéristiques (Zheng et Casari, 2018), ce qui comprend l'extraction, la génération, le regroupement et la réduction des caractéristiques des données afin de mieux représenter le problème sous-jacent (p. ex., le problème d'être en mesure de prédire si, à un moment donné, un apprenant ou une apprenante souhaite ou non faire une lecture plus approfondie sur une plateforme d'apprentissage) pour un apprentissage automatique plus approfondi. Après le nettoyage et le regroupement de données, les développeurs et développeuses d'IA conçoivent ou appliquent des algorithmes d'apprentissage automatique pour entraîner les modèles logiciels et les systèmes qui peuvent soutenir l'apprentissage-l'enseignement de toute matière. Après avoir conçu et testé un système d'IAED, il est alors nécessaire de le mettre en œuvre afin qu'il soit fonctionnel et accessible aux utilisateurs finaux et aux utilisatrices finales (les apprenants et apprenantes) et d'effectuer une maintenance et des mises à niveau régulières.

CONTRIBUTEURS ET CONTRIBUTRICES À UN SYSTÈME D'IAED

Pendant la conception d'un système d'IAED, les développeurs et développeuses doivent choisir l'approche d'IA à mettre en œuvre, ce qui consiste à choisir entre une IA fondée sur les données et une IA fondée sur un modèle. Ce choix détermine la relation entre les données et l'intelligence d'un système d'IA une fois entraîné. L'approche fondée sur les données se concentre sur la construction d'un système capable de trouver quelle est la bonne réponse après avoir « vu » un grand nombre d'exemples de paires question-réponse et après avoir été « entraîné » à trouver la bonne réponse. Ce type d'IA est un grand

consommateur de données. Certains systèmes d'IA sont assez puissants pour être en mesure de faire des généralisations à partir d'un nombre limité de données d'entraînement et de trouver par eux-mêmes des ensembles de caractéristiques exploitables et des critères de décision, mais plusieurs approches d'apprentissage automatique (dont l'apprentissage profond) requièrent un nombre très important de données pour produire des résultats significatifs, tandis que certains exigent leurs propres types d'experts et d'expertes pour les mettre en œuvre.

Contrairement à l'IA fondée sur les données, qui dépend presque entièrement sur les données (soit leur collecte et leur analyse) pour alimenter sa prise de décisions, l'IA fondée sur un modèle saisit les connaissances et permet la prise de décisions par l'entremise d'une représentation et de règles claires qui sont alimentées par les connaissances et la science du domaine de problème pour lequel le système d'IA est conçu. Cependant, les modèles (la science de l'apprentissage dans le cas de systèmes d'IAED) évoluent sans cesse, souvent grâce à des recherches menées sur des données empiriques recueillies dans des contextes d'apprentissage qui proviennent, une fois de plus, d'un grand nombre d'apprenants individuels et d'apprenantes individuelles. Il est important de noter que le processus de renforcement des connaissances englobe les efforts collaboratifs à l'échelle de l'ensemble de l'entreprise scientifique, comme le renforcement des connaissances au sein d'une communauté (Hong et Scardamalia, 2014 ; Scardamalia et Bereiter, 2006) qui s'appuie sur la contribution d'idées provenant de plusieurs disciplines pour résoudre un problème complexe.

Pour ce qui est du processus de développement d'un système d'IAED décrit plus haut, il est évident qu'un système d'IAED ne peut être construit sans données d'ICA qui possèdent les propriétés d'un bien collectif. Par ailleurs, plusieurs systèmes d'IAED (particulièrement les systèmes fondés sur un modèle) dépendent de modèles et de connaissances du ou des domaines, ce qui peut aussi être largement considéré comme un bien collectif. Or, nous constatons aussi qu'un système d'IAED ne peut être construit sans les contributions des tâches de collecte de données, de gestion des données, d'ingénierie des données, de développement d'algorithme, de mise en œuvre et de maintenance. Par conséquent, en plus de la propriété du bassin de données d'ICA, il serait crucial de discuter de la propriété des algorithmes et des systèmes d'IAED ainsi que des contributions des entités publiques et privées à leur développement. Cependant, dans ce chapitre, nous circonscrivons la discussion à la propriété et à la gestion des données d'ICA.

GÉRER L'ICA EN TANT QUE PATRIMOINE COMMUN : PRINCIPES CLÉS ET CADRE DIRECTEUR

La plupart des travaux de recherche sur les communs de connaissances se limitent aux secteurs de la médecine et de la santé publique (Chatterjee *et al.*, 2022). Il y a peu de recherches sur les communs de connaissances dans le domaine de l'éducation. La littérature et l'expérience en matière d'ICA et de gouvernance de ces données se font encore plus rares. La trajectoire actuelle suggère l'émergence d'un marché faussé face à des monopoles et une très faible appropriation de ces données par les apprenants eux-mêmes et apprenantes elles-mêmes et de l'utilisation de ces données à leur profit.

En définissant le terme « commun de connaissances », Gyuris (2014) soutient que « les connaissances en tant que ressource partagée » nécessitent que l'information soit accessible et doive permettre aux récepteurs potentiels et réceptrices potentielles d'internaliser l'information en connaissances. Par conséquent, les connaissances ne peuvent être une ressource partagée sans qu'un ensemble complexe d'établissements et de pratiques offrent aux récepteurs potentiels et réceptrices potentielles la possibilité d'acquérir les compétences et la préparation nécessaires. De même, pour traiter les données d'ICA en tant que patrimoine commun de l'humanité, nous devons mettre en place une structure de gouvernance

à laquelle participent les apprenants et apprenantes et les entités auxquelles appartiennent les plateformes d'apprentissage pouvant transcender les frontières nationales et faire en sorte que l'ICA est regroupée à l'échelle mondiale au profit de tous les apprenants et de toutes les apprenantes du monde entier. Dans les discussions subséquentes, nous proposons des orientations, des directives et des questions qui peuvent être utiles dans l'établissement d'une telle structure de gouvernance.

Principes pour gérer un commun

À la lumière des travaux d'Ostrom et de ces collègues (Hess et Ostrom, 2007 ; Ostrom, 2005), un régime international gérant les données d'ICA en tant que patrimoine commun et les algorithmes d'IA serait nécessaire pour satisfaire les huit principes de base suivants et veiller à ce que les trois « E », c'est-à-dire l'efficacité, l'efficience et l'équité, soient atteints (Hess et Ostrom, 2007 ; Ostrom, 2005) :

1. Définir des limites de groupe claires ;
2. Corréler les règles gouvernant l'utilisation des biens collectifs aux conditions et aux besoins locaux ;
3. Veiller à ce que ceux et celles qui sont concernés par les règles puissent participer à leur modification ;
4. S'assurer que des autorités externes veillent au respect des droits d'établissement de règles des membres de la communauté ;
5. Créer un système pour surveiller le comportement des membres qui est mis en œuvre par les membres de la communauté ;
6. Imposer des sanctions progressives à ceux et celles qui violent les règles ;
7. Fournir un moyen accessible et peu coûteux pour résoudre les différends ;
8. Renforcer la responsabilité de la gouvernance de la ressource collective en niveaux imbriqués, du niveau le plus bas jusqu'à l'ensemble du système interconnecté.

La force derrière l'utilisation du cadre d'Ostrom réside dans le fait qu'il reconnaît l'échelle, une multitude d'acteurs et d'actrices, un processus participatif parmi les diverses parties prenantes et un système de sanctions progressives pour assurer la reddition de compte et la responsabilité dans le but de minimiser le mauvais usage des données regroupées. Il est important de souligner qu'un commun ne réfère pas à la ressource, mais bien à la gouvernance d'une ressource et, en particulier, d'une ressource partagée. Dans ce cas, le partage de données regroupées offre la meilleure solution pour optimiser l'expérience d'apprentissage des apprenantes et apprenantes.

En plus des principes du cadre d'Ostrom, Frischmann *et al.* (2014) proposent un guide utile pour explorer et établir la structure de gouvernance d'un commun pour ce qu'ils appellent un commun de connaissances, ce qui, dans notre cas, serait le commun des ICA. Leur commun de connaissances s'appuie sur le cadre de développement et d'analyse institutionnels (Ostrom, 2005 ; Hess et Ostrom, 2007) et ils proposent les étapes suivantes. Par chaque étape, nous présentons une liste de questions ou de considérations pertinentes pouvant être cruciales dans l'établissement d'une structure de gouvernance pour le commun des ICA.

La définition d'ICA

a. Quel est le contexte (juridique, culturel, économique) du commun des ICA ? De quelle manière sont actuellement recueillies, regroupées et utilisées les ICA ? Y a-t-il des différences entre la façon dont sont recueillies, regroupées et utilisées les ICA dans les différents pays en raison de différences socio-économiques ou culturelles ?

b. Quel est l'état de propriété actuel des ICA (brevetées, sous droit d'auteur, ouvertes ou autre) ?

Les caractéristiques de la ressource

a. Quelle est la nature de la ressource qui sera regroupée et de quelle manière est-elle créée ou obtenue ?

b. Quelles sont les caractéristiques des ICA ? Sont-elles rivales ou non rivales, excluables ou non excluables, tangibles ou intangibles ? Y a-t-il une infrastructure partagée ?

c. Quelles sont les technologies et les compétences requises pour créer, obtenir, maintenir et utiliser les ICA ?

La communauté

a. Qui sont les membres et quels sont leurs rôles ?

b. Quels sont le degré et la nature de l'ouverture de chaque type de membres de la communauté et du grand public ?

Buts et objectifs

a. Quels sont les buts et les objectifs du commun des ICA et de ses membres, y compris les obstacles ou les dilemmes qui devront être surmontés ?

b. Quel est l'historique de l'utilisation des ICA et y a-t-il des structures de gouvernance particulières surveillant le regroupement et l'utilisation des ICA ?

Résultats

a. Quelles sont les retombées pour les membres et autres (p. ex., innovation et résultats créatifs, production, partage et diffusion à un public plus large, interactions sociales découlant du commun) ?

b. Quels sont les coûts et les risques associés au commun, dont ceux qui sont externes et négatifs ?

Governance

La conceptualisation de la gouvernance du commun exige que nous répondions à plusieurs questions à plusieurs facettes. Ceci comprend le fait de réfléchir aux domaines d'action pertinents et à la façon dont ils sont liés aux buts et aux objectifs du commun ainsi qu'aux relations entre les divers types de participants et participantes et avec le grand public. Un autre aspect important est le mécanisme de gouvernance, ce qui inclut l'établissement de règles d'adhésion, les normes et les exigences de la contribution ou de l'utilisation de l'ICA, les mécanismes de résolution de conflits, les sanctions pour la violation d'une règle, etc. Par exemple, la mise en place d'ententes multilatérales et bilatérales entre les pays où les données sont recueillies, et stockées si cela est requis en vertu des lois locales en matière de confidentialité des données, et l'entité compétente ou l'établissement où le bassin de données sera situé et le pays où il sera utilisé par l'utilisateur final ou l'utilisatrice finale (école, université, etc.).

Les décideurs et décideuses et le processus de leur sélection sont des aspects essentiels d'une structure de gouvernance. Par ailleurs, les organisations et les infrastructures technologiques qui encadrent et gouvernent la prise de décisions sont aussi essentielles à la viabilité du commun. Par exemple, la manière dont gère une organisation le bassin sera, dans toute la mesure du possible, fondée sur un algorithme ou sur l'IA (qui est fondée sur des règles), ce qui, par conséquent, réduira les coûts et favorisera un traitement rapide. Sur le plan opérationnel, le principal élément de coût sera la maintenance.

Les règles qui permettront l'usage de l'ICA sont un autre aspect important qui doit être considéré. Toute personne recevant des données, qu'elle soit un abonné ou une abonnée ou un client ponctuel ou une cliente ponctuelle, par exemple, doit signer un contrat limitant l'utilisation des données à des fins éducatives. Une telle utilisation doit être gratuite pour l'apprenant ou l'apprenante et l'établissement d'enseignement et une interdiction expresse de tout usage commercial direct ou indirect doit être incluse. Une clause de pénalité dissuasive devrait aussi être comprise, en plus de l'arbitrage.

La structure de gouvernance devrait aussi comprendre des règles décrivant la manière grâce à laquelle les non-membres interagissent avec le commun. Quelles organisations gouverneront ces interactions ? Un exemple serait d'adopter le modèle de l'ICANN (Christie, 2002), soit une organisation n'appartenant pas à un groupe, à un gouvernement ou à une entreprise en particulier. Le financement des services, à faibles coûts, pourrait se faire selon un modèle d'abonnement ou sur une base transactionnelle, en exigeant, par exemple, des frais pour le transfert de données qui pourraient être calculés selon le volume des données.

Les normes informelles qui gouvernent le commun ainsi que les structures juridiques sont, entre autres, d'autres questions importantes devant être considérées en ce qui a trait à la structure de gouvernance. Par exemple, comment la propriété intellectuelle, les subventions, les contrats, les licences, les taxes et les lois antitrust s'appliquent-ils à cette structure ?

SITUATION ACTUELLE

En ce qui a trait au partage des données, l'initiative de la Commission européenne est actuellement la plus importante. Le 25 novembre 2020, la Commission européenne a soumis une proposition d'acte sur la gouvernance des données (Commission européenne, 2020) et ce dernier est considéré comme étant le projet de règlement le plus ambitieux pour ce qui est de sa portée et des obligations relatives au partage des données. Il traite abondamment du principe de partage des données et du cadre procédural et institutionnel favorisant un partage efficace. L'idée fondamentale est de créer des bassins de partage

de données et des intermédiaires de partage de données réglementés qui veilleront à ce que les droits des Européens et Européennes concernés soient respectés dans le cadre de ces opérations. À moins que de fortes raisons ne justifient le contraire, les données seront partagées de manière anonyme.

Nonobstant l'importance du projet de réglementation, ce dernier ne traite pas du partage de données provenant de sources privées. Il se limite plutôt au secteur public et est largement inspiré par l'idée d'un gouvernement ouvert. L'omission des sources privées est évidemment le principal inconvénient, mais elle s'explique par les conséquences économiques majeures qu'entraînerait une obligation de partage des données, ce qui ne pourrait pas être mis en œuvre sans offrir de compensations.

La proposition d'un acte de gouvernance des données peut, à cet égard, être considérée comme une première étape. Aux États-Unis, il n'existe pas d'initiative fédérale semblable et il en est de même en Inde où aucune loi sur le partage de données n'existe. La loi sur la protection des données personnelles, maintenant la loi sur la protection des données (Lok Sabha, n.d.) qui sera probablement adoptée prochainement en Inde n'inclut pas de disposition semblable à celle du projet de l'Union européenne.

CONCLUSION

> La plupart des organisations œuvrant dans le domaine de l'IA et de l'éthique en IA n'ont pas encore entamé officiellement une discussion sur la gestion de l'ICA. L'attention a en grande partie plutôt porté sur l'établissement d'un code de déontologie relativement au développement d'algorithmes et à l'utilisation des données par les fournisseurs de données respectifs. Or, les enjeux liés à la confidentialité de l'ICA, à la propriété de cette information et à son utilisation d'une manière collective n'ont pas encore été considérés. Le système actuel voulant que chaque plateforme d'apprentissage offre ces services donne des résultats qui ne sont pas optimaux et a le potentiel de favoriser des résultats inéquitables pour les apprenants et apprenantes. Une approche de commun pourrait, quant à elle, offrir une solution pour satisfaire les trois « E » que sont l'efficacité, l'efficience et l'équité au sein du secteur de l'éducation.

RÉFÉRENCES

Beales, H., Brito, J., Davis, J. K. Jr., DeMuth, C., Devine, D., Dudley, S., Mannix, B., et McGinnis, J. O. 2017. *Government Regulation : The Good, The Bad, & The Ugly*. The Regulatory Transparency Project of the Federalist Society. https://regproject.org/wp-content/uploads/RTP-Regulatory-Process-Working-Group-Paper.pdf

Borgman, C. L. 2015. *Big Data, Little Data, No Data : Scholarship in the Networked World*. Cambridge, MA : The MIT Press.

Castiglen, D. 2015. Introduction to the logic of social cooperation for mutual advantage : The democratic contract. *The Political Studies Review*. Vol. 13, No. 2, pp. 161-175.

Friend, C. n.d. Social Contract Theory. Internet encyclopedia of Philosophy. https://iep.utm.edu/soc-cont/

Chatterjee, A., Kuiper, M., et Swierstra, T. 2022. *Dealing with different conceptions of pollution in the Gene Regulation Knowledge Commons*. Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms, Vol. 1865, No. 1. https://doi.org/10.1016/j.bbagrm.2021.194779

Christie, A. 2002. The ICANN Domain-Name Dispute Resolution System as a model for resolving other intellectual property disputes on the internet. *Journal of World Intellectual Property*, Vol. 5, No. 105, pp. 105-117.

Ernst, E., Merola, R. et Samaan, D. 2019. Economics of artificial intelligence : Implications for the future of work. *IZA Journal of Labor Policy*, Vol. 9, No. 1, pp. 1-35.

European Commission. 2020. Proposal for a Regulation of the European Parliament and of the Council on European Data Governance (*Data Governance Act*). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0767&from=EN

Frischmann, B., M., Madison, M. J. et Strandburg, K. J. (eds.). 2014. *Governing Knowledge Commons.* Oxford : Oxford University Press.

Gyuris, F. 2014. Basic education in communist Hungary. A commons approach. International *Journal of the Commons*, Vol. 8, No. 2, pp. 531-553.

Hess, C. et Ostrom, E. 2005. A framework for analyzing the knowledge commons. In : *Understanding Knowledge as a Commons : From Theory to Practice*, eds C. Hess et E, Ostrom, 2007. Cambridge, MA : MIT Press.

Hess, C., et Ostrom, E. (eds.). 2007. *Understanding Knowledge as a Commons : From Theory to Practice*. Cambridge, MA : MIT Press.

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B., (...) et Koedinger, K. R. 2021. Ethics of AI in education : Towards a community-wide framework. *International Journal of Artificial Intelligence in Education*, pp. 1-23. https://doi.org/10.1007/s40593-021-00239-1

Hong, H. Y., et Scardamalia, M. 2014. Community knowledge assessment in a knowledge building environment. *Computers & Education*, No. 71, pp. 279-288.

Jeni, L. A., Cohn, J. F., et De La Torre, F. 2013. Facing imbalanced data – Recommendations for the use of performance metrics. *2013 Humaine association conference on affective computing and intelligent interaction,* IEEE, pp. 245-251.

Lok Sabha. n.d. The Personal Data Protection Bill, 2019. Bill No. 373 of 2019. http://164.100.47.4/BillsTexts/LSBillTexts/Asintroduced/373_2019_LS_Eng.pdf

Margolis, R., Derr, L., Dunn, M., Huerta, M., Larkin, J., Sheehan, J. et al. 2014. The National Institutes of Health's Big Data to Knowledge (BD2K) initiative : Capitalizing on biomedical big data. *Journal of the American Medical Informatics Association.* Vol. 21, No. 6, pp. 957–958. DOI : https://doi.org/10.1136/amiajnl-2014-002974

Ostrom, E. 2005. *Understanding Institutional Diversity Princeton*. N.J. : Princeton University Press.

Ostrom, V. et Ostrom, E. 1977. Public Goods and Public Choices. In E. S. Savas (ed.), *Alternatives for Delivering Public Services : Toward Improved Performance*, pp. 7-49. Boulder, CO : Westview Press.

Rogati, M. 2017. The AI hierarchy of needs. Hacker Noon. https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007

Scardamalia, M. et Bereiter, C. 2006. Knowledge building : Theory, pedagogy, and technology. In K. Sawyer (ed.), *Cambridge Handbook of the Learning Sciences,* pp. 97-118. New York : Cambridge University Press.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., (...) et Mons, B. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, Vol. 3, No. 1, pp. 1-9.

Zheng, A., et Casari, A. 2018. *Feature Engineering for Machine Learning : Principles and Techniques for Data Scientists*. O'Reilly Media, Inc.

ARMES AUTONOMES ET HYPERTRUCAGES : LES RISQUES DE L'ACTUELLE MILITARISATION DE L'IA ET L'URGENT BESOIN DE RÉGLEMENTATION

BRANKA MARIJAN

Chercheuse principale à Project Ploughshares. Membre du conseil d'administration de l'Association canadienne des études sur la paix et les conflits (PACS-Can). Elle est titulaire d'un doctorat de la Balsillie School of International Affairs.

WANDA MUÑOZ

Consultante internationale en droits humains et désarmement. Elle est titulaire d'un Master en affaires internationales de l'Université Columbia et de Sciences Po Paris et est membre du Réseau de sécurité humaine en Amérique latine et dans les Caraïbes (SEHLAC), du Réseau de recherche féministe sur l'IA et du Partenariat mondial sur l'IA.

ODD 3 - Bonne santé et bien-être
ODD 5 - Égalité entre les sexes
ODD 9 - Industrie, innovation et infrastructure
ODD 10 - Inégalités réduites
ODD 11 - Villes et communautés durables
ODD 13 - Mesures relatives à la lutte contre les changements climatiques
ODD 15 - Vie terrestre
ODD 16 - Paix, justice et institutions efficaces
ODD 17 - Partenariats pour la réalisation des objectifs

ARMES AUTONOMES ET HYPERTRUCAGES : LES RISQUES DE L'ACTUELLE MILITARISATION DE L'IA ET L'URGENT BESOIN DE RÉGLEMENTATION

RÉSUMÉ

La technologie fondée sur l'intelligence artificielle (IA) pourrait bientôt servir à déclencher des guerres ou à promouvoir des crimes haineux. Plus de 130 systèmes militaires sont déjà en mesure de localiser des cibles de manière autonome. L'IA s'implante de plus en plus dans des armes semi-autonomes ainsi que dans des technologies sophistiquées d'hypertrucage qui pourraient mener à une escalade de conflits et à une instabilité à l'échelle mondiale.

Les dangers réels associés à la militarisation de telles technologies restent malgré tout exclus des discussions qui se tiennent à un niveau national et international relativement aux usages éthiques et responsables de l'IA. Quand le sujet est abordé dans des forums sur le contrôle des armes et le désarmement, certaines parties prenantes semblent plus intéressées à acquérir de telles armes qu'à encadrer ou bannir leur usage.

Par chance, il n'est pas trop tard pour s'assurer que l'IA est mise à profit pour la majorité de la population mondiale, et non pour servir l'intérêt d'oppresseurs et d'autocrates. Un nombre croissant de chercheurs et chercheuses, d'analystes stratégiques et de membres de la société civile sont soucieux d'élaborer et de promouvoir des mesures qui mèneront à une réglementation efficace, et ils sont aptes à le faire. Il importe particulièrement que les chercheurs et chercheuses du domaine de l'IA se fassent entendre pour assurer la mise au point de nouvelles technologies selon une certaine éthique.

Quelles sont les priorités ? 1) L'adoption d'un instrument juridiquement contraignant visant à interdire les armes dont l'usage nécessite une trop faible intervention humaine ainsi que les armes qui cibleraient des êtres humains, et réglementant toutes les autres armes autonomes ; 2) Des solutions techniques faisant en sorte que les hypertrucages et les contenus manipulés sont signalés ; 3) Une réglementation qui protège les droits humains et interdit les applications qui soutiennent la violence sexiste ou des crimes haineux.

INTRODUCTION

Des drones capables de cibler des personnes sans que des humains les commandent, des technologies d'identification numérique sophistiquées aux mains de groupes armés non étatiques ou violant les droits humains, des vidéos manipulées diffusant les déclarations que des dirigeants et dirigeantes politiques n'ont jamais faites…

Il y a quelques années encore, pareils scénarios auraient été, et ont effectivement été, considérés comme de la science-fiction ou de l'alarmisme. Pourtant, de telles utilisations apparemment dystopiques des nouvelles technologies se sont toutes produites sous une forme ou une autre, et elles sont susceptibles de se produire encore et de prendre de l'ampleur si le développement jusque-là incontrôlé ne fait pas l'objet de mesures de gouvernance judicieuses. En d'autres termes, l'intelligence artificielle (IA) se transforme en une arme, qu'emploient déjà tant des institutions militaires et des organes de sécurité que, de manière malveillante, des acteurs armés non étatiques. Les systèmes d'IA se répandent de plus en plus dans le secteur de la défense ainsi que dans la collecte et l'analyse de renseignements. En outre, les outils d'IA permettant de manipuler des images, des vidéos et du matériel audio se sont perfectionnés et deviennent accessibles, ce qui suscite des inquiétudes par rapport à l'érosion de la confiance du public et au risque qu'ils représentent quant à la stabilité internationale (UNIDIR, 2021). À l'heure d'écriture de ce chapitre, en 2021, il existe un décalage évident en matière de gouvernance pour faire face aux nombreux risques liés à la militarisation de l'IA. Ce décalage requiert l'attention urgente de la communauté de l'IA, de la société civile, des gouvernements et des organisations internationales.

La technologie a longtemps devancé la réglementation, et les applications basées sur l'IA ne font pas exception. Une IA militarisée semble trop souvent une menace lointaine. Le fait que de nombreuses avancées dans le domaine de l'IA aient lieu dans le secteur privé, au sein d'entreprises technologiques de premier plan à la pointe de la recherche et du développement dans ce domaine, peut également représenter un défi en matière de réglementation, puisque cette activité reste loin des regards et des investissements publics. De plus, la nature multifonctionnelle de la technologie masque parfois les incidences potentiellement malveillantes de ses applications. Il y a une concurrence mondiale grandissante entre les principales armées, ce qui place l'IA au centre de la capacité militaire à venir et entraîne par conséquent une augmentation du financement de la recherche et développement à des fins militaires (Keller, 2021). Malgré cette concurrence de plus en plus admise dans le domaine de l'IA militaire, les applications de l'IA liées aux forces armées et à la sécurité sont généralement exclues des discussions générales et des engagements concernant l'IA éthique et responsable. Ainsi, tant l'Organisation des Nations unies pour l'éducation, la science et la culture (UNESCO) que l'Organisation de coopération et de développement économiques (OCDE) et la Commission européenne excluent les applications militaires de leurs mandats an matière d'IA. Bien que certains forums internationaux et les Nations Unies y portent une attention croissante, les appels à une réglementation et à l'adoption de politiques efficaces ont d'abord été lancés par la société civile et des spécialistes techniques.

Dans ce contexte, et compte tenu des progrès continus accomplis en IA et dans des domaines connexes tels que la robotique, la militarisation de l'IA mérite une attention nettement accrue. En particulier, comme le mentionnent Burton et Soare (2019, p. 4), la militarisation de l'IA comporte principalement deux aspects : « (a) la manière dont l'IA s'incorpore ou peut s'incorporer dans les systèmes et les plateformes d'armes ; et (b) la manière dont les technologies de l'IA peuvent servir à des fins malveillantes afin de causer des préjudices sur la scène internationale ». À ce sujet, la première partie de ce chapitre fait ressortir les principales préoccupations concernant la mise au point de systèmes d'armes autonomes, au sens de systèmes dans lesquels aucun humain n'assure, au moment de recourir à la force, les fonctions critiques de la sélection et de l'attaque des cibles. La seconde partie du chapitre se concentre sur les hypertrucages (*deepfakes*), soit des contenus dans lesquels images, vidéos, matériel audio et même textes sont artificiellement manipulés de diverses manières pour faire croire à l'auditoire qu'ils sont réels. Ces deux dimensions de l'IA militarisée sont mises en avant pour souligner des enjeux particuliers, mais

aussi y proposer de possibles solutions, bien nécessaires. Chacun de ces champs de perfectionnement de l'IA soulève des préoccupations similaires par rapport à l'escalade des conflits, à la menace accrue du recours à la force, à l'instabilité mondiale et à la grande difficulté qu'ont les victimes civiles à accéder à la justice. Contrairement à ce qu'en disent d'autres publications (voir Burton et Soare, 2019), certaines utilisations de l'IA militarisée ne sont pas inéluctables. Ce chapitre choisit donc de mettre l'accent sur les occasions propices de réguler ces technologies et de prévenir les abus.

SYSTÈMES D'ARMES AUTONOMES

Il est question dans cette partie des préoccupations que soulève la mise au point des systèmes d'armes autonomes. Les discussions diplomatiques en cours sur cette question sont aussi abordées, y compris certaines des façons dont les connaissances communes au sujet des armes autonomes ont progressé ainsi que les défis qui ont empêché l'adoption d'un instrument de réglementation. Cette partie fait également référence au manque de cohérence entre les discussions sur les armes autonomes et les efforts fournis en vue de parvenir à une IA éthique. Enfin, on y explique pourquoi un traité international sur les systèmes d'armes autonomes s'avère essentiel dans la solution multilatérale apportée à cet enjeu.

Que sont les systèmes d'armes autonomes ?

L'autonomie basée sur l'IA n'est plus qu'une simple possibilité. Des systèmes d'armes qui, une fois activés, sélectionnent et attaquent des cibles, et recourent à la force par eux-mêmes, sans aucune intervention humaine, sont déjà en cours de perfectionnement (Nations Unies, Conseil de Droits de l'Homme, 2013). Dans son rapport de 2013 présenté au Conseil des droits de l'homme de l'Assemblée générale des Nations Unies, le rapporteur spécial Christof Heyns a mentionné les vives inquiétudes que suscitent ces systèmes d'armes, s'interrogeant notamment sur la mesure dans laquelle ils se conforment au droit international en matière de droits humains et au droit international humanitaire.

Des questions concernant la responsabilité et le rôle des humains qui commandent les fonctions critiques de la sélection et de l'attaque des cibles sont au cœur des appels à la réglementation. Les systèmes existants parviennent à repérer des cibles précises, puis à les attaquer avec des explosifs. Certaines technologies sont moins sophistiquées, tels les robots sentinelles SGR-A1 présents dans la zone démilitarisée entre la Corée du Nord et la Corée du Sud. Les sentinelles sont équipées de caméras, de détecteurs de chaleur et de mouvement, ainsi que d'un logiciel de reconnaissance des formes permettant au système de détecter un intrus. Le SGR-A1 peut attaquer une cible avec une mitrailleuse légère à quelque 800 mètres de distance. Actuellement, les drones rôdeurs et les SGR-A1 sont tous commandés par des humains, les armes entièrement autonomes n'existant pas encore. Le rôle des opératrices et opérateurs humains est toutefois variable, et le type et la qualité du contrôle qu'ils exercent sur les systèmes diminuent en raison des innovations techniques et de la pression exercée en vue d'une prise de décisions rapide. Comme Gould (2021) le fait valoir, on assiste à une mise en données de la guerre, celle-ci s'appuyant de plus en plus sur les images captées par des drones, les téléphones satellites, ainsi que la surveillance et la collecte de métadonnées pour catégoriser les comportements normaux et anormaux dans des situations complexes qui ne se réduisent pourtant pas à de tels éléments.

Bien qu'aucune définition ne fasse pour le moment consensus à l'échelle internationale, certaines caractéristiques des systèmes d'armes autonomes tendent à se dégager. Ces systèmes intégreraient des profils de cibles préprogrammés et des indicateurs techniques que reconnaîtraient les capteurs des armes (Moyes, 2019). Ils généreraient des données en fonction de l'environnement, plutôt que d'en recevoir par programmation. Les systèmes traiteraient et analyseraient ces données, puis détermineraient les actions à entreprendre ; ils « emploieraient la force » – en faisant feu ou en lançant un missile, par exemple – si leur analyse conclut au respect de certaines conditions préprogrammées. En 2021, le Comité international

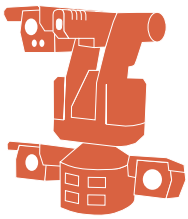
de la Croix-Rouge (CICR) a retenu une définition selon laquelle les systèmes d'armes autonomes sont ceux qui « sélectionnent des cibles et exercent la force contre elles sans intervention humaine », ce qui signifie que « l'utilisateur du système ne choisit pas – et ne connaît même pas –, la ou les cibles spécifiques ni le moment et/ou le lieu précis des frappes » (CICR, 2021, p. 2).

| FIGURE 1 |

Exemples de systèmes existants. Source : Pax Netherlands (2019)

SGR-A1

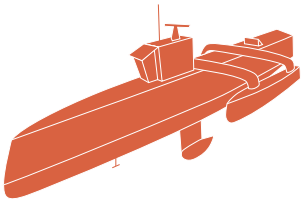
Made by: Hanwha (South Korea)
Sold to: South Korea



This stationary robot, armed with a machine gun and a grenade launcher, operated along the border between North and South Korea. It can detect human beings using infra-red sensors and pattern recognition software. The robot has both a supervised and unsupervised mode available. It can identify and track intruders, with the possibility of firing at them.

SEAHUNTER

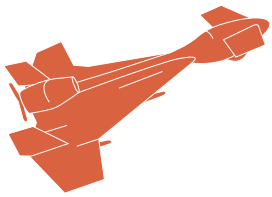
Made by: Pentagon's DARPA (United States)
Sold to: Under development



This 40 m long self-navigating warship is designed to hunt for enemy submarines and can operate without contact with a human operator for 2-3 months at a time. It is currently unarmed. US representatives have said the goal is to arm the Sea Hunters and to build unmanned flotillas within a few years. However, it has been said any decision to use offensive lethal force would be made by humans.

HARPY

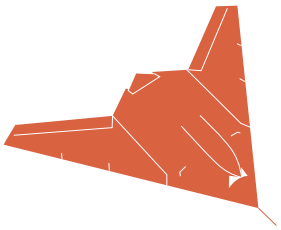
Made by: Israel aerospace industries (Israel)
Sold to: China, India, Israel, South Korea and Turkey



This 2.1 m long "loitering" missile is launched from a ground vehicle. It is armed with a 15 kg explosive warhead. The Harpy can loiter for up to 9 hours at a time, searching for enemy radar signals. It automatically detects, attacks and destroys enemy radar emitters by flying into the target and detonating.

NEURON

Made by: Dassault aviation (France)
Sold to: Under development



This 10 m long stealth unmanned combat aircraft can fly autonomously for over 3 hours for autonomous detection, localization, and reconnaissance of ground targets. The Neuron has fully automated attack capabilities, target adjustment, and communication between systems.

Une telle définition est reconnue par un nombre croissant de personnes demandant un instrument juridiquement contraignant sur les armes autonomes, lequel comprendrait à la fois des interdictions et des obligations (Campaign to Stop Killer Robots, 2021 ; Human Rights Watch, 2021). En même temps, la définition précise de l'expression « armes autonomes » fait encore l'objet de nombreux débats et d'une certaine réticence de la part de certains États qui ne se rallient pas à la perspective d'une majorité de pays et de la société civile. En fait, l'absence de consensus sur une définition offre à certains États une raison – sans doute un prétexte – pour éviter d'entamer une négociation quant à un instrument juridiquement contraignant visant à restreindre la mise au point des systèmes d'armes autonomes (Sauer, 2021). Ce débat occulte le fait qu'il n'est pas nécessaire de s'entendre sur une définition avant de lancer de telles négociations. D'autres processus de désarmement servent d'exemples à cet égard (Devoto *et al.*, 2021), notamment celui qui a conduit à l'interdiction des armes à sous-munitions[80].

Il importe également de mentionner que, même si les discussions à propos de la Convention sur certaines armes classiques (CCAC) font généralement référence aux « systèmes d'armes létaux autonomes » (SALA), cette formulation n'est pas largement acceptée. Plusieurs pays, le CICR et des organisations de la société civile estiment qu'il faudrait mettre de côté le qualificatif *létal*. Comme ils le font remarquer, la létalité n'est pas une caractéristique intrinsèque d'une arme. Il y a violation du droit international humanitaire même quand certaines armes ne tuent pas, mais qu'elles entraînent des blessures injustifiées ou des dommages pour la population civile. Le recours à des armes de défense peut également entraver le droit international humanitaire. Pour ces raisons, nous utilisons « systèmes d'armes autonomes » ou « armes autonomes » quand nous abordons le sujet, et ne parlons de SALA que dans le contexte de la CCAC.

Les multiples inquiétudes suscitées par les armes autonomes

Les inquiétudes que suscitent les systèmes d'armes autonomes peuvent s'analyser sous les angles suivants.

Éthique

Par principe, les décisions de vie ou de mort ne devraient pas incomber à une machine. Les systèmes d'armes autonomes seraient, par définition, dépourvus de la capacité humaine d'analyser les contextes culturels et les situations de conflit, et de comprendre ce que signifie prendre une vie humaine. Permettre aux machines d'en arriver à de telles décisions porte atteinte à la dignité humaine. Comme le souligne Wallach (2013), les armes autonomes devraient être considérées comme un *malum in se*, un mal en soi, étant donné qu'elles « manquent de discernement, d'empathie et de la capacité de poser les jugements nécessaires pour mettre en balance les pertes civiles et l'atteinte d'objectifs militaires. En outre, déléguer des décisions de vie ou de mort à des machines est immoral, car les machines ne peuvent être tenues responsables de leurs actions ». Ainsi, les armes autonomes devraient faire l'objet d'une réglementation, non seulement parce qu'elles pourraient accidentellement tuer des membres de la population civile, mais également parce que, par respect pour la dignité humaine, aucune machine ne devrait mettre en danger la vie de qui que ce soit, même s'il s'agit d'un combattant ou d'une combattante.

Par ailleurs, une approche féministe de l'éthique met au premier plan de tout débat l'expérience vécue des personnes touchées ou potentiellement touchées par le sujet dont il est question (Palmer, s. d.). Selon cette perspective, il est également important de discuter de l'éthique associée aux armes autonomes

80. Pour en savoir plus sur la façon dont a été négociée la Convention sur les armes à sous-munitions, voir : Borrie, J. 2009. *Unacceptable Harm : A History of How the Treaty to Ban Cluster Munitions was Won*. UNIDIR. https://www.unidir.org/publication/unacceptable-harm-history-how-treaty-ban-cluster-munitions-was-won (consulté le 3 septembre 2021).

du point de vue des pays et des populations victimes d'un conflit, qui seraient vraisemblablement les premiers à souffrir de l'utilisation de ces armes. Leurs priorités et leur évaluation de ce qui est éthiquement acceptable ou non seraient certainement très différentes de celles présentées par les États qui sont les principaux producteurs d'armes.

Droit international humanitaire

L'utilisation d'armes autonomes entraînerait certainement des violations du droit international humanitaire (DIH), notamment des principes suivants, qui requièrent un jugement humain : le principe de distinction entre civils et combattants, et entre biens civils et objectifs militaires ; le principe de proportionnalité, qui nécessite d'évaluer si une attaque est susceptible de causer des victimes civiles ou des dommages qui seraient excessifs par rapport à l'avantage militaire direct (ICRC, s. d.).

En outre, plus une arme est autonome, plus il sera difficile d'établir la responsabilité à l'égard de son utilisation et de l'obligation d'offrir recours et réparation aux victimes, de même que de veiller à ce que les personnes ayant commis des violations en matière de DIH en assument les conséquences. Cette responsabilité pourrait incomber à diverses parties prenantes, soit les personnes ou entités qui recueillent les données, font la programmation, commandent le système ou en assurent le fonctionnement. La manière d'attribuer la responsabilité n'est pas évidente sachant que de nombreuses personnes contribuent à mettre au point ou à faire fonctionner de tels systèmes. Si un système prend des décisions sans intervention humaine, il est peu probable qu'un être humain soit tenu responsable des résultats. Voilà qui pose un défi supplémentaire pour les victimes qui cherchent à exercer leurs droits et qui font déjà face à des obstacles de taille[81]. D'un point de vue humanitaire, nous devons également tenir compte des répercussions psychologiques et économiques qu'aurait pour les populations ciblées – déjà traumatisées par les conflits – le fait d'être *aussi* attaquées par des armes autonomes, compte tenu des effets bien documentés de la guerre à distance[82].

Droits humains

S'il y a production d'armes autonomes, celles-ci pourraient être employées non seulement à l'occasion de conflits entre des États, mais aussi, sur la scène nationale, par la police ou les institutions responsables de la sécurité nationale. Cet usage pourrait entraîner des violations des droits humains, tels que le droit à la vie, le droit d'obtenir réparation et le droit au respect de la vie privée. Les arrestations ou les détentions arbitraires ainsi que le fait d'infliger de possibles préjudices à des individus identifiés par des systèmes autonomes ne sont que quelques scénarios envisageables. Les inquiétudes concernant une éventuelle mauvaise utilisation des technologies de reconnaissance faciale par les institutions responsables de la sécurité de même que les incidences en matière de droits humains découlant de leur utilisation ont conduit des entreprises telles qu'Amazon, IBM et Microsoft à demander ou à établir des moratoires sur leur utilisation par les forces de police (Dastin, 2021 ; Allyn, 2020 ; Greene, 2020). L'utilisation potentiellement abusive des technologies de reconnaissance faciale par les services de police a mis en évidence la nécessité d'une réglementation stricte là où les préoccupations et les risques sont criants.

81. Pour en apprendre davantage sur les défis que posent les systèmes d'armes autonomes en matière de responsabilité, consulter : Human Rights Watch. 2015. *Mind the Gap : The Lack of Accountability for Killer Robots.* https://www.hrw.org/report/2015/04/09/mind-gap/lack-accountability-killer-robots

82. Voir, par exemple, Sharkey (2019) ou la publication suivante au sujet des effets de la guerre à distance sur la santé mentale. SaferWorld. S. d. *Warpod episode 8 : Remote Warfare : Interdisciplinary Perspectives* https://www.saferworld.org.uk/multimedia/saferworldas-warpod-episode-8-remote-warfare-interdisciplinary-perspectives.

Préjugés sociaux

Le fait que des préjugés imprègnent des applications d'IA telles que la reconnaissance faciale est bien documenté. Une étude a par exemple révélé un taux d'erreur supérieur (dépassant parfois les 34 %) quand il s'agit de reconnaître des femmes à la peau foncée plutôt que des hommes à la peau claire (Buolamwini et Gebru, 2018). Il a par ailleurs été démontré que des applications d'IA – pas seulement des technologies de reconnaissance faciale – font preuve de préjugés sociaux et en amplifient, particulièrement en ce qui a trait à la « race » ou au genre, et ce, dans des domaines tels que l'éducation, la santé, l'emploi, le logement social ou la prévention de la criminalité. Il n'y a pas lieu de croire que les armes autonomes seraient exemptes de tels préjugés, ce qui pourrait avoir une incidence vitale[83] (Díaz et Muñoz, 2019 ; Ramsay-Joynes, 2019). Comme l'a fait valoir Horowitz (2020), la plupart des armées risquent d'adopter des applications polyvalentes fonctionnant au moyen d'algorithmes commerciaux. Par conséquent, les préoccupations liées aux préjugés s'appliquent également au contexte militaire.

Outre les violations potentielles en matière de droit international humanitaire et de droit international relatif aux droits humains, les armes autonomes qui cibleraient des personnes et dont le fonctionnement dépend de telles applications pourraient avoir un effet particulièrement marqué sur des populations qui sont déjà marginalisées, comme les femmes, ou les personnes racisées, handicapées ou LGBTIQ+. Recourir à de tels outils en situation de conflits risque de mener au ciblage de femmes racisées ou à d'autres bévues. Par un mauvais repérage, il se peut qu'une béquille soit par exemple erronément assimilée à une arme. Les ensembles de données servant à la programmation de systèmes d'armes autonomes ne tiennent probablement pas compte du fait que les personnes utilisant un fauteuil roulant, une canne ou un déambulateur se trouvent plus près du sol et que leur démarche est lente. Ils ne prévoient sans doute pas la réaction des civils. La rapidité de la prise de décisions fondée sur l'IA pourrait aussi entraîner le repérage d'un nombre exagéré de comportements suspects.

Il importe également de tenir compte du fait qu'une forte proportion des personnes qui meurent aux mains de policiers aux États-Unis (un tiers, peut-être même la moitié, selon les estimations) sont des hommes racisés souffrant d'une déficience (Abrams, 2020). Il ne s'agit pas d'une coïncidence : le racisme et le capacitisme mènent à une discrimination intersectionnelle. Quand il est question d'armes autonomes, les conséquences d'une telle discrimination systémique deviennent un enjeu de vie ou de mort.

Sécurité internationale

Certains États décrivent les armes autonomes – et d'autres méthodes de guerre à distance – comme de plus en plus précises. De telles affirmations doivent être analysées à la lumière des répercussions qu'ont les méthodes actuelles de guerre à distance, comme les frappes de drones. Après une tragédie survenue très récemment, une enquête du *New York Times* a suggéré que les États-Unis avaient ciblé un travailleur humanitaire dans une frappe de drone ayant tué neuf autres personnes, dont sept enfants, par suite d'une analyse incorrecte qui les avait conduits à considérer ses activités comme des « mouvements suspects » (Koettl et al., 2021). L'évaluation de Knowles et Watson (2018) selon laquelle la guerre à distance n'est certainement pas précise ni moins horrible et traumatisante pour les victimes est également pertinente en ce qui a trait aux armes autonomes. Elle mène par ailleurs au « paradoxe de la guerre à distance », ce qui signifie que les pays disposant d'armes autonomes pourraient facilement entrer en guerre, puisque leurs propres pertes seraient limitées, sans égards aux effets sur les victimes. Une fois cette technologie mise au point, elle pourrait être reproduite, et son usage étendu aux groupes armés non étatiques illégaux. En outre, le piratage et les attaques adverses s'avéreraient

83. Pour une analyse de l'effet potentiel des armes autonomes sur les populations marginalisées, en particulier en Amérique latine, consulter Díaz et Muñoz (2019), accessible au https://bit.ly/ArmasInterseccionalidad.

potentiellement dangereux, car il est difficile de contenir les effets de ces armes. Comme l'a fait valoir Russel (2021), les armes autonomes pourraient facilement devenir des armes de destruction massive parce qu'elles ne nécessitent aucune supervision humaine ; appuyer sur un simple bouton suffit à lancer une attaque massive au moyen d'un réseau de milliers ou de millions d'armes autonomes.

Équilibre des pouvoirs

Accepter une autonomisation de certaines fonctions cruciales de l'armement aurait aussi une incidence sur l'équilibre des pouvoirs et procurerait aux États militarisés l'avantage d'exercer un important pouvoir géopolitique. Comme le soutient Bengio (2019), « l'IA est essentiellement un outil qui peut être utilisé par les personnes au pouvoir pour conserver ce pouvoir et pour l'accroître ». Les États qui seraient parmi les premiers à mettre au point les technologies et à bénéficier d'un avantage concurrentiel auraient une mainmise démesurée sur la sécurité mondiale. Horowitz (2020) le fait remarquer : l'une des manières d'obtenir un tel avantage est de concevoir « un algorithme général qui pourrait engendrer d'autres algorithmes, fonctionner dans plusieurs domaines et contourner le problème de l'oubli catastrophique (l'oubli des précédents apprentissages après l'acquisition de nouvelle information dans un autre domaine) ». L'attrait d'un tel avantage suscite des investissements importants en vue de concevoir ces technologies au sein des grandes puissances, de la Chine et des États-Unis, mais également de bon nombre de pays ayant une armée modeste (Horowitz, 2020).

Les armes autonomes ne constituent pas qu'un enjeu militaire ou technique qu'il faut laisser à des spécialistes techniques militaires. Même si le point de vue de ces spécialistes se fait davantage entendre, la discussion concernant les armes autonomes doit tendre à l'inclusivité, compte tenu de la grande incidence qu'ont celles-ci sur la sécurité mondiale (Marijan, 2018). Étant donné l'importance que prend l'IA et l'avantage militaire que des États comptent tirer de la technologie, il est nécessaire que des parties expriment à l'extérieur de la sphère militaire certaines craintes d'être mises de côté ou désavantagées. Fait intéressant, la peur d'accuser un retard ou de voir des pays mettre au point des technologies avant que ne le fassent de puissants États suscite également des investissements et favorise l'innovation technologique, ce qui crée un cercle vicieux.

La solution internationale inadéquate quant aux systèmes d'armes autonomes

C'est précisément l'action de nombreux groupes qui a accentué l'attention portée aux risques que posent les systèmes d'armes autonomes. Parmi eux figurent notamment la campagne Stopper les robots tueurs (regroupant des organisations de la société civile de plus de 70 pays), des lauréates du prix Nobel de la paix (Nobel Women's Initiative, 2014), le Parlement européen (2018), le secrétaire général des Nations Unies (Bugge, 2018), l'Alliance pour le multilatéralisme (Alliance for Multilateralism, 2019), et des milliers de spécialistes en IA, en éthique et en sécurité internationale, par exemple : Future of Life Institute et International Committee for Robot Arms Control. Fait important, la Déclaration de Montréal pour un développement responsable de l'intelligence artificielle (2010), ratifiée par 187 organisations, stipule au principe 9(3) : « La décision de tuer doit toujours être prise par des êtres humains et la responsabilité de cette décision ne peut être transférée à un [système d'IA]. »

De plus, les Nations Unies discutent depuis 2014 des systèmes d'armes létaux autonomes dans le contexte de la Convention sur l'interdiction ou la limitation de l'emploi de certaines armes classiques qui peuvent être considérées comme produisant des effets traumatiques excessifs ou comme frappant sans discrimination, telle qu'elle a été modifiée le 21 décembre 2001 (également connue sous le sigle CCAC :

Convention sur certaines armes classiques)[84]. Les Hautes Parties contractantes se réunissent régulièrement depuis sept ans et ont adopté en 2019 une liste de principes directeurs qui résume les ententes et les accords conclus au cours des dernières années (CCW, 2019).

Alors qu'un nombre croissant de pays ont demandé une interdiction des systèmes d'armes létaux autonomes[85], d'autres soutiennent qu'il serait prématuré de lancer de telles négociations étant donné que ces systèmes n'existent toujours pas. Selon eux, l'orientation que prend l'évolution de ces systèmes n'est pas encore claire, et il serait ainsi trop tôt pour élaborer un régime réglementaire applicable à leur utilisation. Cette position fait fi des risques déjà connus de la guerre à distance, de ceux liés aux travers de l'IA et aux enjeux de la responsabilisation, et du précédent que constitue l'interdiction des lasers aveuglants, adoptée à titre préventif en 1995.

Au cours de sept ans de discussions au sujet de la CCAC, seuls quelques pays ont fait référence aux avancées concrètes touchant diverses technologies liées à l'IA et à la manière dont celles-ci pourraient s'intégrer aux systèmes d'armes. Par exemple, des véhicules terrestres sans équipage, des véhicules aériens ainsi que des drones rôdeurs de plus en plus autonomes ont fait l'objet d'une attention étonnamment limitée. L'actuelle absence de cadre réglementaire international fait en sorte que le développement technologique se poursuit sans indications claires quant à ce qui est légal et ce qui est moralement acceptable.

Qui plus est, l'absence de solution réglementaire de la part de la communauté internationale semble en décalage avec les cadres internationaux et régionaux sur l'IA auxquels de nombreux pays ont adhéré jusqu'à maintenant. Ces cadres comprennent (Muñoz, 2020) :

- La **Recommandation de l'UNESCO sur l'éthique de l'intelligence artificielle**, qui indique que « [d]ans les scénarios où les décisions sont considérées comme ayant un impact irréversible ou difficile à renverser, ou qui pourraient impliquer des décisions de vie et de mort, la décision finale devrait être prise par l'homme » et que « [d]e manière générale, les décisions de vie et de mort ne devraient pas être abandonnées à des systèmes d'IA » (UNESCO, 2021, p. 9, 11).
- La **résolution 473 de la Commission africaine des droits de l'homme et des peuples** (2021), qui **« [a]ppelle** les États parties à s'assurer que toutes les technologies de l'intelligence artificielle, la robotique et les autres technologies nouvelles et émergentes qui ont des conséquences de longue portée pour les humains [restent] sous un contrôle humain effectif, en vue de garantir que la menace qu'elles représentent pour les droits fondamentaux de l'homme est écartée. La norme émergente relative au maintien d'un contrôle humain effectif des technologies de l'intelligence artificielle, [de] la robotique et d'autres technologies nouvelles et émergentes doit être codifiée en tant que principe des droits de l'homme ».
- Les **principes de l'IA centrée sur l'humain du G20**, inspirés de ceux contenus dans la **Recommandation du Conseil sur l'intelligence artificielle de l'OCDE** (2019), qui stipule que les « acteurs de l'IA devraient respecter l'état de droit, les droits de l'homme et les valeurs démocratiques, [qui] comprennent la

84. Même si le mandat de la CCAC a trait aux systèmes d'armes « létaux » autonomes, les auteures emploient, pour faire référence à ces armes, « systèmes d'armes autonomes » parce que la létalité ne devrait pas servir de critère pour les définir. La létalité n'est pas une caractéristique définie dans le droit international humanitaire (DIH), et conserver le qualificatif « létaux » constituerait un précédent regrettable et irait à l'encontre du DIH. Pour en savoir davantage, consulter : Muñoz, W. 2021. *It's about more than autonomous weapons systems*. International Human Rights Clinic, Harvard Law School. https://humanitariandisarmament.org/2021/08/30/it-is-about-more-than-autonomous-weapons-systems/ (consulté le 31 août 2021).

85. Depuis août 2021, 31 pays ont demandé une interdiction des systèmes d'armes létaux autonomes. Source : Human Rights Watch. 2021. *Killer Robots : Urgent Need to Fast-Track Talks Shared Vision Forms Sound Basis for Creating a New Ban Treaty*. https://www.hrw.org/news/2021/08/02/killer-robots-urgent-need-fast-track-talks (consulté le 5 octobre 2021).

liberté, la dignité et l'autonomie, la protection de la vie privée et des données, la non-discrimination et l'égalité, la diversité, l'équité, la justice sociale, ainsi que les droits des travailleurs reconnus à l'échelle internationale ».

- La **proposition de règlement de la Commission européenne** (2021) établissant de nouvelles règles axées sur l'excellence et la confiance en matière d'IA, qui propose une interdiction des pratiques comprenant « tous les systèmes d'IA dont l'utilisation est considérée comme inacceptable car contraire aux valeurs de l'Union, par exemple en raison de violations des droits fondamentaux ».
- D'autres cadres de référence régionaux et nationaux tels qu'une **charte sur l'éthique des technologies émergentes dans la région arabe** (UNESCO, 2019), qui vise à repérer des moyens de « guider la science et la technologie vers la bonne voie, en les éloignant des tendances et pratiques contraires à l'éthique qui sont nuisibles aux humains et au contexte environnant », de même qu'une **déclaration de la République de Corée** sur une éthique de l'IA axée sur l'harmonie et visant à ne laisser personne pour compte (Ministry of Foreign Affairs, Republic of Korea, 2020).

En outre, le Partenariat mondial sur l'IA mène d'excellentes initiatives dans des domaines tels que la réduction des changements climatiques, l'intervention relative à la pandémie et l'atteinte des objectifs de développement durable des Nations Unies. Il n'aborde toutefois pas la question de l'IA dans l'armement. De même, la proposition européenne sur l'IA indique que le règlement « ne s'applique pas aux systèmes d'IA développés ou utilisés exclusivement à des fins militaires » (Commission européenne, 2021, p. 45). Il semble donc que la communauté internationale laisse exclusivement à la CCAC le soin de se pencher sur la légalité et la légitimité de déléguer la prise de décisions, notamment sur la vie, à des fonctions autonomes des systèmes d'armes, et ce, même si l'enjeu comporte une incidence éthique pour l'humanité entière.

En septembre 2021 s'est tenue une conférence intitulée Safeguarding Human Control over Autonomous Weapons Systems, organisée par le ministre australien des Affaires étrangères. Les discussions au sujet de la nécessité de laisser aux humains la maîtrise des systèmes d'armes autonomes ont réuni des spécialistes du domaine militaire et de la diplomatie, des éthiciens et éthiciennes, des spécialistes de l'IA issus de la société civile, du secteur privé et de milieux universitaires – dont certains et certaines ont contribué à la proposition de la Commission européenne –, des scientifiques, le CICR et l'UNESCO.

L'URGENT BESOIN D'UN INSTRUMENT INTERNATIONAL JURIDIQUEMENT CONTRAIGNANT QUANT AUX SYSTÈMES D'ARMES AUTONOMES

Un traité international s'impose pour calmer les diverses inquiétudes que suscitent les systèmes d'armes autonomes. Pour être efficace, un tel traité devrait inclure une interdiction des armes autonomes antipersonnel et des armes pouvant être utilisées sans grande intervention humaine, de même que des obligations quant aux autres utilisations des systèmes d'armes autonomes. Un instrument juridiquement contraignant à l'échelle internationale s'inscrit au cœur des interventions qui seront nécessaires, en matière de politiques, pour pallier les inquiétudes relatives aux systèmes d'armes basés sur l'IA. Des codes de conduite, des déclarations, des principes directeurs, un recueil de bonnes pratiques ou l'examen d'armes ne représentent pas une solution suffisante aux enjeux que posent les systèmes d'armes autonomes, car ces mesures n'ont pas le même poids qu'un instrument juridiquement contraignant. Essentiellement, des instruments non contraignants ne peuvent garantir la transparence, l'imputabilité et la responsabilité requises à l'égard de ces armes ; les États n'ont pas l'obligation légale de respecter de telles mesures. Par ailleurs, ces instruments ne constitueraient pas la norme internationale stricte

ayant été établie, par exemple, par la Convention sur l'interdiction des mines antipersonnel, qui a conduit 32 pays qui n'y étaient pas encore parties à se conformer *de facto* à la plupart des obligations qu'elle contient (ICBL, 2021).

Que des armes autonomes en viennent à faire partie de l'arsenal de n'importe quelle force militaire ou policière est déjà préoccupant. À partir du moment où des systèmes d'armes autonomes sont conçus, ils sont susceptibles de se retrouver entre les mains d'acteurs non étatiques et d'être détournés vers des destinataires non autorisés – qui acquièrent ou livrent les armes illégalement –, comme c'est le cas d'autres armes conventionnelles[86]. Considérons par exemple la situation en Afghanistan, où les talibans contrôlent un armement valant des milliards de dollars américains (Cohen et Liebermann, 2021) ainsi que les données biométriques du personnel afghan ayant travaillé pour les forces américaines et celles de l'OTAN, données qu'ils utilisent déjà pour « traquer les Afghans et Afghanes qui ont aidé les forces américaines et alliées », en utilisant pour cela de l'équipement et des données de source américaine (Roy et Miniter, 2021). Un tel risque d'assister à un détournement d'armes, même involontaire, vers des groupes non autorisés à les employer est une préoccupation commune dans le commerce mondial des armes.

Échouer à adopter un traité international sur les systèmes d'armes autonomes signifierait que la communauté internationale accepte *de facto* de déléguer des décisions concernant la vie humaine à des systèmes autonomes. Cet échec pourrait à son tour avoir un effet négatif sur d'autres règlements liés à l'IA et aux technologies émergentes, par exemple en ce qui concerne les décisions relatives à la vie humaine en contexte de soins de santé. En fait, si le droit à la vie doit être cédé à des applications d'IA, pourquoi ne pas céder aussi d'autres droits ?

Les discussions sur les armes autonomes ont déjà duré sept ans en ce qui a trait à la CCAC. Certains des pays qui investissent dans la recherche sur les armes autonomes allèguent eux-mêmes qu'il faut poursuivre les discussions et la recherche. Selon l'organisation Human Rights Watch (2020), l'État d'Israël, la Russie et les États-Unis font partie des pays qui investissement massivement dans la mise au point de divers systèmes d'armes autonomes. Ces mêmes États ont qualifié de « prématurée » l'adoption d'un instrument juridiquement contraignant (Amnesty International, 2021). Une telle position n'est pas neutre ; si les pays qui participent aux pourparlers sur la CCAC acceptent de poursuivre, sans mandat de négociation, les discussions au sujet d'un instrument juridiquement contraignant, leur décision profitera aux pays et aux industries qui mettent déjà au point des armes autonomes.

Sur une note positive, de plus en plus de voix s'élèvent dans les pays du Sud – en provenance notamment de la société civile, des scientifiques et du Groupe d'experts gouvernementaux – pour soutenir un traité sur les armes autonomes. Ils comprennent des États actuellement touchés par des conflits, des pays en ayant connu par le passé, des pays s'étant engagés en faveur du désarmement humanitaire, et d'autres ne s'étant pas prononcés sur la question. Ainsi, les pays du Sud, y compris ceux qui sont touchés par des conflits – et qui seraient éventuellement un terrain d'essai des armes autonomes – expriment leur mécontentement devant l'absence de progrès en matière de traité international et ils présentent des propositions claires quant au cadre juridique international qui s'impose de toute urgence. Dans les discussions sur le désarmement et le contrôle des armes, les pays du Sud ont manifesté qu'ils seraient non seulement preneurs de normes, mais aussi créateurs de celles-ci (Bode, 2019).

86. Voir par exemple : Kirkham, E. 2017. *International efforts to prevent diversion of arms and dual-use goods transfers : challenges and priorities.* SaferWorld. https://www.saferworld.org.uk/resources/publications/1112-international-efforts-to-prevent-diversion-of-arms-and-dual-use-goods-transfers-challenges-and-priorities (consulté le 16 septembre 2021).

Les traités de contrôle des armements et de désarmement humanitaire fonctionnent. Ils ont un impact direct sur la prévention des décès, des blessures, des souffrances humaines et des effets négatifs à long terme sur la vie et les moyens de subsistance de générations de milliers de personnes.[87]

Les dernières rencontres relatives à la CCAC ont soulevé des questions valables quant à une possible intervention multilatérale en vue d'encadrer ces systèmes et au rôle des armées de pointe. En décembre 2021 ont eu lieu deux rencontres déterminantes. D'abord, le Groupe d'experts gouvernementaux sur les systèmes d'armes létaux autonomes s'est réuni, mais n'a pas produit un rapport consensuel contenant des recommandations en vue d'adopter un mandat plus ambitieux. Bien que plusieurs délégations aient demandé un mandat de négociation au sujet d'un instrument juridiquement contraignant, une minorité de pays ont freiné une avancée. À la fin de la rencontre du Groupe d'experts gouvernementaux, la proposition initiale du représentant belge d'élaborer « un instrument » (pas nécessairement contraignant, ce qui servait de compromis pour les pays militarisés) avait perdu toute sa force, ce qui n'a pas empêché les délégations de ces pays de la refuser. La décision a ensuite été remise entre les mains des participants et participantes de la conférence d'examen de la CCAC, une rencontre quinquennale qui s'est tenue à la mi-décembre 2021. La conférence a mené à l'adoption d'un faible mandat qui engage simplement la CCAC à continuer les pourparlers en 2022, durant dix jours, sans objectif clair, surtout pas à l'égard de négociations sur un instrument juridiquement contraignant.

La principale question est maintenant de savoir s'il y a une volonté politique d'examiner l'enjeu des armes autonomes au sein d'un autre forum. C'est ce qui a été fait, par exemple, dans le cas de la Convention sur l'interdiction des mines antipersonnel et de la Convention sur les armes à sous-munitions (Herby, 1998). La situation soulève plus largement une série de questions sur la capacité des initiatives multilatérales à encadrer l'enjeu de la militarisation des technologies émergentes. Une distorsion du « consensus » s'observe alors que quelques pays militarisés ont investi les forums internationaux. Étant donné que la communauté scientifique et l'industrie liées à l'IA jouent un rôle dans le développement de certaines technologies qui auront des applications militaires décisives, il faudra voir comment elles réagiront sans une entente internationale. Il est certainement nécessaire d'accroître l'attention et les pressions afin de parvenir à des ententes et à des normes permettant de prévenir les usages malveillants. Les États doivent tenter de parvenir à des accords ayant une incidence majeure en matière de sécurité internationale.

HYPERTRUCAGES

La mise au point d'armes autonomes montre que l'IA se militarise déjà et qu'elle s'implante dans les systèmes d'armes. Elle peut également servir à créer de l'instabilité ou nourrir l'escalade de conflits de manières dont on ne saisit pas encore l'ampleur. Alors que les armes autonomes ont été assimilées à la prochaine révolution majeure dans la guerre, il est de plus en plus reconnu que les hypertrucages transforment la sécurité globale et l'environnement politique, et qu'ils touchent à des enjeux relatifs à la sécurité mondiale et nationale, à la démocratie, à la violence sexiste ainsi qu'au droit à la protection de la vie privée.

Mais que sont les hypertrucages ? Les hypertrucages (*deepfakes*) consistent essentiellement en des images, des vidéos et du matériel sonore manipulés ou fabriqués. Bien que les manipulations de ce genre existent depuis de nombreuses années, l'accessibilité à la technologie et la rapidité de celle-ci ont considérablement modifié la réalité des créateurs et créatrices de contenus crédibles. Il importe

87. Pour plus d'information, voir : United Nations, Office of Disarmament Affairs. *Landmines*. https://www.un.org/disarmament/convarms/landmines/
The Convention on Cluster Munitions. *Achievements*. https://www.clusterconvention.org/achievements/

également de souligner que ces trucages se distinguent des *cheapfakes* (trucages bas de gamme), c'est-à-dire du ralentissement ou de l'accélération d'une séquence pour faire valoir un point particulier (Venema, 2020). Les avancées de l'IA ont transformé les façons de manipuler des contenus. Les hypertrucages font intervenir deux aspects de l'apprentissage automatique : les réseaux neuronaux et les réseaux antagonistes génératifs (Pantserev, 2020, pp. 39-40). Un réseau neuronal veille à ce que le matériel sonore et vidéo produits soient aussi précis que possible grâce au téléchargement de nombreux contenus qu'il tente de synthétiser. Un autre réseau neuronal, le discriminateur, tente de déterminer si le contenu produit est réel ou fictif. S'il établit qu'une vidéo est fausse, il essaiera alors d'analyser et de pallier les failles qu'il a détectées (Pantserev, 2020, pp. 39-40). Ainsi, la qualité et la précision des hypertrucages s'améliorent sans cesse, ce qui rend la détection de ceux-ci de plus en plus difficile à mesure que le processus se répète.

Selon la jeune pousse Deeptrace, le nombre d'hypertrucages sur le Web a augmenté de 330 % d'octobre 2019 à juin 2020 (Wiggers, 2021). Pourtant, la plupart des gouvernements et des entreprises ne sont pas préparés aux effets généralisés que les hypertrucages peuvent avoir sur la société. À mesure que la technologie s'améliore, le public a du mal à discerner le contenu réel du contenu fabriqué (O'Brien, 2019). Cette difficulté accroît l'incertitude quant à la véracité de l'information, et elle contribue à miner la confiance et la culture citoyenne en ligne, y compris dans les sociétés démocratiques. Étonnamment, les avancées technologiques conduisent également à une remise en question de contenus réels, les hypertrucages servant parfois à réfuter de vraies déclarations. Par ailleurs, ceux-ci ont essentiellement eu tendance à cibler les femmes – quelque 90 % des hypertrucages étant à ce jour des contenus pornographiques –, ce qui justifie la crainte que la technologie répande la violence fondée sur le genre (Venema, 2020). De plus, des vidéos manipulées montrant des dirigeants politiques ont suscité des inquiétudes quant à une possible escalade des conflits ou des malentendus qui, dans les sociétés où ces conflits ont cours, pourraient avoir de vastes répercussions. En matière d'engagement militaire, l'une de ces inquiétudes a trait au rôle central des technologies de l'information et de la communication maîtrisant le discours et les réactions, et donc à l'effet qu'un hypertrucage pourrait éventuellement avoir sur l'escalade d'un conflit. Ainsi, le manque d'adéquation entre la législation et les solutions stratégiques concernant ces trucages nécessite une attention soutenue de la part des technologues, des gouvernements, de l'industrie, des organisations internationales et de la société civile.

Quelles sont les préoccupations avec les hypertrucages ?

La conception d'hypertrucages dépend particulièrement de l'apprentissage automatique et de l'apprentissage profond, qui les rendent convaincants. Comme le fait remarquer Adee (2020), les studios de Hollywood, par exemple, avaient auparavant besoin d'un an et d'une équipe de spécialistes pour « repiquer » un acteur ou une actrice dans une vidéo dans laquelle ils avaient joué. Aujourd'hui, les technologies de trucage permettent rapidement d'insérer des individus dans des images ou dans des scénarios auxquels ils n'ont pas pris part. Des images de synthèse de base prolifèrent aussi largement. Par exemple, des applications pour téléphones intelligents telles que Avatarify ou Zao App permettent d'animer des visages ou de changer le visage de personnes dans des vidéos et des images (Fowler, 2021 ; Meskys *et al.*, 2021). Des innovations technologiques plus sophistiquées encore permettent également désormais à des non-spécialistes des technologies de créer à la maison des contenus truqués convaincants. Ces utilisations sont possibles en raison de certains codes sources ouverts et du fait que peu d'images sont nécessaires pour créer des faux de bonne qualité. Comme les systèmes « apprennent » à partir d'un nombre accru d'images, la qualité s'améliorera également, ce qui soulève la question du repérage de contenus fabriqués. Comme l'affirment Kietzmann et autres (2020), la crédibilité des faux contenus ainsi que notre tendance à croire en des preuves photographiques, et surtout en du matériel audio et vidéo, complexifient l'enjeu des contenus manipulés.

Comme les techniques d'hypertrucage servent souvent au divertissement et à la création de contenus amusants, il est parfois difficile de percevoir certains usages abusifs. Des cas d'utilisation malveillante

ont effectivement été observés. Des célébrités mais aussi des personnes ordinaires sont apparues dans des images particulièrement inappropriées, sans y consentir ou sans en avoir connaissance. Les femmes ont quant à elles été la cible, de manière disproportionnée, d'hypertrucages pornographiques conçus sans leur consentement. Les gouvernements doivent donc aborder les enjeux liés au genre en matière de technologies. Comme le souligne Venema, des femmes de différents coins du monde ont subi les conséquences de la pornographie truquée ; certaines ont perdu leur emploi ou n'ont pu en trouver un. Venema (2020) donne l'exemple de la journaliste indienne Rana Ayyub, qui a été la cible d'une campagne de diffamation par hypertrucage alors qu'elle réclamait justice après le viol d'une jeune fille. Ayyub a également été victime de la divulgation de données personnelles la concernant. Son cas n'est pas unique et montre comment du contenu truqué entraîne des problèmes de sécurité réels pour les femmes. Compte tenu d'une protection inégale des personnes selon leur genre et en fonction des pays, la diffusion de cette technologie pourrait avoir des conséquences désastreuses pour les femmes dont l'image est reprise à des fins malveillantes. Il importe donc que les spécialistes en technologies ainsi que les plateformes de partage de contenus prennent soigneusement en considération les répercussions propres à chaque contexte et à chaque pays.

À mesure que la technologie prolifère, on reconnaît également de plus en plus le rôle que jouent les hypertrucages dans des campagnes de désinformation et dans l'éventuelle escalade ou éclosion de conflits. En 2018, une vidéo de l'ancien président des États-Unis Barack Obama publiée par BuzzFeed et partagée dans les médias sociaux a contribué à cette prise de conscience quant à l'utilisation de contenus truqués dans les campagnes politiques. Dans la vidéo, Obama lance une boutade au sujet du président américain d'alors, Donald Trump, ce qui ne lui ressemble pas du tout. Il le reconnaît d'ailleurs lui-même en s'adressant à la caméra. Obama poursuit en disant que c'est quelqu'un comme Jordan Peele, un réalisateur de films, qui ferait de telles remarques sur le président Trump. On constate alors, en écran partagé, que c'est Peele qui parle et que son équipe a utilisé l'IA pour faire croire qu'il s'agissait des paroles d'Obama. Cette vidéo devait être un avertissement public de l'annonceur mettant en garde contre les dangers des hypertrucages, mais elle a inquiété une partie de l'auditoire. Elle était tout simplement trop réussie. L'image et les vidéos d'Obama sont facilement accessibles, et il a fallu au système d'apprentissage automatique quelque 56 heures d'entraînement pour parvenir à un tel résultat (Romano, 2018). Malgré tout, la vidéo d'Obama a mis en lumière les préoccupations relatives aux faux contenus potentiellement utilisés pour diffuser de la désinformation contre des dirigeants politiques, ou pour provoquer une crise et une escalade de conflits.

Dans les zones déjà touchées par des conflits, pas besoin d'une grande imagination pour se rendre compte des dangers potentiels de la circulation d'une vidéo ou d'un enregistrement audio d'un dirigeant politique incitant à la violence ou proférant des menaces contre d'autres communautés (Citron et Chesney, 2019). La diffusion de fausses informations par les médias sociaux, par exemple en zones de conflit, a montré comment l'appel à la violence dans le monde virtuel peut conduire à la violence dans le monde réel. L'utilisation de Facebook dans l'incitation à la violence au Myanmar a fait l'objet d'une grande attention au fil des ans (Asher, 2021). Des messages Facebook ont servi à cibler la communauté minoritaire des Rohingyas pendant des années, y compris lors de la crise de 2018 qui a conduit au déplacement de quelque 800 000 Rohingyas. En 2014, une publication virale ciblant la communauté musulmane du Myanmar a été à l'origine de deux décès ; un attroupement en réaction à la publication avait entraîné des violences. Ce cas montre comment des vidéos et des images manipulées peuvent être utilisées en situation de fragilité pour cibler des groupes et des communautés en particulier.

Il est certes possible d'imaginer de nombreux scénarios hypothétiques mais, pour certains pays, il s'agit déjà de la réalité. Fait important, c'est dans les situations de fragilité politique et dans les pays en développement ayant un faible niveau de littératie numérique que les hypertrucages représentent des dangers potentiels. Prenons l'exemple du Gabon, qui a connu en 2019 un coup d'État militaire provoqué par ce que l'on croyait être une vidéo du président malade (Breland, 2019). Le président gabonais avait reçu des soins médicaux à l'extérieur du pays et, alors que les demandes d'apparition publique

se multipliaient, le gouvernement a publié un message vidéo. Celui-ci a cependant semblé confirmer les inquiétudes concernant le président, et la vidéo a paru étrange. Le coup d'État a finalement été un échec. La vidéo a été considérée comme un trucage ; elle a fait ressortir que les dangers réels de la technologie se font surtout sentir dans les situations de fragilité et les contextes où la population a été exclue ou où elle n'a pas eu accès à la littératie numérique. Si les hypertrucages touchent plus fortement les communautés marginalisées, il n'en demeure pas moins que la communauté internationale en subit également les effets. Leur utilisation pourrait avoir toutes sortes de retombées quant à la stabilité internationale. Citron et Chesney citent plusieurs exemples de telles retombées en matière de sécurité nationale et mondiale ainsi que de relations diplomatiques. Ils font remarquer que du « faux matériel audio pourrait de manière convaincante mettre en scène des responsables américains « admettant » en privé avoir un plan pour commettre un acte de violence à l'étranger en vue de perturber une initiative diplomatique importante » (Citron et Chesney, 2019, p. 176). Tous les hypertrucages n'auraient pas une visée destructrice, mais certains pourraient être planifiés pour influer sur l'issue de sommets diplomatiques et pour freiner d'éventuelles négociations entre différents pays.

Vers une solution réglementaire pour les hypertrucages

Il y a eu proposition d'un certain nombre de contre-mesures technologiques liées aux hypertrucages ainsi que de solutions réglementaires. Un débat est en cours pour déterminer s'il faut interdire la technologie permettant l'hypertrucage. Des sites Web comme Reddit ont banni la pornographie truquée, et Facebook a également interdit les trucages. La politique du média social interdit plus précisément les vidéos conçues ou modifiées par l'IA et dont on ne sait pas qu'elles ont ainsi été manipulées. Facebook continue d'autoriser les contenus parodiques ou satiriques. Il s'agit là d'un élément clé des discussions sur la réglementation : trouver un moyen d'interdire certains usages tout en laissant place à la liberté d'expression, aux œuvres artistiques et aux contenus produits à des fins de divertissement et effectivement reconnus comme des hypertrucages.

La technologie permettant l'hypertrucage nécessite manifestement un encadrement réglementaire et normatif. L'interdiction de certains usages de la technologie doit être normalisée par les pays, les technologues et la société civile. Par exemple, la pornographie non consensuelle truquée et tous les hypertrucages créés sans le consentement de la personne en cause devraient être interdits. Les administrations nationales et régionales doivent adopter des mesures de protection de la vie privée. Il faudra aussi veiller à combler les lacunes juridiques et parvenir à un large accord mondial qui conduirait à l'adoption d'une norme.

Des avertissements et des notifications indiquant qu'un contenu jugé acceptable a fait l'objet de manipulation s'avèrent nécessaires. Il importe de les transmettre clairement et de les rendre visibles à l'auditoire. Un vague message ou l'apparition d'une fenêtre contextuelle facilement ignorée ne suffisent pas. Il faut réfléchir à la conception des systèmes pour s'assurer que les utilisateurs et utilisatrices sont totalement conscients des hypertrucages et que le média gère adéquatement ceux-ci (Olejnik, 2021).

Et pour la suite ? La prolifération d'hypertrucages ne fait que commencer. À défaut de solutions réglementaires appropriées et adoptées à temps, les dangers de cette technologie – déjà observés dans des situations réelles – ne feront que croître. Malheureusement, les hypertrucages tendent à toucher plus particulièrement des groupes déjà vulnérables et privés de leurs droits, ont des retombées variables en fonction du genre, et soulèvent de graves préoccupations en situation de fragilité politique et dans les pays en développement. En réalité, aucun pays n'est immunisé contre les effets de ces trucages. Les inquiétudes liées à la désinformation prédominent dans les démocraties libérales, et elles sont bien réelles pour tous les pays. Des spécialistes techniques ont proposé des solutions, par exemple d'avoir recours à l'IA pour détecter et éliminer les contenus manipulés. Les États pourraient envisager un certain nombre de mesures réglementaires, notamment l'obligation de désigner les contenus manipulés et le retrait des contenus malveillants (Pantserev, 2020). Pour commencer, il importe que tous les pays

portent davantage attention à l'emploi malicieux des hypertrucages et aux préoccupations que soulèvent la technologie. Les principales entreprises technologiques ont une responsabilité quant au retrait rapide de certains contenus, et il faut par ailleurs qu'une législation précise leur rôle et leurs obligations. Tenir des discussions à l'échelle internationale au sujet de normes sur la conduite responsable des États dans le cyberespace serait également une occasion d'encadrer certaines utilisations des hypertrucages, notamment en contexte électoral. Le recours à des tactiques relevant de zones grises, par exemple les campagnes de désinformation, rappelle que, quand ils sont employés à des fins malicieuses, les hypertrucages peuvent avoir des conséquences bien réelles en matière de sécurité.

CONCLUSION

Les systèmes d'armes autonomes et les hypertrucages soulèvent le fait que l'IA peut se transformer en arme, ce qui suscite des inquiétudes parmi les parties prenantes au sein des gouvernements et à l'extérieur. Il existe d'évidentes préoccupations éthiques, juridiques, humanitaires et relatives aux droits humains en ce qui concerne la mise au point et l'utilisation potentielle de systèmes d'armes de plus en plus autonomes et d'hypertrucages. La mise en place de normes pour régir leur développement et freiner leur prolifération pourrait améliorer la stabilité internationale, compte tenu des risques de piratage ou d'accidents ainsi que des fausses perceptions découlant d'un fonctionnement sans réelle intervention humaine. L'occasion est encore propice pour s'attaquer à ces préoccupations, mais pourrait ne pas le rester longtemps. La concurrence croissante entre des armées d'avant-garde signifie également qu'on pourrait précipiter le recours à des systèmes d'IA pas tout à fait au point ou en déployer. Comme l'histoire l'a démontré, les pays déjà aux prises avec un conflit et ceux du Sud seraient probablement les premiers touchés par une telle militarisation.

Vu la faiblesse des mesures limitant l'accès à ces technologies, il importe d'élargir le bassin des parties prenantes qui s'intéressent à ces enjeux au-delà de celui qui s'est précédemment penché sur d'autres questions en matière de sécurité mondiale et nationale. La sensibilisation accrue du grand public au sujet des effets de ces technologies nécessite aussi urgemment une consolidation de la littératie numérique. Une législation internationale fondée sur les droits humains et le droit humanitaire, la littératie numérique ainsi qu'un public informé s'avèrent des éléments essentiels pour faire face aux menaces très réelles que représentent ces technologies. Un engagement diplomatique multilatéral est également nécessaire pour s'assurer que l'IA et les technologies émergentes restent un outil au service du bien social. En fin de compte, ce sont les gouvernements qui doivent et peuvent élaborer des politiques en vue de protéger les citoyens et citoyennes et de garantir une stabilité internationale.

RÉFÉRENCES

Abrams, A. 2020. *Black, Disabled and at Risk : The Overlooked Problem of Police Violence Against Americans with Disabilities*. Time. https://time.com/5857438/police-violence-black-disabled/

Adee, S. 2020. What Are Deepfakes and How Are They Created ? IEEE Spectrum. https://spectrum.ieee.org/what-is-deepfake

Alliance for Multilateralism. 2019. *Declaration by the Alliance for Multilateralism on Lethal Autonomous Weapons Systems (LAWS)*. https://multilateralism.org/wp-content/uploads/2020/04/declaration-on-lethal-autonomous-weapons-systems-laws.pdf

Allyn, B. 2020. *IBM Abandons Facial Recognition Products, Condemns Racially Biased Surveillance. NPR*. https://www.npr.org/2020/06/09/873298837/ibm-abandons-facial-recognition-products-condemns-racially-biased-surveillance

Altmann, J. 2009. Preventive Arms Control for Uninhabited Military Vehicles. In Capurro, R. and Nagenborg, M. (eds.). *Ethics and Robotics*. Heidelberg : AKA Verlag. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.569.9017&rep=rep1&type=pdf

Amnesty International. 2021. *A critical opportunity to ban killer robots while we still can*. https://www.amnesty.org/en/latest/news/2021/11/global-a-critical-opportunity-to-ban-killer-robots-while-we-still-can/

Asaro, P. 2012. On Banning Autonomous Lethal Systems : Human Rights, Automation and the Dehumanizing of Lethal Decision-making. Special Issue on New Technologies and Warfare. *International Review of the Red Cross*, Vol. 94, No. 886, pp. 687-709. https://www.cambridge.org/core/journals/international-review-of-the-red-cross/article/on-banning-autonomous-weapon-systems-human-rights-automation-and-the-dehumanization-of-lethal-decisionmaking/992565190BF2912AFC5AC0657AFECF07

Asher, S. 2021. *Myanmar coup : How Facebook became the 'digital tea shop'*. BBC News. https://www.bbc.com/news/world-asia-55929654

Bengio, Y. 2019. *AI pioneer : The dangers of abuse are very real.* Nature. https://www.nature.com/articles/d41586-019-00505-2

Bode, I. 2019. Norm-making and the Global South : Attempts to Regulate Lethal Autonomous Weapons Systems. *Global Policy,* No. 10, No. 3, pp. 359-364.

Breland, A. 2019. *The Bizarre and Terrifying Case of the "Deepfake" Video that Helped Bring an African Nation to the Brink*. Mother Jones. https://www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo/

Bugge, A. 2018. *U.N.'s Guterres urges ban on autonomous weapons.* Reuters. https://www.reuters.com/article/us-portugal-websummit-un-idUSKCN1NA2HG

Burton, J. et Soare, S. R. 2019. *Understanding the Strategic Implications of the Weaponization of Artificial Intelligence*. 11th International Conference on Cyber Conflict : Silent Battle T. Minárik, S. Alatalu, S. Biondi, M. Signoretti, I. Tolga, G. Visky (eds.) NATO CCD COE Publications, Tallinn. https://www.ccdcoe.org/uploads/2019/06/Art_14_Understanding-the-Strategic-Implications.pdf

Buolamwini, J. et Gebru, T. 2018. *Gender Shades : Intersectional Accuracy Disparities in Commercial Gender Classification.* MIT. https://www.media.mit.edu/publications/gender-shades-intersectional-accuracy-disparities-in-commercial-gender-classification/

Campaign to Stop Killer Robots. 2021. *Clear momentum towards a new legal framework on autonomous weapons*. https://www.stopkillerrobots.org/2021/08/clear-momentum-towards-a-new-legal-framework-on-autonomous-weapons/

CCW. 2019. *Meeting of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects.* CCW/MSP/2019/9 (voir les principes directeurs à l'annexe 3). https://documents-dds-ny.un.org/doc/UNDOC/GEN/G19/343/64/PDF/G1934364.pdf?OpenElement

CICR. 2021. *Position du CICR sur les systèmes d'armes autonomes.* Genève, CICR. https://www.icrc.org/en/download/file/178612/french_-_icrc_position_and_backgroud_paper.pdf

Citron, D. K. et Chesney, R. 2019. *Deep Fakes : A Looming Challenge for Privacy, Democracy, and National Security*. California Law Review (107 :1753). https://scholarship.law.bu.edu/faculty_scholarship/640

Cohen, Z. and Liebermann, O. 2021. *Rifles, Humvees and millions of rounds of ammo : Taliban celebrate their new American arsenal*. CNN. https://edition.cnn.com/2021/08/21/politics/us-weapons-arsenal-taliban-afghanistan/index.html

Commission africaine des droits de l'homme et des peuples. 2021. Résolution sur la nécessité d'élaborer une étude sur les droits de l'homme et des peuples et l'intelligence artificielle (IA), la robotique et d'autres technologies nouvelles et émergentes en Afrique – CADHP/Rés.473(XXXI) 2021. https://www.achpr.org/fr_sessions/resolutions?id=504

Commission européenne. 2021. *Proposition de règlement du Parlement européen et du Conseil établissant des règles harmonisées concernant l'intelligence artificielle (législation sur l'intelligence artificielle) et modifiant certains actes législatifs de l'Union.* https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206

Dastin, J. 2021. *Amazon extends moratorium on police use of facial recognition software.* Reuters. https://www.reuters.com/technology/exclusive-amazon-extends-moratorium-police-use-facial-recognition-software-2021-05-18/

Déclaration de Montréal pour un développement responsable de l'intelligence artificielle. 2018. https://www.declarationmontreal-iaresponsable.com/la-declaration

Devoto, M., Janssen, E. et Muñoz, W. 2021. A*nálisis de las propuestas y declaraciones de países del Sur Global sobre el marco normativo y operativo en el área de Sistemas de Armas Autónomas.* SEHLAC. https://bit.ly/2WD8ELQ

Díaz, M. et Muñoz, W. 2019. *Los riesgos de las armas autónomas : una perspectiva interseccional latinoamericana.* SEHLAC. https://bit.ly/ArmasInterseccionalidad

Dreifus, C. 2019. Toby Walsh, A.I. Expert, Is Racing to Stop the Killer Robots. *The New York Times.* https://www.nytimes.com/2019/07/30/science/autonomous-weapons-artificial-intelligence.html

Fowler, G. A. 2021. *Anyone with an iPhone can now make deepfakes. We aren't ready for what happens next*. Washington Post. https://www.washingtonpost.com/technology/2021/03/25/deepfake-video-apps/

Future of Life Institute. 2015. *Autonomous weapons : an open letter from AI & robotics researchers.* https://futureoflife.org/open-letter-autonomous-weapons/

Gould, L. 2021. *Remote Warfare : Interdisciplinary perspectives*. Safer World (podcast episode). https://www.saferworld.org.uk/multimedia/saferworldas-warpod-episode-8-remote-warfare-interdisciplinary-perspectives

GPAI. 2021. https://gpai.ai/projects/

Greene, J. 2020. *Microsoft won't sell its facial-recognition technology, following similar moves by Amazon and IBM*. The Washington Post. https://www.washingtonpost.com/technology/2020/06/11/microsoft-facial-recognition/

G20. 2019. *G20 Ministerial Statement on Trade and Digital Economy*. https://www.mofa.go.jp/files/000486596.pdf

Herby, M. 1998. *An international ban on anti-personnel mines : History and negotiation of the "Ottawa treaty"*. ICRC. https://www.icrc.org/en/doc/resources/documents/article/other/57jpjn.htm

Horowitz, M. C. 2020. *AI and the Diffusion of Global Power.* Centre for International Governance Innovation. https://www.cigionline.org/articles/ai-and-diffusion-global-power/

Human Rights Watch. 2012. *Losing Humanity : The Case against Killer Robots.* https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots

----. 2020. *Stopping Killer Robots : Country Positions on Banning Fully Autonomous Weapons and Retaining Human Control*. https://www.hrw.org/report/2020/08/10/stopping-killer-robots/country-positions-banning-fully-autonomous-weapons-and

----. 2021. *Areas of Alignment. Common Visions for a Killer Robots Treaty.* https://www.hrw.org/news/2021/08/02/areas-alignment

ICBL. 2021. *Treaty Status*. http://www.icbl.org/en-gb/the-treaty/treaty-status.aspx

IKV Pax Christi. 2011. *Does Unmanned Make Unacceptable ? Exploring the Debate on using Drones and Robots in Warfare.* https://paxvoorvrede.nl/media/download/does-u-make-ulowspreads_0.pdf

Keller, J. 2021. *Pentagon to spend $874 million on artificial intelligence (AI) and machine learning technologies next year*. Military Aerospace. https://www.militaryaerospace.com/computers/article/14204595/artificial-intelligence-ai-dod-budget-machine-learning

Kietzmann, J., Lee, L. W., McCarthy, I. P., Kietzmann, T. C. 2020. Deepfakes : Trick or treat ? *Business Horizons*, Vol. 63, No. 2, pp. 135-146, ISSN 0007-6813. https://doi.org/10.1016/j.bushor.2019.11.006.

Knowles, E. et Watson, A. 2021. *Remote Warfare : Lessons Learned from Contemporary Theatres*. SaferWorld. https://www.saferworld.org.uk/resources/publications/1280-remote-warfare-lessons-learned-from-contemporary-theatres

Koettl, C., Hill, E., Aikins, M., Schmitt, E., Tiefenthäler, A. et Jordan, D. 2021. *How a U.S. Drone Strike Killed the Wrong Person.* The New York Times, 10 septembre. https://www.nytimes.com/video/world/asia/100000007963596/us-drone-attack-kabul-investigation.html.

Marijan, B. 2018. Human-less or human more ? *The Ploughshares Monitor*, Vol. 39, No. 2, pp. 5-7.

Meskys, E., Liaudanskas, A., Kalpokiene, J. and Jurcys, P. 2021. Regulating deep fakes : legal and ethical considerations. *Journal of Intellectual Property Law & Practice,* Vol. 15, No. 01. pp. 24-31. DOI :10.1093/jiplp/jpz167.

Moyes, R. 2012. *Autonomous weapons – the risks of a management by 'partition.'* Article 36. https://article36.org/updates/autonomous-weapons-the-risks-of-a-management-by-partition/

Moyes, R. 2019. *Target profiles*. Article 36. http://www.article36.org/wp-content/uploads/2019/08/Target-profiles.pdf

Muñoz, W. 2021a. *Autonomous weapons systems : an analysis from human rights, humanitarian and ethical artificial intelligence perspectives*. https://bit.ly/SEHLACAI-AWS

----. 2021b. *It's about more than autonomous weapons systems*. Armed Conflict and Civilian Protection Initiative of the International Human Rights Clinic, Harvard Law School. https://humanitariandisarmament.org/2021/08/30/it-is-about-more-than-autonomous-weapons-systems/

Nations Unies. Conseil de Droits de l'Homme. 2013. *Rapport du Rapporteur spécial sur les exécutions extrajudiciaires, sommaires ou arbitraires, Christof Heyns.* A/HRC/23/47 (9 avril). https://www.ohchr.org/Documents/HRBodies/HRCouncil/RegularSession/Session23/A-HRC-23-47_fr.pdf (consulté le 4 octobre 2021).

Nobel Women's Initiative. 2014. *Nobel Peace Laureates call for Preemptive Ban on "Killer Robots".* https://nobelwomensinitiative.org/nobel-peace-laureates-call-for-preemptive-ban-on-killer-robots/

O'Brien, M. 2019. *Why 'deepfake' videos are becoming more difficult to detect*. PBS NewsHour. https://www.pbs.org/newshour/show/why-deepfake-videos-are-becoming-more-difficult-to-detect

OCDE. 2019. *Recommandation du Conseil sur l'intelligence artificielle*. https://legalinstruments.oecd.org/fr/instruments/OECD-LEGAL-0449 (consulté le 5 octobre 2021).

OECD.AI Policy Observatory. n.d. *OECD AI principles overview.* https://www.oecd.ai/dashboards/ai-principles/P6

Olejnik, L. 2021. *TechLetters #23 – Deepfakes OK ? Vulnerable IoTs. SolarWind hacks in Europe. Cyber sanctions, Russia twice. Hacked cheese*. https://techletters.substack.com/p/techletters-23-deepfakes-ok-vulnerable?s=r

Palmer, C. s.d. *Ethical Theory and Philosophical Method Feminist Ethic*s. Lancaster University. https://www.lancaster.ac.uk/users/philosophy/awaymave/401/feminist.htm

Pantserev, K. A. 2020. The Malicious Use of AI-Based Deepfake Technology as the New Threat to Psychological Security and Political Stability. In Jahankhani, H. et al. *Cyber Defence in the Age of AI, Smart Societies and Augmented Humanity*, pp. 37-55. Springer.

Pax Netherlands. 2019. *Killer robots : what are they and what are the concerns.* https://paxforpeace.nl/media/download/pax-booklet-killer-robots-what-are-they-and-what-are-the-concerns.pdf

Ramsay-Jones, H. 2019. *Racism and Fully Autonomous Weapons*. https://www.ohchr.org/Documents/Issues/Racism/SR/Call/campaigntostopkillerrobots.pdfReaching Critical Will. 2021. *CCW Report*. https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2021/gge/reports/CCWR9.4.pdf

République de la Corée. 2020. *ROK and UNESCO co-organize Virtual Asia-Pacific Consultation on UNESCO Recommendation on the Ethics of Artificial Intelligence.* Ministry of Foreign Affairs. https://www.mofa.go.kr/eng/brd/m_5676/view.do?seq=321173&srchFr=&amp%3BsrchTo=&amp%3BsrchWord=&amp%3BsrchTp=&amp%3Bmulti_itm_seq=0&amp%3Bitm_seq_1=0&amp%3Bitm_seq_2=0&amp%3Bcompany_cd=&amp%3Bcompany_nm=

Roy, S. et Miniter, R. 2021. *Taliban kill squad hunting down Afghans — using US biometric data*. New York Post. https://nypost.com/2021/08/27/taliban-kill-squad-hunting-afghans-with-americas-biometric-data/

Russel, S. 2021. *It's time to ban autonomous killer robots before they become a threat.* Financial Times. https://www.ft.com/content/04a07148-d963-4886-83f6-fcaf4889172f

Sauer, F. 2021. Stepping back from the brink : Why multilateral regulation of autonomy in weapons systems is difficult, yet imperative and feasible. ICRC. https://international-review.icrc.org/articles/stepping-back-from-brink-regulation-of-autonomous-weapons-systems-913

Sharkey, A. 2019. Autonomous weapons systems, killer robots and human dignity. *Ethics and Information Technology.* Vol. 21, pp. 75–87.

Sharkey, N. 2007. *Robot Wars are a Reality*. The Guardian. https://www.theguardian.com/commentisfree/2007/aug/18/comment.military

UNESCO. 2019. *Charter of Ethics of Science and Technology in the Arab Region*. https://unesdoc.unesco.org/ark:/48223/pf0000372169

———. 2020. *Outcome document : first draft of the Recommendation on the Ethics of Artificial Intelligence*. SHS/BIO/AHEG-AI/2020/4 REV.2. https://unesdoc.unesco.org/ark:/48223/pf0000373434

———. 2021. *Élaboration d'une Recommandation sur l'éthique de l'intelligence artificielle*. https://fr.unesco.org/artificial-intelligence/ethics (consulté le 5 octobre 20021).

UNIDIR. 2020. *Modernizing Arms Control.* https://unidir.org/publication/modernizing-arms-control

UNIDIR. 2021. *The 2021 Innovations Dialogue : Deepfakes, Trust And International Security*. https://unidir.org/events/2021-innovations-dialogue

Venema, A. E. 2020. Deepfakes as a Security Issue : Why Gender Matters. WIIS Global. https://wiisglobal.org/deepfakes-as-a-security-issue-why-gender-matters/

Wallach, W. 2013. *Terminating the Terminator*. Science Progress. http://bit.ly/2mjl2dy

Wiggers, K. 2021. *Fewer than 30 % of business have a plan to combat deepfakes, survey finds*. VentureBeat. https://venturebeat.com/2021/05/24/less-than-30-of-business-have-a-plan-to-combat-deepfakes-survey-finds/

ÉTHIQUE DU *CARE* ET INTELLIGENCE ARTIFICIELLE : LA NÉCESSITÉ D'INTÉGRER UNE APPROCHE NORMATIVE FÉMINISTE

PAULINE NOISEAU

Chercheuse en éthique de l'intelligence artificielle (IA) à l'Université de Montréal.

ODD 5 - L'égalité entre les hommes et les femmes

ODD 10 - Réduction des inégalités

ODD 11 - Villes et communautés durables

ODD 16 - Paix, justice et institutions fortes

ODD 17 - Partenariats pour les objectifs

ÉTHIQUE DU *CARE* ET INTELLIGENCE ARTIFICIELLE : LA NÉCESSITÉ D'INTÉGRER UNE APPROCHE NORMATIVE FÉMINISTE

RÉSUMÉ

Depuis quelques années, nous assistons à un florilège de dispositifs vis-à-vis de l'IA : chartes, déclarations, ensemble de principes éthiques, etc. Ces structures s'inspirent directement de philosophies morales dominantes pareillement au déontologisme, ou encore à l'éthique conséquentialiste. Ainsi, la dimension éthique de certains usages de l'IA est déterminée par ces philosophies morales dont les prémisses et les représentations qu'ils contiennent ne sont que très rarement remises en question. Il s'agira donc dans cet article de bouleverser ce que nous appelons « la saturation éthique de l'IA » en faisant appel à une théorie réellement alternative à savoir l'éthique du *care* en tant qu'elle nous invite à changer de regard et à adopter de nouvelles perspectives critiques sur l'IA. Les éthiques du *care* marquent une rupture nette en posant comme critère de moralité de l'action le soin apporté à autrui, la considération de l'interdépendance du vivant et la prise en charge de la vulnérabilité comme caractère inhérent à l'espèce humaine (Brugère, 2011). Les éthiques du *care* sont jugées féministes puisqu'elles mettent au cœur du politique et de l'action collective le *care* comme socle du vivant, effectué traditionnellement et majoritairement par les femmes et les personnes historiquement marginalisées dans notre société, de manière gratuite ou peu reconnue (Gilligan, 2011). Nous nous poserons les questions suivantes : comment expliquer, dès lors, que les éthiques du *care* constituent un véritable angle mort des réflexions en éthique de l'IA ? L'IA est-elle éthique, si nous adoptons une perspective d'éthique du *care* ? Cette proposition est importante, car elle permet d'une part d'élargir nos perspectives éthiques et critiques vis-à-vis de l'IA et d'autre part, d'améliorer les politiques publiques en la matière afin de les rendre plus justes et plus inclusives.

INTRODUCTION

> To understand what is at stake, we must focus less on ethics and more on power. AI is invariably designed to amplify and reproduce the forms of power it has been deployed to optimize. Countering that requires centering the interests of the communities most affected. Instead of glorifying company founders, venture capitalists, and technical visionaries, we should begin with the lived experiences of those who are disempowered, discriminated against and harmed by AI systems.
>
> Kate Crawford (2021). *Atlas of AI*. p. 224-225

> Mais une éthique du *care*, avec les exigences morales d'attention et de responsabilité qui l'accompagnent, pourrait permettre de dévoiler la manière dont les puissants tentent de fausser la compréhension des besoins pour maintenir leurs privilèges et leurs positions de pouvoir.
>
> Joan Tronto (2009). *Un monde vulnérable. Pour une politique du care*. p.198

Nous pouvons définir l'intelligence artificielle (IA) comme l'ensemble des systèmes informatiques qui permettent de simuler et de reproduire certaines fonctions de l'intelligence humaine pareillement à la mémorisation, à l'apprentissage ou encore au calcul (Boden, 1990). Si l'IA existe depuis bien longtemps, elle fait parler d'elle comme technique particulièrement saisissante depuis les travaux sur l'apprentissage profond (Goodfellow et al., 2016). Cette dernière prouesse technique permet notamment à travers un réseau de neurones – directement inspiré du cerveau humain – d'atteindre des degrés de capacités complexes encore jamais égalées en termes de rapidité et d'exactitude. L'IA est ainsi utilisée pour remplacer certaines tâches qui pourraient, en étant suppléées par celle-ci, libérer du temps et de l'énergie à l'être humain. Les bienfaits sont donc principalement de l'ordre de l'opérationnalisation, c'est-à-dire, le fait de rendre plus productif un système et ainsi de gagner en termes quantitatifs et qualitatifs. Certaines utilisations ou usages de l'IA sont largement reconnus comme étant positifs en tant qu'ils facilitent, en d'autres termes rendent plus accessibles certaines actions, augmentant de fait le pouvoir des êtres humains, entendu ici comme capacité à faire et à agir.

Toutefois, certains usages de l'IA ont été identifiés comme coûteux à différents niveaux : sociaux, politiques ou encore environnementaux. Nous pensons notamment à l'empreinte carbone liée à la maintenance des algorithmes, aux biais algorithmiques, aux mégadonnées de la surveillance, mais encore aux atteintes à l'émancipation humaine à travers des procédures de *nudging* et d'orientations de choix.

C'est à ce moment-là que l'éthique est entrée en scène comme discipline permettant de déterminer la valeur morale de certaines utilisations afin d'en limiter ou d'en maîtriser le développement. Nous pouvons définir l'éthique de l'IA comme le domaine qui « (...) tente de réfléchir, d'identifier et de proposer une utilisation de l'IA qui soit en accord avec une *manière d'être commune*, c'est-à-dire un ensemble de valeurs et de principes qui sont spécifiques à une société » (Noiseau, Mörch et al., 2021).

Malgré l'implication de l'éthique pour guider et orienter de manière responsable le développement de l'IA, force est de constater que ces propositions normatives n'ont pas permis de faire en sorte que le développement de l'IA soit en accord avec les enjeux collectifs et sociaux auxquels nous faisons face. De plus, on observe une grande similarité dans les analyses proposées par les éthiques de l'IA, tant au niveau formel qu'au niveau conceptuel. Au niveau formel, ces analyses éthiques prennent la forme d'un ensemble de recommandations, principes ou valeurs éthiques à appliquer directement sur des cas d'usage ou des pratiques déjà existantes. Au niveau conceptuel, les contenus se rejoignent largement en soulevant pour la plupart, les mêmes enjeux et mêmes valeurs (Jobin et al., 2019). Cette homogénéisation normative peut s'expliquer de deux manières.

Premièrement, elle rend compte d'une domination de certaines théories morales au profit d'autres qui expriment une même manière de percevoir et de comprendre le monde. Deuxièmement, elle n'existe que parce que la discipline de l'éthique de l'IA n'use pas d'une théorie morale suffisamment alternative pour bouleverser les prémisses philosophiques, la forme du raisonnement moral et les conclusions éthiques qui s'en suivent. En effet, l'éthique de l'IA semble s'être arrêtée historiquement à des structures vieillissantes, directement inspirées de la modernité, qui partent d'une conception libérale de l'être humain, en d'autres mots, rationnel, autonome et indépendant. Or, si les éthiques de l'IA sont peu capables de cibler ou d'identifier les ramifications négatives de certains usages, voire même de remettre en cause la pertinence et ainsi l'existence de cette technologie dans notre écosystème global de vie, c'est parce qu'elles passeraient, à notre sens, à côté d'une activité essentielle de l'être humain, à savoir le *care*[88].

Nous pouvons rassembler les activités de soin par le terme de *care* qui englobe une grande part des pratiques qui permettent de maintenir la vie. En effet, Tronto (1993) dans un ouvrage publié sous le titre original *Moral Boundaries. A Political Argument for an Ethics of Care* distingue quatre phases au *care* : se soucier de (*caring about*) ; prendre en charge (*taking care of*) ; prendre soin (*care giving*) ; recevoir le soin (*care receiving*) (Tronto, 2009, p. 147). Explicitons ces différentes étapes.

Il s'agit dans un premier temps d'être capable de reconnaître l'existence d'un besoin (1). La première disposition et pratique du *care* est donc de l'ordre de l'attention, c'est-à-dire l'attitude à s'engager consciemment avec l'autre, à percevoir des signes explicites ou non d'un besoin. Par exemple, reconnaître le besoin pour une personne à parler d'un évènement traumatisant ou d'être pris en charge par un.e professionnelle de la santé. Négativement, ne pas faire attention à l'autre – ne pas s'en soucier – aboutit ainsi à un échec du *care* car la première étape en saurait manquée.

Ensuite, après avoir conscientisé et reconnu le besoin de l'autre, la deuxième étape consiste à prendre en charge cette nécessité (2). La personne engagée se doit d'assumer la responsabilité qui lui incombe et d'identifier la réponse à apporter à la présente situation. La prise en charge peut se solder par l'apport d'une structure matérielle spécifique permettant la réponse au besoin. Nous disons par exemple qu'un père prend en charge sa famille en travaillant, en lui apportant les ressources matérielles nécessaires pour sa survie.

Delà, apparaît la troisième étape qui est celle du prendre soin, autrement dit, la réponse concrète et directe au besoin (3). Il faut ici distinguer la prise en charge du prendre soin. Si le médecin prend en charge la personne patiente en lui administrant un traitement, ce sont les personnes infirmières qui prennent soin de la personne patiente en réalisant concrètement le traitement et en engageant la relation. Si un père prend en charge les besoins de la famille en ramenant un salaire à la maison, il ne prend pas soin de celle-ci, car il faut bien matériellement préparer les repas, laver les corps, écouter les enfants, etc.

Enfin, le *care* se termine par l'étape de la réception du soin (4). En d'autres termes, il s'agit de vérifier si le soin qui a été effectué correspondait bien au besoin initial. Le soin est caduc s'il n'est pas reconnu comme étant complet par le ou la bénéficiaire. La dernière étape permet notamment d'assurer une responsabilité temporelle de la part du pourvoyeur ou de la pourvoyeuse du soin. Une institution qui aurait la prétention de prendre soin d'une catégorie de la population sans valider la réception et l'adéquation du soin apporté au besoin identifié aurait échoué dans son travail de *care*.

88. Le terme de « care » sera conservé dans notre étude pour des raisons épistémiques dans la mesure où le concept de care en anglais couvre une plus grande panoplie d'interprétations relatives à l'action même de soin, et parce que le *care* englobe des actions qui dépassent la signification du concept de soin en français. Si le concept de *care* en anglais est largement accepté par les chercheurs et chercheuses en éthique, on pourra toutefois trouver quelques exceptions françaises pareillement aux termes de soin, ou encore de sollicitude. Nous pensons notamment à l'ouvrage de Fabienne Brugère, *Le sexe de la sollicitude* publié en 2008 aux Éditions du Seuil.

Ces quatre phases permettent de dire ce que le *care* est et ce qu'il n'est pas. Autrement dit, il est à la fois une disposition et un travail (Tronto, 2009, p.145). Affirmer qu'il est un travail permet d'évacuer l'esthétique de la vocation de celui-ci et sa dimension *a priori* sentimentale (Paperman, 2021). Pensons notamment au discours assurant que les femmes ou les personnes immigrantes seraient par nature plus enclines à prendre soin des autres. Au contraire, dire que le *care* est un travail permet de montrer de quelle manière les personnes privilégiées se seraient déchargées de leur responsabilité de *care* à l'égard des autres membres de la société en faisant de celui-ci une activité sans valeur aucune (Gilligan et al., 2013). Certaines situations de vie forcent en effet l'étonnement : comment est-il aujourd'hui possible de travailler plus de 50h par semaine dans une entreprise et avoir une famille de trois enfants, tout en étant responsable de la vie des autres ? Nous pouvons naturellement nous poser les questions suivantes : qui prend soin des enfants ? Du foyer ? Qui lave les draps ? Qui fait le ménage ? Qui achète la nourriture ? Qui prépare les repas ? Bref, qui fait en sorte que la vie soit vie ?

Dès lors, on peut dire que les activités de *care* sont à la fois partout et nulle part. Partout, car ce sont bien ces actions qui permettent de nous maintenir en vie et qui soutiennent le système social et économique (secteurs de la santé, de l'éducation, de la propreté, des besoins de première nécessité) ; nulle part, car dévalorisées et relayées à certaines catégories de la population qui effectuent ce travail dans l'ombre pendant que d'autres continuent à jouir de leurs privilèges de ne pas se soucier des autres[89].

Depuis les années 1980, ces pratiques de *care* font l'objet de recherches importantes dans différents milieux pareillement aux sciences sociales et humaines. Ce qu'on appelle aujourd'hui l'éthique du *care* est apparu après la publication d'un ouvrage devenu célèbre par sa portée révolutionnaire, *In a Different Voice* de Carol Gilligan publié dans sa version originale en 1982 aux États-Unis. Dans ce livre, l'autrice remet en cause la hiérarchie du développement moral proposée par la psychologie d'alors, dont Lawrence Kohlberg, pour qui le summum de la maturité morale équivaut à la capacité pour un être humain à formuler activement des principes de justice abstraits et universels. Selon lui et ses collègues du domaine, le raisonnement moral se comprendrait en termes de droits et de devoirs, à l'aune d'une conception de la justice extérieure à son milieu de vie particulier. Carol Gilligan, dans son travail, va profondément dépasser cette éthique en distinguant une autre forme de raisonnement moral. Cette autre pensée éthique adviendrait à partir d'une expérience spécifique, qui serait liée au genre, puisque les femmes en seraient détentrices. Effectivement, elle montre que les femmes n'ont pas la même compréhension, réflexion et réponse lorsqu'il s'agit de bien agir face à un dilemme moral. Les femmes, selon Gilligan, comprendraient l'agir moral en termes de responsabilité apportée à autrui dans le but d'entretenir un écosystème de vie. Cette autre forme de raisonnement moral qu'elle appelle l'éthique du *care*, s'opèrerait donc, non pas en termes d'adéquation à des règles abstraites et universelles, mais se réaliserait en vue de garantir un réseau relationnel interdépendant et inscrit dans le temps. L'indifférence à l'égard des relations et des besoins des autres deviendrait dès lors, à partir d'une perspective du *care*, une carence, et non le symbole d'une maturité morale.

En posant comme critère de moralité de l'action la préservation d'un écosystème interdépendant au niveau relationnel et délimité dans un contexte particulier en vue de préserver la vie et ses qualités, les éthiques du *care* bouleversent radicalement les théories morales. En effet, les éthiques du *care* permettent d'attribuer une dimension morale à ce qui auparavant ne faisait l'objet d'aucune préoccupation, c'est-à-dire, les pratiques ordinaires de soin, réalisées dans le silence et dans

89. La situation pandémique de la Covid-19 est tout à fait révélatrice de cette inégalité quant à la responsabilité de *care* dans notre société. Le confinement généralisé a montré que certains travaux étaient essentiels au fonctionnement de la société : nous pensons ici aux personnels de la santé, de l'éducation, aux caissiers et caissières, aux agent.e.s d'entretien, aux secteurs de l'alimentation, etc. Nous pouvons également faire référence ici aux travaux de Silvia Federici et spécifiquement son fameux ouvrage *Le capitalisme patriarcal*, publié en 2019 aux éditions La Fabrique. Dans cet ouvrage, l'autrice y montre de quelle manière le système économique capitaliste s'est construit sur le travail gratuit des femmes dans la sphère privée notamment à travers ce qu'elle appelle l'invention de la ménagère (voir p. 125).

l'indifférence générale, car cantonnées à la sphère privée, du sentimental et du relationnel (Tronto, 2009, voir ici le deuxième chapitre « Contre la « moralité des femmes » »). Ce qui était reconnu comme la forme la plus élevée de la réflexion morale, en d'autres termes la coïncidence avec des principes de justice éloignés des structures matérielles et relationnelles d'existence, devient d'un coup, grâce aux éthiques du *care*, *une* forme de raisonnement moral, et non la seule. L'éthique du *care* vient donc botter en touche l'éthique de la justice et ses fondements en valorisant une approche particulariste, relationnelle et contextuelle de l'action morale. Devient éthique, non pas celui ou celle qui agit conformément à des règles désincarnées, mais au contraire, celui ou celle qui permet, à travers des pratiques de *care*, de préserver concrètement et clairement un écosystème vivant.

Si le *care* peut naturellement se rapporter aux pratiques exercées par les proches aimants et aidants, il s'étend à d'autres sphères, comme le définissent Tronto et Fisher (2009, [1991], p. 40) :

> Au niveau le plus général, nous suggérons que le « prendre soin » (caring) soit considéré comme une activité générique qui comprend tout ce que nous faisons pour maintenir, perpétuer et réparer notre « monde », de sorte que nous puissions y vivre aussi bien que possible. Ce monde comprend nos corps, nous-mêmes et notre environnement, tous éléments que nous cherchons à relier en un réseau complexe, en soutien à la vie.

Rappelons également que ce qui pourrait s'apparenter à une forme d'essentialisme moral ne l'est pas. Effectivement, il ne s'agit pas d'affirmer qu'il existerait une morale des femmes, mais à l'inverse, que celle-ci émanerait de conditions politiques et matérielles propres à un sujet (Tronto, 2009). Si les femmes ont été amenées à opérer un raisonnement moral basé sur le *care* c'est parce qu'elles ont été historiquement prédisposées à réaliser ce travail de prise en charge des besoins des autres dans l'enceinte familiale.

L'éthique du *care* s'inscrit dans une perspective féministe puisqu'elle dépasse la dualité en morale et ainsi la hiérarchie morale qui la sous-tend (Gilligan, 2010). Il s'agit également de reconnaître la dimension éthique des activités réalisées par les femmes et les personnes historiquement marginalisées dans notre société. Ce que Joan Tronto appelle notamment « le pouvoir des pauvres » est absolument essentiel au fonctionnement des institutions et de la mise en marche des activités puisque sans *ces petites mains qui prennent soin de nous*, c'est l'ensemble du système qui s'écroule. Le *care* est donc un brillant outil, critique et révolutionnaire, qui permet de remettre les pendules à l'heure en identifiant les vrais besoins d'une société d'une part, et qui permet de voir d'autre part, de quelle manière les puissants et les privilégiés se dé-saisissent des questions liées à la charge du *care* pour ne pas être responsables du soin des autres (Hamrouni, 2015 ; Tronto, 2009).

Dès lors, comment expliquer la quasi-absence de perspectives critiques sur l'IA à partir des éthiques du *care*[90] ? Qu'y aurait-il *à perdre ou à gagner* à questionner certains usages de l'IA à partir du *care* ? Si nous vivions dans une structure éthique du *care*, entendue comme forme de vie, l'IA trouverait-elle toujours sa raison d'être ? Pour le dire autrement, à quoi ressemblerait une analyse éthique sur l'IA à partir du *care* ? L'IA pourrait-elle être juste si nous adoptions une *manière d'être commune* basée sur une conception de l'être humain comme naturellement vulnérable, dont les degrés de besoins de soin sont variables tout au long de l'existence, engagé dans un réseau interdépendant et relationnel du vivant ?

90. Nous pensons toutefois ici aux travaux de Vanessa Nurock. Je tiens par ailleurs à la remercier pour nos brefs échanges qui ont été forts utiles pour la rédaction de cet article.

    Notons également l'analyse qui avait été proposée par le laboratoire de recherche Algora Lab – Université de Montréal et Mila – Institut Québecois d'intelligence artificielle de la délibération internationale – dont j'ai été la coordonnatrice scientifique – sur l'instrument normatif proposé par l'UNESCO. Dans ce rapport, nous soulevions le manque d'une conception éthique du *care* dans la version préliminaire de la Recommandation. Voir spécifiquement le point 2.5 à la page 18 du rapport d'analyse (Algora Lab, 2020).

Nous proposerons dans ce chapitre d'analyser les enjeux éthiques et sociaux de l'IA à partir d'une approche normative féministe qui reconnaît le *care* comme étant au cœur de notre vie humaine. Nous verrons dans un premier temps que l'ontologie qui sous-tend le développement de l'IA se rapporte à une conception libérale de l'autonomie et de l'être humain. Dans un deuxième temps, nous proposerons une voie de sortie à ce que nous appellerons la *saturation éthique en IA* en cherchant à voir autrement et au-delà des considérations actuelles. Enfin, nous montrerons que l'approche du risque devrait être dépassée en ce qui concerne l'IA en vue d'épouser une approche basée sur la responsabilité et l'attention dans un monde en crise[91].

CONTRE L'INDIVIDU AUGMENTÉ : VULNÉRABILITÉ ET INTERDÉPENDANCE

Une des premières choses à prendre en compte lorsqu'il s'agit d'adopter une perspective éthique et critique à propos d'un objet quelconque est de définir ou d'identifier la conception de l'humanité ou de la vie que cette théorie morale supporte. Toute philosophie morale part d'un ensemble de considérations sur l'espèce humaine. Effectivement, pour bâtir une théorie sur la qualité de la vie, une forme de vie ou manière de vivre, bref une éthique, faut-il déjà définir ce qu'on entend par humanité. En d'autres termes, on ne saurait proposer une configuration à un sujet sans que celui-ci n'ait fait l'objet d'une définition au préalable. Pour parler d'éthique, il faut d'abord et avant tout détenir à l'origine une conception de l'humanité, qui fait office de prémisse première à la réflexion ou raisonnement moral qui s'ensuit. Ces représentations doivent être soumises à la critique. Qu'il s'agisse de l'éthique des vertus d'Aristote, de l'éthique déontologique de Kant, mais encore des éthiques conséquentialistes, chacune prend racine au sein d'une définition de l'être humain, qu'il est nécessaire d'interroger.

Si les perspectives éthiques sur l'IA font florès, elles reposent toutes sur des théories morales dominantes qui partent d'une même conception de l'être humain, autrement dit, rationnel, autonome et indépendant. Ce que critique l'éthique du *care*, c'est spécifiquement la prémisse libérale sur laquelle ces théories morales reposent.

En effet, à partir de la modernité, ce qui a été premièrement reconnu comme étant la principale caractéristique de l'être humain, c'est sa rationalité. Pour le dire autrement, l'humain serait profondément et essentiellement un être doué de raison. Or, la raison a été opposée aux sentiments, aux émotions, aux relations, en résumé : à l'ensemble des choses qui nous ramènent matériellement à la vie, pour au contraire, viser une abstraction et une représentation conceptuelle de la réalité. L'objectivité serait donc le lieu visé d'une disposition proprement rationnelle dans laquelle l'on traiterait d'un objet extérieur à soi, de manière neutre et impartiale. La morale a également été associée à cette rationalité et cette impartialité. On ne peut être moral que si nous agissons en vue d'un bien universel. Rappelons ici que pour Kohlberg, la maturité morale est associée à la capacité pour un individu à formuler des principes de justice abstraits qui conviennent à tous/tes. Il faudrait donc, pour être moralement juste, porter le voile, pour le dire en termes rawlsiens, c'est-à-dire, se désubjectiver, détenir ce qu'on appelle « un point de vue moral », un non-lieu-politique et social, bref *être un inconnu au milieu de nulle part*. La moralité se penserait donc ici à partir d'un individu isolé du monde, autonome, rationnel et indépendant vis-à-vis des autres.

91. Lorsque nous utilisons les termes « monde en crise », nous faisons explicitement référence à la crise climatique (GIEC, 2021). Voir les travaux de : Debourdeau, 2013 ; Servigne et Stevens (2015) ; les écoféministes (Hache, 2016, Starhawk 2019). Ces recherches et données sur l'environnement nous invitent à changer de paradigme dans tous les domaines, y compris l'éthique. C'est là que l'éthique du care prend tout son sens aujourd'hui

C'est spécifiquement cette conception de l'être humain que les éthiques du *care* critiquent. Brugère dans *Le sexe de la sollicitude* examine la représentation fantasmée dans la littérature philosophique d'un être humain qui n'a besoin de personne pour se construire en reconnaissant notamment que « L'individu indépendant est bien l'une des grandes fictions théoriques de notre mythologie occidentale. » (Brugère, 2008, p. 50). En effet, la vulnérabilité, c'est-à-dire la capacité pour une personne à être heurtée et blessée est complètement évacuée des théories morales dominantes puisqu'elle est associée à une forme de fragilité qui pourrait porter atteinte à la formulation d'un jugement indépendant et neutre, et en ce sens juste. Or, ce qu'affirment les éthiciennes du *care*, c'est le caractère essentiel et non exceptionnel de la vulnérabilité humaine (Paperman, 2011). Nous avons tous et toutes fait l'expérience de la vulnérabilité dans la mesure où nous avons eu besoin de recevoir du soin pour rester en vie. Que ce soit le nourrisson, l'adolescent, le malade, l'endeuillé, ils/elles ont besoin, pour poursuivre leur vie, de *care*. Il est par conséquent nécessaire, et non contingent. Cette première reconnaissance ontologique nous permet d'affirmer d'autres éléments. En effet, si nous avons intrinsèquement besoin de *care*, alors nous sommes de manière absolue reliés aux autres, sous la forme d'une interdépendance ou d'une interconnexion (Perreault, 2015). Le vivant dans sa globalité, incluant ici les animaux non humains et les territoires vivants sont également vulnérables (Laugier, 2012). De ce fait, il faut comprendre le monde comme un vaste réseau de vulnérabilités, variables et particulières, qui, à travers des liens et des relations de *care*, permettent de maintenir et de préserver ce que nous avons en commun, à savoir la vie.

Dès lors, nous avons en main un outil critique des plus puissants pour questionner non pas les conclusions éthiques sur l'IA, mais bien les prémisses sur lesquelles elles reposent, les rendant de fait, caduques, voire complètement nulles. En effet, nous pourrions tout à fait remettre en cause toute perspective actuelle en éthique sur l'IA dans la mesure où elle se fonderait sur une conception artificielle de l'être humain, voire complètement fantasmée. On pourrait donc affirmer ici que toute éthique qui omettrait les activités de *care* dans son raisonnement louperait sa cible car elle passerait à côté de l'activité essentielle de la vie humaine, ce qui lui permet d'être.

Dès lors, comment peut-on croire que ces raisonnements moraux puissent être à la hauteur de nos exigences civilisationnelles et technologiques, en tant qu'ils ne reconnaissent *même* pas le propre de la nature humaine, à savoir son appartenance à la communauté du vivant et son besoin de *care*. Après avoir critiqué les prémisses ontologiques des éthiques dominantes, nous nous interrogerons sur la forme de leur raisonnement moral et donnerons des pistes de sortie de la *saturation éthique en IA*.

SORTIR DE LA SATURATION ÉTHIQUE : DU PRINCIPE À L'ORDINAIRE

Nous observons depuis quelques années déjà, ce que nous appellerons une saturation éthique dans le domaine de l'IA. Nous pouvons définir la saturation éthique comme l'effet d'une redondance accrue de certains concepts, jugements et propositions vis-à-vis de l'IA, rendant de ce fait compte d'une même vision du monde et de ses valeurs fondamentales. Cette saturation est d'autant plus visible lorsque nous observons la quantité astronomique de cadres normatifs développés en vue d'encadrer l'IA qui sont pour la plupart tous similaires en termes de formes et de contenus (Voarino, 2019, p. 170). On observe en ce sens une forme de consensus quant aux principes à appliquer dans le domaine de l'IA (Jobin, Ienca, Vayena, 2019). Cette saturation éthique peut s'expliquer de deux manières.

D'abord, ces cadres normatifs s'inspirent du raisonnement propre aux théories de la justice. En effet, une perspective de justice vise avant tout à identifier des principes abstraits jugés supérieurs, car *transcendant* toute forme de spécificité ou particularité matérielle. Le juste ou le bon se validerait donc à l'aune d'une adéquation à des principes universels, éloignés de toute vie ordinaire. Cet ensemble de règles pourrait, selon une perspective de justice, permettre le bon fonctionnement du monde.

De plus, les théories de la justice usent d'un procédé qui permet de s'affranchir des conditions politiques d'existence, c'est ce qu'on appelle « point de vue moral ». Ce point de vue moral est utilisé comme l'œil ou le regard de celui ou de celle qui serait capable de penser en dehors du monde et d'adopter un point de vue extérieur, pour, justement, savoir ce qu'il est bon ou non de faire.

Toutefois, la performativité de ces principes dans le monde, en d'autres termes, la manière dont ceux-ci se déploient dans le champ de l'existence reste à créer, à déterminer, bref à imaginer pour les sujets et sujettes de ce monde. Si la délibération peut être un outil intéressant pour penser le processus qui va de l'abstraction des principes de justice à leur matérialisation sociale et culturelle (Noiseau et al., 2021), il reste que, à l'aune des vertus du modèle de *care*, cette forme de fonctionnement moral est incomplète. Présentons plusieurs aspects pertinents de l'éthique du *care* qui permettent de répondre à ces incomplétudes.

Premièrement, comme l'a montré Gilligan dans ses différents travaux et spécifiquement à partir d'un ensemble d'entretiens réalisés auprès de femmes ayant subi un avortement (1982), agir de façon juste ne consisterait pas *uniquement* à suivre un ensemble de règles abstraites, en prenant en compte ses droits et ses devoirs, mais à inscrire *différemment* sa *voix* dans un contexte de vie marqué par les relations et les émotions : « Les femmes perçoivent le dilemme moral comme un problème de responsabilités et de préoccupations du bien-être (*care*) de l'autre, et non comme une question de droits et de règles. » (Gilligan, 2019, p. 117). En d'autres termes, il y a renversement du paradigme en morale puisqu'il ne s'agit plus de partir du haut vers le bas, du principe vers l'action réelle, mais d'opérer un va-et-vient responsable entre soi et les autres. Pour le dire d'une autre façon, il s'agit de se poser systématiquement les questions suivantes : de quelle manière puis-je prendre à la fois soin des autres et de moi ? Suis-je en train de maintenir et de poursuivre la vie ? Ai-je été responsable vis-à-vis des autres ? Comment puis-je conserver les liens ? L'éthique du *care*, devient ainsi un autre idéal type normatif à distinguer de celui des théories de la justice (Clement, 1996 ; Perreault, 2015). Cet idéal type se focalise sur le contexte de vie du sujet plutôt que sur un non-lieu désincarné et abstrait ; privilégie les liens de connexion entre les membres d'un écosystème à la place d'une posture indépendante vécue comme séparation vis-à-vis du monde ; et enfin, reconnaît le critère de moralité d'une action comme la capacité à maintenir des relations humaines au lieu de garantir un principe d'égalité formelle et juridique.

Deuxièmement, l'éthique du *care* permet de situer la création théorique en morale dans le monde matériel. En effet, comme le montre Joan Tronto dans un ouvrage publié sous le titre original *Moral Boundaries. A Political Argument for an Ethics of Care* (1993), « La théorie morale n'est pas indépendante des conditions sociales et historiques. » (2009, p. 93). En d'autres mots, il s'agit d'affirmer le caractère profondément politique de qualification de ce qui est juste et de ce qui est bon à faire dans le monde. En effet, Tronto reconnaît l'incapacité pour les philosophes à répondre aux enjeux actuels auxquels nous faisons face[92] par le fait que ces théories se sont volontairement désinscrites des conditions qui permettaient de soutenir la vie de ce monde : « L'ironie en est que c'est précisément la force de la théorie morale universaliste, son détachement du monde, qui la disqualifie pour résoudre les types de problèmes moraux qui se présentent actuellement. » (2009, p. 202).

92. La liste des enjeux actuels auxquels nous faisons face sont nombreux. Pour n'en citer qu'un, et pas des moindres, nous mentionnerons ici l'état écologique de la planète. Voir ici le dernier rapport de WWF « Rapport. Planète Vivante 2016. Risque et résilience dans l'Anthropocène. » (WWF, 2016). Nous pourrions formuler des liens entre les philosophies qui promeuvent une conception libérale de l'individu et les répercussions des activités humaines sur la planète. En détachant l'individu de son contexte de vie (incluant ici les écosystèmes humains et non humains), en l'érigeant en espèce supérieure vis-à-vis des autres espèces présentes sur Terre, l'être humain a été intérieurement incapable de mesurer les conséquences de ces actions car convaincus d'être le seul être véritablement de valeur sur la planète.

Enfin, l'éthique du *care* permet de bouleverser les conceptions épistémologiques en morale puisqu'elle ne part pas d'un point de vue moral abstrait, mais s'inscrit d'abord et avant tout à l'intérieur d'un sujet situé dans le monde, conception largement inspirée des épistémologies féministes (Harding, 1993). Il y aurait donc ici un renversement total de la perspective puisqu'au lieu de s'extraire du monde dans lequel le sujet se trouve, il s'y enracine entièrement afin de reconnaître la place qu'il occupe dans l'ensemble des représentations et symboliques qui constituent le monde. C'est notamment ce que Perreault (2015) appelle le processus de subjectivation située. Ce processus vise à se reconnaître dans l'espace social, comme marqueur de différenciation sexuelle et de genre : « La différence entre le *care* et la justice n'est pas qu'une différence morale ou langagière ; elle implique aussi une différence dans l'expérience sensible d'individus sexués ou sexualisés, situés dans un territoire symbolique qui les distingue les uns des autres. » (2015, p. 46). Cette distinction expérientielle rend notamment compte d'une manière d'être avec les autres et avec soi-même. Se voir et se reconnaître comme étant dépendant vis-à-vis des autres, c'est accepter d'être autrement avec l'autre, car conscient du lien qui nous lie. Dans son aspiration à l'autonomie et à l'indépendance, la perspective de justice participe à nier également ce qui constitue psychiquement le sujet dans le monde : « La différence sexuelle propre au système patriarcal exigerait à cet égard un travail de négation des liens fondamentaux qui réunissent les sujets dans l'espace commun. Contribuant à la négation d'autrui, la séparation qui la fonde s'articule de même à la négation de soi. » (Perreault, 2015, p. 47). User d'une approche formelle et abstraite en morale aurait donc des conséquences spirituelles et philosophiques bien plus importantes qu'on ne l'imagine puisqu'en niant son enracinement au monde, l'individu rejette également la caractéristique fondatrice de son existence, c'est-à-dire le fait d'être *un sujet qui vit*.

Dès lors, quelles sont les conclusions que l'on peut tirer à propos de notre analyse éthique sur l 'IA ? D'une part, il semblerait que les encadrements à propos de son développement ou de son déploiement ne soient que peu constructifs puisqu'ils reposent sur une conception de l'être humain fragmentaire et inachevée. D'autre part, elles usent, comme nous venons de le montrer d'une perspective en morale, celle de la théorie de la justice qui ne prend pas en compte les facteurs réels et concrets qui permettent à un écosystème de poursuivre son *crédo* évident, à savoir la vie. L'usage du point de vue moral et de la perspective de justice dans les analyses éthiques sur l'IA ne peuvent rendre compte de la globalité des enjeux puisqu'il ne s'agit ici *que* d'adopter une forme de raisonnement moral, celle de la justice, dont, nous le rappelons, se fonde sur une conception erronée de l'être humain. L'ensemble des cadres normatifs tels que les principes, les règles, les chartes, les recommandations sur l'IA deviennent inachevées, voire incomplètes, car ne comprenant pas les tenants et les aboutissants d'un contexte de vie, dont les acteurs et actrices sont engagées dans des réseaux de relations particulières – comprenant ici les personnes humaines, animaux et territoires vivants – ils en établissent des normes qui ne correspondent finalement à rien de réel. Par ailleurs, l'utilisation d'une seule forme de raisonnement moral, par volonté ou ignorance de la part des décideurs et intellectuels, a des conséquences importantes sur le monde. En effet, encadrer le développement de l'IA en usant de principes de justice abstraits et universels, c'est se servir de cadres éthiques qui nous ont spécifiquement amenés à cette situation de crise généralisée dans laquelle nous nous trouvons. Comment pourrait-on en ce sens croire que pour nous sauver des méfaits d'une technologie, nous devrions nous servir des outils éthiques qui ont été peu capables de les prévenir ? Nous y reviendrons dans la section suivante.

Ensuite, s'agissant de l'*a priori* détermination morale intégrée dans de l'IA, celle-ci se trouve profondément bousculée. En effet, l'idée d'une IA éthique, d'un robot vertueux ou d'algorithmes justes, toutes ces propositions usent de la perspective de la théorie de la justice et donc de ses infondées philosophiques. Parmi les nombreux ouvrages traitant de la possibilité de l'insertion d'un critère de moralité de l'action dans les algorithmes, les perspectives d'éthique du *care* sont *tout simplement* mises de côté. La raison de cet oubli est-elle due à une ignorance quant aux éthiques alternatives et féministes ? Ou au contraire, y a-t-il un intérêt structurel à ne pas analyser l'IA à partir d'une éthique qui remet entièrement en question nos présupposés ontologiques et normatifs ? L'idée d'un robot vertueux (Gibert, 2020), bon

ou juste ne ferait aucun sens dans une perspective de *care*. En effet, le bon ou le juste, selon le *care*, ne pourrait être reconnu à l'intérieur d'un principe descendant, puisqu'il se trouve de manière immanente, dans un contexte de vie marqué par des relations d'interdépendance au vivant dans sa globalité. Il faudrait dès lors, si nous voulions que l'IA soit juste d'un point de vue du *care*, intégrer une disposition visant à agir de telle sorte qu'elle préserve un réseau relationnel dans le but de maintenir la vie, ce qui demanderait une grande intelligence émotionnelle et relationnelle à cette IA. Si la disposition de *care* reste possible à intégrer car de l'ordre de l'intériorité et donc de la conscience ; le bât blesse lorsque nous envisageons la dimension pratique du *care*. En effet, une IA pourrait-elle prendre soin des individus si le *care* est un travail ? Une pratique ? Une action ? Par ailleurs, si nous voulons qu'une IA soit éthique selon le *care*, comment pourrait-elle l'être puisqu'étant soumise à la programmation, elle est nécessairement décontextualisée et déracinée de son lieu de vie. Les enjeux de la conscience de la temporalité et du *care* comme pratique matérielle semblent aujourd'hui difficiles à dépasser. Il semblerait donc que nous soyons dans une sorte d'impasse : une IA éthique ne semble être possible *que* dans un contexte de valorisation de la théorie de la justice et non de l'éthique du *care*.

Il ne serait donc pas farfelu d'affirmer que *l'IA aurait un genre* dont l'existence nous révèlerait notre histoire patriarcale (Nurock, 2019). En s'édifiant neutre et impartial, l'IA s'inspirerait et poursuivrait au contraire ce point de vue moral que nous avons largement remis en cause. Nurock identifie un danger important à savoir la potentielle artificialisation de l'éthique, qu'elle définit comme « (…) l'implantation dans une technologie (ironiquement supposée neutre ou impartiale) de structures héritées historiquement qui soutiennent des dynamiques de domination moralement et politiquement injustifiées. » (Nurock, 2019). Après avoir montré que les éthiques sur l'IA partent d'un présupposé ontologique falsifié et qu'elles usent de formes morales déracinées, nous verrons que l'approche de l'encadrement de l'IA va contre l'impératif de responsabilité et d'attention dont l'éthique du *care* est la représentante.

AU-DELÀ D'UN ENCADREMENT DES RISQUES : RESPONSABILITÉ ET ATTENTION

On pourrait formuler l'hypothèse selon laquelle l'éthique de l'IA trouverait son origine dans les méfaits de son objet. En d'autres termes, c'est parce que nous aurions identifié des implications *potentiellement* ou *actuellement* négatives de l'IA sur le réel et les individus qui le constituent que son éthique est apparue. Pour le dire autrement, l'éthique de l'IA serait pertinente en tant qu'elle serait un outil d'encadrement et de gestion normative des implications jugées néfastes pour le monde. Si l'IA n'avait aucun effet déplorable, il n'y aurait donc pas d'éthique de l'IA puisqu'elle serait jugée tout simplement bonne pour l'humanité. Le doute quant au caractère acceptable de son usage permet le développement de son éthique.

On remarque notamment que dans la plupart des délibérations citoyennes sur les usages de l'IA —par exemple, les processus citoyens de la Déclaration de Montréal pour un développement responsable de l'IA (2018) et du Dialogue inclusif sur l'éthique de l'IA (2020) —, tendent à évaluer ce qu'on nomme les « enjeux éthiques de l'IA », en d'autres termes, ses risques potentiels ou actuels sur le bon fonctionnement d'une société. Dès lors, l'intégration de l'IA dans le champ de l'existence humaine apporte, ce faisant, une acceptation de ses risques. La structure éthique dans laquelle nous vivons aurait donc accepté l'idée d'une société du risque, c'est-à-dire, un espace d'interactions humaines où il y aurait nécessairement des aléas ou des dangers[93].

93. Cette forme de vie créatrice de risques et de dangers est d'autant plus visible lorsqu'elle permet de renforcer des structures d'oppression et de violence à l'égard des femmes. Je pense ici à la robotique sexuelle et à ses effets symboliques, matériels et politiques. Voir ici mes travaux de recherche de mémoire de maîtrise (2018).

Tronto (2012) dans un ouvrage intitulé *Le risque ou le care ?* interroge deux conceptions de gestion des affaires humaines : le risque et le *care*. La société du risque, théorisée notamment par l'auteur Ulrich Beck, accepterait, voire revendiquerait l'idée selon laquelle nous ne serions plus en mesure de contenir les effets inattendus de certaines actions, en d'autres mots, nous serions dans une société du « hors contrôle » (Beck, 1998). Cette société du risque émanerait de la modernité et serait celle d'une décharge collective de la responsabilité face au développement industriel et technologique. Tronto remet largement en question cette conception sociale. En effet, pour l'autrice, cette vision du monde est liée à une expérience genrée puisque le masculin a largement été associé à la protection face au danger : « La société du risque crée l'image d'un monde « risqué », ce qui induit une compréhension du monde social comme dangereux et lié à la tâche humaine de protection et d'administration. La société du risque opère ainsi dans un univers métaphoriquement masculin. » (Tronto, 2012, p. 25). C'est parce que le monde est risqué que nous devons l'administrer, protéger et gouverner les individus. Or, ce que nous dit Tronto, c'est qu'à l'inverse, une société du *care* poserait d'abord et avant tout, la responsabilité du soin vis-à-vis d'autrui : « (...) le *care* suppose que les individus deviennent autonomes et capables d'agir d'eux-mêmes à travers un processus complexe de croissance, de développement, à travers lesquels ils sont les uns et les autres interdépendants et transformés dans leur vie. » (Tronto, 2012, p. 33). La gestion des risques se trouverait sans fondement dans une société du *care* puisqu'ils seraient contenus et prévenus par le fait que les besoins seraient pris en charge collectivement et de manière responsable dès le départ : « La société du *care* présuppose que les personnes vivent dans un monde où elles composent en permanence avec la vulnérabilité et le besoin – en faisant parfois aussi l'expérience de la joie. » (Tronto, 2012, p. 46). Par conséquent, l'intention de l'éthique de l'IA visant à encadrer son développement responsable partirait d'une conception sociale basée sur le risque.

L'éthique du *care* nous invite ainsi à adopter une posture spécifiquement attentionnelle vis-à-vis du monde (Garrau, 2014). Prêter attention signifie ici *écouter* et *entendre la voix* de ceux ou de celles qui sont au cœur d'un écosystème particulier afin de déterminer si le *care* est adéquat ou non (Garrau, 2014). Cette considération de la parole des principaux intéressés par un dispositif quelconque, ici, d'un usage de l'IA est absolument nécessaire pour déterminer si cette prise en charge du besoin est adaptée et appropriée à la situation et aux points de vue des acteurs/rices impliquées. Autrement dit, la détermination des besoins sociaux et politiques ne doivent être élaborée qu'à travers l'inclusion des personnes qui réalisent les activités essentielles de notre société, sans quoi, les injustices structurelles continueront de se perpétuer : « Si cette délibération n'a pas lieu ou si elle a lieu sans que les voix des pourvoyeurs de *care* soient prises en compte en effet, les inégalités qui structurent les relations de *care* dans les sociétés libérales contemporaines ne pourront manquer de se reproduire. » (Garrau, 2014, p. 66). La question est donc de savoir : les pourvoyeurs et pourvoyeuses de *care* ont-elles été consultées avant de développer certaines applications de l'IA ? À ce jour, aucune consultation spécifique n'a été menée auprès des personnes dites « de première ligne », c'est-à-dire aux personnes travaillant dans les métiers de *care* (nous pensons ici aux domaines de l'éducation, du ménage, de l'agroalimentaire, des sciences infirmières, etc).

Enfin, je tâcherai ici de proposer quelques marches de manœuvre pour le développement *possible* d'une IA éthique, ici, entendue, selon le *care*. Une des propositions, élaborée par Nurock et al. (2021) est de repenser la structure de l'éthique dès la conception, ou communément appelée *ethics by design* en incluant le *care* comme l'un des critères déterminants. La procédure serait de répondre à quatre questions, permettant ainsi de faire de l'IA éthique selon le *care* dès la conception. Ces préoccupations sont les suivantes : *1) De quoi nous soucions-nous ? 2) Pour quoi ou pour qui le faisons-nous ? 3) Prenons-nous soin ? 4) Prenons-nous soin avec l'autre ?*[94]. Ces différentes interrogations,

94. Ces quatre questions sont suivies de reformulations tout aussi pertinentes les unes des autres : 1) *What is important to us in the development of AI ?* 2) *Have we attented to the most vulnerable ?* 3) *Have we taken care to safeguard user's choices and integrated their requirements, rights, needs, etc. in the system ?* 4) *How do we govern AI democratically and remain mindful of the transformations that AI is capable of bringing about in our democratic institutions and in the public arena ?* (Nurock et al., 2021)

si répondues de bonne foi et de manière responsable, c'est-à-dire en étant engagé activement dans le processus, permettraient dès lors de déterminer et de reconnaître le caractère *profondément* éthique de cet usage de l'IA dans un monde où les êtres sont vulnérables, interdépendants, et ont donc besoin avant tout de *care* pour vivre, voire survivre.

Par ailleurs, il est important de noter et de reconnaître qu'il existe certains développements de l'IA qui ne se présentent *que comme étant* au service de causes communes importantes. C'est ce qu'on appelle « l'IA pour l'humanité », « l'IA pour le bien commun » ou encore « l'IA pour la planète ». Si ces IA sont déployées à destination d'un meilleur vivre-ensemble, il reste à confirmer si elles répondent adéquatement et pertinemment aux questions posées par l'éthique du *care* dès la conception, telles que présentées précédemment.

CONCLUSION

Dans un ouvrage intitulé *Courage Call to Courage Everywhere*, Winterson (2018) se demandait si l'IA n'était pas la pire chose qu'il puisse arriver aux femmes. Nous nous sommes posé une autre question : l'IA peut-elle être en accord avec une éthique féministe, démocratique et inclusive, soit l'éthique du *care* ? Pour y répondre, nous avons, à partir des éthiques du *care*, commencé par questionner les fondements ontologiques sur lesquels les principales éthiques utilisées fondent leur jugement moral vis-à-vis de l'IA. Ensuite, nous avons adopté une perspective éthique en déboulonnant la forme du raisonnement moral utilisé dans les éthiques dominantes, à savoir le point de vue moral propre à la théorie de la justice en nous servant des critiques soulevées par les philosophes du *care*. Enfin, nous avons montré que l'éthique de l'IA trouvait sa pertinence dans une société du risque et non du *care*. En ce sens et après notre réflexion, nous pouvons établir que le *care* revêt un caractère critique et révolutionnaire des plus puissants pour questionner les éthiques de l'IA et leurs conclusions sur le monde. L'objectif de cet article était donc de sortir de la saturation éthique en IA afin d'élargir notre regard sur les différentes formes de vie que nous pouvons formuler ensemble. À partir de cette nouvelle perspective éthique, il s'agira pour nous de déterminer si, à partir de l'éthique du *care*, l'IA reste pertinente ou non.

L'éthique du *care* représente une chance extraordinaire pour la terre, les êtres vivants et les humains de se découvrir autrement, de nouer de nouvelles relations et de construire un monde juste et équitable en vue d'une chose assez simple, à savoir *la préservation de la vie sur Terre.* Reste à savoir quelle place l'IA *devrait* ou *pourrait* avoir dans ce nouveau monde.

BIBLIOGRAPHIE

Algora Lab – Université de Montréal et Mila et Institut Québécois d'intelligence artificielle. 2020. *Le dialogue inclusif sur l'éthique de l'IA. Contribution à la recommandation de l'UNESCO sur l'éthique de l'intelligence artificielle*, https://opendialogueonai.com.

Beck, U. 1998. « Le conflit des deux modernités et la question de la disparition des solidarités : liens personnels, liens collectifs », dans *Lien social et politiques*, RIAC, (39), p. 15-25.

Brugère, F. 2008. *Le sexe de la sollicitude*, Paris, Seuil.

Brugère, F. 2011. *L'éthique du « care »*, Paris, Presses Universitaires de France.

Boden, M.A. 2005. *The Philosophy of Artificial Intelligence*, Oxford, Oxford University Press.

Clement, G. 1996. *Care, Autonomy, and Justice. Feminism and The Ethic of Care*, Boulder, Westview Point.

Crawford, K. 2021. *Atlas of AI*. New Haven,Yale University Press.

Debourdeau, A. 2013. *Les grands textes fondateurs de l'écologie*, Paris, Flammarion.

Dilhac, M., Christophe, A. et Voarino, N. 2018. *Rapport de la Déclaration de Montréal pour un développement responsable de l'intelligence artificielle*. Montréal, Université de Montréal.

Fisher, B. et Tronto, J. 1991. « Toward a feminist theory of care », dans Abel, B. et Nelson, M. (dir.), *Cicles of Care : Work and Identity in Women's Lives*, New York, State University of New York Press.

Garrau, M. 2014. *Care et attention*, Paris, Presses universitaires de France.

Gibert, M. 2020. *Faire la morale aux robots. Une introduction à l'éthique des algorithmes*. Montréal, Atelier 10.

GIEC (Groupe d'experts intergouvemental sur l'évolution du climat), 2021. *Climate Change 2021. The Physical Science Basis*. https://www.ipcc.ch/report/ar6/wg1/downloads/report/IPCC_AR6_WGI_Full_Report.pdf.

Gilligan, C. 1982. *In a Different Voice : Psychological Theory and Women's Development*, Cambridge, Harvard University Press.

Gilligan, C. 2010. « Une voix différente. Un regard prospectif à partir du passé », dans Nurock, Vanessa (dir.) *Carol Gilligan et l'éthique du care*. Paris, Presses universitaires de France.

Gilligan, C. 2011. « Une voix différente. Un regard prospectif à partir du passé », dans Laugier, S. et Paperman, P. (dir.), *Le souci des autres. Éthique et politique du care*. Paris, Éditions de l'École des hautes études en sciences sociales.

Gilligan, C., Hochschild, A. et Tronto, J. 2013. *Contre l'indifférence des privilégiés : à quoi sert le care ?* Paris, Payot.

Gilligan, C. 2019. *Une voix différente. La morale a-t-elle un sexe ?* (traduit par Annick Kwiatek), Paris, Flammarion.

Goodfellow, I., Bengio, Y. et Courville, A. 2016. *Deep Learning*, MIT Press.

Hache, E. 2016. *Reclaim. Recueil de textes écoféministes*. Paris, Éditions Camourakis.

Hamrouni, N. 2015. *Le care invisible : genre, vulnérabilité et domination*. Thèse de doctorat. Université de Montréal et Université catholique de Louvain.

Harding, S. 1993. « Rethinking Standpoint Epistemology : What is "Strong Objectivity" ? », dans Alcoff, L. et Potter, E. (dir.), *Feminist Epistemologies*. New York, Routledge.

Jobin, A., Ienca, M. et Vayena. E. 2019. « The global landscape of AI ethics guidelines », dans *Nature Machine Intelligence*.

Laugier, S. 2012. *Tous* vulnérables *? Le care, les animaux, l'environnement*. Paris, Payot.

Noiseau, P. Lanteigne, C., Flores Echaiz, L., Gomez Salazar, F.G., Mai, V., Dilhac, M-A et Mörch, C-M. 2021. « Le dialogue inclusif sur l'éthique de l'IA : délibération en ligne citoyenne et internationale pour l'UNESCO », dans *Communication Technologies et développement.*

Noiseau, P. 2019. *Les enjeux éthiques de la robotique sexuelle : une perspective critique féministe*. Mémoire de maîtrise, Université de Montréal.

Nurock, V. 2019. L'intelligence artificielle a-t-elle un genre ? Perspectives philosophiques sur l'artificialisation de l'éthique, du social et du politique. *Cités*, 80, pp. 61-74.

Nurock, V., Chatila, R. et Parizeau, M-H. 2021. « What does « Ethical by Design » Mean ? Dans Braunschweig, B. et Ghallab, M. (dir.), *Reflections on Artifical Intelligence for Humanity*, Vol 12600, Springer International Publishing.

Paperman, P. 2011. « Les gens vulnérables n'ont rien d'exceptionnel », dans Paperman, P. et Laugier, S. (dir.) *Le souci des autres. Éthique et politique du care*. Paris, Éditions de l'École des hautes études en sciences sociales.

Paperman, P. 2021. « D'une voix discordante : désentimentaliser le *care*, démoraliser l'éthique, dans Molinier, P., Paperman, P. et Laugier, S. (dir.) *Qu'est-ce que le care ? Souci des autres, sensibilité, responsabilité*, Paris, Payot.

Perreault, J. 2015. « Renégocier la « voix différente » : retour sur l'œuvre de Gilligan », dans Bourgault, S. et Perreault, J. (dir.) *Le care : éthique féministe actuelle*. Montréal, Les éditions du remue-ménage.

Servigne, P. et Stevens, R. 2015. *Comment tout peut s'effondrer*, Paris : Seuil.

Starwak. 2019. *Quel monde voulons-nous ?* (traduit par Isabelle Stengers), Paris, Éditions Cambourakis.

Tronto, J. 1993. *Moral Boundaries. A Political Argument for an Ethic of Care*. New York, Routledge.

Tronto, J. 2009. *Un monde vulnérable. Pour une politique du care* (traduit par Hervé Maury). Paris, Éditions La découverte.

Tronto, J. 2012. *Le risque ou le care ?* (traduit par Fabienne Brugère). Paris, Presses universitaires de France.

Voarino, N. 2019. *Systèmes d'intelligence artificielle et santé : les enjeux d'une innovation responsable*. Thèse de doctorat, Université de Montréal.

WWF. 2016. *Rapport. Planète vivante 2016. Risque et résilience dans l'Anthropocène*.

L'intelligence artificielle (IA) a d'ores et déjà un effet profond sur nos sociétés. Alors que les avancées scientifiques et technologiques se succèdent à un rythme soutenu, il est essentiel de poursuivre voire d'accélérer la conversation internationale portant sur leur développement et leur gouvernance. Dans ce contexte, Mila et UNESCO unissent leurs forces pour mener à la publication de cet ouvrage collectif de 18 chapitres issus d'un appel à contributions lancé mondialement en 2021.

Les chapitres sélectionnés traversent les frontières disciplinaires et géographiques. Ils incluent les perspectives d'universitaires, de membres de la société civile et d'innovateur.rice.s. avec pour objectif de faire glisser notre regard de ce que nous savons déjà vers ce qui nous échappe : les angles morts de la gouvernance de l'IA.

Avec cette publication, Mila et UNESCO espèrent offrir aux membres des milieux politiques, de la recherche, de l'innovation et de la société civile des perspectives utiles pour faire face à l'immense tâche qui nous incombe : assurer un développement de l'IA qui ne laisse personne derrière. Ceci signifie qu'il faut collectivement s'atteler à sa gouvernance afin que celle-ci soit centrée sur l'humain, inclusive, éthique, durable et qu'elle soutienne la pleine réalisation des droits humains et de l'État de droit.

9 789232 002853