Université de Montréal

Systèmes d'intelligence artificielle et santé : les enjeux d'une innovation responsable

Une analyse des craintes et des attentes citoyennes face aux défis de l'exercice de la responsabilité

Par

Nathalie Voarino

Faculté de médecine

Thèse présenté(e) en vue de l'obtention du grade de Doctorat (PhD)

en Sciences biomédicales, option bioéthique

Septembre 2019

Université de Montréal

Département de médecine sociale et préventive, Faculté de Médecine

---

Cette thèse intitulée

Systèmes d'intelligence artificielle et santé : les enjeux d'une innovation responsable

Une analyse des craintes et des attentes citoyennes face aux défis de l'exercice de la responsabilité

Présentée par

Nathalie Voarino

A été évaluée par un jury composé des personnes suivantes

Isabelle Ganache
Présidente-rapporteure

Béatrice Godard
Directrice de recherche

Ghislaine Cleret de Langavant
Codirectrice

Catherine Régis
Membre du jury

Geneviève Dubois-Flynn
Examinatrice externe

Résumé

L'avènement de l'utilisation de systèmes d'intelligence artificielle (IA) en santé s'inscrit dans le cadre d'une nouvelle médecine « haute définition » qui se veut prédictive, préventive et personnalisée en tirant partie d'une quantité inédite de données aujourd'hui disponibles. Au cœur de l'innovation numérique en santé, le développement de systèmes d'IA est à la base d'un système de santé interconnecté et auto-apprenant qui permettrait, entre autres, de redéfinir la classification des maladies, de générer de nouvelles connaissances médicales, ou de prédire les trajectoires de santé des individus en vue d'une meilleure prévention. Différentes applications en santé de la recherche en IA sont envisagées, allant de l'aide à la décision médicale par des systèmes experts à la médecine de précision (*ex.* ciblage pharmacologique), en passant par la prévention individualisée grâce à des trajectoires de santé élaborées sur la base de marqueurs biologiques.

Des préoccupations éthiques pressantes relatives à l'impact de l'IA sur nos sociétés émergent avec le recours grandissant aux algorithmes pour analyser un nombre croissant de données relatives à la santé (souvent personnelles, sinon sensibles) ainsi que la réduction de la supervision humaine de nombreux processus automatisés. Les limites de l'analyse des données massives, la nécessité de partage et l'opacité des décisions algorithmiques sont à la source de différentes préoccupations éthiques relatives à la protection de la vie privée et de l'intimité, au consentement libre et éclairé, à la justice sociale, à la déshumanisation des soins et du patient, ou encore à la sécurité. Pour répondre à ces enjeux, de nombreuses initiatives se sont penchées sur la définition et l'application de principes directeurs en vue d'une gouvernance éthique de l'IA. L'opérationnalisation de ces principes s'accompagne cependant de différentes difficultés de l'éthique appliquée, tant relatives à la portée (universelle ou plurielle) desdits principes qu'à la façon de les mettre en pratique (des méthodes inductives ou déductives).

S'il semble que ces difficultés trouvent des réponses dans la démarche éthique (soit une approche sensible aux contextes d'application), cette manière de faire se heurte à différents défis. L'analyse des craintes et des attentes citoyennes qui émanent des discussions ayant eu lieu lors de la coconstruction de la Déclaration de Montréal relativement au développement responsable de l'IA permet d'en dessiner les contours. Cette analyse a permis de mettre en évidence trois principaux défis relatifs à l'exercice de la responsabilité qui pourrait nuire à la mise en place d'une

gouvernance éthique de l'IA en santé : l'incapacitation des professionnels de santé et des patients, le problème des *mains multiples* et l'agentivité artificielle. Ces défis demandent de se pencher sur la création de systèmes d'IA capacitants et de préserver l'agentivité humaine afin de favoriser le développement d'une responsabilité (pragmatique) partagée entre les différentes parties prenantes du développement des systèmes d'IA en santé. Répondre à ces différents défis est essentiel afin d'adapter les mécanismes de gouvernance existants et de permettre le développement d'une innovation numérique en santé responsable, qui doit garder l'humain au centre de ses développements.

Mots-clés : intelligence artificielle, données massives, innovation responsable, innovation numérique, éthique, bioéthique, santé connectée, santé numérique, agentivité, gouvernance algorithmique

Abstract

The use of artificial intelligence (AI) systems in health is part of the advent of a new "high definition" medicine that is predictive, preventive and personalized, benefiting from the unprecedented amount of data that is today available. At the heart of digital health innovation, the development of AI systems promises to lead to an interconnected and self-learning healthcare system. AI systems could thus help to redefine the classification of diseases, generate new medical knowledge, or predict the health trajectories of individuals for prevention purposes. Today, various applications in healthcare are being considered, ranging from assistance to medical decision-making through expert systems to precision medicine (e.g. pharmacological targeting), as well as individualized prevention through health trajectories developed on the basis of biological markers.

However, urgent ethical concerns emerge with the increasing use of algorithms to analyze a growing number of data related to health (often personal and sensitive) as well as the reduction of human intervention in many automated processes. From the limitations of big data analysis, the need for data sharing and the algorithmic decision 'opacity' stems various ethical concerns relating to the protection of privacy and intimacy, free and informed consent, social justice, dehumanization of care and patients, and/or security. To address these challenges, many initiatives have focused on defining and applying principles for an ethical governance of AI. However, the operationalization of these principles faces various difficulties inherent to applied ethics, which originate either from the scope (universal or plural) of these principles or the way these principles are put into practice (inductive or deductive methods).

These issues can be addressed with context-specific or bottom-up approaches of applied ethics. However, people who embrace these approaches still face several challenges. From an analysis of citizens' fears and expectations emerging from the discussions that took place during the coconstruction of the Montreal Declaration for a Responsible Development of AI, it is possible to get a sense of what these difficulties look like. From this analysis, three main challenges emerge: the incapacitation of health professionals and patients, the *many hands* problem, and artificial agency. These challenges call for AI systems that empower people and that allow to maintain human agency, in order to foster the development of (pragmatic) shared responsibility among the various stakeholders involved in the development of healthcare AI systems. Meeting these challenges is essential in order to adapt existing governance mechanisms and enable the

development of a responsible digital innovation in healthcare and research that allows human beings to remain at the center of its development.

Keywords: artificial intelligence, big data, responsible innovation, digital innovation, ethics, bioethics, digital health, agency, algorithmic governance

Table des matières

Liste des tableaux

Liste des schémas

Liste des sigles et abréviations

ADN : Acide désoxyribonucléique

AI HLEG : High-Level Expert Group on Artificial Intelligence

APOLLO : Adaptive Patient-Oriented Longitudinal Learning And Optimization

ATM : Appropriate Technology Movement

CCNE : Comité consultative national d'éthique (France)

CDT : Center for Democracy and Technology (États-Unis)

CERNA : Commission de réflexion sur l'Éthique de la Recherche en sciences et technologies du Numérique d'Allistene (France)

CIFAR : Institut canadien de recherches avancées

CNIL : Commission nationale de l'informatique et des libertés (France)

CNN : Convolutional Neural Networks

CRR : Conduite responsable en recherche

EHR : Electronic Health Records

EMR : Electronic Medical Records

ÉPTC2 : Énoncé de politique des trois Conseils, version 2

FDA : Food And Drug Administration (États-Unis)

FRQ : Fonds de recherche du Québec

GAFAM : Google, Apple, Facebook, Amazon et Microsoft

GANs : Generative Adversarial Networks

IA : Intelligence Artificielle

IEEE : Institute of Electrical and Electronics Engineers

INESSS : Institut national d'excellence en santé et services sociaux

INSPQ : Institut national de santé publique du Québec

IRCAD : Institut de recherche contre les cancers de l'appareil digestif (France)

IRR : Innovation et recherche responsable

ISO : Organisation internationale de normalisation

ITI : Information Technology Industry Council (États-Unis)

JSAI : The Japanese Society for Artificial Intelligence

LPRP : Loi sur la protection des renseignements personnels (Canada)

LPRPDE : Loi sur la protection des renseignements personnels et les documents électroniques (Canada)

NHS : National Health Service (Royaume-Uni)

NIH : National Institutes of Health (États-Unis)

NLP : Natural language processing

NoC : global Network of Internet & Society Centers

OBVIA : Observatoire international sur les impacts sociétaux de l'IA et du numérique

OCDE : Organisation de coopération et de développement économiques

OMS : Organisation mondiale de la santé

OSTP : Office of Science and Technology Policy (États-Unis)

PHR : Personal Health Records

RGPD : Règlement général sur la protection des données (Europe)

SVM: Support Vector Machine

UNESCO : Organisation des Nations unies pour l'éducation, la science et la culture

WEF : The World Economic Forum

À Claire, Jeanine et Marie-France

Remerciements

Nombreuses sont les personnes qui m'ont accompagnée dans le travail de ces quatre dernières années, je regrette de ne pouvoir toutes les nommer. Je remercie d'abord Ghislaine Cleret de Langavant, sans qui je n'aurais pas terminé ce doctorat, pour la pertinence de ses conseils, sa confiance, la qualité de son encadrement et son soutien qui se sont avérés indispensables. Je remercie également Béatrice Godard pour la qualité de son encadrement, son soutien et sa grande réactivité, tout particulièrement dans les dernières étapes. Je souhaite également exprimer ma reconnaissance à Louis Chartrand, Camille Vézy et Ariane Mauriello, pour leurs révisions, leurs conseils ainsi que nos nombreuses discussions très inspirantes. Je remercie Jean-Christophe Bélisle-Pipon, mon collègue et ami, pour m'avoir éclairée à de (très) nombreuses reprises, mais aussi pour m'avoir offert nombreuses des opportunités qui m'ont amenée à me dépasser. Je tiens aussi à remercier Marc-Antoine Dilhac, pour toujours me soutenir ainsi que pour avoir grandement contribué à ma formation. Je souhaite également remercier toute l'équipe de la Déclaration de Montréal ainsi que le Vice-Rectorat à la recherche, à la création et à l'innovation pour leur confiance et sans qui se projet n'aurait pu se réaliser. Je tiens à remercier plus particulièrement Anne Marie-Savoie et Isabelle Bayard pour leur patience. Je remercie également Virginie pour son écoute précieuse; Valentine et Victoria, mes partenaires de thèse dans la joie comme dans la douleur; ainsi que Vincent, pour ses précieux conseils. Je remercie également Chloé, Maly, Guillaume, Coco, Chani, Yohan, Andrea, Adeline, Caro, Bibiane, Catherine, Laetitia, mes parents et mon frère pour leur soutien inconditionnel – quelle qu'en soit la forme – mais surtout pour avoir su égayer les différentes étapes de ce long processus. Je remercie aussi les membres de mon jury pour leurs commentaires pertinents. Je souhaite enfin remercier *Thèsez-vous*, pour les tomates collectives qui m'ont grandement aidé à terminer la rédaction de cette thèse; ainsi que l'Institut de recherche en santé publique de l'Université de Montréal (IRSPUM), les Programmes de bioéthique et la Faculté des études supérieures et postdoctorale (FESP) pour le soutien financier.

Avant-Propos

Marseille. (Probablement) septembre 2008. C'est la première fois de ma vie que j'ai une expérience intime avec une ligne de code. Dans mon programme de licence en neurosciences, nous sommes dans mes souvenirs une des premières cohortes de biologistes à qui on offre des cours d'informatique. Les bêta-testeurs d'un pont entre l'informatique et le biologique, essentiel nous dit-on considérant les développements à venir dans le domaine.

Je ne peux qu'imaginer le niveau de déception de notre professeur, qui nous annonce lors de la première rencontre l'objectif principal : nous allons apprendre à coder des réseaux de neurones. Échec cuisant. La seule chose que nous avons réussi à coder, mon binôme de l'époque et moi-même, c'est une boucle qui demande inlassablement à l'utilisateur s'il aime le couscous, et qui ne s'arrête (en le félicitant) que lorsqu'il répond oui. L'examen final s'est transformé en questionnaire à choix multiples en place et lieu d'évaluer nos lignes de codes en langage *Python*, et toutes les notes ont été augmentées de quelques points, car beaucoup trop d'étudiants auraient échoué le cours. Bien que j'eusse une affection particulière pour mes colocs développeurs (quand ils ne criaient pas trop fort à 2h du matin « Pourquoi ça compile pas ? ») ils travaillaient pour moi sur des choses abstraites, que personne ne comprend et dont je ne saisissais ni la portée, ni l'utilité.

Montréal. Septembre 2017. Je ressors des cartons mon manuel « Apprendre à coder en Python » que m'a offert mon coloc il y a trois ans et que je n'ai jamais ouvert. C'est que « l'intelligence artificielle » a le vent en poupe. Un chercheur qu'il n'est plus nécessaire de nommer vient de se voir décerner la « plus grosse subvention de l'histoire de l'Université de Montréal ». Ça s'excite en bioéthique et les bruits de couloir circulent : 10% du budget serait probablement consacré à l'éthique et aux sciences sociales.

C'est à peu près au même moment que je tente de m'identifier à une discipline, notamment suite à mes discussions avec mon amie sociologue (et française) qui m'explique que bioéthicien, ce n'est pas vraiment une profession. Je réalise également que pour la plupart de mes proches, je travaille sur des choses abstraites, que personne ne comprend et dont ils ne saisissent ni la portée, ni l'utilité. Je décide à l'époque de tout arrêter, mais celle qui deviendra ma codirectrice me convainc du contraire. J'opte alors pour une tactique mixte : je me présente comme « plutôt philosophe » aux

sociologues, « plutôt sociologue » aux philosophes (tactique qui ne fonctionne que lors de situations où les chercheurs des deux disciplines ne se retrouvent pas côte à côte).

Alors que je suis en quête d'identité académique, un des membres du jury de mon examen de synthèse lance l'idée de la création d'une Déclaration de principes éthiques pour le développement responsable de l'IA. Il m'invite à assister aux premières réunions du projet. L'équipe n'est alors pas encore constituée et la mission est floue. Je ne comprends pas exactement ce que je fais-là, mais puisque ça n'a l'air de déranger personne, je décide de suivre le mouvement.

Hochelaga. Septembre 2019. J'ai passé l'été à rédiger ma thèse et mon manuel « Apprendre à coder en Python » a pris la poussière (non, je ne l'ai toujours pas ouvert). Près de 500 citoyens consultés plus tard, le projet a abouti à une Déclaration de principes lancée il y a presque un an et endossée par près de 1000 signataires.

Après avoir appris par ma directrice, avec beaucoup de tristesse, qu'une analyse « semi »-inductive (à défaut d'une approche semi-inductive) ça n'existe pas, j'ai finalement trouvé une posture dans laquelle je suis confortable pour analyser une seconde fois les discussions issues de cette consultation citoyenne. J'ai également compris que je n'étais ni philosophe, ni sociologue mais, peu importe le titre, je n'ai jamais été aussi certaine que ma place se situe quelque part entre science et société.

Aujourd'hui, à tous les développeurs qui se sentiraient « moralement surchargés » face aux impératifs éthiques d'une innovation responsable, je ne peux que humblement leur recommander de suivre les conseils de la thèse d'une bioéthicienne qui manifeste l'envie équivoque d'apprendre à coder depuis près de dix ans sans jamais pourtant s'y mettre : *c'est l'intention qui compte.*

Introduction

Devant la présence croissante de la numérisation au sein de nos sociétés, plusieurs auteurs annoncent l'avènement d'une 4ème révolution industrielle (Schwab 2016; Lahlou 2015). Cette numérisation touche tous les secteurs, incluant celui de la santé. Au cœur de l'innovation numérique en santé, se retrouvent de nombreuses applications relevant de l'intelligence artificielle (IA) et des données massives. [1] Dans la dernière décennie, le domaine de l'IA a en effet connu des avancées majeures grâce, entre autres, à la sophistication des outils informatiques et à la disponibilité d'un nombre de données croissant (Cardon, Cointet, et Mazières 2018). Cependant, ce à quoi réfère « l'IA » demeure ambigu et parfois contesté.

Le 3 juin 2018, lors d'une entrevue à la radio intitulée « l'intelligence artificielle n'existe pas »[2], le sociologue Yves Gingras soulignait une certaine négligence sémantique relativement à l'usage du terme, l'intelligence n'étant pas clairement définie dans le contexte de l'IA. Des propos similaires sont également soutenus par l'informaticien Luc Julia, cocréateur de Siri, dans son livre « L'intelligence artificielle n'existe pas » paru en 2019 (Julia 2019). L'auteur défend que le terme, utilisé pour la première fois en 1956 par l'informaticien McCarthy lors de la conférence de Dartmouth, a été particulièrement mal choisi car il alimente nombreux des fantasmes et des craintes qui entourent l'IA et ses usages.

Si plusieurs s'accordent à dire que l'objectif de l'IA est de reproduire certains des comportements humains (jugés « intelligents »)[3], l'ambiguïté entourant la définition de l'intelligence est notable[4]. Il existe de nombreuses interprétations de l'intelligence, en fonction du

[1] L'innovation numérique en santé ne saurait cependant se résumer à l'intelligence artificielle et aux données massives. Elle peut par exemple aussi référer à différentes approches de télémédecine ou de e-santé (ex. des consultations médicales à distance) (Global observatory for eHealth 2016).
[2] Voir : https://ici.radio-canada.ca/premiere/emissions/les-annees-lumiere/segments/chronique/74731/science-critique-gingras-intelligence-artificielle-n-existe-pas
[3] Objectif qui s'observe également dans la manière d'évaluer si une machine est intelligente : comme par exemple le test de Turing qui définit le comportement intelligent d'une machine selon ses similitudes avec le comportement d'un humain (Proudfoot 2011; Devillers 2017).
[4] L'ambiguïté renvoie d'ailleurs également au terme « artificiel », qui ne fait pas l'objet non plus d'une définition consensuelle. Si l'artificiel peut se définir dans son opposition au naturel, cette opposition ne saurait déterminer de

domaine depuis lequel elle est étudiée (ex. philosophie, psychologie, biologie) mais également en fonction des différentes manières de l'évaluer, qui sont autant de manières d'en définir les critères (Sternberg 1982; Gardner 1997). Selon les conceptions, l'intelligence (humaine) renvoie à une pluralité de formes et peut référer à la plasticité cérébrale, à la façon de traiter l'information, à la connaissance et ses représentations, aux capacités d'interaction entre individus, ou encore être déterminée par des normes culturelles et sociales (Sternberg 1982; Gardner 1997). Face à ces multiples conceptions, difficile de définir stricto sensu ce qui constitue un comportement dit « intelligent ».

Toujours dans l'entrevue du 3 juin 2018, Gingras soulignait l'importance de cette lacune : les algorithmes ne seraient pas « intelligents » à proprement parler car ils représentent seulement des propriétés ou des règles de calcul. La perception d'intelligence de l'IA relèverait alors plutôt du fantasme (Julia 2019; Ganascia 2017), d'une anthropomorphisation des machines *via* différentes projections d'intentions (humaines) (Devillers 2017) voire, comme le mentionne Gingras, d'un phénomène de mode. Il n'est cependant pas nécessaire de déterminer si les « comportements de l'IA » sont véritablement intelligents pour reconnaître son existence, en tant qu'objet social, dans l'imaginaire (moral) du public comme des experts. L'usage du terme n'a en effet pas cessé depuis 1956, fait l'objet d'un certain battage médiatique, et est repris dans les films de science-fiction[5] tout comme dans les demandes de subvention de recherche. Toutefois, ce qui relève d'une « négligence sémantique » relativement au terme demande une attention particulière, une perception exagérée des compétences de l'IA pouvant conduire à en surestimer les risques et les bénéfices.

L'ambiguïté relative à la perception d'intelligence de l'IA alimente en effet nombreuses des craintes et des promesses associées à l'usage de ces technologies. Certains discours, qui relèveraient du catastrophisme, alimentent la crainte d'une perte de contrôle des humains sur les

manière claire ce qui relève du naturel ou de l'artificiel. Notamment, parce que la nature se voit de plus en plus artificialisée et également, car l'artificiel - issue de la main humaine - peut être considéré comme naturel, faisant partie pour certains de « l'ordre biologique » (Larrère et Larrère 2015).

[5] Par exemple, le film *A.I. Intelligence Artificielle* de Spielberg (2001); *Her* de Jonze (2013) ou *Ex Machina* de Garland (2015).

machines. Cette perte de contrôle apparaîtrait avec ce que plusieurs auteurs nomment l'« explosion intelligente » (Shulman, Jonsson, et Tarleton 2009) ou la « singularité », qui réfère au moment où l'IA dépassera l'intelligence humaine (Ganascia 2017; Muehlhauser et Helm 2012). Cette crainte est par exemple soulignée par Hamet et Tremblay (2017) dans leur revue de la littérature des applications de l'IA en médecine :

> The biggest apprehension we have is that AI will become so sophisticated that it will surpass human brain capabilities and eventually will take control over our lives (p. 39).

Si certains prennent ces scénarios dystopiques (ou utopiques, selon la perspective) très au sérieux (ce que présente Ganascia, 2017), d'autres considèrent que ces préoccupations relèvent plutôt de la science-fiction. Ganascia (2017) parle bien d'un « mythe de la singularité » et plusieurs soutiennent que nous sommes loin de ces avancées (Devillers 2017; Ganascia 2017; Julia 2019). L'entretien de ce catastrophisme risquerait alors de détourner l'attention des risques et enjeux éthiques qui accompagnent les usages actuels de l'IA. Gibert questionne dans ce sens s'il ne faudra pas plutôt « avoir peur de la peur de l'IA » (Gibert 2019). L'avènement de l'IA et de l'usage des données massives, en santé comme ailleurs, soulève de nombreuses préoccupations imminentes qui ne nécessitent pas l'apparition d'une explosion intelligente pour survenir. Par exemple, l'automatisation du travail de l'humain sur l'humain (Lahlou 2015) soulève, comme lors des précédentes révolutions industrielles, des préoccupations relatives au remplacement des humains par les machines et au risque de chômage qui l'accompagne (Campolo et al. 2017; Russell, Dewey, et Tegmark 2015). Différents auteurs s'inquiètent des conséquences sur le respect de la vie privée, dans un monde où de plus en plus de données sur les individus sont collectées, stockées et analysées (Azencott 2018; Chow-White et al. 2015; Villani 2018; Déclaration de Montréal IA Responsable 2018) ou d'une perte de contrôle plus subtile, liée à la gouvernance algorithmique, soit comment l'IA nous gouverne considérant que nous accordons de plus en plus d'autorité aux recommandations qui émanent de ses analyses (Déclaration de Montréal IA Responsable 2018; Cardon 2018; Musiani 2013).

Inversement, s'observe un certain engouement autour de l'IA et de ses développements qui pourrait conduire à en surestimer les avantages. Il est alors possible de questionner si les bénéfices de l'innovation numérique ne relèvent pas d'un phénomène de mode ou d'un certain « hype » qui

renvoie, comme le mentionne Gibert (2019), au « parfum d'hyperbole » qui entoure l'IA et ses promesses, risquant de provoquer un enthousiasme possiblement exagéré. L'hyperbole qui entoure le développement technologique n'est pas le propre de l'IA et a souvent conduit, comme le présentent Bell, Lucke et Hall (2012) dans le contexte du développement de nouveaux médicaments, à un engouement possiblement démesuré lors de l'avènement d'une nouvelle technologie, qui conduit irrémédiablement à en surestimer les bénéfices[6] (Bell, Lucke, et Hall 2012).

En santé, l'IA promet d'améliorer la gestion du système de santé, notamment car elle détient le potentiel d'automatiser certaines tâches répétitives afin d'augmenter l'efficacité du système et d'en rentabiliser les coûts (Chartrand et al. 2017; Alanazi, Abdullah, et Qureshi 2017). La recherche sur l'IA en santé démontre des avenues prometteuses pour le traitement et la compréhension des cancers, les maladies du système nerveux ou les maladies cardiovasculaires (Jiang et al. 2017). L'IA participe également au développement d'une médecine dite de précision, de plus en plus ciblée sur les besoins et les caractéristiques des patients (Bayer et Galea 2015; Jameson et Longo 2015; Ashley 2015). Concernant les données massives en santé, elles offrent le potentiel d'effectuer des études à grande échelle et à faible coût ou d'améliorer la qualité des soins de santé (Rumbold et Pierscionek 2017). Également, l'utilisation de l'IA en santé publique permettrait de faire progresser la surveillance, la prévention et le contrôle de maladies émergentes, comme le font par exemple les différents programmes de détection numérique de maladies (*ex.* Health Map) (Brownstein, Freifeld, et Madoff 2009) qui peuvent notamment aider la prise de décision en vue du développement de politiques de santé (Bourguet et al. 2013).

Le potentiel de l'IA en santé a également été identifié par différentes firmes internationales telle que Google, Facebook, Apple ou Microsoft qui investissent massivement dans le secteur, contribuant ainsi à l'engouement qui entoure ces technologies. Nombreux des groupes rattachés à Google se centrent par exemple sur le développement d'applications médicales de l'IA, tel que

[6] Un engouement similaire s'est par exemple observé face aux innovations en neurosciences, ce qui a conduit à la description de l'apparition d'une certaine « neuromania » (Legrenzi et Umilta 2011).

Google Health[7] qui investit dans les domaines de la santé, du bien-être et des sciences de la vie notamment *via* les équipe de Google Genomics[8] (une base de données génétiques accessible à tous en vue de « donner de l'ampleur aux travaux de recherche »), ou Google Research[9] (qui se centre sur les applications de santé et travaille en partenariat avec plusieurs universités de renom). C'est également le cas d'autres compagnies qui appartiennent au groupe Alphabet comme Calico[10] (centré principalement sur le vieillissement), Verily[11] ou DeepMind[12]. Apple Health[13] et Microsoft Health[14] investissent également massivement dans le secteur, avec le développement d'applications mobiles de santé ou de différents objets connectés, à destination des professionnels de santé comme des particuliers.

L'IA détiendrait donc le potentiel, pour certains auteurs, de « révolutionner le système de santé » de différentes manières (Helbing 2015; Perry 2017; Mettler 2016; Raghupathi 1997). Cependant, selon Panch, Mattie, et Celi (2019), la « vérité qui dérange » est que les algorithmes qui font l'objet de nombreuses recherches en santé ne sont, en pratique, souvent pas exécutables. C'est par exemple le cas dans le contexte clinique, en raison de différents facteurs politiques et économiques ou de l'organisation fragmentée de l'écosystème de la santé qui, la plupart du temps, ne possède pas les infrastructures nécessaires ni les données pertinentes – ce qui limite l'idée d'une transformation réelle et durable du système de santé par l'IA (Panch, Mattie, et Celi 2019).

Ainsi, entre catastrophisme et engouement exagérés, il apparaît important de se pencher sur les réels bénéfices de l'avènement de l'IA en santé et de déterminer les véritables risques et enjeux éthiques qui sont associés à son utilisation en vue d'une innovation responsable. L'innovation responsable est essentielle à l'appropriation technologique soit, une innovation *pour* et *par* la société, qui se veut respectueuse des valeurs éthiques et sociales en vue d'assurer la pertinence et

[7] Voir : https://health.google/
[8] Voir : https://cloud.google.com/genomics/
[9] Voir : https://research.google.com/teams/brain/healthcare/
[10] Voir : https://www.calicolabs.com/
[11] Voir : https://verily.com/
[12] Voir : https://deepmind.com/applied/deepmind-health/
[13] Voir : https://www.apple.com/ca/healthcare/
[14] Voir : https://www.microsoft.com/en-us/microsoft-health

la durabilité de l'innovation (Barré, 2011). Dans cette perspective, il est essentiel de se pencher sur les technologies, les méthodes et les techniques qui relèvent de l'IA afin de diminuer l'ambiguïté associée à l'usage du terme. Il est également nécessaire de se détacher, pour une évaluation appropriée, de l'hyperbole associée à l'utilisation de l'IA et des technologies apparentées qui pourrait conduire à en surestimer la portée disruptive. Les algorithmes sont en effet présents dans nos sociétés depuis longtemps et en régissent de nombreux aspects, notamment en santé comme le reconnaissent Shameer et al. (2018) dans le cadre de la cardiologie :

> The application of machine learning technologies in cardiovascular medicine is not new. Scientists have long used computers and early techniques drawn from AI to analyse and interpret cardiovascular phenotyping data, for example, automated analysis of ECGs and imaging systems (p. 1156).

Les débats sur les implications éthiques de l'avènement de l'IA ont quant à eux débutés dès les années 60 (Morley et al. 2019).

Ainsi, en vue d'identifier les enjeux liés à une innovation numérique en santé responsable, cette thèse se penche sur différents aspects essentiels à considérer pour préserver la confiance en l'IA et dans les institutions de santé en vue de s'assurer de la mise en place d'un encadrement éthique effectif et pertinent. Dans le Chapitre 1 sont décrites la méthodologie et les différentes conceptions qui guident les analyses réalisées dans la présente thèse. Dans le Chapitre 2 sont présentées les techniques, les méthodes et les analyses qui relèvent de l'IA ainsi que leur lien inhérent aux données massives. Ce chapitre se penche également sur les différentes avenues prometteuses associées à leurs usages en santé et les différents points de rupture qui invitent à réfléchir à leur portée disruptive. Le Chapitre 3 fait quant à lui état des différentes limites de l'usage de l'IA et des données massives, mais également des différents risques et enjeux éthiques qui émergent des utilisations qui en sont faites. Dans le Chapitre 4 sont présentées les différentes initiatives qui ont tenté de définir des principes directeurs pour une gouvernance éthique de l'IA et les différentes difficultés de leur mise en application. Une analyse des discussions ayant eu cours lors de la coconstruction de la Déclaration de Montréal pour un développement responsable de l'IA est présentée dans le Chapitre 5. Regroupant des experts de différents domaines mais également un public non-expert, les perspectives citoyennes exprimées lors de ces discussions sont particulièrement pertinentes en vue d'explorer les défis d'une innovation responsable, notamment

car elles permettent de dresser un portrait de leurs craintes et de leurs attentes dans une perspective prospective. Cette analyse a ainsi permis de mettre en évidence les différents défis de l'exercice de la responsabilité dans le cadre de l'utilisation de l'IA et des données massives en santé. Enfin, dans le Chapitre 6, les différents constats qui émanent des chapitres précédents sont discutés afin de dégager différentes pistes de réflexion en vue de répondre aux enjeux d'une innovation numérique en santé responsable.

Références bibliographiques

Alanazi, Hamdan O., Abdul Hanan Abdullah, et Kashif Naseer Qureshi. 2017. « A Critical Review for Developing Accurate and Dynamic Predictive Models Using Machine Learning Methods in Medicine and Health Care ». *Journal of Medical Systems* 41 (4): 69. https://doi.org/10.1007/s10916-017-0715-6.

Ashley, Euan A. 2015. « The Precision Medicine Initiative: A New National Effort ». *JAMA* 313 (21): 2119-20. https://doi.org/10.1001/jama.2015.3595.

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): 20170350. https://doi.org/10.1098/rsta.2017.0350.

Barré, Rémi. 2011. « Des concepts à la pratique de l'innovation responsable : à propos d'un séminaire franco-britannique ». *Natures Sciences Societes* Vol. 19 (4): 405-9.

Bayer, Ronald, et Sandro Galea. 2015. « Public Health in the Precision-Medicine Era ». *The New England Journal of Medicine* 373 (6): 499-501. https://doi.org/10.1056/NEJMp1506241.

Bell, Stephanie K., Jayne C. Lucke, et Wayne D. Hall. 2012. « Lessons for Enhancement From the History of Cocaine and Amphetamine Use ». *AJOB Neuroscience* 3 (2): 24-29. https://doi.org/10.1080/21507740.2012.663056.

Bourguet, Jean-Rémi, Rallou Thomopoulos, Marie-Laure Mugnier, et Joël Abécassis. 2013. « An artificial intelligence-based approach to deal with argumentation applied to food quality in a public health policy ». *Expert Systems with Applications* 40 (11): 4539-46. https://doi.org/10.1016/j.eswa.2013.01.059.

Brownstein, John S., Clark C. Freifeld, et Lawrence C. Madoff. 2009. « Digital Disease Detection — Harnessing the Web for Public Health Surveillance ». *New England Journal of Medicine* 360 (21): 2153-57. https://doi.org/10.1056/NEJMp0900702.

Campolo, Alex, Madelyn Sanfilippo, Meredith Whittaker, et Kate Crawford. 2017. « AI Now 2017 Report ». https://ainowinstitute.org/AI_Now_2017_Report.pdf.

Cardon, Dominique. 2018. « Le pouvoir des algorithmes ». *Pouvoirs* N° 164 (1): 63-73.

Cardon, Dominique, Jean-Philippe Cointet, et Antoine Mazières. 2018. « La revanche des neurones ». *Reseaux* n° 211 (5): 173-220.

Chartrand, Gabriel, Phillip M. Cheng, Eugene Vorontsov, Michal Drozdzal, Simon Turcotte, Christopher J. Pal, Samuel Kadoury, et An Tang. 2017. « Deep Learning: A Primer for Radiologists ». *RadioGraphics* 37 (7): 2113-31. https://doi.org/10.1148/rg.2017170077.

Chow-White, Peter A., Maggie MacAulay, Anita Charters, et Paulina Chow. 2015. « From the Bench to the Bedside in the Big Data Age: Ethics and Practices of Consent and Privacy for Clinical Genomics and Personalized Medicine ». *Ethics and Information Technology* 17 (3): 189-200. https://doi.org/10.1007/s10676-015-9373-x.

Déclaration de Montréal IA Responsable. 2018. « Rapport de la Déclaration de Montréal pour un Développement Responsable de l'Intelligence Artificielle ». https://docs.wixstatic.com/ugd/ebc3a3_d806f109c4104c91a2e719a7bef77ce6.pdf.

Devillers, Laurence. 2017. *Des robots et des hommes: Mythes, fantasmes et réalité*. Plon.

Ganascia, Jean-Gabriel. 2017. *Le Mythe de la Singularité : faut-il craindre l'intelligence artificielle ?* Science ouverte. Le Seuil.

Gardner, Howard. 1997. *Les formes de l'intelligence (français)*. Sciences. Odile Jacob.

Gibert, Martin. 2019. « Faut-il avoir peur de la peur de l'IA ? » *La Quatrième Blessure* (blog). 11 janvier 2019. https://medium.com/@martin.gibert/faut-il-avoir-peur-de-la-peur-de-lia-1687abc35342.

Global observatory for eHealth. 2016. « Global diffusion of eHealth : Making universal health coverage achievable ». World Health Organization. http://apps.who.int/iris/bitstream/10665/252529/1/9789241511780-eng.pdf?ua=1

Hamet, Pavel, et Tremblay, Johanne. 2017. « Artificial intelligence in medicine ». *Metabolism*, Insights Into the Future of Medicine: Technologies, Concepts, and Integration, 69 (Supplement): S36-40. https://doi.org/10.1016/j.metabol.2017.01.011.

Helbing, Dirk. 2015. « Societal, Economic, Ethical and Legal Challenges of the Digital Revolution: From Big Data to Deep Learning, Artificial Intelligence, and Manipulative Technologies ». SSRN Scholarly Paper ID 2594352. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=2594352.

Jameson, J. Larry, et Dan L. Longo. 2015. « Precision Medicine--Personalized, Problematic, and Promising ». *The New England Journal of Medicine* 372 (23): 2229-34. https://doi.org/10.1056/NEJMsb1503104.

Jiang, Fei, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, et Yongjun Wang. 2017. « Artificial intelligence in healthcare: past, present and future ». *Stroke and Vascular Neurology*. https://doi.org/10.1136/svn-2017-000101.

Julia, Luc. 2019. *L'intelligence artificielle n'existe pas*. First. ISBN: 9782412046746

Lahlou, Saadi. 2015. « Un monde numérique : le renversement du miroir ». Dans . Vol. 53. Variances.

Larrère, Catherine, et Raphaël Larrère. 2015. « Le naturel et l'artificiel ». *Sciences humaines*, 153-74.

Legrenzi, Paolo, et Carlo Umilta. 2011. *Neuromania: On the limits of brain science*. Oxford University Press.

Mettler, M. 2016. « Blockchain technology in healthcare: The revolution starts here ». Dans *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)*, 1-3. https://doi.org/10.1109/HealthCom.2016.7749510.

Morley, Jessica, Luciano Floridi, Libby Kinsey, et Anat Elhalal. 2019. « From What to How. An Overview of AI Ethics Tools, Methods and Research to Translate Principles into Practices ». *arXiv:1905.06876 [cs]*, mai. http://arxiv.org/abs/1905.06876.

Muehlhauser, Luke, et Louie Helm. 2012. « The Singularity and Machine Ethics ». Dans *Singularity Hypotheses: A Scientific and Philosophical Assessment*, édité par Amnon H. Eden, James H. Moor, Johnny H. Søraker, et Eric Steinhart, 101-26. The Frontiers Collection. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-32560-1_6.

Musiani, Francesca. 2013. « Governance by algorithms ». *Internet Policy Review*, 2(3).

Panch, Trishan, Heather Mattie, et Leo Anthony Celi. 2019. « The "Inconvenient Truth" about AI in Healthcare ». *Npj Digital Medicine* 2 (1): 1-3. https://doi.org/10.1038/s41746-019-0155-4.

Perry, Philip. 2017. « How artificial intelligence will revolutionize healthcare ». *Big Think* (blog). 2017. http://bigthink.com/philip-perry/how-artificial-intelligence-will-revolutionize-healthcare.

Proudfoot, Diane. 2011. « Anthropomorphism and AI: Turing's much misunderstood imitation game ». *Artificial Intelligence*, Special Review Issue, 175 (5): 950-57. https://doi.org/10.1016/j.artint.2011.01.006.

Raghupathi, W. 1997. « Health Care Information Systems ». Communications of the ACM. 1 août 1997. https://link.galegroup.com/apps/doc/A20036637/AONE?sid=lms.

Rumbold, John M. M., et Barbara K. Pierscionek. 2017. « A Critique of the Regulation of Data Science in Healthcare Research in the European Union ». *Bmc Medical Ethics* 18 (avril): 27. https://doi.org/10.1186/s12910-017-0184-y.

Russell, Stuart, Daniel Dewey, et Max Tegmark. 2015. « Research Priorities for Robust and Beneficial Artificial Intelligence ». *AI Magazine* 36 (4): 105-14.

Schwab, Klaus. 2016. *La quatrième révolution industrielle*. Dunod. Suisse: World Economic Forum.

Shameer, Khader, Kipp W. Johnson, Benjamin S. Glicksberg, Joel T. Dudley, et Partho P. Sengupta. 2018. « Machine Learning in Cardiovascular Medicine: Are We There Yet? » *Heart* 104 (14): 1156-64. https://doi.org/10.1136/heartjnl-2017-311198.

Shulman, Carl, Henrik Jonsson, et Nick Tarleton. 2009. « Machine ethics and superintelligence ». Dans , 95–97. Tokyo, Japan: Carson Reynolds and Alvaro Cassinelli.

Sternberg, Robert J., éd. 1982. *Handbook of Human Intelligence*. USA: Cambridge University Press.

Villani, Cédric. 2018. « Donner un sens à l'intelligence artificielle. Pour une stratégie nationale et européenne. » https://www.aiforhumanity.fr/pdfs/9782111457089_Rapport_Villani_accessible.pdf.

Chapitre 1 – Méthodologie

1. Problème de recherche

La présente thèse porte sur les enjeux d'une innovation responsable en ce qui a trait au développement de l'intelligence artificielle (IA) et de ses applications en santé. Plus particulièrement, il s'agit d'identifier quels sont les risques et les enjeux éthiques qui accompagnent ce développement et quels sont les défis de l'exercice de la responsabilité en vue d'une gouvernance éthique effective et pertinente.

Déterminer si une innovation est responsable demande de se pencher sur les bénéfices potentiels qui accompagnent l'utilisation de l'innovation en question (ici, les systèmes d'IA) afin d'évaluer la pertinence de leur utilisation (ici, dans le secteur de la santé). Les différents systèmes d'IA présentent en effet autant d'opportunités de valoriser les données massives relatives à la santé. Ils connaissent différentes applications prometteuses, allant de l'aide à la décision médicale par des systèmes experts (Kononenko 2001) à la médecine de précision (*ex.* ciblage pharmacologique) (Hamet et Tremblay 2017), en passant par la prévention individualisée grâce à des trajectoires de santé élaborées sur la base de marqueurs biologiques (Torkamani et al. 2017).

Cependant, des préoccupations éthiques pressantes relatives à l'impact de l'IA sur nos sociétés émergent avec le recours grandissant aux algorithmes pour analyser un nombre croissant de données relatives à la santé ainsi que la réduction de la supervision humaine de nombreux processus automatisés (Floridi et Taddeo 2016; Cath et al. 2016). Ces préoccupations sont relatives à la protection de la vie privée et de la confidentialité (Lahlou 2008; Azencott 2018; Stahl et Wright 2018), au respect du consentement libre et éclairé (Chow-White et al. 2015; Jones, Kaufman, et Edenberg 2018; Christen et al. 2016), au manque de transparence au regard des décisions algorithmiques (Ananny et Crawford 2018; Selbst et Barocas 2018), à la déshumanisation du soin (Coeckelbergh 2015), aux biais que pourraient perpétuer les algorithmes (Kim 2016; Altman, Wood, et Vayena 2018; Friedler, Scheidegger, et Venkatasubramanian 2016) ou encore à l'utilisation sécuritaire des systèmes d'IA et des potentiels mésusages (Brundage et al. 2018). Ces

technologies peuvent en effet représenter un risque, selon la capacité de la société à les gouverner et à les adopter de manière responsable, qu'elles concernent les risques qui émergent de leur utilisation en vue de bénéfices en santé ou lorsqu'utilisées à des fins qui n'étaient pas imaginées ou désirées lors de leur conception (Brundage 2016).

Face à ces préoccupations, de nombreuses initiatives nationales et internationales se sont penchées sur la définition de principes éthiques directeurs, de lignes directrices et de recommandations de politiques publiques en vue d'un développement responsable de l'IA (Amnesty International 2018; Déclaration de Montréal IA Responsable 2018; IEEE 2017; Villani 2018; AI HLEG 2019). L'importance d'une recherche socialement responsable a été mise de l'avant (Brundage 2016; Cath et al. 2016; Floridi et Taddeo 2016). Elle suppose que les chercheurs et les ingénieurs du domaine se soucient des impacts bénéfiques futurs de leurs travaux en demeurant attentifs aux dangers et aux enjeux éthiques potentiels qui y sont associés (Russell, Dewey, et Tegmark 2015; Brundage 2016; Moor 2006; Sharkey 2008). Le secteur de la santé n'échappe pas à cette nécessité. Une innovation numérique en santé responsable demande alors de limiter les risques tout en assurant les bénéfices de l'utilisation des systèmes d'IA. Cependant, si définir les principes éthiques directeurs du développement de l'IA a fait l'objet de nombreuses initiatives, des difficultés apparaissent relativement à leur mise en pratique (Mittelstadt 2019; Jobin, Ienca, et Vayena 2019), soit relativement à la manière de les opérationnaliser, mais également à l'identification des acteurs responsables de le faire.

L'innovation numérique en santé s'accompagne en effet de différents défis en ce qui concerne l'attribution de la responsabilité. Selon Noorman (2016), l'informatisation croissante de la société s'accompagne d'une érosion systématique de la responsabilité, compliquant la prévention des dommages et des risques associés à l'utilisation de systèmes d'IA. Des questionnements relatifs à la possibilité de tenir pour responsables les développeurs des conséquences qui accompagnent l'utilisation des algorithmes qu'ils développent apparaissent, considérant la nature de plus en plus autonome de ces algorithmes (Noorman 2016). De plus, l'implication d'une multitude d'acteurs dans l'utilisation des systèmes d'IA en santé, de la génération des données à l'application médicale (ex. développeurs, scientifiques des données, professionnels de santé, patients), demande de se

pencher sur le partage des responsabilités entre ces différentes parties prenantes, et non seulement celle des personnes à l'origine de leur conception (Chartrand, 2017; Noorman, 2016). Également, l'implication des secteurs à la fois public et privé dans ces développements, aux intérêts et aux normes potentiellement différents, complique la mise en place d'un encadrement éthique effectif en ce qui a trait à l'innovation responsable (Sharon, 2016).

Les bénéfices potentiels majeurs, les avancées rapides dans le domaine et l'hyperbole dont bénéficie actuellement l'IA ne laissent pas de doute quant au recours croissant à ces technologies, qui semble s'accompagner de transformations majeures et prometteuses pour le système de santé. Cependant, ces transformations demandent de répondre aux différentes préoccupations éthiques et sociales qui accompagnent l'utilisation de systèmes d'IA en santé. Il est ainsi essentiel de déterminer quels sont les défis de l'exercice et du partage des responsabilités entre les différentes parties prenantes du développement des systèmes d'IA et des données massives en santé en vue d'une innovation responsable.

2. Question, objectifs et propositions de recherche

2.1. Question de recherche

La présente thèse est structurée autour de la question de recherche suivante :

> Quels sont les défis du partage et de l'exercice de la responsabilité face aux risques et enjeux éthiques soulevés par l'utilisation de systèmes d'intelligence artificielle en santé en vue d'informer une innovation responsable ?

2.2. Objectifs de la thèse

Cette thèse comprend trois principaux objectifs de recherche :

1. Identifier les bénéfices, les risques et enjeux éthiques qui accompagnent l'utilisation de systèmes d'intelligence artificielle en santé.

2. Identifier les défis de la mise en place d'une gouvernance éthique adaptée, notamment en ce qui a trait à l'exercice de la responsabilité des parties prenantes de l'innovation numérique en santé.
3. Informer la mise en place d'un encadrement éthique pertinent et effectif pour une innovation responsable en ce qui a trait au développement des systèmes d'intelligence artificielle en santé.

2.3. Propositions de recherche

Le terme « proposition de recherche » est ici préféré à celui « d'hypothèse de recherche » car il est plus adapté à l'approche choisie pour les fins de cette analyse (cf. Section 5.1. du présent chapitre) et est plus fidèle au processus itératif qui a mené à leur formulation. La pertinence de ce terme est ici défendue par D'Amboise et Audet (1996) :

> Qu'il soit novice ou expérimenté, le chercheur aura une opinion quant à la réponse la plus probable aux questions de recherche. Il exprimera ces opinions sous forme de propositions de recherche, de telles propositions tenant place d'hypothèses de recherche. Comme dans certaines études à caractère plus inductif elles ne sont pas à proprement parler vérifiées au cours du processus de recherche, nous préférons le terme 'proposition' à celui d''hypothèse', évitant ainsi toute confusion. Ces propositions expriment les relations les plus probables entre les variables d'intérêt. S'inspirant de ses connaissances limitées du phénomène à observer, le chercheur élabore des propositions plausibles, susceptibles d'être modifiées ultérieurement à la lumière de faits nouveaux (D'Amboise et Audet 1996, p. 80).

Trois propositions de recherche guident la présente thèse :

1. Le développement des systèmes d'intelligence artificielle en santé s'accompagne de bénéfices et de risques et par conséquent d'enjeux éthiques qu'il est nécessaire de considérer d'un point de vue de l'innovation responsable.
2. La nature disruptive des technologies en jeu vient défier l'application des lignes directrices et des cadres éthiques existants, en particulier en ce qui a trait au partage et à l'exercice de la responsabilité des parties prenantes du système de santé (*ex.* institutions privées et institutions publiques, professionnels de santé, patients).
3. Une innovation responsable demande de répondre aux défis de l'exercice de la responsabilité dans le contexte de l'utilisation des systèmes d'intelligence artificielle en santé en vue d'adapter l'encadrement éthique existant.

3. Cadre de référence théorique

Le cadre de référence donne ici quelques repères interprétatifs vis-à-vis de l'analyse réalisée dans la présente thèse, inspiré de la vision défendue par Paquette (2007) qui préconise l'utilisation d'un cadre de référence théorique large et souple, multidisciplinaire et composé de connaissances générales à propos du phénomène à étudier.

Paquette (2007) défend l'intérêt d'un cadre théorique multiréférencé pour la recherche monographique constructiviste utilisant une approche analytique inductive[15]. Si la présente thèse ne constitue pas une monographie, le cadre de référence théorique développé par l'auteur reste un modèle pertinent, notamment car il est conçu en vue d'une analyse inductive, permet l'alliance de différents types de théories (ex. institutionnelles et organisationnelles) et guide une recherche visant à produire de nouvelles connaissances sur la base de propositions théoriques (Paquette, 2007).

Plus précisément, le cadre de référence de Paquette (2007) est composé de quatre pôles :

1) Un pôle paradigmatique qui rend compte des préconceptions du chercheurs, lesquelles orientent notamment ses choix méthodologiques et théoriques. Dans le contexte de son étude, le pôle paradigmatique de Paquette renvoie principalement aux paradigmes constructiviste et relativiste.

2) Un pôle stratégique, qui présente la stratégie retenue pour réaliser la recherche.

3) Un pôle technique, qui explique le choix des méthodes de collecte de données.

4) Un pôle théorique, qui « présente les orientations privilégiées quant à l'approche de traitement et d'analyse des données, d'interprétation et de validation des résultats » (Paquette, 2007, p. 7).

[15] Le cadre développé par Paquette (2007) est décrit dans le contexte d'une étude qui vise contribuer à l'élaboration conceptuelle de la notion de transaction sociale sur la base de l'étude d'évènements transactionnels.

Dans le cadre du présent chapitre, les éléments qui relèvent du pôle paradigmatique tel que conçu par Paquette se retrouvent dans la section 7.2. *Limites relatives à la posture de recherche*. Les pôles « stratégique » et « technique » sont présentés en détail dans les sections 4 et 5 relatives à la collecte et à l'analyse des données. Le pôle théorique du cadre de Paquette est enfin développé autour de quatre principaux types de « repères interprétatifs » dans les sections qui suivent, permettant d'identifier les éléments à prendre en compte en vue de répondre à la question de recherche. Ces quatre types de repères interprétatifs sont construits autour de la notion **d'innovation responsable**, perspective selon laquelle les enjeux éthiques de l'utilisation des systèmes d'IA en santé et les discussions de la coconstruction de la Déclaration de Montréal (voir section 4 : *Collecte des données* du présent chapitre) sont explorés. Ces référents théoriques portent ainsi sur : 1) la définition de l'innovation responsable et les notions qui y sont rattachées; 2) les éléments en lien avec la responsabilité face au risque technologique, qu'ils soient issus de la littérature en éthique et en bioéthique ou de lignes directrices institutionnelles; 3) la conception de la responsabilité retenue pour l'analyse de la présente thèse et 4) le contexte général de l'innovation responsable, soit un bref aperçu macrosociologique qui donne quelques repères quant au développement technologique.

3.1. Quelques notions relatives à l'innovation responsable

L'innovation responsable renvoie au terme d'innovation et recherche responsable (IRR),[16] qui connaît un regain d'intérêt ces dernières années, en particulier en Europe (Owen, Macnaghten, et Stilgoe 2012) depuis, entre autres, que la Commission européenne s'est penchée sur la définition du terme (Barré 2011; Owen, Macnaghten, et Stilgoe 2012). Selon la définition de René von Schomberg de la Commission Européenne :

> L'innovation responsable est un processus transparent et interactif par lequel les acteurs sociaux, les chercheurs et les innovateurs collaborent pour l'acceptabilité éthique, la durabilité et la pertinence sociétale (*societal desirability*) de l'innovation – permettant ainsi l'insertion des avancées des sciences et des techniques dans la société (tirée d'une conférence, dans Barré, 2011, p. 406)

[16] Pellé et Reber (2016) distinguent la recherche et l'innovation essentiellement sur la base de leur temporalité : la recherche est plus longue que l'innovation, notamment car elle ne répond pas à des impératifs de mise sur le marché des technologies développées et qu'elle nécessite beaucoup de temps pour produire des connaissances robustes. Parce-qu'il semble de plus en plus difficile de cloisonner distinctement les deux et qu'elles sont interdépendantes (*cf.* Chapitre 6), le terme « innovation » sera utilisé tout au long de la thèse, pouvant référer soit à la recherche, soit à l'innovation.

L'IRR tente ainsi de répondre aux préoccupations relatives à l'apparition de risques (éthiques) non anticipés, aux limitations de leur gestion, et à l'orientation de l'innovation vers des objectifs sociaux appropriés en vue du bien commun (Barré 2011; Owen, Macnaghten, et Stilgoe 2012). L'IRR est un processus en devenir pour aligner la recherche et l'innovation avec les valeurs, besoins et attentes de la société. Celui-ci en appelle à différents niveaux de responsabilité scientifique, celle-ci pouvant prendre plusieurs formes : une responsabilité qui relève plutôt de la déontologie ou plutôt de la responsabilité sociale de la science (*broader impact)* (Davis and Laas, 2014).

Relativement à la responsabilité qui relève de la « déontologie », elle fait écho au domaine de la conduite responsable en recherche (CRR), approche principalement nord-américaine qui fait joindre l'éthique de la recherche (principalement centrée sur les enjeux de recherches avec des participants) et l'intégrité scientifique (principalement centrée sur le chercheur), dans une vision globale de conduite éthique de la recherche (Voarino et al. 2019). Ce domaine s'intéresse particulièrement à la responsabilité et à l'intégrité des chercheurs et est souvent centré sur la définition et l'identification de manquements (comme la fraude ou la faute)[17], régulièrement associé à trois méconduites classiques : la falsification, la fabrication de résultats et le plagiat (Resnik 2003; Steneck et Bulger 2007; Bouter et al. 2016). Ces domaines ont ainsi pour vocation de définir les déterminants d'une recherche dite « responsable », mais restent souvent cloisonnés à la définition de conduite exemplaire dans la pratique quotidienne de la recherche et s'adresse principalement aux chercheurs, étudiants et personnel de la recherche.

Cependant, la raison d'être des impératifs de la CRR, de l'éthique de la recherche, et de l'intégrité scientifique relève bien d'une certaine responsabilité sociale : ces impératifs permettent de préserver la confiance du public en la science (Koepsell 2017; Resnik 2017; Resnik 2011), et

[17] Différentes lignes directrices de CRR visent cependant des approches positives, comme par exemple le Cadre de référence des trois organismes sur la conduite responsable de la recherche (2016) qui défend une approche éducative et vise à favoriser la sensibilisation, ou la Politique sur la conduite responsable en recherche des Fonds de recherche du Québec (2014) qui promeut une vision positive de la conduite responsable et définit des pratiques exemplaires.

définissent la base d'un contrat entre science et société selon d'importantes valeurs morales et sociales et différentes obligations réciproques :

> It is incumbent upon scientists to communicate with the public, and to interact in ways that are both educational and ethical because science and the public stand in mutually beneficial relationships to one another, and are also mutually dependent (Koepsell 2017 p. 85).

Owen, Macnaghten, et Stilgoe (2012) décrivent trois principales caractéristiques émergentes de l'IRR :

> 1) la démocratisation de la gouvernance d'intention (une science *pour* la société) qui appelle à la délibération inclusive sur ce que la science doit apporter à la société, au-delà de simplement définir ce qu'elle ne doit pas faire ;
>
> 2) l'institutionnalisation de la capacité à satisfaire les attentes de la société (une science *avec* la société) qui met l'accent sur l'intégration de ces mécanismes de délibération à l'intérieur et autour des processus de recherche ; et
>
> 3) la redéfinition de la responsabilité scientifique selon de nouvelles obligations qui vont au-delà de celles de la CRR, qui ne se limitent pas à celles des chercheurs, et qui nécessitent une réflexion sur la constitution, le financement et la mise en œuvre de programmes d'innovation.

Parce-qu'elle est centrée sur « *the consequences of techno-scientific applications to a deliberate and continuous ex ante consideration of what we collectively want science and innovation to do or not to do*" (Lehoux et al. 2018 p. 277), la mission de l'IRR revêt ainsi une dimension éthique. Le terme recouvre cependant également une dimension juridique et politique relativement aux interprétations de la responsabilité (Barré 2011). C'est à la dimension éthique de l'IRR que s'intéresse la présente thèse.

L'engagement des parties prenantes de la société civile par l'entremise de processus de délibération inclusifs à toutes les étapes de l'innovation et de la recherche étant au cœur de la mission de l'IRR (Lehoux et al. 2018; Barré 2011; Pellé et Reber 2016), les discussions ayant eu lieu lors de la coconstruction citoyenne de la Déclaration de Montréal pour un développement

responsable de l'IA peuvent être mise à profit pour réfléchir à l'IRR dans le contexte de la présente thèse (voir section 4 : *Collecte des données* du présent chapitre).

3.2. Théories éthiques de la responsabilité face au risque technologique

Parce-qu'elle revêt une dimension éthique et prospective, l'innovation responsable invite à se pencher sur les théoriques éthiques de la responsabilité face au risque technologique. Afin de situer l'analyse de la thèse, il est nécessaire de se baser sur une théorie éthique déterminant les bases de la responsabilité face à l'innovation technologique qui ne soit ni trop permissive ni trop restrictive afin de répondre aux enjeux de l'innovation responsable des systèmes d'IA en santé. La gestion des risques dans le cadre l'innovation responsable fait en effet écho à une approche théorique à mi-chemin entre un impératif technologique et le principe de responsabilité de Hans Jonas. Le premier implique l'essai de tout le possible technoscientifique, car celui-ci est gage de progrès et qu'il s'agit de respecter la liberté et l'indépendance de la science (Hottois, 1990). Cet impératif n'implique cependant aucune responsabilité face aux risques qui nous intéressent. Au contraire, avec l'avènement d'une société du risque, Ulrick Beck décrit, en 1986, l'existence d'une certaine irresponsabilité organisée (Beck, 1986). Celle-ci découle d'un certain consensus sur le progrès technologique, qui est perçu comme gage de bien-être social et fait état de normes au sein de nos sociétés occidentales (Beck, 1986). L'ensemble de ces facteurs amène à une « non-responsabilité » de la science, notamment car celle-ci est rendue indicible par l'idée d'une irréversibilité du progrès (Beck, 1986). À l'opposé, le principe de responsabilité de Hans Jonas (1979) présuppose une responsabilité envers les générations futures, nécessitant de toujours préférer le scénario du pire afin de prévenir les risques qui pourraient accompagner l'innovation (Jonas, 1979). Cette « heuristique de la peur » a cependant été souvent critiquée pour son catastrophisme réducteur et l'immobilisme anti-progrès qu'elle risque d'entraîner (Godard, 2000).

Entre ces deux extrêmes (un refus ou un essai systématique de toutes innovations technologiques), l'idée d'une « *voie moyenne* » proposée par Hottois (1990) semble ici appropriée, afin de tirer profit des perspectives prometteuses que nous offre l'IA tout en minimisant les risques associés à son développement. Celle-ci prescrit d'accompagner le développement technologique d'une pensée réflexive et critique respectant « l'essai de certains possibles technoscientifiques en fonction de critères à déterminer » (Hottois, 1990).

De telles approches supposent ainsi d'agir face à l'existence de la probabilité d'un risque. Le caractère prospectif des enjeux qui font l'objet de cette analyse demande de renseigner la mise en place d'une innovation responsable selon un principe de précaution[18]. Ce principe invite à agir même dans l'incertitude, en l'absence de connaissances, aux vues de la gravité et de l'irréversibilité des dommages encourus (Bourg et Papaux 2008). Il permet la mise en place de mesures révisables en fonction de l'avancée des connaissances (Bourg et Papaux, 2008), la preuve ne venant plus de la démonstration de la présence de risques mais de l'absence de risque (Vineis, 2005). Cet aspect permet de justifier la mise en place de mesures pour répondre aux risques et enjeux éthiques qui accompagnent l'utilisation des systèmes d'IA en santé qu'ils soient avérés ou non. Le principe de précaution demande cependant d'agir dans des conditions où la probabilité de l'apparition du risque est raisonnable. Cette non-absence doit être confrontée aux bénéfices afin de mettre en place une action adaptée (Lecourt, 2007). Si le principe de précaution est parfois vu comme possiblement immobiliste, l'absence de consensus sur sa définition tout comme le caractère révisable des mesures qu'il implique en font un principe flexible (Kuhlau et al. 2011). Le principe de précaution fait ainsi écho à l'obligation de « décision dans l'ignorance » de l'IRR (Barré 2011) et à l'approche proportionnelle de l'éthique appliquée.

Une approche proportionnelle demande de tenir compte d'un équilibre entre risques et bénéfices, limitant les mécanismes de gestion dans le cas de risques de niveau « minimal », soit ceux des recherches « où la probabilité et l'ampleur des préjudices éventuels découlant de la participation à la recherche ne sont pas plus grandes que celles des préjudices inhérents aux aspects de la vie quotidienne du participant qui sont associés au projet de recherche. » (ÉPTC2, 2018, p. 23). L'INSPQ préconise également un système de gestion qui s'adapte au niveau de risques, en les pondérant avec les bénéfices, considérant qu'il y a des risques dont le niveau est « acceptable », soit *«* suffisamment faible pour ne pas nécessiter de mesure de contrôle supplémentaire bien que ces mesures puissent malgré tout être mises en place sur une base volontaire pour réduire encore davantage le risque » (Institut national de santé publique du Québec 2016, p. XII). Lorsque le

[18] Le principe de responsabilité de Hans Jonas, parce-qu'il implique une responsabilité prospective, est parfois considéré comme étant à l'origine du principe de précaution.

risque dépasse cette limite, il peut être équivalent aux bénéfices potentiels de la recherche (risques moyens) ou largement supérieur (représentant un risque élevé ou sérieux). Quand le risque n'est pas minimal, il est nécessaire de confronter celui-ci aux bénéfices en vue d'une approche proportionnelle. Considérant les promesses de l'utilisation des systèmes d'IA en santé (*cf.* Chapitre 2) mais également les limites et enjeux éthiques qui accompagnent leur développement (*cf.* Chapitre 3), une approche proportionnelle est un bon référent en ce qui a trait à la description des enjeux d'une innovation responsable en santé.

3.3. Une certaine conception de la responsabilité (morale)

Pour répondre à la question de recherche, il est également nécessaire de se pencher brièvement sur différentes conceptions de la responsabilité (morale), afin de réfléchir aux défis de son exercice mais aussi car la responsabilité est un des concepts clés de l'innovation responsable – la responsabilité morale étant, pour Pellé et Reber (2016), la plus appropriée pour guider les réflexions relatives à l'IRR.

Il existe de nombreuses conceptions de la responsabilité et plus particulièrement de la responsabilité morale. Métayer (2001) en décrit deux grands courants d'interprétation : la responsabilité formelle et la responsabilité sollicitude. Selon la première, la responsabilité est interprétée comme un constituant essentiel de l'agir intentionnel, elle est rétrospective et situe la responsabilité dans la liberté du sujet autonome. Elle constitue le pôle négatif de la responsabilité morale (déterminant par exemple les critères de l'individu à blâmer). La responsabilité sollicitude correspond quant à elle au pôle positif de l'exigence morale. Elle est prospective et situe la responsabilité dans l'appel à l'aide d'un autrui vulnérable.

Selon Métayer (2001), ces visions « absolutisantes » de la responsabilité issues de la philosophie morale souffrent d'un « déficit de conceptualisation » :

> En cherchant à fonder la responsabilité morale dans des termes absolus, elles lui donnent une extension illimitée qui la rend diffuse, insaisissable, sans prise sur la réalité sociale et historiquement située de la vie morale (p. 19).

Pour « rompre avec l'approche abstraite des éthiques normatives », Métayer propose une approche pragmatique de la responsabilité qui tente de raccorder la responsabilité morale à la vie sociale, « éloignée de toutes prétentions fondationnelles et vouée plutôt à l'analyse des pratiques sociales d'attribution des responsabilité » (Métayer, 2001, p. 23). Cette conception renvoie au pragmatisme

philosophique, qui définit en effet des engagements éthiques en écartant les questions relatives aux fondements conceptuels ou à l'authenticité d'une norme (Ralph 2018). Une vision pragmatique de la responsabilité demande ainsi de ne pas considérer seulement lesdits fondements (ceux qui supportent la responsabilité) mais également les attitudes, les expériences vécues ou les normes dont elle est issue ou qu'elle façonne (Smiley, 1992; Ralph 2018).

Le pragmatisme, dans sa conception classique, met l'accent selon Ralph (2018) sur des compétences telles que le jugement pratique ou l'habileté à décider (par exemple, en évaluant les conséquences d'une action) dans le contexte d'un engagement à améliorer un problème social partagé. Ce problème peut survenir face à l'évolution de la pratique indépendamment des normes existantes, voire au détriment de celles-ci (Ralph 2018). Une responsabilité peut alors apparaître comme une réponse pragmatique au problème créé, qui prend son sens « pour et dans la pratique » (Ralph 2018). C'est ainsi selon une conception pragmatique de la responsabilité que sont conduites les analyses de la présente thèse, qui n'a pas pour objectif de proposer une nouvelle conception de la responsabilité morale ou de préciser les fondements théoriques qui la supportent mais bien d'identifier quels sont les enjeux d'une innovation responsable dans les contextes d'application des systèmes d'IA – qui peut être considéré ici comme le problème social partagé.

Dans ce sens, Smiley (1992) situe la signification de la responsabilité morale non pas dans le discours rationnel mais dans différentes caractéristiques de la pratique sociale (ex. les coutumes, habitudes et croyances partagées par les membres d'une communauté particulière ou les règles de conduites institutionnalisées). C'est ainsi dans des contextes d'interaction circonscrits qu'il est possible de préciser les exigences de la responsabilité (qu'elle soit positive ou négative) (Métayer, 2001). Une approche pragmatique ne suppose pas cependant de nier l'existence des fondements théoriques qui supportent la norme :

> « Pragmatic constructivism recognizes as real the ideational (epistemic and normative) structures of a particular 'community of practice', but it values the norms of that community only to the extent critical inquiry establishes their ability to effect practical consequences that ameliorate lived social problems" (Ralph 2018 p. 175).

La responsabilité pragmatique permet ainsi de se détacher des critères de la responsabilité formelle ou des exigences de la responsabilité sollicitude, bien qu'elle autorise la reconnaissance de leur valeur.

C'est dans la même logique que Le Moigne (2006) défend l'utilité d'une responsabilité pragmatique et solidarisante (plutôt que partagée) qui se situe entre deux heuristiques qui tentent de définir la responsabilité : celles de la peur (avec le principe de responsabilité de Jonas) et celle de l'espérance (avec le Principe Espérance de Bloch), et demande de mobiliser la responsabilité éthique de chacun. Dans le cadre de l'attribution de la responsabilité morale aux « systèmes intelligents » (notamment, différents systèmes d'IA ou robots), Crnkovic et Persson (2008) défendent également une approche pragmatique, où la responsabilité est vue comme un mécanisme de régulation sociale qui vise à renforcer les actions considérées comme « bonnes », tout en minimisant celles considérées comme « mauvaises ».

L'identification des défis d'une innovation numérique en santé responsable est ainsi réalisée selon la structure de ce que Métayer nomme les « interpellations responsabilisantes », qui demandent de considérer la responsabilité et son partage comme une pratique de responsabilisation et qui supposent que les positions morales et les définitions de la responsabilité des acteurs qui prévalent ne cessent d'évoluer (Métayer 2001; Ralph 2018; Smiley 1992). La responsabilité pragmatique permet ainsi de se dégager des conceptions de la responsabilité morale traditionnelle et d'utiliser une conception de la responsabilité plus proche du sens commun, en considérant (mais en ne se restreignant pas) aux critères spécifiques de la responsabilité formelle (ex. la conscience des conséquences, le libre arbitre des agents responsables ou encore la causalité entre l'action de l'agent et ses conséquences (Chartrand 2017; Métayer 2001; Noorman 2016) ou aux exigences de la responsabilité positives (ex. la prospectivité ou le souci de l'autre) (Pellé et Reber 2016; Métayer 2001). Pour les fins de la présente thèse, une conception pragmatique de la responsabilité autorise ainsi une certaine flexibilité dans l'analyse. Notamment, cette conception a permis de considérer l'ensemble des défis potentiels relatifs à l'exercice de la responsabilité, tels que soulignés par les citoyens, et ce quelle qu'en soit leur(s) conception(s).

Bien que les objectifs de la présente thèse ne soient pas orientés vers une contribution théorique sur les fondements de la responsabilité, il a été nécessaire de se pencher sur quelques notions qui relèvent de la philosophie morale – en particulier en ce qui a trait à la responsabilité formelle – ne serait-ce que pour comprendre et délimiter les défis de l'exercice de la responsabilité (pragmatique). Il s'agit essentiellement de la notion d'agentivité (*cf.* Chapitre 5 et 6) qui se trouve être au cœur des défis qui ressortent du discours citoyen, mais aussi dans une moindre mesure de l'autonomie, de la responsabilité morale et du libre arbitre. Il n'aurait pas été possible de décrire les craintes et les attentes citoyennes sans les nommer, considérant la nature de ces dernières. C'est autour de ces notions[19] que les discussions citoyennes ont ainsi pu être mises en perspective. À défaut d'un apport théorique, ces notions – riches, complexes et contestées - de la philosophie morale sont ici mises en lien avec une conception pragmatique de la responsabilité, le but étant d'identifier les défis de l'exercice de la responsabilité et de les mettre en perspective en vue d'informer les réflexions sur la mise en place d'un encadrement éthique effectif en ce qui a trait à l'utilisation des systèmes d'IA, en recherche comme dans la pratique clinique.

3.4. Un aperçu du contexte général de l'innovation responsable

Il est également nécessaire de citer les théories qui se penchent sur le contexte général dans lequel se développe l'innovation numérique en santé, lequel n'est pas sans conséquences sur l'attribution de la responsabilité et la conduite éthique de la recherche. Ce contexte nous est donné par différentes théories sociales du champ des études sur les sciences (*Sciences studies*). Il est notamment reconnu que la science et la recherche s'inscrivent aujourd'hui dans un nouveau mode de production de la connaissance (*New production of Knowledge*), qui est passé d'un mode 1 (où les connaissances sont produites selon les intérêts de la communauté académique et de manière relativement indépendante) à un mode 2 (où les connaissances sont produites en fonction des contextes d'application) (Gibbons et al., 1994). Dans le même registre, les études décrivant l'apparition d'un néolibéralisme scientifique (Lave, Mirowski, et Randalls 2010) sont aussi pertinentes : elles mettent en évidence une rupture qui se manifeste par le passage de la science considérée comme bien public à une science perçue comme une source de compétitivité

[19] Il s'agirait plutôt de concepts, mais le terme « notions » est ici préféré car c'est bien comme notions seulement qu'elles sont utilisées dans la présente thèse.

économique (Bonneuil et Joly, 2010). Trois grandes caractéristiques du néolibéralisme ont été relevées par Bonneuil et Joly (2013) (reprenant Lave, Mirowski et Randalls 2010):

1) « Le financement public de la recherche est dépassé par le financement privé ». On parle ainsi d'une « financiarisation de l'innovation »;
2) On observe « une extension des droits de propriété intellectuelle sur la connaissance scientifique » avec une remise en cause de la science ouverte;
3) Il existe aujourd'hui une emprise croissante des approches empruntées au management des entreprises sur le système de production des connaissances (notamment, le financement public indexé sur la base de la performance des chercheurs).

Ces trois caractéristiques permettent de prendre en comptes certains des intérêts qui pourrait guider la conduite de la recherche et de l'innovation (ex. la performance du chercheur, les intérêts économiques), potentiellement en conflit avec ceux de l'avancée des connaissances pour le seul bien commun. Ces considérations demandent de porter une attention particulière au contexte du développement responsable de l'IA, où s'observe effectivement une imbrication étroite du secteur public et privé. Faisant écho au *mode 2* de Gibbons, les connaissances sont produites selon un impératif d'utilité dès le début de la résolution du problème, qu'importe que cette utilité soit celle du gouvernement, de l'industrie ou de la société en général (Gibbons 1994). Si les connaissances selon ce *mode 2* répondent à des facteurs d'offre et de demande, cela se fait au-delà des considérations commerciales (le marché n'est plus seulement économique mais devient un marché de la connaissance) (Gibbons 1994). Ainsi, la diversité du financement et de sa distribution entre organisations et institutions différentes (*ex.* firmes, industries, universités) augmente le nombre de sites où la recherche se déroule, conduisant à une massification de l'éducation et de la recherche (Gibbons 1994) et ainsi à un plus grande compétition. Si la transdisciplinarité du *mode 2* est dynamique et conduit à faire émerger de potentielles avancées plus grandes, le lieu d'utilisation de cette massification de la connaissances et comment elle se développe devient alors aussi difficile à prédire que pour les recherches plus disciplinaires du *mode 1* (Gibbons 1994), et par là même complexifie grandement la tâche de la prévention des risques et de leur encadrement.

De plus, Beck (1986) a mis en évidence un processus de double-inversion entre la science et son environnement (ou phénomène de scientifisation secondaire), celle-ci intégrant les craintes

associées aux effets non-intentionnels de ses utilisations (Beck, 1986) et jouant ainsi dans la société du risque « un rôle fondamentalement ambivalent, source à la fois de bien-être et de danger, à la fois contestée et omniprésente » (Bonneuil et Joly 2013 citant Beck, 1986 p. 23). Ce rôle ambivalent de la science fait écho au concept de double-usage de la recherche (qui concerne les recherches aux usages considérés à la fois « bons » et « mauvais »). De ce concept découle un dilemme de nature éthique qui présente un conflit entre les valeurs défendables de protection de la santé et de la sécurité publique *versus* la promotion du progrès scientifique (Selgelid 2009b; 2009a). Il s'agit en effet, d'une part, de protéger la liberté académique afin d'assurer l'avancée des connaissances et, d'autre part, d'empêcher le mésusage potentiel et de prévenir ou gérer les risques issus de son développement (Miller et Selgelid 2007; Selgelid 2009a; Resnik 2009). Outre sa dimension éthique et prospective, il est intéressant de nommer ce concept dans le contexte de l'innovation responsable, car la littérature sur le double-usage accorde aux scientifiques une responsabilité face aux utilisations problématiques de leurs recherches, et ce malgré la non-intentionnalité qui leur est attribuée. Ce dernier correspond à une obligation morale de ne pas nuire, et donc de prévenir le mésusage associé aux recherches dans le cadre des capacités et habilités du chercheur, dans la mesure où ce mésusage est raisonnablement prévisible (Kuhlau et al. 2008). Cette responsabilité implique un engagement actif des scientifiques à évaluer les risques potentiels de leurs recherches (Kuhlau et al. 2008). La responsabilité scientifique selon ces théories s'étendrait donc au-delà de la simple conduite de la recherche et, faisant échos au concept d'IRR ou au principe de précaution, reconnaissent une responsabilité prospective de la science et des chercheurs face aux conséquences potentiellement négatives de leurs travaux.

Ainsi, l'innovation responsable implique de se pencher sur les différentes valeurs morales qui guident le développement d'une science *avec* et *pour* la société, ouvrant la porte à la redéfinition de la responsabilité scientifique. Cette responsabilité de la science face à la société est reprise dans les théories éthiques de la responsabilité face au risque technologique, qui demandent de prévenir les risques tout en favorisant les bénéfices des recherches et innovations selon un principe de précaution ou une approche proportionnelle. Concept central de l'innovation responsable, la responsabilité dans sa conception pragmatique permet de guider la réflexion de la présente thèse afin de ne pas restreindre l'analyse à des critères formels de responsabilité morale. Différentes

théories des études sur les sciences invitent à tenir compte du contexte de l'innovation responsable, des différents intérêts qui pourraient guider les avancées de la science ainsi que des facteurs qui pourrait complexifier la gestion des risques et l'attribution de la responsabilité. Partant de ces différents constats et pour répondre aux différents objectifs précédemment mentionnés, l'analyse de la présente thèse vise à mettre en relation différents écrits avec les données collectées dans le cadre de la coconstruction de la Déclaration de Montréal pour un développement responsable de l'IA.

4. Collecte des données

4.1. Méthodes de collecte

4.1.1. Le projet de la Déclaration de Montréal pour un développement responsable de l'intelligence artificielle

Le projet de la Déclaration de Montréal est une initiative de l'Université de Montréal visant à faire évoluer la première version d'une déclaration de principes éthiques en vue du développement responsable de l'IA, lancée le 3 novembre 2017 lors du Forum IA responsable au palais des congrès de Montréal (Déclaration de Montréal IA Responsable 2018). Le projet de la collecte initiale[20] visait ainsi à définir des principes éthiques directeurs et des recommandations de politiques publiques pour un développement responsable de l'IA. Afin d'encadrer le débat entourant le développement de l'IA et les valeurs qu'il mobilise, dans la première phase de la Déclaration, sept valeurs ont été identifiées : bien-être, autonomie, justice, vie privée, connaissance, démocratie et responsabilité (Déclaration de Montréal IA Responsable 2018).

Afin de mobiliser « l'intelligence collective », de nombreuses activités de coconstruction ont été réalisées et plus de 500 citoyens ont été consultés. Les consultations ont mobilisé les citoyens autour des développements de l'IA dans cinq secteurs d'activité clés : éducation, monde du travail, système judiciaire et police prédictive, santé et ville intelligente et objets connectés. L'analyse de ces données a conduit à la réalisation d'une cartographie de l'ensemble des risques et enjeux éthiques associés au développement de l'IA (*cf.* Déclaration de Montréal IA Responsable

[20] Le projet de la collecte initiale est un projet de mobilisation des connaissances et non un projet de recherche.

2018 p. 117) mais également à mettre en évidence des pistes de solutions envisagées par les citoyens pour répondre à ces enjeux. Ainsi, **12 catégories de risques et enjeux éthiques** ainsi que **11 catégories de pistes de solution** ont été mises en évidence (*cf*. Déclaration de Montréal IA Responsable 2018, p.106 et 107), sur la base de l'analyse de *post-it* rédigés par les participants, d'affiches de synthèses rédigées collectivement et de comptes rendus issus des membres de l'équipe de la Déclaration.

L'ensemble de ces analyses a permis d'aboutir à la rédaction d'une Déclaration de **10 principes éthiques** (*cf*. Tableau 7 du Chapitre 5 de la présente thèse) pour un développement responsable de l'IA, eux-mêmes divisés en nombreux sous-principes d'applications. Cette analyse a également permis la formulation de **8 recommandations en vue de politiques publiques** (*cf*. Déclaration de Montréal IA Responsable 2018, p. 311 à 318).

4.1.2. Données issues de la coconstruction de la Déclaration de Montréal

Les données analysées dans la présente thèse relèvent des discussions qui ont porté sur le thème de la santé uniquement *(cf.* Annexe 2 : *Informations sur les tables de coconstruction)*, et diffèrent de celles analysées dans le cadre du projet de la Déclaration de Montréal. En effet, l'analyse réalisée dans le cadre de la présente thèse se base sur la retranscription des enregistrements disponibles – et non sur les comptes rendus, *post-its* et affiches. S'il existe forcément quelques redondances avec les analyses réalisées une première fois pour la Déclaration de Montréal (ex. la mise en évidence que les citoyens demandent la création de lois), une analyse des retranscriptions s'est avérée complémentaire, étant plus ciblée et ayant permis d'explorer en profondeur les discussions citoyennes relativement à l'utilisation des systèmes d'IA dans le secteur de la santé. La grille d'analyse a été conçue de manière à répondre aux objectifs de recherche (qui diffèrent des objectifs de la collecte initiale) et ne vise donc pas à identifier des catégories de risques et enjeux éthiques ou des pistes de solution mais la perception citoyenne sur les défis de l'exercice de la responsabilité.

Concernant les consultations en présentiel qui font l'objet de l'analyse de la présente thèse[21], elles se sont déroulées du 16 février 2018 au 6 avril 2018. Ces consultations se déclinent en trois principales activités : 1. Cafés citoyens; 2. Journées de coconstruction et 3. Groupes de

[21] Ce projet a été approuvé par le Comité d'éthique de la recherche en santé (CERES) de l'Université de Montréal en date du 23 juillet 2018 (Certificat 18-106-CERES-D, *cf*. Annexe 1).

discussion. La coconstruction a réuni un total de 45 tables dont 12 sur le thème de la santé (six cafés-citoyens; cinq tables de coconstruction et un groupe de discussion).

Les cafés citoyens étaient des ateliers d'une durée de trois heures, qui se sont déroulés dans différentes bibliothèques de la ville de Montréal. La participation se basait sur la prise de connaissance de scénarios prospectifs se déroulant en 2025 et mettant en scène différents développements potentiels de l'IA en santé ainsi que certains des enjeux éthiques associés (*cf.* Annexe 3 : *Scénarios* et Tableau 1 qui résume ces scénarios). Les citoyens participants étaient invités à réagir aux enjeux soulevés dans les scénarios sur la base des principes de la Déclaration.

Tableau 1. – Les quatre scénarios utilisés pour stimuler les discussions ayant eu lieu autour du thème de la santé.

Titre	**Résumé**	**Applications d'IA en jeu** (*cf.* Chapitre 2)	**Principaux enjeux éthiques mobilisés** (*cf.* Chapitre 3)
Les jumeaux numériques	Olivier reçoit un pronostic de dépression par le biais d'une notification sur son téléphone intelligent. Le pronostic est établi sur la base du diagnostic de ses jumeaux numériques, qui ont des profils de santé similaires déterminés sur la base des données partagées dans un nuage de santé mondial.	Santé mobile	Déshumanisation du patient Vie privée et confidentialité
L'assurance santé discriminante	Olivier doit répondre à une recommandation algorithmique concernant ses habitudes de vie s'il ne veut pas voir sa franchise d'assurance santé majorée de 10%. Il peut sinon choisir de ne plus partager ses données dans le nuage de santé mondial, au risque d'être défavorisé sur les listes d'attente pour accéder à des soins.	Santé mobile	Justice sociale
Un robot pour maintenir une personne âgée à domicile	Une famille fait l'acquisition d'un robot de soins à domicile, Vigilo, afin d'éviter à leur grand-mère Soline une institutionnalisation dans une maison de retraite. Le robot est responsable d'évaluer périodiquement l'évolution de la démence de Soline, de sa prise de médicaments et est capable de	Robots de soins	Déshumanisation des soins Consentement libre et éclairé Vie privée et confidentialité

	converser avec elle. Il transfert régulièrement les données collectées à une équipe de soin et fait également des suivis à sa famille. La présence du robot conduit Soline à éviter de plus en plus les relations sociales avec d'autres humains; elle finit même par s'attacher à lui.		
Décision thérapeutique à l'hôpital	Soline se rend à l'hôpital après un malaise vagal. Après plusieurs examens, les résultats de l'analyse de l'algorithme d'aide au diagnostic sont contradictoires avec le pronostic du cardiologue de Soline, expert réputé du domaine. Le médecin ne sait alors plus s'il doit suivre son intuition ou respecter les recommandations du système expert. Il hésite à informer Soline de la situation.	Systèmes experts	Consentement libre et éclairé Sécurité

Chaque **café citoyen** était organisé en tables thématiques regroupant chacune cinq à douze personnes. Les discussions se sont déroulées en deux phases : la première visait l'identification d'enjeux éthiques, la seconde la formulation de recommandations. Les réactions, commentaires et idées émanant de chacune des tables étaient synthétisés par les participants (avec l'aide de l'animateur) sur une affiche par le biais de *Post-its* dans un premier temps (phase 1) et sur une affiche mettant en évidence les enjeux et recommandations à adresser de manière prioritaire dans un second temps (phase 2). Cette seconde phase se terminait par la rédaction d'une première page (une) de journal fictive mettant en scène la recommandation identifiée comme prioritaire par la table. Les phases 1 et 2 de l'activité se déroulaient sur 1h30 au total. Une session introductive de 45 minutes posait les bases nécessaires à la compréhension de l'activité, sous la forme de conférences, présentant les principes de la Déclaration et la démarche de coconstruction. Les données ont ainsi été collectées par le biais des affiches rédigées en groupe, des notes des animateurs et auxiliaires de recherche, et de l'enregistrement des discussions.

Les **tables de coconstruction** suivaient la même approche de collecte de données que les cafés citoyens, mais se déroulaient sur une journée complète. Ainsi, la phase 1 se déroulait le matin

(1h30) et la phase 2 l'après-midi (1h30) (soit un total de 3h) suivies d'une session plénière d'une heure.

Des **groupes de discussion** ont également été organisés, récoltant le même type d'informations mais suivant un format plus libre, adapté au contexte. Les données issues du groupe de discussion qui s'est déroulé lors de l'intervention dans le cours « Éthique et politique de santé » le 29 mars 2018 ont été intégrées à la présente analyse (1h30 de discussion sur les enjeux et les recommandations qui concernent le développement de l'IA en santé lors de 3h de cours). Bien que le format de collecte était plus libre que les cafés citoyens et tables de coconstruction, les discussions de ce groupe portaient sur les mêmes scénarios et étaient organisées autour des mêmes phases (enjeux et recommandations). La collecte ne s'est cependant pas centrée autour des principes de la Déclaration et les résultats n'ont pas été synthétisés sur les affiches. Les données récoltées lors de ce groupe de discussion sont cependant complémentaires et demeurent pertinentes pour l'objet de l'analyse de la présente thèse.

4.2. Caractérisation de l'échantillon

Le recrutement des participants a été effectué par différents moyens par l'équipe de la Déclaration (*cf.* Annexe 4 : *Informations sur le recrutement*). Deux types de recrutement ont été réalisés pour les fins de la coconstruction : 1) un recrutement ciblé d'experts et parties prenantes du développement de l'IA pour chacun des secteurs concernés; et 2) un recrutement plus large de citoyens par le biais de différentes publicités. Les discussions analysées dans la présente thèse ont réuni 68 participants. Aucun critère d'exclusion n'a été appliqué (genre, groupe d'âge, milieu socio-professionnel) : toute personne souhaitant participer aux discussions a été acceptée dans la limite des places disponibles (voir la section « Limites » concernant la représentativité de l'échantillon).

Considérant que cette analyse est issue de l'utilisation secondaire d'un projet de mobilisation des connaissances, la collecte de données sociodémographiques (cf. Annexe 5 : *Questionnaire sociodémographique*) en vue de caractériser l'échantillon n'a pas été réalisée de

manière systématique. Sur les plus de 370[22] participants à la coconstruction en présentiel avant le 6 avril 2018, 61% ont accepté de remplir un questionnaire démographique non obligatoire. À titre illustratif, les données en question, issues du rapport de la Déclaration de Montréal (Déclaration de Montréal IA Responsable 2018), sont résumées dans le Tableau 2.

Tableau 2. – Caractérisation de l'ensemble des citoyens ayant participé à la coconstruction en présentiel de la Déclaration de Montréal reproduite telle que présentée dans le rapport de la Déclaration.

Catégorie	Sous-Catégorie	N
Genre (227 citoyens)	Masculin	122
	Feminin	105
Âge (60 citoyens)	12 ans et moins	0
	13-18 ans	1
	19-34 ans	12
	35-44 ans	15
	45-54 ans	11
	55-64 ans	14
	65-74 ans	5
	75 et plus	2
Scolarité (218 citoyens)	Aucun certificat, diplôme ou grade	3
	Diplôme d'études secondaires ou l'équivalent	2
	Titre d'études postsecondaires	6
	Diplôme d'études collégiales	14
	Certificat universitaire inférieur au baccalauréat	9
	Baccalauréat	**56**
	Certificat universitaire supérieur au baccalauréat	**87**
	Diplôme en médecine	**2**
	Doctorat acquis	**39**
Secteurs d'activité (315 citoyens)	Administration publique	17
	Arts, spectacles et loisirs	16
	Autres	**49**
	Commerce de détail	3
	Énergie et ressources	3
	Enseignement	**36**
	Finance et assurances	12

[22] Le rapport indique que la consultation a été effectuée sur plus de 500 participants. Ce chiffre est approximatif car considérant la nature du processus de coconstruction, il n'a pas été possible de déterminer avec précision le nombre de personnes ayant participé (ce chiffre englobe en effet toutes les activités et pas seulement les activités en présentiel). Néanmoins, un rapport préliminaire mentionnait que les activités avaient réuni un total de 436 participants (66 ayant répondu à un questionnaire en ligne qui ne fait pas l'objet de la présente analyse, 370 ayant participé aux activités de coconstruction en personne). La caractérisation ayant été réalisée sur ce premier échantillon, c'est donc le n = 370 qui est utilisé ici.

	Gestion de société et d'entreprises	6
	Hébergement et services de restauration	1
	Information et culture	15
	Recherche (industrielle ou universitaire)	**46**
	Services professionnels, scientifiques et techniques	28
	Soins de santé, biotechnologies et assistance sociale	17
	Technologies de l'information	**63**
	Transport et entreposage	3

Les citoyens ayant répondu au questionnaire n'ont pas systématiquement répondu à toutes les questions. Dans le Tableau 2, le nombre de citoyens répondants est indiqué à côté de chaque catégorie. Il est à noter que pour la dernière catégorie (Secteur d'activité), 34% des répondants ont indiqué plus d'un secteur (le chiffre indiqué ne correspond ainsi pas au nombre de citoyens ayant répondu). Sur la base du chiffre le plus élevé (soit celui de la catégorie Genre), 227 participants ont rempli le questionnaire sociodémographique (soit environ 61% considérant les 370 citoyens consultés lors de la période mentionnée).

Conformément aux données collectées, une distribution relativement paritaire (122 hommes, 105 femmes) des citoyens ayant participé à la coconstruction de la Déclaration de Montréal est observée. La grande majorité des participants ayant rempli le questionnaire (184) détenaient un Baccalauréat ou un diplôme de niveau supérieur, l'échantillon représentant ainsi une population relativement scolarisée. Une grande proportion de citoyens participants a mentionné exercer dans le domaine des technologies de l'information (63), en recherche (46) ou en enseignement (36).

Il n'a pas été possible d'accéder spécifiquement aux données sociodémographiques des 68 participants aux discussions analysées dans la présente thèse. Cependant, les tours de table[23] de cinq des neuf tables analysées ont été enregistrés (soit les informations relatives à 38 participants

[23] Les tours de table correspondent au moment, en début d'activité, ou chaque participant se présente.

sur 68). Également, si le tour de table n'était pas disponible, les neuf participants du groupe de discussion étant des étudiants en bioéthique, leur domaine d'expertise a pu être déterminé, ainsi que celui de P49. Sur la base de ces informations, une caractérisation complémentaire a ainsi pu être réalisée. Les informations relatives à cette caractérisation sont résumées dans le Tableau 3.

Tableau 3. – Caractérisation complémentaire des 68 citoyens ayant participé aux discussions des tables santé.

Catégorie	Sous-Catégorie	N
Genre (68/68 citoyens)	Masculin	36
	Féminin	32
Domaines d'expertise[24] (48/68 citoyens)	Éthique, bioéthique et expertises associées	16
	Santé et expertises associées	11
	IA et expertises associées	10
	Autres	11

La quasi-parité homme-femme s'observe également dans l'échantillon des 68 citoyens qui ont participé aux discussions analysées dans la présente thèse (soit 36 hommes et 32 femmes). Parmi eux, les citoyens qui ont un domaine d'expertise relatif à l'éthique ou la bioéthique sont les plus nombreux (n =16), suivis par ceux qui ont une expertise relative à la santé (n = 11) ou une expertise autre (n=11). Il est à noter que certaines des expertises se chevauchent (ex. la bioéthique relève tant de la santé que de l'éthique; un représentant d'une compagnie qui utilise l'IA pour la médecine de précision a été considéré dans la catégorie « IA » alors que son expertise relève aussi de la santé).

Considérant le recrutement effectué, est entendu par « citoyens » dans la présente thèse les parties prenantes identifiées aux domaines concernés (ex. experts en droit de la santé, acteurs du secteur privé dans le domaine de l'IA en santé, patients-partenaires), qui possèdent une expertise reliée au développement de l'IA, à la bioéthique et/ou à la santé, et d'autres individus aux expertises variées. Bien qu'il n'ait pas été possible de les caractériser avec précision en vue de l'analyse, cet échantillon hybride a permis une confrontation des points de vue et expertises qui demeure intéressante considérant que les activités ont fait l'objet d'une approche consensuelle.

[24] Il n'a pas été possible de déterminer avec précision le secteur d'activité ou le niveau d'étude car les citoyens participants se sont présentés en fonction du rôle qu'ils exercent et qui est pertinent à considérer relativement à leur participation. Par exemple, les citoyens qui se sont présentés comme « patients-partenaires » ont été considérés dans la catégorie « Santé et expertises associées », quelle que soit la profession qu'ils exercent.

5. Analyse des données

5.1. Approche holistico-inductive et analyse thématique

Afin de mettre en évidence la vision des citoyens participants relativement aux défis de l'exercice de la responsabilité face à l'utilisation des systèmes d'IA en santé, les neuf enregistrements disponibles sur les 12 tables de coconstruction réalisées sur le thème de la santé ont été intégralement retranscrits avec le logiciel ExpressScribe et analysés par le biais du logiciel NVivo. Les notes, compte-rendus des discussions et retranscriptions des affiches réalisées par les participants ont servi de support à la retranscription et à l'analyse mais n'ont pas été codés dans le logiciel NVivo.

N'ont pas été retranscrites les explications relatives à l'activité (ex. consignes, lecture des scénarios) ni les longues hésitations. Afin de rester au plus proche de la parole citoyenne, le reste des discussions a été retranscrit fidèlement, incluant parfois les erreurs de grammaire ou les répétitions. Certains mots ou phrases dans les enregistrements étaient inaudibles. Pour refléter les passages dont la retranscription est incertaine, les mots ont été mis [entre crochets]. Afin de respecter l'anonymat des participants, leur nom a été remplacé par un identifiant unique qui commence par la lettre P pour « participant » ainsi qu'un chiffre déterminé en fonction de l'ordre de prise de parole (ex. P1, P2…P68).

Les extraits des retranscriptions ont été catégorisés selon une analyse thématique (Braun et Clarke 2006) suivant une approche holistico-inductive afin de permettre l'émergence de nouveaux sous-thèmes au fil de l'analyse. L'approche holistico-inductive est décrite par D'Amboise et Audet (1996) comme suit :

> Le terme « holistico » fait allusion au fait que le chercheur porte son attention sur l'ensemble du phénomène d'intérêt, c'est-à-dire qu'il cherche à comprendre ou à décrire en profondeur le phénomène dans son contexte et son environnement général. […] Quant au terme « inductif », il réfère à un raisonnement qui va du particulier au général, plus précisément qui débute par l'observation de phénomènes particuliers pour ensuite essayer de dégager une théorie plus générale de ces observations (p. 76).

L'approche holistico-inductive demande de s'intéresser à l'aspect qualitatif des données – et non pas à l'aspect quantitatif, à la signification des phénomènes ainsi que de déduire les idées à partir

des données elles-mêmes. Elle suppose une démarche itérative qui permet de préciser les questions, les propositions et le cadre de la recherche au fur et à mesure de l'analyse, selon un « va-et-vient continuel entre les faits observés, la théorie existante et la théorie émergente » (D'Amboise et Audet 1996). Cette approche laisse davantage de place à l'interprétation du chercheur, qui sélectionne des « unités » en fonction de la richesse de l'information qui pourra en être retirée.

5.2. Grille d'analyse

La grille d'analyse a été établie sur la base d'une étude pilote réalisée sur une table (soit environ 10% de l'échantillon). En accord avec une approche holistico-inductive, cette grille a évolué au fil de l'analyse et été révisée plusieurs fois. L'approche holistico-inductive demande qu'une certaine structure minimale soit établie au préalable, tout en laissant assez de flexibilité à l'induction (D'Amboise et Audet 1996). L'identification de certains défis de l'exercice de la responsabilité dans la littérature a permis d'établir une première grille d'analyse. Ainsi, les craintes et attentes citoyennes relatives au partage des responsabilités entre parties prenantes du développement responsable de l'IA qui ont émergées des délibérations ont été regroupées selon trois principales catégories tirées de la littérature scientifique sur le sujet. Les catégories ont été créées de manière à être mutuellement exclusives pour éviter le double codage (sauf celles des différents acteurs dans le cadre du défi des *mains multiples*, *cf.* Chapitre 5).

La grille d'analyse finale présente trois grandes catégories de défis relativement à l'exercice de la responsabilité (définie selon le cadre conceptuel comme une responsabilité pragmatique) : préserver les capabilités humaines, le problème des mains multiples (*many hands)* et l'agentivité artificielle (*cf*. Tableau 4). Considérant la démarche itérative, une quatrième catégorie a été ajoutée à mi-chemin entre la préservation des capabilités et l'agentivité artificielle. Pour chacune de ces catégories ont été répertoriées les craintes et les attentes citoyennes. Les craintes réfèrent soit aux défis de l'exercice de la responsabilité eux-mêmes, soit aux conséquences de ceux-ci. Les attentes réfèrent aux attentes normatives des citoyens relativement à ces défis ou à ces conséquences. Chaque catégorie regroupe de nombreuses sous-catégories. L'ensemble de ces catégories est présenté en détail dans le Chapitre 5 et discuté dans le Chapitre 6. Le Tableau 4 ci-dessous présente une version simplifiée des catégories qui sont ressorties des discussions citoyennes.

Tableau 4. – Présentation simplifiée de la grille d'analyse.

Notions mobilisées	Craintes	Attentes
Préserver les capabilités humaines	Technologies incapacitantes	Technologies capacitantes
Le problème des *mains multiples*	Déresponsabilisation et conséquences sur les soins et la santé	Responsabilité partagée sur la base d'un contrat social
Agentivité artificielle	Les systèmes d'IA sont des agents	Les systèmes d'IA doivent être des outils
Préserver les capabilités humaines – Agentivité artificielle	Remplacement des humains par les machines	Coopération humains-machines

Il est important de noter que cette analyse ne saurait rendre compte de l'ensemble des discussions ayant eu lieu autour des tables de coconstruction. Différentes catégories qui ne répondaient pas directement aux questions de recherche n'ont pas été inclues.

6. Notes sur le corpus de textes et les références

Alors que le Chapitre 5 présente les résultats empiriques de l'analyse des données issues de la coconstruction de la Déclaration de Montréal, cette analyse est mise en parallèle avec différentes sources (articles scientifiques ou littérature grise) qui ont servi de base pour la rédaction des autres chapitres. En majorité, les textes ont été identifiés par le biais d'une méthode « boule de neige » (ou *snowball methods*) (Pawson et al. 2005), incluant dans le corpus les références pertinentes identifiées au fil des lectures.

Afin de répondre aux questions de recherche, il a été nécessaire d'explorer la littérature de différents domaines. Des articles de littérature scientifique du domaine biomédical, de l'intelligence artificielle, des données massives, de la santé, de l'éthique ou de la bioéthique ont été sélectionnés afin d'identifier les bénéfices et les risques qui accompagnent l'utilisation des systèmes d'IA en santé. Également, différents rapports, déclarations de principes, politiques et lignes directrices relatifs à l'intelligence artificielle, aux données massives, à la bioéthique, à l'innovation responsable ou à l'éthique de la recherche ont été explorés en vue de comprendre les

enjeux de l'innovation numérique responsable. De plus, de nombreux sites web ont été examinés. Il s'agit essentiellement de sites d'entreprises qui développent des applications d'IA – notamment en santé, ou d'initiatives de partage de données (incluant les données de santé).

Si la lecture de la littérature et des références n'a pas fait l'objet d'une analyse systématique, une certaine saturation a été observée en ce qui a trait aux différents risques et enjeux éthiques qui accompagnent l'utilisation des systèmes d'IA (Chapitre 3) ou aux perspectives de leur utilisation en santé (Chapitre 2).

7. Limites

L'analyse réalisée dans la présente thèse connait trois principales catégories de limites, qui sont issues : 1) des conditions de la collecte initiale des données, 2) de la posture de recherche, et 3) de la conception du projet en lui-même.

7.1. Limites relatives aux conditions de la collecte initiale

Premièrement, concernant les limites relatives aux conditions de la collecte initiale, il est nécessaire de revenir sur le fait qu'il n'a pas été possible de faire une caractérisation systématique et complète des participants. La caractérisation partielle donne cependant une bonne idée de la constitution de l'échantillon, sans pouvoir affirmer de sa représentativité. En réponse à cette limite, aucune généralisation à l'ensemble de la population n'est opérée relativement à l'échantillon : il ne s'agit pas de refléter l'opinion d'une partie définie de la population vivant au Québec mais seulement de tirer des pistes de réflexions des discussions – par ailleurs hautement pertinentes - qui ont eu lieu lors de cette consultation. Ces limites sont d'ailleurs celles de toutes approches holistico-inductive : les résultats obtenus connaissant peu de validité externe, ils sont difficilement généralisables à une population; la validité du résultat dépend de la compétence du chercheur (D'Amboise et Audet 1996).

Deuxièmement, il est à noter que les scénarios ont été créés pour mettre en scène des risques et enjeux éthiques relatifs à des technologies qui seraient potentiellement utilisées dans un avenir proche, créant ainsi une apparente induction. L'analyse effectuée ne saurait en effet mettre en évidence quels sont les risques du développement de l'IA identifiés selon les citoyens mais plutôt

leur perception des risques et enjeux éthiques, préalablement identifiés en vue de développer les scénarios, et leur réaction face à ces risques. Également, considérant l'objectif du projet de la collecte initiale (développer des principes directeurs et recommandations de politiques publiques), les discussions ont été principalement orientées autour des risques (puisqu'il était question de les limiter) et non des bénéfices associés au développement de l'IA en santé. Si les citoyens ont à quelques reprises abordé lesdits bénéfices, ces discussions ont été limitées. D'autres consultations mériteraient donc d'être menées en ce sens afin de pouvoir tirer des conclusions relativement à la réelle acceptabilité sociale des technologies en question.

7.2. Limites relatives à la posture de recherche

Concernant les limites relatives à la posture de recherche, ce sont celles de toutes démarches inductives d'analyse qualitative. Sans s'étendre dans le détail d'un véritablement pôle paradigmatique tel que dans le cadre de Paquette (2007), il a été jugé pertinent de préciser brièvement que la posture de recherche adoptée pour la présente thèse se situe dans un courant plutôt constructiviste, posture qui suppose que la connaissance est construite selon une certaine représentation de la réalité (Hacking, 1995; Avenier, 2011) considérée comme incertaine, diversifiée et subjective (Anadón 2006). Comme le défend Paquette (2007) relativement à l'ontologie constructiviste des recherches inductives, la réalité est ainsi construite pour répondre aux nécessités du chercheur et de ses intentions :

> Ce que produit le chercheur n'est pas le reflet de la réalité, mais un construit susceptible d'expliquer temporairement une réalité à partir des actions quotidiennes des acteurs, de leurs interactions, de leurs interrogations et de leurs transactions (p. 8).

Plus précisément, Avenier (2011) considère qu'il existe deux hypothèses fondatrices communes aux approches constructivistes :

> 1) Il n'est pas possible de séparer le système observant du système observé dans le processus de production des connaissances (Avenier, 2011). Comme dans toute approche de recherche similaire, il n'est jamais possible de parer complètement aux biais implicites inhérents à tout chercheur (Bourdieu 2004). Dans le cadre de la présente thèse, il est ainsi nécessaire de mentionner que l'induction de l'analyse réalisée n'est que relative. En effet, les tables de coconstruction sur la santé ont été animées et les données analysées une première fois les fins du rapport de la Déclaration de Montréal par l'auteure de la présente thèse. Ainsi, l'analyse de

cette thèse a démarré avec une idée assez claire des éléments qui pourraient ressortir des discussions citoyennes.

2) L'élaboration de connaissances est considérée comme « un acte de construction de représentations forgées par des humains pour donner sens aux situations dans lesquelles ils interviennent » (Avenier, 2011, p. 376). La genèse de sens se fait notamment par des formes communes de recherche qualitative, comme la catégorisation (Mucchieli 2007). C'est ici la rigueur du codage et les catégories créées qui permettent de parer la limite issue de la posture de recherche. De plus, il est à noter que la première vague d'analyse (celle pour la Déclaration) n'avait pas été faite sur la base des enregistrements.

Enfin, l'ontologie constructiviste selon Paquette (2007) permet de considérer que la construction de la connaissance se fait sur la base de valeurs qui apparaissent dans les contradictions et les consensus. Cette vision est particulièrement appropriée considérant la démarche délibérative de coconstruction de la Déclaration de Montréal, ou les discussions citoyennes reflètent les valeurs des participants qui se sont exprimés au travers de leurs craintes et de leurs attentes, en visant des propositions consensuelles.

7.3. Limites relatives au projet en lui-même

Concernant les limites relatives au projet en lui-même, il est nécessaire de mentionner que le détail de l'analyse réalisée ne saurait rendre compte de l'ensemble des discussions ayant eu lieu autour des tables de coconstruction sur le thème de la santé. Différentes catégories n'ont pas été abordées, car elles ne répondent pas directement aux objectifs de recherche susmentionnés, telle que les discussions relatives à l'universalité d'accès aux soins, à la justice sociale ou à la transformation de la relation médecin-patient – quand elles n'avaient pas trait à la responsabilité. Cette limite est aussi inhérente à l'approche holistico-inductive :

> Il [le chercheur] laisse venir à lui toutes les informations susceptibles de jeter un éclairage sur le phénomène d'intérêt, quitte à les éliminer plus tard si elles ne s'avèrent pas utiles (D'Amboise et Audet 1996).

Également, il est essentiel de tenir compte du fait que les concepts mobilisés pour comprendre les craintes et les attentes citoyennes (ex. agentivité, responsabilité morale, capabilités

– *cf*. Chapitre 5) sont des concepts qui réfèrent à une littérature extrêmement riche – et non consensuelle – en philosophie morale. Il n'est pas nécessaire pour les fins de cette thèse d'explorer tous ces concepts ni de prendre position quant à leur définition, mais plutôt de voir comment leur conception pragmatique informe les réflexions bioéthiques. Cependant, une analyse approfondie de ces concepts à la lumière des craintes et des attentes citoyennes est sans nul doute d'intérêt et mériterait des recherches subséquentes.

De plus, tel que défendu dans le Chapitre 4 et dans le Chapitre 6, une approche pragmatique invite à trouver des réponses dans les contextes d'applications de l'utilisation des systèmes d'IA, selon les spécificités inhérentes à chaque situation. Ceci demanderait, en vue de formuler des réponses précises aux défis identifiés dans la présente thèse et de valider la pertinence pratique d'une conception pragmatique de la responsabilité, de réaliser des études de cas. Cependant, l'apport des analyses qui suivent se cantonne à la dimension « macro » de l'IRR telle que définit par Barré (2011), soit celle « des débats sociétaux et des visions de plus long terme d'élaboration des règles et codes de conduite » (p. 407). La présente thèse se limite ainsi à l'utilisation d'une conception pragmatique pour : 1) défendre l'utilité d'une approche contextuelle et 2) identifier les défis qui pourraient survenir lors de l'application d'une telle approche, sur la base de la littérature et des discussions citoyennes. Les deux autres dimensions présentées par Barré (la dimension méso des agences de financement en charge des priorités scientifiques et la dimension micro du « chercheur dans son laboratoire »), si elles sont considérées au fil des chapitres, mériteraient d'être explorées de manière approfondie lors d'études futures.

Enfin, il est important de souligner (bien que cela reste implicite) que l'orientation de la recherche, des analyses et des conclusions a été majoritairement réalisée selon une approche de l'éthique plutôt occidentale, avec un intérêt particulier pour les développements (éthiques et techniques) ayant cours au Canada, aux États-Unis et en France. Comme abordé lors du Chapitre 4, le domaine de l'éthique de l'IA (en santé comme dans d'autres secteurs) mériterait plus de recherche sur la base d'approches qui réfèrent à d'autres systèmes de valeurs.

Références bibliographiques

AI HLEG, (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.

Altman, M., A. Wood, et E. Vayena. 2018. « A Harm-Reduction Framework for Algorithmic Fairness ». *IEEE Security Privacy* 16 (3): 34-45. https://doi.org/10.1109/MSP.2018.2701149.

Amnesty International, Access Now. 2018. « The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems ». RightsCon Toronto. https://www.accessnow.org/toronto-declaration.

Anadón, M. 2006. La recherche dite « qualitative »: de la dynamique de son évolution aux acquis indéniables et aux questionnements présents. *Recherches qualitatives*, 26(1), 5-31.

Ananny, Mike, et Kate Crawford. 2018. « Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability ». *New Media & Society* 20 (3): 973-89. https://doi.org/10.1177/1461444816676645.

Avenier, M. 2011. Les paradigmes épistémologiques constructivistes : post-modernisme ou pragmatisme ?. *Management & Avenir*, 43(3), 372-391. doi:10.3917/mav.043.0372.

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): 20170350. https://doi.org/10.1098/rsta.2017.0350.

Barré, Rémi. 2011. « Des concepts à la pratique de l'innovation responsable : à propos d'un séminaire franco-britannique ». *Natures Sciences Societes* Vol. 19 (4): 405-9.

Bourdieu, Pierre. 2004. *Esquisse pour une auto-analyse*. Raisons d'agir.

Bourg, Dominique, et Alain Papaux. 2008. « Des limites du principe de précaution: OGM, transhumanisme et détermination collective des fins ». *Économie publique/Public economics*, nº 21. http://economiepublique.revues.org/pdf/7932.

Bouter, Lex M., Joeri Tijdink, Nils Axelsen, Brian C. Martinson, et Gerben ter Riet. 2016. « Ranking major and minor research misbehaviors: results from a survey among participants of four World Conferences on Research Integrity ». *Research Integrity and Peer Review* 1 (novembre): 17. https://doi.org/10.1186/s41073-016-0024-5.

Braun, Virginia, et Victoria Clarke. 2006. « Using thematic analysis in psychology ». *Qualitative Research in Psychology* 3 (2): 77-101. https://doi.org/10.1191/1478088706qp063oa.

Brundage, Miles. 2016. « Artifical intelligence and responsible innovation ». Dans *Fundamental issues of artificial intelligence*, Springer International Publishing Switzerland, 543-53. V.C. Müller.

Brundage, Miles, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, et Bobby Filar. 2018. « The malicious use of artificial intelligence: Forecasting, prevention, and mitigation ». *arXiv preprint arXiv:1802.07228*.

Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, Social Sciences and Humanities Research Council of Canada.2016. Tri-Agency Framework: Responsible Conduct of Research. *Secretariat on Responsible Conduct of Research*. https://rcr.ethics.gc.ca/eng/documents/Framework2016-CadreReference2016_eng.pdf

Cath, Corinne J. N., Sandra Wachter, Brent Mittelstadt, Mariarosaria Taddeo, et Luciano Floridi. 2016. « Artificial Intelligence and the "Good Society": The US, EU, and UK Approach ». SSRN Scholarly Paper ID 2906249. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=2906249.

Chartrand, Louis. 2017. « Agencéité et responsabilité des agents artificiels ». *Éthique publique. Revue internationale d'éthique sociétale et gouvernementale*, nº vol. 19, n° 2 (novembre). https://doi.org/10.4000/ethiquepublique.3068.

Chow-White, Peter A., Maggie MacAulay, Anita Charters, et Paulina Chow. 2015. « From the Bench to the Bedside in the Big Data Age: Ethics and Practices of Consent and Privacy for Clinical Genomics and Personalized Medicine ». *Ethics and Information Technology* 17 (3): 189-200. https://doi.org/10.1007/s10676-015-9373-x.

Christen, Markus, Josep Domingo-Ferrer, Bogdan Draganski, Tade Spranger, et Henrik Walter. 2016. « On the Compatibility of Big Data Driven Research and Informed Consent: The Example of the Human Brain Project ». Dans *The Ethics of Biomedical Big Data*, édité par Brent Daniel Mittelstadt et Luciano Floridi, 199-218. Law, Governance and Technology Series. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-33525-4_9.

Coeckelbergh, Mark. 2015. « Artificial Agents, Good Care, and Modernity ». *Theoretical Medicine and Bioethics* 36 (4): 265-77. https://doi.org/10.1007/s11017-015-9331-y.

Crnkovic, Gordana Dodig, et Daniel Persson. 2008. « Sharing Moral Responsibility with Robots: A Pragmatic Approach. » Dans *Frontiers in Artificial Intelligence and Applications* Volume 173, édité par Anders Holst, Per Kreuger, et Peter Funk. IOS Press Books.

ÉPTC2 : Conseil de recherches en sciences humaines du Canada, Conseil de recherches en sciences naturelles et en génie du Canada, et Instituts de recherche en santé du Canada. 2018. « Énoncé de politique des trois Conseils : Éthique de la recherche avec des être humains ». http://www.ger.ethique.gc.ca/fra/policy-politique_tcps2-eptc2_2018.html.

D'Amboise, G., et J. Audet. 1996. *Le projet de recherche en administration. Un guide général à sa préparation. Chapitre 4. L'approche holistico-inductive*. Université Laval, Faculté des sciences de l'administration.

Déclaration de Montréal IA Responsable. 2018. « Rapport de la Déclaration de Montréal pour un Développement Responsable de l'Intelligence Artificielle ». https://docs.wixstatic.com/ugd/ebc3a3_d806f109c4104c91a2e719a7bef77ce6.pdf.

Floridi, Luciano, et Mariarosaria Taddeo. 2016. « What Is Data Ethics? » *Philosophical Transactions of the Royal Society A-Mathematical Physical and Engineering Sciences* 374 (2083): 20160360. https://doi.org/10.1098/rsta.2016.0360.

Fonds de recherche du Québec (FRQ). 2014. Politique sur la conduite responsable en recherche. FRQ. Disponible sur : www.frq.gouv.qc.ca/hxtNx87eSZkT/wp-content/uploads/Politique-sur-la-conduite-responsable-en-recherche_FRQ_sept-2014.pdf

Friedler, Sorelle A., Carlos Scheidegger, et Suresh Venkatasubramanian. 2016. « On the (im)possibility of fairness ». *arXiv:1609.07236 [cs, stat]*, septembre. http://arxiv.org/abs/1609.07236.

Gibbons, Michael. 1994. *The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies*. SAGE.

Hamet, Pavel, et Johanne Tremblay. 2017. « Artificial intelligence in medicine ». *Metabolism*, Insights Into the Future of Medicine: Technologies, Concepts, and Integration, 69 (Supplement): S36-40. https://doi.org/10.1016/j.metabol.2017.01.011.

Hacking, I. 1995 ; Entre science et réalité : la construction sociale de quoi ? traduit de l'anglais par Baudouin Jurdant. Paris : Éditions La Découverte 2001

IEEE, Institute of Electrical and Electronics Engineers. 2017. « Ethically aligned design - Version 2 - For Public Discussion ». https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf.

Institut national de santé publique du Québec. 2016. « La gestion des risques en santé publique au Québec : cadre de référence ». https://www.inspq.qc.ca/pdf/publications/2106_gestion_risques_sante_publique.pdf.

Jobin, Anna, Marcello Ienca, et Effy Vayena. 2019. « Artificial Intelligence: the global landscape of ethics guidelines ». *arXiv:1906.11668 [cs]*, juin. http://arxiv.org/abs/1906.11668.

Jones, M. L., E. Kaufman, et E. Edenberg. 2018. « AI and the Ethics of Automating Consent ». *IEEE Security Privacy* 16 (3): 64-72. https://doi.org/10.1109/MSP.2018.2701155.

Kim, Pauline T. 2016. « Data-Driven Discrimination at Work ». *William & Mary Law Review* 58: 857-936.

Koepsell, David. 2017. « Duties of Science to Society (and Vice Versa) ». Dans *Scientific Integrity and Research Ethics*, 85-95. SpringerBriefs in Ethics. Springer, Cham. https://doi.org/10.1007/978-3-319-51277-8_8.

Kononenko, Igor. 2001. « Machine learning for medical diagnosis: history, state of the art and perspective ». *Artificial Intelligence in Medicine* 23 (1): 89-109. https://doi.org/10.1016/S0933-3657(01)00077-X.

Kuhlau, Frida, Stefan Eriksson, Kathinka Evers, et Anna T. Höglund. 2008. « Taking Due Care: Moral Obligations in Dual Use Research ». *Bioethics* 22 (9): 477-87. https://doi.org/10.1111/j.1467-8519.2008.00695.x.

Kuhlau, Frida, Anna T. Höglund, Kathinka Evers, et Stefan Eriksson. 2011. « A Precautionary Principle for Dual Use Research in the Life Sciences ». *Bioethics* 25 (1): 1-8. https://doi.org/10.1111/j.1467-8519.2009.01740.x.

Lahlou, Saadi. 2008. « Identity, Social Status, Privacy and Face-Keeping in Digital Society ». *Social Science Information* 47 (3): 299-330. https://doi.org/10.1177/0539018408092575.

Lave, Rebecca, Philip Mirowski, et Samuel Randalls. 2010. « Introduction: STS and Neoliberal Science ». *Social Studies of Science* 40 (5): 659-75. https://doi.org/10.1177/0306312710378549.

Le Moigne, Jean-Louis. 2006. « L'expérience de la responsabilité appelle l'éthique, qui appelle l'épistémique, qui appelle la pragmatique… ». *Postface, in J.-J. Rosé, ed., Responsabilité sociale de l'entreprise. Pour un nouveau contrat social, Brussels: De Boeck*.

Lehoux, Pascale, Fiona A. Miller, Dominique Grimard, et Philippe Gauthier. 2018. « Anticipating Health Innovations in 2030–2040: Where Does Responsibility Lie for the Publics? » *Public Understanding of Science* 27 (3): 276-93. https://doi.org/10.1177/0963662517725715.

Métayer, Michel. 2001. « Vers une pragmatique de la responsabilité morale ». *Lien social et Politiques*, nº 46: 19-30. https://doi.org/10.7202/000320ar.

Miller, Seumas, et Michael J. Selgelid. 2007. « Ethical and Philosophical Consideration of the Dual-Use Dilemma in the Biological Sciences ». *Science and Engineering Ethics* 13 (4): 523-80. https://doi.org/10.1007/s11948-007-9043-4.

Mittelstadt, Brent. 2019. « AI Ethics – Too Principled to Fail? » SSRN Scholarly Paper ID 3391293. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=3391293.

Moor, J. H. 2006. « The Nature, Importance, and Difficulty of Machine Ethics ». *IEEE Intelligent Systems* 21 (4): 18-21. https://doi.org/10.1109/MIS.2006.80.

Mucchielli, A. 2007. Les processus intellectuels fondamentaux sous-jacents aux techniques et méthodes qualitatives. Recherches qualitatives, 3, 1-27.

Noorman, Merel. 2016. « Computing and Moral Responsibility ». Dans *The Stanford Encyclopedia of Philosophy*, édité par Edward N. Zalta, Winter 2016. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2016/entries/computing-responsibility/.

Hottois, Gilbert. 1990. *Le paradigme bioéthique. Une éthique pour la technoscience*. Bruxelles, ERPl Science.

Owen, Richard, Phil Macnaghten, et Jack Stilgoe. 2012. « Responsible Research and Innovation: From Science in Society to Science for Society, with Society ». *Science and Public Policy* 39 (6): 751-60. https://doi.org/10.1093/scipol/scs093.

Paquette, Danielle. 2007. « Le rôle du cadre de référence théorique dans une recherche monographique constructiviste ». *Recherches qualitatives* 27 (1): 3–21.

Pawson, Ray, Trisha Greenhalgh, Gill Harvey, et Kieran Walshe. 2005. « Realist Review--a New Method of Systematic Review Designed for Complex Policy Interventions ». *Journal of*

Health Services Research & Policy 10 Suppl 1 (juillet): 21-34. https://doi.org/10.1258/1355819054308530.

Pellé, Sophie, et Bernard Reber. 2016. *Ethique de la recherche et innovation responsable*. ISTE editions. Vol. 2. Innovation et recherche responsables.

Ralph, Jason. 2018. « What Should Be Done? Pragmatic Constructivist Ethics and the Responsibility to Protect ». *International Organization* 72 (1): 173-203. https://doi.org/10.1017/S0020818317000455.

Resnik, David. 2017. « What Is Ethics in Research & Why Is It Important? » National Institute of Environmental Health Sciences. 2017. https://www.niehs.nih.gov/research/resources/bioethics/whatis/index.cfm.

Resnik, David B. 2003. « From Baltimore to Bell Labs: Reflections on Two Decades of Debate about Scientific Misconduct ». *Accountability in Research* 10 (2): 123-35. https://doi.org/10.1080/08989620300508.

———. 2009. « What is "dual use" research? A response to Miller and Selgelid ». *Science and engineering ethics* 15 (1): 3–5.

———. 2011. « Scientific Research and the Public Trust ». *Science and Engineering Ethics* 17 (3): 399-409. https://doi.org/10.1007/s11948-010-9210-x.

Russell, Stuart, Daniel Dewey, et Max Tegmark. 2015. « Research Priorities for Robust and Beneficial Artificial Intelligence ». *AI Magazine* 36 (4): 105-14.

Selbst, Andrew D., et Solon Barocas. 2018. « The Intuitive Appeal of Explainable Machines ». SSRN Scholarly Paper ID 3126971. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=3126971.

Selgelid, Michael J. 2009a. « Governance of dual-use research: an ethical dilemma ». *Bulletin of the World Health Organization* 87 (9): 720-23. https://doi.org/10.1590/S0042-96862009000900017.

———. 2009b. « Dual-Use Research Codes of Conduct: Lessons from the Life Sciences ». *NanoEthics* 3 (3): 175-83. https://doi.org/10.1007/s11569-009-0074-y.

Sharkey, Noel. 2008. « The Ethical Frontiers of Robotics ». *Science* 322 (5909): 1800-1801. https://doi.org/10.1126/science.1164582.

Sharon, Tamar. 2016. « The Googlization of Health Research: From Disruptive Innovation to Disruptive Ethics ». *Personalized Medicine* 13 (6): 563-74. https://doi.org/10.2217/pme-2016-0057.

Smiley, Marion. 1992. Moral Responsibility and the Boundaries of Community. Power and Accountability from a Pragmatic Point of View. Chicago, The University of Chicago Press. 296

Stahl, B. C., et D. Wright. 2018. « Ethics and Privacy in AI and Big Data: Implementing Responsible Research and Innovation ». *IEEE Security Privacy* 16 (3): 26-33. https://doi.org/10.1109/MSP.2018.2701164.

Steneck, Nicholas H., et Ruth Ellen Bulger. 2007. « The History, Purpose, and Future of Instruction in the Responsible Conduct of Research ». *Academic Medicine: Journal of the Association of American Medical Colleges* 82 (9): 829-34. https://doi.org/10.1097/ACM.0b013e31812f7d4d.

Torkamani, Ali, Kristian G. Andersen, Steven R. Steinhubl, et Eric J. Topol. 2017. « High-Definition Medicine ». *Cell* 170 (5): 828-43. https://doi.org/10.1016/j.cell.2017.08.007.

Villani, Cédric. 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., V. Couture, S. Mathieu-Chartier, J. C. Bélisle-Pipon, E. St-Hilaire, B. Williams-Jones, F. J. Lapointe, C. Noury, M. Cloutier, et P. Gauthier. 2019. « Mapping responsible conduct in the uncharted field of research-creation: A scoping review ». *Accountability in Research* 26 (5): 311-46. https://doi.org/10.1080/08989621.2019.1620607.

Chapitre 2 – Les promesses de l'utilisation des systèmes d'intelligence artificielle en santé

Le développement de systèmes d'intelligence artificielle (IA) en santé fait partie intégrante d'une nouvelle médecine « haute définition » (Torkamani et al. 2017) qui se veut prédictive, préventive et personnalisée en tirant profit de la quantité inédite de données aujourd'hui disponibles. Cette nouvelle ère de la médecine, quel qu'en soit le champ d'application, est caractérisée par l'utilisation de données massives analysées par le biais de différentes méthodes d'IA en vue de les valoriser. Cette alliance permet d'envisager de nombreuses avenues prometteuses, qu'elles concernent la médecine de précision, l'aide au diagnostic par le biais de systèmes experts, ou la prévention et l'accompagnement des patients par le biais d'agents de soin artificiels. Dans l'optique d'une innovation numérique en santé responsable, il est essentiel de considérer les bénéfices potentiels majeurs de l'avènement de ces technologies, mais aussi de se pencher sur les transformations que les systèmes d'IA et les données massives sont susceptibles d'engendrer en ce qui a trait à la santé et sa gestion.

1. L'intelligence artificielle et les données massives au cœur de l'innovation numérique en santé

1.1. Une quantité inédite de données relatives à la santé

Eric Schmidt, directeur général de Google, annonçait en 2010 lors de la conférence *Techonomy* que nous produisons aujourd'hui autant d'information tous les deux jours que tout ce que nous avons pu produire depuis le début de notre civilisation jusqu'à 2003 (Brown et al. 2018). Si certains remettent cette affirmation en question (ex. Moore 2011), il est avéré que la quantité de données produites et stockées chaque jour croît de manière exponentielle, particulièrement depuis les deux dernières décennies (Mittelstadt et Floridi 2016b). L'information disponible par la génération et le stockage de ces données est sans précédent, tant en ce qui concerne l'échelle de grandeur que la variété des données collectées. Ce phénomène est communément appelé le *Big Data*, traduit ici par « données massives », et fait référence à « un environnement de données dans

lequel les architectures évolutives répondent aux exigences d'application analytique qui traitent, à grande vitesse, des volumes élevés de données aux formats variés » (Isitor et Stanier 2016, p. 4, traduction libre). Généralement, le *Big Data* réfère soit au processus d'analyse d'ensembles de données massives, soit à l'aspect massif des données elles-mêmes (Mittelstadt et Floridi 2016b). Considérant la croissance exponentielle des données disponibles, référer à la complexité procédurale de l'analyse plutôt qu'à une échelle de grandeur semble plus pertinent quand il s'agit de définir le *Big Data*, car ce qui est massif aujourd'hui risque de ne plus être considéré comme tel demain (Mittelstadt et Floridi 2016b).

Les données massives sont communément caractérisées par « les 5 V », référant en anglais aux 5 caractéristiques suivantes : *volume, velocity, variety, veracity, value* (Brown et al. 2018; Mittelstadt et Floridi 2016b). Le volume – ou l'échelle de grandeur – correspond à l'immense quantité de données générées. La variété réfère quant à elle à la grande diversité de données disponible, notamment concernant leur format, qu'il s'agisse de données structurées ou celles, non-structurées, qui requièrent un « prétraitement » supplémentaire (Brown et al. 2018). La liste présentée par Shafqat (2018) illustre cette diversité en ce qui concerne les données biomédicales :

> The healthcare Big Data involves all the clinical data from Computerized Physician Order Entry (CPOE) and clinical decision support systems—physicians compiled reports, prescriptions, medical imaging, laboratory, pharmacy, insurance and other administrative data; electronic patient records (EPRs); machine generated/sensor data, from monitoring vital signs; social media posts including Twitter feeds, blogs, Web sites, Facebook updates and other platforms; and minimal patient care data including emergency care data, news feeds, and medical journals (p. 2).

La vélocité réfère à la vitesse à laquelle de nouvelles données sont rendues disponibles (certaines sont statiques tandis que d'autres sont mises à jour régulièrement) (Brown et al. 2018) ou à la vitesse à laquelle les données diffusées sont analysées (Mittelstadt et Floridi 2016b). La véracité réfère à la crédibilité que l'on peut accorder à ces données, considérant l'incertitude des informations qu'elles contiennent ou leur niveau de précision (Brown et al. 2018; Mittelstadt et Floridi 2016b). Un des enjeux étant alors de réussir à gérer de manière efficiente ces données produites à une échelle « inimaginable » (Bizer et al. 2012), une 5ème dimension est parfois ajoutée, celle de la valeur – l'accès aux données massives n'étant pertinent que s'il est possible de donner du sens à l'information qui en découle – notamment pour justifier l'effort nécessaire à l'analyse (Brown et al. 2018).

Parmi ces données massives, les données biomédicales ne sont pas des moindres. Plus de 30% des données stockées aujourd'hui dans le monde concernent la santé (Brouard 2017, citant le rapport du *Internet of thing market* de 2017). Les données massives considérées comme des données de santé proviennent de sources hétérogènes. D'abord, avec la numérisation croissante de tous les secteurs d'activité, différents systèmes de santé à travers le monde produisent eux-mêmes des données massives. De nombreuses initiatives gouvernementales incitent à la numérisation et au développement de bases de données harmonisées qui mutualisent les données de santé des populations, qu'elles soient publiques ou privées – plus particulièrement au travers des *electronic health records* (EHRs)[25] (Blumenthal et Tavenner 2010; Blumenthal 2009; Villani 2018; Devillier 2017b; OMS 2016). Les EHRs sont de différentes nature, allant des signes vitaux des patients à leurs données diagnostiques, en passant par leurs données démographiques (Blumenthal et Tavenner 2010).

Les systèmes de collecte et stockage des EHRs ont pour objectif d'emmagasiner des données spécifiques aux individus qui seront notamment essentielles au développement d'une médecine de précision prédictive (Mirnezami, Nicholson, et Darzi 2012) ou pour la recherche en santé par le biais de biobanques ou de la création de larges cohortes d'individus qui acceptent de partager leurs EHRs (Ashley 2015; Mittelstadt et Floridi 2016b; Lipworth et al. 2017). Les données massives en santé peuvent également provenir de firmes de santé privées comme, par exemple, la firme *23andme*, lancée dans le but de fournir des tests génétiques directement au consommateur (patient) et qui a accumulé ces dernières années les données génétiques de près d'un million d'individus contactables et prêts à offrir leurs données à la recherche (Ashley 2015; Sharon 2016). Cette compagnie possède ainsi la plus grosse base de données ADN dans le monde, utilisable par les chercheurs, ayant déjà conduit à une quarantaine de publications scientifiques (Sharon 2016).

[25] L'organisation mondiale de la santé (OMS) différencie les dossiers médicaux électroniques (*Electronic medical records* ou EMRs) des dossiers de santé électroniques (*Electronic health records* ou EHRs) des dossiers de santé personnels (*Personal health records* ou PHRs) (OMS 2012). Les EHRs sont des dossiers médicaux numérisés utilisés pour saisir, stocker et partager des informations entre les prestataires de soins de santé au sein d'une organisation. Les EMRs réfèrent aux mêmes types de données partagées entre les différentes organisations de santé. Ils peuvent par exemple inclurent des données démographiques, l'historique médical du patient, les données relatives à la médication et aux allergies. Ils ont été développés pour soutenir la dispensation de soins au-delà des frontières géographiques. Ils peuvent également être utilisés par les patients dans le but d'avoir un rôle plus actif dans la gestion de leur propre santé. Les PHRs, sont des dossiers médicaux informatisés créés et maintenus par une personne proactive dans la gestion de sa santé. Le dossier peut être privé ou mis à la disposition des prestataires de soins.

Elles peuvent également provenir d'institutions gouvernementales, comme aux États-Unis où la Food and Drug Administration (FDA) collecte en continu les données de facturation d'actes médicaux dans le cadre de son projet pilote *Sentinel Initiative*[26], qui vise à faire un suivi anonyme de 125 millions de patients dans l'optique de mettre en place une surveillance proactive d'effets secondaires potentiels de médicaments approuvés.

Les données massives relatives à la santé proviennent non seulement des systèmes de santé et de la recherche biomédicale mais également de sources externes (Shafqat et al. 2018) ou de sources d'information « informelles » relatives au style de vie, au bien-être, à l'environnement ou aux facteurs socio-économiques (Cano et al. 2017). Elles peuvent par exemple provenir des téléphones intelligents et divers objets connectés (comme des capteurs); communément appelé « l'internet des objets » (au travers duquel les objets connectés communiquent entre eux) (Brouard 2017). Également, un intérêt croissant s'est observé ces dernières années en santé pour les données issues des médias sociaux et du web, données relativement différentes mais complémentaires des données scientifiques traditionnelles (Peek et al. 2015). Leur qualité est cependant variable, notamment parce-que les données issues d'Internet peuvent être dépassées, conflictuelles ou intentionnellement erronées (Bizer et al. 2012; Peek et al. 2015).

Ces données, utiles à la santé mais collectées en dehors du cadre formel d'une prise en charge médicale, amènent le Comité consultatif national d'éthique (CCNE) français à l'appellation « données relatives à la santé », qui « inclut nécessairement aussi celles qui – sans être en elles-mêmes qualifiées de données de santé – le deviennent, soit par leur croisement avec d'autres données qui permet de tirer une conclusion sur l'état de santé ou le risque pour la santé d'une personne, soit 'par destination' (parce qu'elles sont utilisées dans un parcours de soin) » (CCNE 2019, p. 20). Les données les plus considérées dans la littérature sur l'IA en santé demeurent cependant des données biomédicales « classiques » : il s'agit des images diagnostiques, des données génétiques et des données électro diagnostiques (Jiang et al. 2017). Selon les objectifs de recherche ou de soins, il peut s'agir également de données phénotypiques, moléculaires, issues d'essais cliniques ou d'études populationnelles, ou encore de dispositifs de mesures de signes vitaux (Shafqat et al. 2018; Shameer et al. 2018).

[26] Voir : https://www.fda.gov/safety/fdassentinelinitiative/ucm2007250.htm

Toutes ces données – et notamment, leur mutualisation – représentent un intérêt majeur pour la médecine comme pour la recherche en santé (Rial-Sebbag 2017; Lipworth et al. 2017). Leur stockage promet l'amélioration de la qualité des soins (ex. aide à la décision médicale, surveillance de l'émergence de pathologies, gestion de la santé des populations) et la réduction des coûts (Shafqat et al. 2018). Cependant, un des principaux défis relatifs au données massives se situe dans leur interprétation, soit de trouver, de traiter et de combiner les informations qui en sont issues de manière à leur donner un sens (Bizer et al. 2012). En effet, sans un usage significatif, les données massives sont inutiles (Bizer et al. 2012). La disponibilité d'un nombre croissant de données de plus en plus diversifiées ouvre la porte à de nouvelles approches en recherche médicale (comme l'apprentissage automatique, ou autres approches statistiques) (Azencott 2018) et plus particulièrement d'IA. Ces méthodes permettent par différents moyens d'extraire de l'information pertinente de cet environnement de données qu'il est quasi impossible aujourd'hui d'analyser « manuellement » (Shafqat et al. 2018), afin de faire avancer la médecine comme la recherche en santé (Chen, Elenee Argentinis, et Weber 2016). La nécessaire complémentarité entre algorithmes et données massives est très bien illustrée par la Commission nationale informatique et libertés (CNIL) française (2017) : « L'algorithme sans données est aveugle. Les données sans algorithmes sont muettes » (p. 18). L'IA représentent alors différentes manières de valoriser les données massives.

1.2. Intelligence artificielle : différentes manières de valoriser les données massives

1.2.1. Intelligence artificielle et systèmes d'intelligence artificielle

Le terme « intelligence artificielle » est apparu en 1956, nommé ainsi pour la première fois par l'informaticien John McCarthy (Cardon, Cointet, et Mazières 2018). Il n'existe pas de consensus relativement à la définition du terme. Il peut être considéré que l'objectif des recherches en IA « est centré sur le problème de la modélisation artificielle de la gamme « complète » des capacités cognitives, interactives et sociales des humains ; avec un accent particulier sur l'apprentissage et l'interaction autonome » (Baillie 2016, p. 416, traduction libre). L'apprentissage étant largement considéré comme une des conditions requises de l'intelligence, l'apprentissage automatique *(machine-learning)* est une des branches majeures de l'intelligence artificielle (Kononenko 2001). Si l'IA peut parfois référer au développement d'algorithmes à « haut niveau

d'intelligence » (capables de la plupart des capacités cognitives humaines typiques) voire parfois de « superintelligents » (intellect qui excède grandement les capacités humaines) (Müller et Bostrom 2016; Bostrom et Yudkowsky 2011), le domaine est cependant encore loin d'avoir atteint cet objectif. Considérant l'ambiguïté entourant la définition des termes « intelligence » et « artificielle », il est difficile de prime abord de comprendre à quoi réfère concrètement l'IA. Afin d'identifier les technologies en jeu, il est nécessaire de se pencher sur l'historique de ce qui peut relever du « champs de l'IA ».

L'IA s'est historiquement développée autour de deux principales façons de concevoir et programmer le fonctionnement intelligent : les **approches connexionnistes**[27], qui réfèrent aux techniques d'apprentissage utilisant des réseaux de neurones et les **approches symbolique**s qui réfèrent elles à différentes méthodes de calculs sur la base de symboles qui correspondent à une réalité matérielle (Cardon, Cointet, et Mazières 2018). Le regain d'intérêt pour l'IA des dernières années, et les discussions éthiques associées, sont principalement issus des récentes avancées majeures des approches connexionnistes (soit, celles qui fonctionnent sur la base de réseaux de neurones) qui souffraient d'un mal de popularité depuis plusieurs décennies (Cardon, Cointet, et Mazières 2018). Plus particulièrement, c'est dans le domaine de l'apprentissage profond (ou *deep learning*) qui, grâce à la fois au nombre massif de données disponibles et à la puissance de calcul des ordinateurs, a pu mettre en application avec succès des modèles développés dans les années cinquante (qui relevaient du champ de la cybernétique) et qui n'avaient pu jusqu'alors démontrer leur efficacité du fait des limitations techniques de l'époque (Cardon, Cointet, et Mazières 2018).

C'est pourquoi le discours qui entoure aujourd'hui l'IA (incluant le discours éthique) réfère très souvent (sans toujours le nommer explicitement) aux méthodes et techniques qui relèvent des réseaux de neurones. Il est cependant important de noter que l'IA ne se résume pas seulement à la modélisation connexionniste. Il existe d'autres façon de modéliser l'intelligence, comme les approches probabilistes ou les approches à base de règles[28]. Certaines applications d'IA en santé qui connaissent relativement populaires relèvent d'ailleurs plutôt des approches symboliques

[27] Telle est la distinction présentée par Cardon, Cointet et Mazières (2018). Il est également possible de distinguer les approches statistiques (qui comprennent mais ne se limitent pas aux approches connexionnistes) des approches symboliques.

[28] Il est à noter qu'il existe d'autre manière de distinguer les différentes approches de modélisation, comme par exemple celle de Meunier (2017) qui propose une catégorisation sur la base de trois grands modèles : 1) le modèle formel ; 2) les modèles matériels ; et 3) le modèle conceptuel. Pour les fins de cette thèse, la distinction entre IA symbolique et connexionniste demeure suffisante et seule l'approche connexionniste sera présentée plus en détails.

comme, par exemple, les systèmes experts qui, basés sur l'encodage de connaissances humaines, font des prescriptions ou des recommandations de traitements de « la même manière » que le ferait un expert médical (Kattan 2001).

Pour les fins de cette thèse, le terme « systèmes d'IA » sera utilisé, définit comme « tout système informatique utilisant des algorithmes d'IA, que ce soit un logiciel, un objet connecté ou un robot » (Déclaration de Montréal IA Responsable 2018, p. 20). Si ce choix n'est peut-être pas, d'un point de vu informatique, le plus pertinent, il l'est d'un point de vu de l'éthique de l'IA, permettant de tenir compte de l'ambiguïté qui entoure le terme et de référer à un ensemble de techniques, de méthodes ou d'approches de modélisation, sans toutefois que celles-ci soient bien définies. Le terme systèmes d'IA permet de prendre en compte un large éventail de pratiques mais également d'explicitement nommer le terme « intelligence artificielle » qui, bien que polysémique, renvoie à un objet social présent dans l'imaginaire moral collectif, l'objectif étant de se pencher sur l'éthique de l'IA et le partage des responsabilités face à sa gestion plus que de résoudre les problèmes sémantiques associés à l'usage du terme. L'appellation « systèmes » revêt également un aspect dynamique et englobant, les systèmes d'IA évoluant dans le temps et notamment en fonction des données sur lesquelles ils apprennent (ce qui permettra de référer, dans les pages suivantes, tantôt aux enjeux relatifs aux données massives, tantôt à ceux relatifs à l'IA, les deux considérés comme indissociables en ce qui a trait à leur utilisation en santé).

Il est cependant essentiel de démystifier brièvement quelques-unes des méthodes, techniques et analyses à la base des avenues prometteuses associées à l'utilisation des systèmes d'IA en santé, ne serait-ce que pour bien comprendre les enjeux et préoccupations qui accompagnent ce troisième printemps fulgurant de l'IA. Ainsi, sera d'abord présenté l'exploration de données (ou *data mining*), champ qui a évolué en marge de l'IA sans se soucier du débat entre approches symboliques et connexionnistes, mais utilise des techniques d'apprentissage artificiel pour répondre aux nouveaux problèmes d'ingénierie (donner du sens aux ensembles de données massives) apparus avec la numérisation croissante de la société (Cardon, Cointet, et Mazières 2018). Seront ensuite présentés les principaux types de tâches d'apprentissage automatique (ou *machine learning*); soit l'apprentissage supervisé, l'apprentissage par renforcement et l'apprentissage non-supervisé; régulièrement mentionné dans les écrits qui ont trait à l'IA en santé. L'emphase sera ensuite mise sur la présentation de l'apprentissage profond, car il représente le

renouveau des approches connexionnistes – et par extension de l'IA – et se trouve à la source de nombreuses préoccupations éthiques subséquemment présentées.

1.2.2. L'exploration de données (*data mining*)

Tirer de l'information pertinente des ensembles gigantesques et complexes de données massives est aujourd'hui un défi majeur. C'est cependant l'objectif des méthodes de *knowledge discovery in databases*, dont le principal champ est le *data mining* (Maimon et Rokach 2010b). Comme le mentionnent Maimon et Rokach (2010):

> Data Mining is the core of the (knowledge discovery in databases) process, involving the inferring of algorithms that explore the data, develop the model and discover previously unknown patterns. The model is used for understanding phenomena from the data, analysis and prediction (p. 1).

Les avancées récentes en exploration de données ont permis le développement de différentes méthodes particulièrement efficaces pour l'analyse automatisée d'ensembles de données biomédicales complexes, permettant de faire ressortir de nouvelles connaissances pertinentes pour le domaine de la santé, qu'il s'agisse de la recherche, de la prestation de soins ou de la santé publique (Luo et al. 2016).

Il existe deux principales branches de *data mining* (Maimon et Rokach 2010b) : celle orientée vers la vérification (le système vérifie l'hypothèse de l'utilisateur ou d'une autre source, comme par exemple un expert) et celle orientée vers la découverte (le système découvre des nouvelles règles et nouvelles tendances (*patterns)* de manière relativement autonome) (Maimon et Rokach 2010b). Les méthodes de vérification incluent les méthodes issues des statistiques traditionnelles les plus communes (ex. t-tests, ANOVA) (Maimon et Rokach 2010b), mais la plupart des problèmes de *data mining* consistant à faire émerger des hypothèses à partir des données, cette approche connaît moins d'engouement que celle orientée vers la découverte (Maimon et Rokach 2010b).

Les approches orientées vers la découverte sont, pour la plupart, basées sur l'apprentissage inductif, c'est-à-dire que le modèle est construit, explicitement ou implicitement, en se basant sur une généralisation d'un nombre suffisant d'exemples afin de devenir applicable à de nouveaux

exemples (Maimon et Rokach 2010b). Ces approches peuvent elles-mêmes se diviser en 2 champs (notamment en médecine) : les méthodes visant la prédiction et les méthodes visant la description (Maimon et Rokach 2010b; Lavrač et Zupan 2010). Les méthodes descriptives visent la découverte de tendances (*patterns*) ou regroupements *(clusters)* intéressants dans un ensemble de données (Lavrač et Zupan 2010). Les méthodes prédictives visent pour leur part à faire émerger un modèle prédictif à partir des données et qui pourra être réutilisable avec d'autres données (Lavrač et Zupan 2010). Selon Maimon et Rokach (2010), les méthodes qui visent la prédiction en *data mining* sont également connues sous le nom « d'apprentissage supervisé » pour la communauté d'apprentissage automatique (Maimon et Rokach 2010b), bien qu'il existe des algorithmes d'apprentissage machine utilisés à des fins descriptives (notamment, les algorithmes non-supervisés).

1.2.3. Les différentes tâches d'apprentissage automatique

L'apprentissage automatique est une des méthodes largement utilisées dans le domaine biomédical pour tirer parti des données massives en santé (Jiang et al. 2017; Kononenko 2001). Selon Peek *et al.* (2015), il s'agit de l'un des champs les plus dynamiques de l'IA en médecine ces 30 dernières années. L'apprentissage automatique détient l'avantage de permettre la compréhension d'ensembles de données larges, complexes et hétérogènes (Shameer et al. 2018).

Le traitement automatique du langage naturel (ou *natural language processing* – NLP) est un exemple de champs d'application de l'apprentissage automatique. Le NLP consiste en un ensemble de méthodes, approches et techniques d'analyse du langage tant au niveau sémantique qu'aux niveaux syntaxique et lexical (Flasiński 2016). Les *chatbots* (qui simulent les agents conversationnels humains) sont des exemples de succès de ce champ d'application (Flasiński 2016). Le NLP permet d'extraire des informations utiles de textes narratifs pour, par exemple, assister les décisions cliniques (Jiang et al. 2017). Un intérêt croissant en NLP s'observe pour l'analyse de données produites par les médias sociaux, notamment pour le bénéfice du système de santé (Peek et al. 2015). Le NLP s'inscrit également dans le courant du *web mining*, qui tente d'appliquer les techniques de *data mining* et d'apprentissage automatique aux données et documents issus du web, afin d'identifier dans cet ensemble de données de l'information pertinente (Fürnkranz 2005). Le NLP s'est notamment montré particulièrement prometteur pour des domaines comme la psychiatrie, permettant d'analyser de manière pertinente les données issues des rapports

sur les patients (Lovejoy, Buch, et Maruthappu 2019) ou la surveillance en santé publique par le biais, entre autres, de l'analyse de sentiments dans les discours issus des médias sociaux (ex. sur les vaccins ou la perception du système de santé) (Paul et al. 2016).

Il existe trois principaux types d'apprentissage automatique (qui correspondent, factuellement, aux types de tâches qui seront assignées aux algorithmes) : 1) l'apprentissage supervisé; 2) l'apprentissage par renforcement et 3) l'apprentissage non-supervisé.

L'apprentissage supervisé représente peut-être la forme la plus emblématique d'apprentissage automatique et est réalisé grâce à des algorithmes qui fonctionnent sur la base d'un grand nombre d'exemples concrets et explicitement pondérés – c'est-à-dire, des données étiquetées, comme celles qui concernent les précédents diagnostiques (Kononenko 2001; LeCun, Bengio, et Hinton 2015; Chartrand et al. 2017; Alanazi, Abdullah, et Qureshi 2017). Les machines à vecteur de support (ou *support vector machine* - SVM), exemple de méthodes d'apprentissage supervisé, ont gagné en popularité par leur solides assises théoriques (Shmilovici 2010). Elles ont démontré d'excellentes performances de classification, notamment en santé (Lavrač and Zupan 2010). En médecine, ces modèles de classification peuvent être utilisés pour le diagnostic, le pronostic ou la planification de traitements (Lavrač et Zupan 2010; Zhang 2010).

L'apprentissage automatique peut également référer à un **apprentissage par renforcement**. Contrairement à l'apprentissage supervisé (où le système d'IA apprend en mesurant son erreur par rapport aux résultats escomptés), le modèle d'apprentissage par renforcement découvre le « comportement approprié » en utilisant certains critères de « récompense » pour orienter la fonction décisionnelle (soit, le modèle favorise les comportements récompensés sans recevoir d'indications concernant le résultat attendu) (Shameer et al. 2018; Alanazi, Abdullah, et Qureshi 2017; Devillers 2017). Aujourd'hui, l'apprentissage par renforcement est par exemple utilisé pour l'analyse d'images médicales, le dépistage de pathologies comme en cardiologie, ou la sélection personnalisée de prescriptions (Shameer et al. 2018). Ces systèmes ont la particularité, une fois entrainés, d'être les plus à même d'apprendre en continu sans intervention humaine directe (Devillers 2017).

L'apprentissage non-supervisé est la troisième grande catégorie de tâches qui relèvent de l'apprentissage automatique. Dans ce cas, les données qui servent d'exemples pour l'apprentissage ne sont pas étiquetées (par exemple, les images sur lesquels les algorithmes apprennent ne sont pas annotées) mais le modèle apprend directement à partir des données brutes, et vise à trouver des organisations dans les données de façon encore plus autonome, indépendamment d'un objectif de prédiction ou de renforcement (LeCun, Bengio, et Hinton 2015; Maimon et Rokach 2010b; Alanazi, Abdullah, et Qureshi 2017; Devillers 2017), en regroupant par exemple des images sur la base de leur variabilité inhérente (Chartrand et al. 2017). L'apprentissage non-supervisé présente des avantages majeurs et est présenté comme « le graal » des chercheurs par Devillers (2017). Apprenant directement à partir des données brutes, il permet, entre autres, de limiter les erreurs ou le manque de clarté de l'étiquetage manuel nécessaire à l'apprentissage supervisé et de rentabiliser les coûts associés à cet étiquetage, notamment pour certains types de données difficiles à expliquer (Chartrand et al. 2017; Alanazi, Abdullah, et Qureshi 2017). L'apprentissage non-supervisé n'est cependant possible ou pertinent que dans certaines conditions bien précises (Harnad 2005). S'il est moins fréquent de retrouver des applications (notamment en santé) d'apprentissage non-supervisé que celles d'apprentissage supervisé, les avancées actuelles laissent croire qu'elles deviendront plus importantes à long terme (Kononenko 2001; LeCun, Bengio, et Hinton 2015; Chartrand et al. 2017). Il est également possible de retrouver l'utilisation de méthodes semi-supervisées, utilisant à la fois un petit nombre de données étiquetées, et un grand nombre de données non-étiquetées (Chartrand et al. 2017; Alanazi, Abdullah, et Qureshi 2017).

1.2.4. L'apprentissage profond

L'apprentissage automatique – en théorie, quel que soit le type de tâche - peut être modélisé par le biais de réseaux de neurones artificiels, modèle computationnel inspiré du fonctionnement des réseaux de neurones biologiques (Alanazi, Abdullah, et Qureshi 2017) aujourd'hui considérés comme des outils standards de *data mining* (Zhang 2010). Les réseaux de neurones propres à l'apprentissage automatique ont l'avantage de détenir une puissante capacité de modélisation et d'être particulièrement adaptatifs. En effet, le modèle est déterminé par des caractéristiques apprises à partir des données elles-mêmes, même lorsque celles-ci contiennent des informations incomplètes ou du bruit (Zhang 2010). Ces modèles sont particulièrement adaptés à l'analyse de

données collectées dans des contextes réels dont les caractéristiques sont difficiles à définir à l'avance, en particulier lorsque ces caractéristiques sont non-linéaires (Zhang 2010). Le réseau de *self organizing maps* de *Kohonen* (SOM), qui a trouvé plusieurs applications en médecine a, par exemple, démontré sa pertinence pour des applications telles que l'analyse de données ophtalmiques, la classification d'enregistrements du son des poumons, l'analyse de similarité moléculaire, ou encore celle d'une base de données sur le cancer du sein (Lavrač et Zupan 2010).

De plus en plus, les applications d'apprentissage automatique connexionnistes utilisent une catégorie de techniques appelée apprentissage profond *(deep learning)* (LeCun, Bengio, et Hinton 2015) - aujourd'hui largement adoptée par la communauté biomédicale et le système de santé (Miotto et al. 2018). Les méthodes d'apprentissage profond utilisent des réseaux de neurones artificiels qui apprennent selon une succession de niveaux correspondant à des niveaux croissants d'abstraction (Dodig-Crnkovic Gordana 2016). Sur la base de modules simples, la représentation d'un niveau (en commençant par l'entrée brute) est transformée en une représentation d'un niveau supérieur, légèrement plus abstrait (LeCun, Bengio, et Hinton 2015). En composant avec suffisamment de transformations d'un niveau à l'autre, des fonctions très complexes peuvent ainsi être apprises (LeCun, Bengio, et Hinton 2015). La différence fondamentale avec les réseaux de neurones classiques (à une seule couche) est le nombre de couches « cachées », leur connexion et leur capacité à apprendre des abstractions significatives à partir des données de manière plus efficace (Miotto et al. 2018). L'aspect essentiel de l'apprentissage profond est que les fonctions des différentes couches ne sont pas conçues par des ingénieurs humains mais apprises à partir des données (LeCun, Bengio, et Hinton 2015). Les algorithmes d'apprentissage profond sont généralisables, soit applicables à de nouvelles combinaisons de valeurs, au-delà de celles apprises durant l'entrainement (LeCun, Bengio, et Hinton 2015).

Les algorithmes d'apprentissage profond ont récemment été développés, entre autres, pour la reconnaissance d'objets, la retranscription des discours en texte, la prédiction de l'activité de molécules médicamenteuses, la reconstruction de circuits neuronaux ou encore la prédiction des effets de mutations de l'ADN sur l'apparition de maladies (LeCun, Bengio, et Hinton 2015). S'ils sont particulièrement prometteurs, ils ont cependant des limites : ils ne sont pas toujours les

algorithmes les plus appropriés (par exemple, d'autres méthodes seront plus adaptées pour l'analyse de données bien structurées et bien définies), ils ont besoin de beaucoup de données pour apprendre et sont considérés comme relativement opaques par rapport aux autres méthodes d'apprentissage automatique (Chartrand et al. 2017).

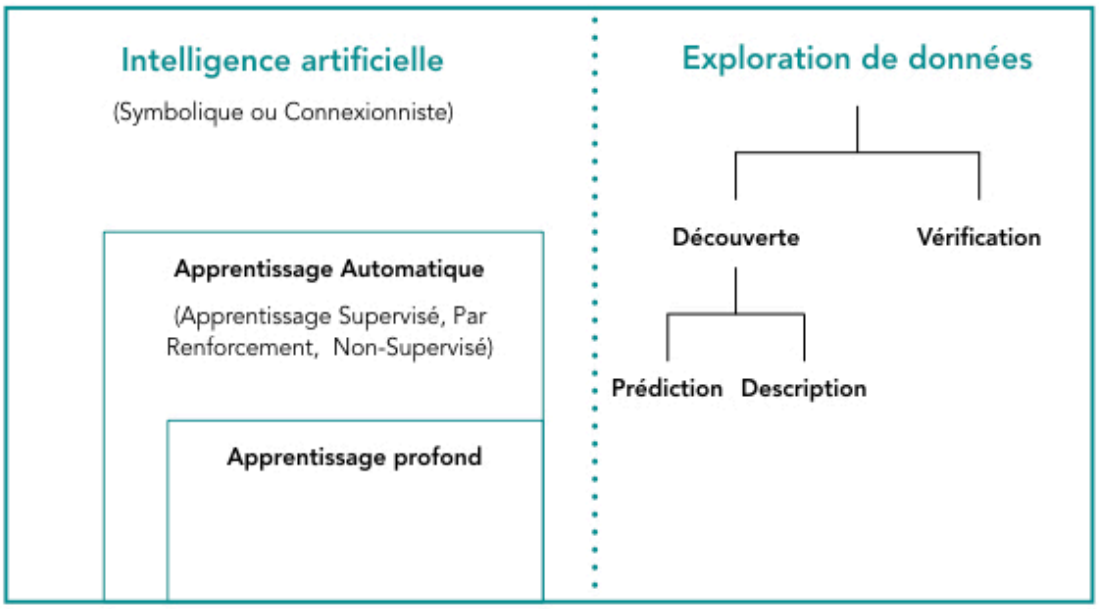
Différentes applications d'apprentissage profond ont démontré leur utilité en santé. Par exemple, les réseaux de neurones convolutifs (ou *convolutional neural networks* - CNN) correspondent à l'approche d'apprentissage profond la plus utilisée pour les tâches de traitement d'images – particulièrement efficace pour l'analyse d'images naturelles comme en radiologie (Chartrand et al. 2017; LeCun, Bengio, et Hinton 2015). Le plus récent succès qu'ait connu cette approche est son utilisation pour des tâches de reconnaissance faciale, mais elle connaît également des applications en langage naturel ou reconnaissance vocale (LeCun, Bengio, et Hinton 2015). De nombreuses compagnies développent aujourd'hui des CNN pour permettre des applications de visualisation en temps réel dans les téléphones intelligents, les appareils photos, les robots ou les voitures autonomes (LeCun, Bengio, et Hinton 2015).

Récemment, une des avancées majeures en apprentissage profond est l'usage de *Generative Adversarial Networks* (GANs), notamment pour leur application en imagerie médicale (Yi, Walia, et Babyn 2018; Máttyus et Urtasun 2018). Ces réseaux sont réputés pour leur capacité à générer des images réalistes et nettes (Máttyus et Urtasun 2018). Les GANs correspondent à un modèle d'apprentissage profond où deux réseaux de neurones sont entraînés simultanément, l'un ayant pour objectif la génération d'images (qui permet d'aider l'exploration et la découverte de la structure sous-jacente des données d'apprentissage et la génération de nouvelles images à partir d'elles) et l'autre sur la discrimination de ces dernières (et peut être utilisé comme détecteur) (Yi, Walia, et Babyn 2018). L'idée est de réaliser l'apprentissage en trompant le discriminateur qui essaie de distinguer les exemples réels des exemples créés par le générateur (Máttyus et Urtasun 2018). Leur force est qu'ils apprennent de manière non-supervisée ou minimalement supervisée (Yi, Walia, et Babyn 2018) bien qu'ils fonctionnent mal pour des tâches supervisées classiques (Máttyus et Urtasun 2018). Dans le domaine de l'imagerie, ils ont démontrés être utiles, entre autres, pour la restauration d'images (notamment lorsque des artéfacts ont nui à leur qualité) –

évitant ainsi la répétition d'examens et permettant la génération automatique de rapports, la suppression d'artéfacts ou la détection d'anomalies (Yi, Walia, et Babyn 2018).

Ces différents ensembles de méthodes, techniques et analyses qui relèvent de l'IA, et particulièrement les algorithmes d'apprentissage automatique, sont ainsi autant d'opportunités de donner un sens aux données massives aujourd'hui disponibles (voir Schéma 1). La valorisation de ces données est à la source de nombreuses avenues prometteuses pour le système de santé, que celles-ci concernent le développement de médecine de précision, de systèmes experts d'aide au diagnostic ou de nouveaux agents de soin.

Schéma 1 - Les différents ensembles de méthodes, techniques et analyse qui relèvent de l'intelligence artificielle.



2. Des avenues prometteuses

2.1. De puissants outils pour soutenir les professionnels de santé

Les systèmes d'IA connaissent de nombreuses applications en santé, ce qui en fait de nouveaux outils numériques pour les professionnels du secteur. Par exemple, différents robots d'assistance se développent pour la chirurgie de précision. L'Institut de recherche contre les cancers de

l'appareil digestif (IRCAD) en France a recours à un dispositif de chirurgie augmentée[29] qui offre des outils complémentaires pour augmenter l'efficacité des pratiques. Ces dispositifs permettent, par l'entremise d'un système de réalité augmentée interactive (qui utilise l'apprentissage automatique pour modéliser les organes du patient sur la base d'images médicales et superpose les image réelles et virtuelles), de bouger les instruments au rythme du mouvement (Guerriero et al. 2018; Quero et al. 2019). Ces avancées permettent, entres autres, d'augmenter la précision de même que de diminuer les complications post-opératoires (Guerriero et al. 2018; Quero et al. 2019).

Les systèmes d'IA sont également d'intéressants outils de triage, permettant de prioriser ou d'orienter les patients vers les services appropriés. Par exemple, au Québec, l'entreprise Bonjour Santé, qui propose aux cliniques québécoises des services de gestion des rendez-vous, a pu, grâce à l'usage d'un algorithme de triage, réduire le temps d'attente moyen dans les cliniques d'environ 4 heures à 45 minutes[30]. Bonjour Santé souhaite réduire ce temps d'attente à 20 minutes dans le cadre d'un projet en partenariat avec le Mila, qui vise à prédire le temps d'attente de manière encore plus précise en couplant et analysant la grande quantité de données non-nominatives cumulées jusqu'à présent - issues des millions de rendez-vous médicaux précédents - grâce à des techniques d'apprentissage profond qui devraient permettre de prédire le temps d'attente de manière encore plus précise. Les systèmes d'IA pourrait également être utilisés afin d'optimiser la mise en place d'essais cliniques en automatisant la sélection des participants à la recherche (CNIL 2017; Collier, Fu, and Yin 2017).

Un des champs majeurs d'application des systèmes d'IA en santé concerne cependant le développement de l'apprentissage automatique pour l'aide à la décision médicale (Alanazi, Abdullah, et Qureshi 2017), communément appelé « systèmes experts » et dont l'application principale est l'aide au diagnostic (Patel et al. 2009; Peek et al. 2015). Ces systèmes sont de

[29] Voir : https://swissdigitalhealth.com/lircad-au-coeur-de-la-chirurgie-augmentee/
[30] Voir https://clinique.bonjour-sante.ca/salle-de-presse/lintelligence-artificielle-au-service-des-patients et https://www.tvanouvelles.ca/2018/05/28/bonjour-sante-estime-pouvoir-regler-lattente-en-clinique-definitivement https://www.journaldemontreal.com/2018/05/28/lintelligence-artificielle-au-secours-de-lattente-en-sante

puissants outils d'aide à la décision pour les professionnels de santé. Ils peuvent être prédictifs (soit, évaluer les résultats) ou prescriptifs (soit, faire des recommandations de traitement) (Kattan 2001). Ils sont conçus pour améliorer le soin en optimisant la prise de décision médicale, en formulant eux-mêmes les recommandations :

> The distinguishing feature of medical expert systems is that they make recommendations based on input data; they are differentiated from decision support systems in that the latter are designed to help clinicians make decisions rather than actually make the recommendation, which is what an expert system does. This recommendation is essentially a prediction (of diagnosis or prognosis) or prescription (i.e., a treatment recommendation) (Kattan 2001).

Un des systèmes experts médicaux le plus célèbre est peut-être le logiciel Watson, développé par IBM[31], qui s'appuie sur des quantités massives de données structurées et non-structurées issues de sources médicales et scientifiques variées (ex. dossiers médicaux, analyses de cas similaires et littérature scientifique disponible). Watson peut ainsi offrir une réponse structurée sur la base des informations les plus pertinentes que contiennent ces données et ainsi aider les médecins dans leurs décisions cliniques (Lee 2014). Grâce à ces centaines d'algorithmes intégrés qui relèvent de différentes techniques d'apprentissage automatique (dont le NLP), Watson peut également émettre des recommandations de traitements de manière pratiquement instantanée (Lee 2014). Ce système expert pourrait également devenir un outil particulièrement intéressant pour les chercheurs en sciences de la vie, une de ses versions permettant d'analyser la littérature médicale, les brevets, les données génomiques, chimiques et pharmacologiques – données typiquement analysées par les chercheurs – de manière beaucoup plus rapide et beaucoup plus précise, permettant de mettre en évidence de nouvelles relations ou de nouvelles hypothèses (Chen, Elenee Argentinis, et Weber 2016).

Watson n'est cependant pas le seul système expert biomédical à se développer. Pour ne citer qu'eux, le logiciel Nuance[32] analyse la documentation médicale des trente dernières années visant ainsi à aider les professionnels de santé dans l'exercice de leur fonction, ou le programme APOLLO[33] (*Adaptive Patient-Oriented Longitudinal Learning and Optimization*) a pour but

[31] Voir : https://www.ibm.com/watson/
[32] Voir : https://www.nuance.com/healthcare/artificial-intelligence.html
[33] Voir : https://www.mdanderson.org/cancermoonshots/research_platforms/apollo.html

d'analyser les données génétiques collectées sur des patients souffrant de cancer afin de soutenir la prise de décision des médecins relativement aux traitements les plus adaptés.

Les systèmes d'IA offrent également la possibilité de traiter de plus en plus rapidement les données cliniques et la documentation médicale (dont la littérature scientifique) pour aider la prise de décision des professionnels de santé tout comme l'analyse des chercheurs. L'apprentissage profond montre ainsi un potentiel important pour le diagnostic précoce de la maladie d'Alzheimer et s'est montré capital dans le traitement des patients afin de prendre des mesures préventives avant la présence de dommages irréversibles au cerveau (Liu et al. 2014). Les méthodes d'IA ont également démontré un fort potentiel dans la détection de cancers, en utilisant des données provenant de différents types de cancer pour former automatiquement des fonctions qui aident à améliorer le diagnostic et la classification en utilisant l'apprentissage supervisé sur des données génomiques (Fakoor et al. 2013). Tel que mentionné dans le cadre de la médecine de précision, ces techniques offrent la possibilité de révolutionner la précision de la classification diagnostique (Torkamani et al. 2017).

La branche de l'IA qui a montré les performances les plus remarquables est certainement celle de la reconnaissance visuelle pour l'analyse et l'interprétation de différents types d'images médicales (Torkamani et al. 2017). L'interprétation de clichés médicaux se fait sur la base de réseaux de neurones profonds entrainés sur des milliers d'images cliniques. Ces méthodes ont démontré, entre autres, des niveaux de performance considérables pour la classification de cancers de la peau (Esteva et al. 2017) ou pour la détection de rétinopathies diabétiques (Gulshan et al. 2016). Les avancées de l'analyse d'images par les techniques de CNN sont particulièrement prometteuses en radiologie, ayant entre autres permis de raffiner la classification et la détection de lésions (dont les lésions malignes) (Chartrand et al. 2017).

Les systèmes d'apprentissage profond pourraient bientôt dépasser les compétences humaines pour certaines tâches d'interprétation d'images (Chartrand et al. 2017) voire les ont surpassées dans certains contextes précis (ex. interprétation de scanners cérébraux (Merkow et al. 2017), détection

de rétinopathies diabétiques (Gulshan et al. 2016) ou de cancer de la peau (Esteva et al. 2017)). Puisqu'ils pourraient à terme devenir plus performants que les humains, notamment en termes de classification, ces systèmes pourraient permettre des avancées majeures et améliorer significativement les prestations de soins.

2.2. Des systèmes d'intelligence artificielle au contact direct avec le patient

L'autonomie croissante des systèmes d'IA s'accompagne de l'avènement de différents dispositifs qui se retrouvent au contact direct avec les patients, dépassant le cadre conventionnel des soins. Ici, le patient interagit directement avec des interfaces ou des « agents artificiels » qui ne nécessitent pas la présence directe d'un professionnel.

Le développement de robots sociaux à visée médicale s'inscrit dans cette lignée. Différents robots de soins ou *carebots* sont actuellement en développement ou d'ores et déjà utilisés pour des fins thérapeutiques. Certaines recherches se penchent par exemple sur le développement de robots conversationnels empathiques, embarquant des systèmes d'IA, capables de détecter les émotions des humains avec qui ils interagissent et de réagir de manière appropriée (Devillers 2017). Ces robots relativement autonomes ont donc parfois la possibilité de simuler l'empathie (ou affects) (Devillers 2017) et connaissent différentes applications thérapeutiques potentielles. Ils fonctionnent sur la base de capteurs qui collectent des données et de programmes qui les analysent, de manière à pouvoir agir sur leur environnement en conséquence (Devillers 2017). Par exemple le robot PARO[34], développé par une entreprise japonaise, est un robot interactif thérapeutique d'assistance aux personnes âgées, lancé au Japon en 2005, utilisé dans les maisons de retraite en Europe depuis 2003 et certifié par la FDA aux États-Unis depuis 2009 (Devillers 2017). Ce robot à l'apparence d'un phoque en peluche et a pour but de donner un retour émotionnel aux personnes âgées, en particulier celles atteintes de la maladie d'Alzheimer, afin de limiter les effets négatifs de leur perte de lien social (Devillers 2017). Les robots de soins peuvent également intervenir dans la thérapie des troubles autistiques (Hamet et Tremblay 2017). Le robot NAO[35], développé par le Centre Hospitalier Universitaire de Nantes, l'association française Robots ! et Stéréolux, est par

[34] Voir : http://www.parorobots.com/
[35] Voir : https://www.stereolux.org/rob-autisme

exemple utilisé par les chercheurs comme médiateur thérapeutique pour les enfants autistes et comme assistant de communication et d'enseignement afin d'améliorer l'apprentissage et le développement social de ces enfants (Devillers 2017).

Dans la même lignée, les *softbots* (avatars psychothérapeutiques émotionnellement sensibles) connaissent différentes perspectives d'usages en santé. Ils sont actuellement développés, entre autres, pour intervenir dans le contrôle de la douleur chez les enfants atteints du cancer, dans la détection précoce de perturbations émotionnelles chez les autochtones américains (incluant les tendances suicidaires) ou encore dans le contrôle des hallucinations paranoïdes (Hamet et Tremblay 2017). Les robots conversationnels (ou *chatbots*) ont également une place en santé. Par exemple, le *chatbot* d'assistance médicale *Melody,* développé par la société chinoise Baidu, collecte des informations sur les patients pour ensuite les relayer aux médecins dans le but de faciliter le diagnostic (Shakhovska 2017). Le *Dr. AI bot* collecte des informations sur les patients (ex. symptômes, historique de maladies, paramètres biologiques) pour ensuite classer les causes des symptômes par sévérité (Shakhovska 2017). L' *Insomno Bot* a quant à lui été développé pour les personnes qui ont différents problèmes de sommeil (Shakhovska 2017). Ces agents conversationnels peuvent aider les professionnels de santé mais également directement améliorer l'expérience de soin du patient. Par exemple, des recherches en santé mentale ont démontré que certains *chatbots* pourraient augmenter l'adhérence des patients aux thérapies cognitivo-comportementales (Lovejoy, Buch, et Maruthappu 2019).

Enfin, avec l'avènement des objets connectés, et plus particulièrement des téléphones intelligents, les applications mobiles peuvent elles aussi devenir de nouvelles interfaces de soins. Celles-ci font partie de ce qui est communément appelé la santé mobile (ou *mobile health*), définie par l'Organisation Mondiale de la Santé (OMS) comme : « services and information provided through mobile technology, such as mobile phones and handheld computers » (OMS 2012, p. 79). L'usage des dispositifs mobiles peut par exemple inclure :

> Data collection for surveillance and public health (e.g. outbreak investigation) ; real-time monitoring of an individual's health ; treatment support, health advice and medication compliance ; health information to practitioners, researchers and patients ; health education and awareness programmes ; diagnostic and treatment support, communication for health-care workers (OMS 2012, p. 79).

De nombreuses applications mobiles se développent ainsi avec un but sanitaire. Par exemple, la compagnie DeepMind Health (Google) a développé en partenariat avec le National Health Services (Royaume-Uni) l'application mobile Streams[36], avec l'objectif annoncé de « Helping clinicians get patients from test to treatment, faster »[37]. L'application AICure[38] a elle pour but de mesurer l'adhérence du patient à son traitement par le biais d'une surveillance via la caméra de son téléphone intelligent. La compagnie Cogito[39], qui développe des applications mobiles actuellement testées au Brigham et au Women's Hospital à Boston, a pour but d'analyser l'activité des individus sur leurs téléphones (ex. les médias sociaux) afin de détecter des profils de dépression sur la base de leurs communications. L'application Companion de cette compagnie a par exemple permis de mettre en évidence de bons prédicteurs de symptômes de dépression et de stress-post traumatique (Lovejoy, Buch, et Maruthappu 2019). Le *chatbot* Mr. Young, un guide personnel de bien-être mental, est développé dans le but d'effectuer de la prévention en santé mentale en éduquant les utilisateurs, en examinant leurs données personnelles, en proposant des ressources et des traitements pertinents et en évaluant les progrès de l'utilisateur[40]. L'algorithme de Cardiogram[41], développé par l'Université de Californie, est quant à lui capable de détecter les battements de cœur anormaux des porteurs de leur montre intelligente. Enfin, iCarbonX[42], entreprise chinoise, a pour objectif affiché de « digitalizing everyone's life information » et de construire un « soi digital » qui, en combinant différents types de données, pourrait recommander des programmes de bien-être personnalisés, des choix alimentaires et éventuellement des médicaments.

Les applications comme MoleScope[43] et MoleMapper[44] permettent quant à elles l'auto-examen et la détection de mélanomes via la caméra du téléphone intelligent des utilisateurs, et pourraient ainsi participer à la prévention du cancer. Elles sont également d'intérêt pour la recherche, comme c'est le cas de MoleMapper, utilisée pour la collecte et le partage des données collectées sur 2 069 participants dans le cadre d'une étude mettant en évidence les associations entre l'apparition de mélanomes et certains risques démographiques (Webster et al. 2017). Ces

[36] Voir : https://deepmind.com/blog/scaling-streams-google/
[37] Voir : https://deepmind.com/applied/deepmind-health/
[38] Voir : https://aicure.com/
[39] Voir : https://www.cogitocorp.com/
[40] Voir : https://www.mryoung.co/
[41] Voir : http://cardiogr.am/about/
[42] Voir : https://www.icarbonx.com/en/
[43] Voir : https://molescope.com/
[44] Voir : http://molemapper.org/

données ont été partagées dans une publication scientifique de la revue *Nature* afin d'engager la collaboration multidisciplinaire de chercheurs pour mieux comprendre et prévenir l'apparition de mélanomes (Webster et al. 2017). Plus que des agents de soins, ces objets connectés deviennent alors de nouveaux outils de recherche et pourraient ainsi permettre d'augmenter drastiquement les cohortes de participants à la recherche (Brouard 2017). Ces données sont particulièrement intéressantes pour la recherche en tant que données collectées dans un contexte réel puisqu'elles sont maintenant obtenues par l'entremise de l'utilisateur lui-même et non du système de santé (Brouard 2017). Leur grande validité écologique comparé aux conditions de laboratoire est ici particulièrement intéressante pour la communauté médicale comme pour celle de la recherche en santé.

Certaines compagnies se développent même avec la double mission recherche-soin clinique. La compagnie AIFred Health[45], au Canada, se penche sur la psychiatrie de précision en développant une technologie pour lutter contre la dépression. Combinant apprentissage profond, données scientifiques (dont médicales) et les données directement en provenance des patients (qui sont réutilisées au fur et à mesure pour l'apprentissage de leurs réseaux de neurones afin d'améliorer leur pouvoir prédictif), leurs applications visent à générer des recommandations personnalisées de traitements, effectuer un suivi des symptômes des patients et, sur la base de toutes les données du patient qui sont conservées, proposer des analyses et des recommandations et mener des recherches cliniques.

La santé mobile est ainsi particulièrement pertinente pour la prévention :

> The potential for disease prevention through discreet, continuous monitoring and real-time, personalized feedback is significant. Moreover, because low physical fitness is a greater risk factor for all-cause mortality than smoking, diabetes, and obesity combined, and because physical activity is readily quantifiable through a mobile or wearable device, the prospect of measuring and testing intervention strategies in a large population in a randomized fashion is appealing (Ashley 2015, p. 2).

Les téléphones intelligents permettent en effet aujourd'hui de collecter des données telles que l'activité physique des individus de manière plus continue et plus précise qu'il n'avait été possible de le faire auparavant (Ashley 2015), ce qui représente un avantage non négligeable pour la santé publique (Brouard 2017). Par exemple, le *Withings Health Observatory*[46], qui analyse les données

[45] Voir : https://aifredhealth.com/
[46] Voir : https://obs.withings.com/us/activity pour la cartographie des États-Unis.

de plus de 100 000 utilisateurs en France, aux États-Unis et au Royaume-Uni, constitue un nouvel outil de veille en cartographiant des mesures telles que l'activité physique, le poids, ou la tension artérielle des individus impliqués (Brouard 2017). Les données sont collectées par le biais de leurs objets connectés et cet observatoire travaille et partage actuellement ses données avec différentes institutions de recherche (ex. *Stanford Medicine*), hôpitaux et compagnies privées[47].

2.3. Vers une médecine de précision

Le nombre grandissant de données disponibles (notamment en pharmacogénomique et pharmacogénétique – Ritchie 2012) et les avancées en apprentissage automatique ont permis le développement d'une médecine de précision aux retombées prometteuses pour la médecine comme pour la recherche en santé (Jameson and Longo 2015). Le terme de médecine de précision a succédé à celui de médecine personnalisée, bien que ces termes soient souvent utilisés de manière interchangeable (Jameson and Longo 2015), notamment suite au lancement de la *Precision medicine initiative*[48] en 2015 aux États-Unis (Ashley 2015; Larry Jameson et Longo 2015; Chambers, Feero, et Khoury 2016). Cette initiative vise au couplage de différents types de données personnelles (i.e. génomiques, cliniques, comportementales, environnementales) afin d'encourager le développement de soins de plus en plus ciblés relativement aux caractéristiques des patients (Chambers, Feero, et Khoury 2016). La médecine de précision est aujourd'hui une stratégie prédominante pour de nombreuses compagnies pharmaceutiques, instituts de biotechnologie ou centres de recherche médicaux (Ritchie 2012; Bayer et Galea 2015).

Si les termes sont parfois utilisés de manière interchangeable, Ashley (2015) présente la distinction entre médecine de précision et médecine personnalisée, faite selon l'auteur dans le rapport du National Research Council américain sur le sujet : « The authors explain that their use of "precision" was intended to avoid the implication that medications would be synthesized personally for single patients. Rather, they [the NCR] hoped to convey a broader concept that

[47] Voir : https://www.withings.com/ca/en/health-institute

[48] Lancée en 2015 par le président Obama et faisant suite à une publication du National Research Council annonçant le terme de médecine de précision (Ashley 2015), cette initiative finance des projets tel que le projet *All of us* du National Institute of Health (NIH) qui a pour objectif de construire une cohorte nationale à grande échelle de participants afin d'étudier différents facteurs (e.g. environnementaux, biologiques, comportementaux) sur un échantillon qui reflète la diversité de la population. Voir : https://allofus.nih.gov/about/about-all-us-research-program

would include precisely tailoring therapies to subcategories of disease, often defined by genomics » (Ashley 2015, p. 2119). Les deux notions se distinguent donc essentiellement par leur portée : la médecine de précision renvoie à une conception plus large que le ciblage approprié des traitements pour l'individu-patient. Le développement de la médecine de précision vise à contrer les conséquences du fait que la plupart des traitements médicaux aient été développés jusqu'à présent pour le « patient moyen » (Jameson and Longo 2015). Les lignes directrices de mise en place d'essais cliniques tendent en effet à normaliser les ensembles de la population qui sont étudiés en appliquant les méthodes de réduction des biais, conduisant à la généralisation parfois excessive à l'ensemble de la population de résultats obtenus sur un petit échantillon d'individus. La médecine de précision pourrait en cette circonstance permettre une plus grande granularité dans l'identification des spécificités de chaque patient et tenir compte d'une plus grande variabilité dans les réponses aux traitements (Peck 2018; Manrai, Patel, et Ioannidis 2018).

La médecine de précision peut donc plus spécifiquement se définir ainsi :

> Treatments targeted to the needs of individual patients on the basis of genetic, biomarker, phenotypic, or psychosocial characteristics that distinguish a given patient from other patients with similar clinical presentations. Inherent in this definition is the goal of improving clinical outcomes for individual patients and minimizing unnecessary side effects for those less likely to have a response to a particular treatment (Jameson and Longo 2015 p. 613).

La médecine de précision pourrait ainsi permettre d'augmenter les réponses à la médication en déterminant les interventions les plus efficaces et en diminuant les effets secondaires (Hamet et Tremblay 2017). Elle vise également à adapter le dosage de médicaments en définissant plus précisément les profils d'individus qui seraient susceptibles de réagir positivement à une molécule (Peck 2018; Manrai, Patel, et Ioannidis 2018). C'est par exemple le cas de certaines études en oncologie de précision qui visent à maximiser l'utilisation de traitements de chimiothérapie existants en ciblant de manière plus précise et plus appropriée les marqueurs génétiques des patients[49]. Utilisant notamment les méthodes de SVM, cette démarche est particulièrement utile lorsqu'il s'agit de thérapies moléculaires (Ding et al. 2018).

[49] En effet certains patients ne sont pas sélectionnés comme des candidats potentiels à certaines thérapies moléculaires car l'identification de leurs marqueurs génétiques – qui permettraient de savoir si la cible thérapeutique est appropriée

L'apprentissage automatique joue un rôle central dans l'adaptation des traitements et diagnostics aux caractéristiques du patient (Azencott 2018; Jameson and Longo 2015) en combinant par exemple les EHRs et les données génomiques (Ashley 2015). Les systèmes d'IA dans ce contexte pourrait également permettre de découvrir de nouvelles cibles thérapeutiques, comme l'a démontré l'utilisation d'algorithmes non-supervisés dans l'étude des interactions entre protéines (Hamet et Tremblay 2017). La médecine de précision vise également l'identification des patients aux caractéristiques rares qui possèdent des symptômes similaires ou la réduction du temps d'identification de nouvelles maladies par le biais de comparaisons entre séquences génomiques (Ashley 2015). En psychiatrie de précision, l'utilisation de réseaux de neurones et de méthodes de SVM est particulièrement pertinente pour la détection précoce des pathologies, l'identification de traitements individualisés, l'ajustement des doses et la classification des maladies mentales (Bzdok et Meyer-Lindenberg 2018). En médecine cardiovasculaire de précision, les algorithmes de prédiction automatisée des risques pourraient permettre d'orienter le soin clinique ou de préciser le phénotypage de maladies complexes (Shameer et al. 2018).

Ainsi, comme le reconnaissent Shameer et al. (2018) en ce qui a trait à la cardiologie de précision :

> Various types of AI algorithms will be essential for understanding the nuanced individual risk factors, behavioural drivers and therapeutic pathways predictive of disease outcomes in specific patient cohorts and also for instituting early therapeutic interventions. The application of machine learning algorithms in prospective clinical trials would allow comparison with current standard of care practices with a goal of implementing precision diagnostics, risk stratification and personalised therapeutics (p. 1163).

Plus que la simple optimisation des traitements, la médecine de précision a parfois permis de définir de nouveaux syndromes (Ashley 2015). En conséquence, elle pourrait permettre de moderniser ou redéfinir la classification diagnostique actuelle, comme celle de l'OMS (Ashley 2015; Mirnezami, Nicholson, et Darzi 2012; Torkamani et al. 2017).

- est souvent trop inexacte (Ding et al. 2018). L'apprentissage automatique pourraient permettre de pallier ce problème en analysant les données génomiques à plus large échelle.

Les applications de systèmes d'IA en santé pourraient de plus soutenir la détection de risques sanitaires et aider à prévenir des maladies sur la base d'un « savoir collectif » (CNIL 2017) issu de l'analyse des données partagées par l'ensemble de la population. Les systèmes d'IA pourraient ainsi doter les professionnels de santé d'outils de prévention ciblée, *via* l'identification de sujets avec un historique familial de maladies héréditaires ou un risque accru de maladies chroniques (Hamet et Tremblay 2017). Ceci pourrait être réalisé par l'analyse de biomarqueurs ou de variants d'ADN liés à certaines pathologies ainsi que par l'étude de leur évolution (Hamet and Tremblay 2017). Dans cette perspective, l'avènement de « jumeaux numériques »[50] permettrait la prédiction des trajectoires de santé, utilisées comme « useful benchmarks for defining the individualized optimal health range for specific health parameters as well as predictors of health outcomes overall » (Torkamani et al. 2017, p. 840).

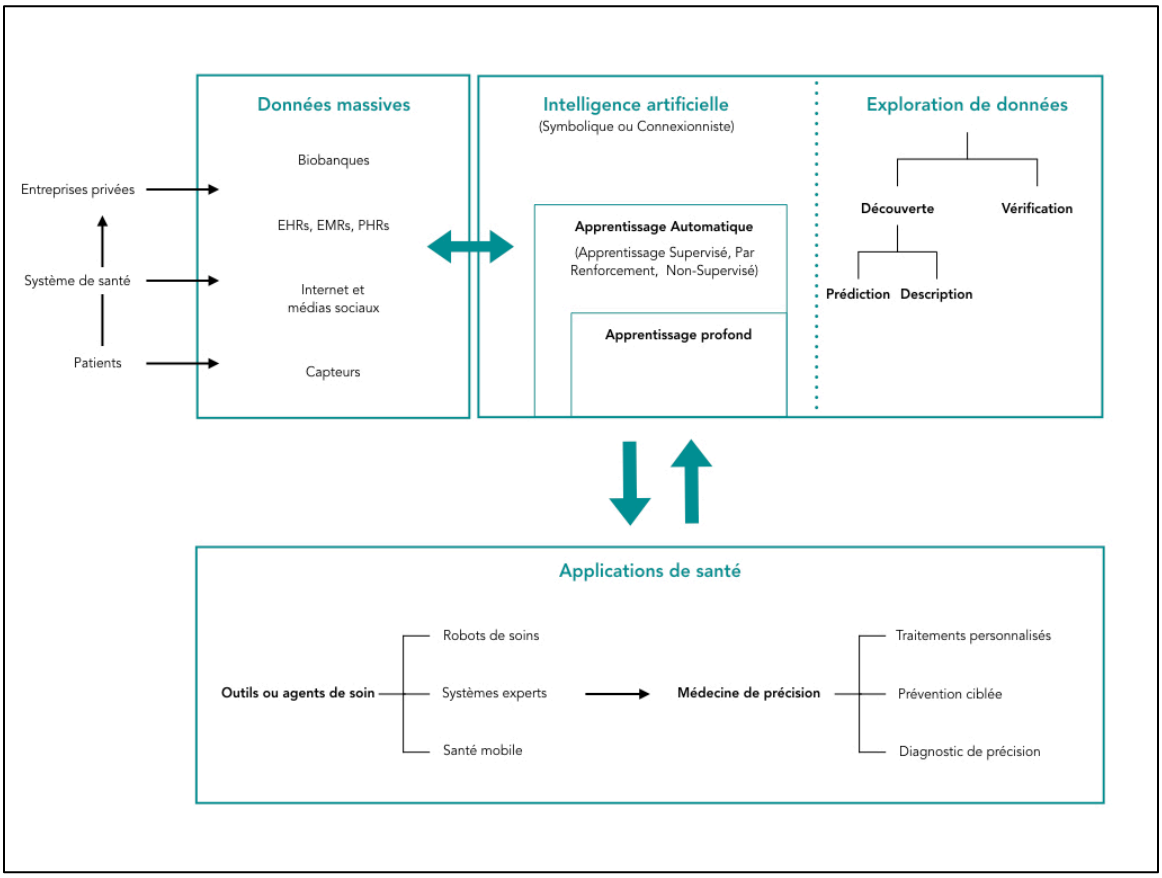
Les systèmes d'IA en santé détiennent ainsi le potentiel de contribuer au développement d'une médecine hautement personnalisée en établissant des profils biologiques (caractérisation génomique ou moléculaire des individus et de leur maladie) pour individualiser les traitements et les stratégies thérapeutiques, avec un éventuel monitorage en continu des paramètres de santé de chaque individu (Torkamani et al. 2017; CNIL 2017; Hamet and Tremblay 2017; Peek et al. 2015).

3. Un système de santé en transition

L'avènement de l'innovation numérique en santé, et plus particulièrement des méthodes utilisées en IA et des données massives, offre donc des perspectives prometteuses pour les systèmes de santé, que celles-ci concernent la prestation de soins ou la recherche (voir Schéma 2).

[50] Bien que le concept de jumeau numérique soit issu de l'ingénierie, Torkamani et al. (2017) décrivent un système où le jumeau numérique serait une autre personne (existante) qui partage différents paramètres de santé avec un individu, et qui peut être utilisé pour déterminer quand celui-ci a un risque élevé de développer une maladie, recommander des comportements pour réduire ce risque, et identifier les signes précoces de maladies (Torkamani et al. 2017).

Schéma 2 - Vision d'ensemble de l'utilisation des données massives et de l'IA en santé



L'ensemble de ces méthodes et leurs usages qui tendent à devenir omniprésents semblent conduire à de potentielles transformations profondes des systèmes de santé. De nombreux auteurs décrivent en effet l'émergence d'une « révolution » dans les méthodes de recherche (ex. Rial-Sebbag 2017) ; voire d'un changement de paradigme scientifique qui transformerait la façon dont la connaissance est produite, selon une science guidée par les données :

> La science-data-driven marquerait l'avènement d'un quatrième paradigme après la science empirique décrivant les phénomènes, la science théorique usant des modèles et des généralisations, et la science computationnelle simulant les phénomènes complexes (Hey et al., 2009 dans Coutellec and Weil-Dubuc 2017 p. 65).

Le phénomène du *Big Data* et la construction des connaissances à partir de ces données massives représenteraient en effet un nouveau paradigme de recherche pour de nombreuses disciplines (Coutellec et Weil-Dubuc 2017; Kitchin 2014; Chang, Kauffman, et Kwon 2014). La forme que prend cette nouvelle épistémologie est cependant contestée :

> Whilst Jim Gray envisages the fourth paradigm of science to be data-intensive and a radically new extension of the established scientific method, others suggest that Big Data ushers in a new era of empiricism, wherein the volume of data, accompanied by techniques that can reveal their inherent truth, enables data to speak for themselves free of theory (Kitchin 2014 p. 3).

Kitchin (2014) met en effet en évidence deux principales tendances concernant la forme que pourrait prendre ce changement de paradigme : celle d'un nouvel empirisme et celle d'une science guidée par les données *(data-driven)*. Selon la vision d'un nouvel empirisme, l'avènement du quatrième paradigme se fait par un mode empirique de production de la connaissance, à partir des données et sans jamais prendre assise sur une théorie. Les données produisent du sens de manière inhérente et les méthodes scientifiques traditionnelles (qui fonctionnent sur la base d'hypothèses préalables) deviennent par conséquent caduques. C'est ainsi la « sérendipité digitale » qui permet, selon cette vision, la création de nouvelles connaissances, et ce de manière objective et hautement simplifiée (Kitchin 2014). Si cette vision a particulièrement gagné en crédibilité en dehors du milieu universitaire – notamment, dans le milieu des affaires – elle est également critiquée pour son manque de validité scientifique et ses retombées qui ne permettraient pas l'avancée des connaissances *per se* (Kitchin 2014).

Selon la vision de l'apparition d'une science guidée par les données, ce nouveau paradigme est centré sur le meilleur moyen de donner du sens aux phénomènes et aux processus, mais les hypothèses et les idées sont générées à partir des données elles-mêmes plutôt qu'à partir de théories (Kitchin 2014; Coutellec et Weil-Dubuc 2017). Cette approche reconfigure alors la version traditionnelle des méthodes scientifiques, comme un nouveau moyen d'élaborer des théories – et n'est donc pas « theory free » comme le nouvel empirisme décrit précédemment. Dans le cadre d'une *evidence-based medecine* (ou la médecine fondée sur les données probantes), la valorisation des données massives par les systèmes d'IA représente en effet une opportunité de rationaliser la production de « preuves » pour aider la décision clinique, qu'elles soient issues de données biomédicales, de la littérature scientifique ou des médias sociaux et du web (Peek et al. 2015). L'épistémologie de la science *data driven* favoriserait l'émergence de nouvelles idées que la science guidée par la connaissance n'aurait pas permise (Kitchin 2014). Cette science guidée par les données pourrait ainsi transformer notre conception des développements aboutissant à la connaissance, notamment parce qu'elle marquerait la fin du modèle « hypothético-déductif » pour

laisser place à un modèle « empirico-inductif » où les données guident la création de connaissance (Coutellec et Weil-Dubuc 2017).

Considérant l'hyperbole (ou « *hype* ») entourant le développement et l'usage de l'IA et des données massives (Emanuel et Wachter 2019; Gandomi et Haider 2015; Fox et Do 2013; Gibert 2019), il est raisonnable de questionner si l'arrivée de ces méthodes représente réellement un changement de paradigme. Kitchin (2014) reconnait que l'avènement de ce nouveau paradigme n'est pas manifeste et nécessite encore un cadre théorique robuste. On peut également questionner si un modèle guidé par les données est une réelle transformation des approches scientifiques traditionnelles, considérant que construire des théories à partir des données est une démarche qui existait avant l'innovation numérique, notamment dans le domaine des sciences sociales[51]. Cependant, le « véritable point de bascule », selon Cardon, Cointet, et Mazières (2018), est la rétropropagation permises par les récentes avancées connexionnistes : les couches additionnelles de réseaux de neurones peuvent apprendre à partir des potentielles erreurs observées en sortie (*outputs*) du réseau, qui peut être propagée vers les entrées. Cela signifie que les réseaux de neurones peuvent maintenant analyser ce qui jusqu'à présent constituait le résultat de leur analyse (Cardon, Cointet, et Mazières 2018). De plus, c'est l'efficacité de la prédiction connexionniste qui marque une rupture importante dans le champ de l'IA : il est possible de faire des prédictions sur des données non-linéaires que l'on retrouve couramment dans des contextes réels en dehors des conditions de laboratoire, les modèles étant de plus en plus tolérants aux potentielles erreurs retrouvées dans les jeux de données (Cardon, Cointet, et Mazières 2018).

Que le changement de paradigme scientifique soit avéré ou non, l'innovation numérique – et donc l'usage des systèmes d'IA – s'accompagne incontestablement de transformations importantes en ce qui a trait aux systèmes de santé, que l'on considère qu'il existe une réelle rupture avec leur fonctionnement traditionnel ou que ces transformations s'opèrent en continuité[52] d'une

[51] C'est par exemple l'approche défendue par différentes méthodes empiriques dites « inductives », comme la théorisation ancrée qui, à la différence des approches traditionnelles où la théorie est créée puis vérifiée, consiste à construire une théorie et à la valider à partir des données (Paillé 1994).
[52] Satava décrit par exemple déjà en 2003 les effets disruptifs de l'avènement de technologies en santé ; incluant l'IA mais ne se limitant pas à celle-ci.

automatisation croissante où l'ampleur de la présence technologique est sans précédent. Selon Schwab, qui défend l'idée de l'apparition d'une quatrième révolution industrielle qui prend source dans l'avènement des nouvelles technologies numériques, ces changements ne sont pas un simple prolongement de la 3ème révolution industrielle pour trois raisons : 1) la rapidité (exponentielle) du déploiement de cette révolution ; 2) l'ampleur et la profondeur des bouleversements qui l'accompagnent – qui ne transforment pas seulement notre manière de faire mais également notre conception de « qui nous sommes » ; et 3) son impact systémique – car cette révolution implique la transformation de systèmes entiers à tous les niveaux de la société (Schwab 2016). Selon Lahlou (2015), cette nouvelle révolution est caractérisée par l'automatisation du travail sur l'humain, qui fait suite à celle du travail sur la nature, sur la matière et sur l'information qui se sont opérées par le biais des révolutions industrielles précédentes.

L'automatisation du travail « sur l'humain » et belle est bien celle qui s'observe en santé. Il est en effet possible d'identifier différents changements relativement au secteur médical, imposés par la transition numérique. En premier lieu, les avenues prometteuses décrites précédemment s'accompagnent d'une certaine « immédiateté dans notre rapport aux données » dont l'accès s'est partiellement affranchi de la médiation humaine, soit celle des professionnels de santé ou des patients (Coutellec et Weil-Dubuc 2017). C'est particulièrement le cas de la santé mobile, où le patient se retrouve de plus en plus responsabilisé dans la gestion de sa propre santé (Devillier 2017b). La portabilité de ces nouveaux dispositifs crée de nouveaux lieux et de nouveaux agents de soins qui peuvent défier la gouvernance actuelle des systèmes de santé :

> Notre système de santé sera profondément remanié dans les prochaines années du fait des innovations thérapeutiques ou diagnostiques à venir. Le passage des actes de l'hôpital vers la ville, de la ville au domicile, du domicile au travail, le transfert des compétences entre professionnels et patients (surveillance de traitement, du diabète, de la pression artérielle…), les innovations de la robotique, de la télémédecine et des Big Data représentent autant de challenges que nous allons devoir prendre en compte et maîtriser pour créer la santé de demain (Brouard, 2017, p. 29).

Concernant la recherche en santé, un changement majeur s'observe concernant la taille potentielle des cohortes de participants ou le temps nécessaire pour faire les études (Brouard 2017). Le ResearchKit d'Apple a par exemple transformé les iPhones en outils de collecte pour la recherche,

transformant les centaines de millions d'utilisateurs de ces téléphones intelligents en participants potentiels. C'est ce que présente Sharon (2016) :

> Early ResearchKit apps have been a success in this sense: Apple likes to boast that within 24 hours of the launch of Stanford's cardiovascular study, 11,000 participants had signed up. Similarly, the Parkinson mPower app was downloaded by 680 people in its first 3 h, and the Icahn School of Medicine's asthma study app enrolled more than 8000 participants within 6 months, all without any direct contact with researchers (p. 565).

S'ensuit une dissociation entre la capacité de collecte et la capacité d'analyse (temporellement et sémantiquement) qui conduit ainsi à un décalage entre le rythme de production des données et la vitesse de l'appropriation du sens qu'elles recèlent (Coutellec et Weil-Dubuc 2017).

Également, ces nouveaux modes de collecte de données (en particulier, les applications mobiles ou les capteurs) offrent l'opportunité à des firmes multinationales (ex. les GAFAM[53]) d'entrer dans la sphère de la recherche biomédicale (Sharon 2016). L'entrée de ces nouveaux acteurs s'accompagne de nouvelles asymétries de pouvoir entre secteurs public et privé entraînant un déséquilibre entre les expertises qui affecte la recherche dans son ensemble (Sharon 2016). En effet, ces firmes internationales, possédant des moyens supérieurs à ceux dont peut disposer la recherche publique, deviennent indispensables à l'existence même d'un système de santé dont le fonctionnement repose sur les données, rendent le système public traditionnel caduque[54] et amènent des préoccupations relatives à l'apparition d'un monopole privé tant au niveau de l'expertise numérique que de celle nécessaire à la gestion des données (Sharon 2016).

L'innovation numérique en santé s'accompagne de répercussions pour les patients, les professionnels de santé et le système de santé à différents niveaux (Thompson et al. 2018). Pour Devillier (2017) c'est l'ensemble du parcours de soin du patient qui est affecté par le partage de données. Les transformations qui l'accompagnent ne sont pas sans conséquence sur l'évaluation

[53] Acronyme faisant références aux 5 principales firmes privées de l'écosystème de l'IA : Google, Amazon, Facebook, Apple et Microsoft.

[54] C'est par exemple ce que présente Sharon (2016), qui décrit que l'Institut National du cancer américain a reconnu que son système traditionnel d'analyse des données n'est plus viable et que l'infrastructure de grands organismes de recherche comme le NIH ne sont plus à la hauteur de celles des entreprises privées internationales qui collectent et conservent des données génomiques (Sharon, 2016).

éthique et scientifique des projets de recherche ou de la pratique clinique (Rial-Sebbag 2017; Thompson et al. 2018). Des implications en termes de consentement libre et éclairé des patients ou des participants (Rial-Sebbag 2017), de protection de la vie privée (Azencott 2018); de sécurité (Brundage et al. 2018) ou encore de justice sociale (Ganascia 2018; Mittelstadt et Floridi 2016) apparaissent. Le CCNE français va jusqu'à décrire une « rupture » qui conduit à un changement dans la perception des enjeux éthiques traditionnels de la recherche et du soin, selon quatre principales caractéristiques majeures des données massives en santé :

> - Un changement d'échelle, tenant à l'augmentation considérable du nombre des données disponibles et de notre capacité à les analyser ;
>
> - Leur pérennité : utiliser les données ne les détruit pas, elles sont donc réutilisables ;
>
> - Leur diffusion rapide, qui permet leur partage, et peut s'opérer au-delà de l'équipe médicale et des frontières nationales ;
>
> - Leur capacité à générer de nouvelles informations (données secondaires) et de nouvelles hypothèses, par l'effet de leur traitement (CCNE 2019, p.15).

Le changement associé au développement desdites technologies n'est alors pas seulement technologique mais également éthique et social. Ce changement tend vers une redéfinition de la relation de soin et une responsabilisation accrue du patient. Les risques associés à l'innovation numérique en santé ne sont pas précisément définis, et il est encore nécessaire de démontrer la pertinence de l'application pratique de ces technologies, soit d'en valider la portée bénéfique (Panch, Mattie, et Celi 2019)[55]. De nouveaux acteurs entrent en ligne de compte et de nombreux éléments amènent à penser qu'une redéfinition des responsabilités telles qu'entendues jusqu'à présent est nécessaire, pour les chercheurs comme pour les cliniciens :

> Le passage d'une recherche basée sur des hypothèses à une recherche basée sur les données a eu de ce fait de grandes implications pour les acteurs de la santé en modifiant la qualité de leurs liens et l'étendue de leurs droits et devoirs (Rial-Sebbag 2017 p. 44).

Considérant les différents points de rupture qui accompagnent les transformations décrites, il est alors possible de considérer les technologies qui relèvent de l'IA et des données massives comme des technologies « disruptives » (Sharon 2016; van den Broek et van Veenstra 2018;

[55] Le manque de connaissance sur les bénéfices et les risques est développé plus en détail dans le Chapitre 3.

Thompson et al. 2018; Howard 2014; Wadhwa 2014). Selon le modèle de Christensen (2003) les technologies (ou innovations) disruptives se définissent en opposition aux technologies incrémentielles (*incremental*) ou durables (*sustaining*) (Christensen, Raynor, et McDonald 2015; Christensen et Overdorf 2000). Alors que le développement des dernières vise l'amélioration de produits déjà existants, les innovations disruptives ont généralement un impact majeur et inattendu sur le marché (Pavie et Egal 2014). Ces innovations modifient radicalement les conditions d'usage et impliquent le plus souvent un changement technique ou technologique radical (Pavie et Egal 2014)[56]. L'anticipation des risques associés aux innovations disruptives ainsi que l'évaluation de leur impact sur la société est plus difficile que pour les autres innovations, selon Pavie et Egal (2014), en raison de deux principaux facteurs : 1) la complexité à anticiper les niveaux d'adoption et donc de gérer l'effet de masse potentiel qui en résulte et 2) le *knowledge gap*, soit la connaissance limitée de l'existence de risques ou autres conséquences imprévisibles :

> « Disruptive innovations often rely on new techniques or technologies, for which scientific knowledge is still limited, and for which all consequences cannot always be foreseen » (Pavie et Egal, p. 58).

Également, la rapidité de l'implémentation des technologies disruptives et la relative lenteur des processus de régulation créent un écart entre le développement technologique et son encadrement (Howard 2014; Satava 2003), risquant de laisser la gouvernance de ces technologies guidée par d'autres impératifs que les exigences éthiques et sociales :

> « Disruptive technology is immediately felt on nately, our political, social, and behavioral systems are too slow to respond, and the moral and ethical implications are either ignored or made subservient to a more pressing (commercial?) need » (Satava 2003 p. 247).

La rupture et les conséquences qui accompagnent le développement de technologies disruptives viennent ainsi défier le respect des règles traditionnelles de gouvernance (notamment éthiques) et présente des enjeux de responsabilité sociale attribuée aux acteurs de la recherche et de l'innovation

[56] Considérer les technologies qui relèvent de l'IA et des données massives comme disruptives selon les théories de Christensen (soit, d'un point de vue managérial) mériterait cependant une analyse plus approfondie. La « disruption » telle que décrite par Christensen (2003), renvoie à un processus par lequel une petite entreprise disposant de peu de ressources est en mesure de remplacer avec succès les entreprises établies en développant des innovations considérées a priori comme inappropriées pour satisfaire les consommateurs selon les métriques d'évaluation en place (Danneels 2004; Christensen, Raynor, et McDonald 2015; Christensen et Overdorf 2000). Cette théorie renvoie à un énoncé basé sur des corrélations (Christensen, Raynor, et McDonald 2015) relatives aux effets perturbateurs de certaines technologies sur la dynamique du marché (Kenagy et Christensen 2002). Ces technologies changent les métriques de performance selon lesquelles les entreprises vont rivaliser (Danneels, 2004). Ici, c'est bien selon la rupture dans les conditions d'usage et la complexité de la gestion des risques tels que décrits par Pavie et Egal (2014) que le terme est utilisé.

(Wadhwa 2014; Pavie et Egal 2014; Howard 2014; Kolko 2012; Satava 2003). De plus, les facteurs d'incertitude décrit par Pavie et Egal (2014) rendent la détermination de la responsabilité des parties prenantes de l'innovation indispensable pour permettre aux humains de se protéger eux-mêmes ainsi que de protéger leur environnement (Pavie et Egal 2014). Il semble alors essentiel de se pencher sur le caractère disruptif de l'innovation numérique en santé d'un point de vue de l'attribution de la responsabilité associée à leur développement. Les transformations qu'amènent ces technologies demandent de prendre expressément en considération leurs conséquences éthiques et sociales en vue d'une innovation responsable.

4. Conclusion

L'innovation numérique en santé se caractérise par la collecte, le stockage et l'analyse d'un nombre croissant de données massives. Pour donner du sens à ces ensembles de données, différentes méthodes relatives à l'IA sont aujourd'hui utilisées, participant au développement d'un système de santé connecté et apprenant. Les méthodes d'IA et les données massives sont indissociables et se potentialisent, les algorithmes d'apprentissage automatique ayant besoin de beaucoup de données pour apprendre et les données massives n'étant que peu utiles sans l'analyse automatisée que permettent les méthodes utilisées en IA.

Une nouvelle ère de la médecine se profile ainsi, où IA et données massives occupent une place prépondérante et pourraient permettre de maximiser l'efficacité des traitements, de rentabiliser les coûts en automatisant certaines tâches, voire de mettre en évidence de nouvelles pistes pour la prestation de soins tout comme pour la prévention. Dans cette lignée, des algorithmes de plus en plus autonomes traitant de problèmes de plus en plus complexes permettent de soutenir les professionnels de la santé dans leurs prises de décisions, aussi importantes que celles liées au diagnostic. Lorsqu'ils sont embarqués dans des robots de soins ou autres dispositifs telles que les applications mobiles, les algorithmes peuvent même se trouver à être en « relation » directe avec le patient.

Cependant, les changements qui accompagnent l'avènement de ces technologies semblent marquer une rupture avec le système traditionnel de soin et de recherche en santé. Cette transformation au sein du système de santé présente alors des défis sur lesquels il est nécessaire de

se pencher dans l'optique d'une utilisation éthique de l'IA et des données massives, considérant la place croissante qu'occupent ces technologies. Ainsi, identifier les déterminants à la source de la transformation du système de santé, les risques et enjeux éthiques qui en découlent ainsi que les moyens d'y remédier est essentiel afin d'assurer un développement responsable de l'innovation numérique en santé.

Références bibliographiques

Alanazi, Hamdan O., Abdul Hanan Abdullah, et Kashif Naseer Qureshi. 2017. « A Critical Review for Developing Accurate and Dynamic Predictive Models Using Machine Learning Methods in Medicine and Health Care ». *Journal of Medical Systems* 41 (4): 69. https://doi.org/10.1007/s10916-017-0715-6.

Ashley, Euan A. 2015. « The Precision Medicine Initiative: A New National Effort ». *JAMA* 313 (21): 2119-20. https://doi.org/10.1001/jama.2015.3595.

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): 20170350. https://doi.org/10.1098/rsta.2017.0350.

Baillie, Jean-Christophe. 2016. « Artificial Intelligence: The Point of View of Developmental Robotics ». Dans *Fundamental issues of artificial intelligence*, Springer International Publishing Switzerland, 555-68. V.C. Müller.

Bayer, Ronald, et Sandro Galea. 2015. « Public Health in the Precision-Medicine Era ». *The New England Journal of Medicine* 373 (6): 499-501. https://doi.org/10.1056/NEJMp1506241.

Bizer, Christian, Peter Boncz, Michael L. Brodie, et Orri Erling. 2012. « The Meaningful Use of Big Data: Four Perspectives – Four Challenges ». *SIGMOD Rec.* 40 (4): 56–60. https://doi.org/10.1145/2094114.2094129.

Blumenthal, David. 2009. « Stimulating the Adoption of Health Information Technology ». *The New England Journal of Medicine* 360 (15): 1477-79. https://doi.org/10.1056/NEJMp0901592.

Blumenthal, David, et Marilyn Tavenner. 2010. « The "Meaningful Use" Regulation for Electronic Health Records ». *The New England Journal of Medicine* 363 (6): 501-4. https://doi.org/10.1056/NEJMp1006114.

Bostrom, Nick, et Eliezer Yudkowsky. 2011. « The Ethics of Artifical Intelligence ». Dans *The Cambridge Handbook of Artificial Intelligence*, 316-35. Cambridge University Press.

Broek, Tijs van den, et Anne Fleur van Veenstra. 2018. « Governance of big data collaborations: How to balance regulatory compliance and disruptive innovation ». *Technological Forecasting and Social Change* 129 (avril): 330-38. https://doi.org/10.1016/j.techfore.2017.09.040

Brouard, Benoît. 2017. « Chapitre 2. Utilisation des Big Data en santé : le cas des objets connectés ». *Journal international de bioethique et d'ethique des sciences* Vol. 28 (3): 27-30.

Brown, Nathan, Jean Cambruzzi, Peter J. Cox, Mark Davies, James Dunbar, Dean Plumbley, Matthew A. Sellwood, et al. 2018. « Chapter Five - Big Data in Drug Discovery ». Dans *Progress in Medicinal Chemistry*, édité par David R. Witty et Brian Cox, 57:277-356. Elsevier. https://doi.org/10.1016/bs.pmch.2017.12.003.

Brundage, Miles, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, et Bobby Filar. 2018. « The malicious use of artificial intelligence: Forecasting, prevention, and mitigation ». *arXiv preprint arXiv:1802.07228*.

Bzdok, Danilo, et Andreas Meyer-Lindenberg. 2018. « Machine Learning for Precision Psychiatry: Opportunities and Challenges ». *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 3 (3): 223-30. https://doi.org/10.1016/j.bpsc.2017.11.007.

Cano, Isaac, Akos Tenyi, Emili Vela, Felip Miralles, et Josep Roca. 2017. « Perspectives on Big Data applications of health information ». *Current Opinion in Systems Biology*, • Mathematical modelling • Mathematical modelling, Dynamics of brain activity at the systems level • Clinical and translational systems biology, 3 (juin): 36-42. https://doi.org/10.1016/j.coisb.2017.04.012.

Cardon, Dominique, Jean-Philippe Cointet, et Antoine Mazières. 2018. « La revanche des neurones ». *Reseaux* n° 211 (5): 173-220.

CCNE. 2019. « Données massives (big data) et santé : une nouvelle approche des enjeux éthiques ». Avis 130. Comité Consultatif National d'Éthique français. https://www.ccne-ethique.fr/sites/default/files/avis_130.pdf.

Chambers, David A., W. Gregory Feero, et Muin J. Khoury. 2016. « Convergence of Implementation Science, Precision Medicine, and the Learning Health Care System: A New Model for Biomedical Research ». *JAMA* 315 (18): 1941-42. https://doi.org/10.1001/jama.2016.3867.

Chang, Ray M., Robert J. Kauffman, et YoungOk Kwon. 2014. « Understanding the paradigm shift to computational social science in the presence of big data ». *Decision Support Systems*, 1. Business Applications of Web of Things 2. Social Media Use in Decision Making, 63 (juillet): 67-80. https://doi.org/10.1016/j.dss.2013.08.008.

Chartrand, Gabriel, Phillip M. Cheng, Eugene Vorontsov, Michal Drozdzal, Simon Turcotte, Christopher J. Pal, Samuel Kadoury, et An Tang. 2017. « Deep Learning: A Primer for Radiologists ». *RadioGraphics* 37 (7): 2113-31. https://doi.org/10.1148/rg.2017170077.

Chen, Ying, JD Elenee Argentinis, et Griff Weber. 2016. « IBM Watson: How Cognitive Computing Can Be Applied to Big Data Challenges in Life Sciences Research ». *Clinical Therapeutics* 38 (4): 688-701. https://doi.org/10.1016/j.clinthera.2015.12.001.

Christensen, Clayton M., et Michael Overdorf. 2000. « Meeting the Challenge of Disruptive Change ». *Harvard Business Review*, 1 mars 2000. https://hbr.org/2000/03/meeting-the-challenge-of-disruptive-change.

Christensen, C. 2003. The innovator's solution. Boston: Harvard Business School Press.

Christensen, Clayton M., Michael E. Raynor, et Rory McDonald. 2015. « What Is Disruptive Innovation? » *Harvard Business Review*, 1 décembre 2015. https://hbr.org/2015/12/what-is-disruptive-innovation.

CNIL (Commission nationale informatique et libertés). 2017. « Comment permettre à l'homme de garder la main ? Les enjeux éthiques des algorithmes et de l'intelligence artificielle ».

Collier, Matthew, Richard Fu, et Lucy Yin. 2017. « Artificial intelligence : healthcare's new nervous system ». *Accenture consulting* (blog). 2017. https://www.accenture.com/us-en/insight-artificial-intelligence-healthcare.

Coutellec, Léo, et Paul-Loup Weil-Dubuc. 2017. « Chapitre 7. Big data ou l'illusion d'une synthèse par agrégation. Une critique épistémologique, éthique et politique ». *Journal international de bioethique et d'ethique des sciences* Vol. 28 (3): 63-79.

Danneels, Erwin. 2004. « Disruptive Technology Reconsidered: A Critique and Research Agenda ». *Journal of Product Innovation Management* 21 (4): 246-58. https://doi.org/10.1111/j.0737-6782.2004.00076.x.

Déclaration de Montréal IA Responsable. 2018. « Rapport de la Déclaration de Montréal pour un Développement Responsable de l'Intelligence Artificielle ». https://docs.wixstatic.com/ugd/ebc3a3_d806f109c4104c91a2e719a7bef77ce6.pdf.

Devillers, Laurence. 2017. *Des robots et des hommes: Mythes, fantasmes et réalité*. Plon.

Devillier, Nathalie. 2017. « Chapitre 6. Les dispositions de la loi de modernisation de notre système de santé relatives aux données de santé ». *Journal international de bioethique et d'ethique des sciences* Vol. 28 (3): 57-61.

Ding, Michael Q., Lujia Chen, Gregory F. Cooper, Jonathan D. Young, et Xinghua Lu. 2018. « Precision Oncology beyond Targeted Therapy: Combining Omics Data with Machine Learning Matches the Majority of Cancer Cells to Effective Therapeutics ». *Molecular Cancer Research* 16 (2): 269-78. https://doi.org/10.1158/1541-7786.MCR-17-0378.

Dodig-Crnkovic Gordana. 2016. « Information, Computation, Cognition. Agency-Based Hierarchies of Levels ». Dans *Fundamental Issues of Artificial Intelligence*, Springer, 141-61. V.C. Müller.

Emanuel, Ezekiel J., et Robert M. Wachter. 2019. « Artificial Intelligence in Health Care: Will the Value Match the Hype? » *JAMA*, mai. https://doi.org/10.1001/jama.2019.4914.

Esteva, Andre, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, et Sebastian Thrun. 2017. « Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks ». *Nature* 542 (7639): 115-18. https://doi.org/10.1038/nature21056.

Fakoor, Rasool, Faisal Ladhak, Azade Nazi, et Manfred Huber. 2013. « Using deep learning to enhance cancer diagnosis and classificiation ».

Flasiński, Mariusz. 2016. « Application areas of AI systems ». Dans *Introduction to artificial intelligence*, 223-35. Springer.

Fox, Stephen, et Tuan Do. 2013. « Getting real about Big Data: applying critical realism to analyse Big Data hype ». *International Journal of Managing Projects in Business* 6 (4): 739-60. https://doi.org/10.1108/IJMPB-08-2012-0049.

Fürnkranz, Johannes. 2005. « Web Mining ». Dans *Data Mining and Knowledge Discovery Handbook*, édité par Oded Maimon et Lior Rokach, 899-920. Boston, MA: Springer US. https://doi.org/10.1007/0-387-25465-X_42.

Ganascia, Jean-Gabriel. 2018. « Éthique, intelligence artificielle et santé ». Dans *Traité de bioéthique*, 527–540. ERES.

Gandomi, Amir, et Murtaza Haider. 2015. « Beyond the hype: Big data concepts, methods, and analytics ». *International Journal of Information Management* 35 (2): 137-44. https://doi.org/10.1016/j.ijinfomgt.2014.10.007.

Gibert, Martin. 2019. « Faut-il avoir peur de la peur de l'IA ? » *La Quatrième Blessure* (blog). 11 janvier 2019. https://medium.com/@martin.gibert/faut-il-avoir-peur-de-la-peur-de-lia-1687abc35342.

Guerriero, Ludovica, Giuseppe Quero, Michele Diana, Luc Soler, Vincent Agnus, Jacques Marescaux, et Francesco Corcione. 2018. « Virtual Reality Exploration and Planning for Precision Colorectal Surgery »: *Diseases of the Colon & Rectum* 61 (6): 719-23. https://doi.org/10.1097/DCR.0000000000001077.

Gulshan, Varun, Lily Peng, Marc Coram, Martin C. Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, et al. 2016. « Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs ». *JAMA* 316 (22): 2402-10. https://doi.org/10.1001/jama.2016.17216.

Hamet, Pavel, et Johanne Tremblay. 2017. « Artificial intelligence in medicine ». *Metabolism*, Insights Into the Future of Medicine: Technologies, Concepts, and Integration, 69 (Supplement): S36-40. https://doi.org/10.1016/j.metabol.2017.01.011.

Harnad, Stevan. 2005. « Cognition is categorization ». Dans *Handbook of Categorization*, édité par Henri Cohen et Claire Lefebvre. Elsevier. http://cogprints.org/3027/.

Howard, Alex. 2014. « Disruptive Technologies Pose Difficult Ethical Questions for Society ». TechRepublic. 22 avril 2014. https://www.techrepublic.com/article/disruptive-technologies-pose-difficult-ethical-questions-for-society/.

Isitor, Emmanuel, et Clare Stanier. 2016. « Defining Big Data ». Dans American University, Bulgaria. http://eprints.staffs.ac.uk/2767/.

Jameson, J. Larry, et Dan L. Longo. 2015. « Precision Medicine--Personalized, Problematic, and Promising ». *The New England Journal of Medicine* 372 (23): 2229-34. https://doi.org/10.1056/NEJMsb1503104.

Jiang, Fei, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, et Yongjun Wang. 2017. « Artificial intelligence in healthcare: past, present and future ». *Stroke and Vascular Neurology*. https://doi.org/10.1136/svn-2017-000101.

Kattan, M. W. 2001. « Expert Systems in Medicine ». Dans *International Encyclopedia of the Social & Behavioral Sciences*, édité par Neil J. Smelser et Paul B. Baltes, 5135-39. Oxford: Pergamon. https://doi.org/10.1016/B0-08-043076-7/00556-8.

Kenagy, John W., et Clayton M. Christensen. 2002. « Disruptive Innovation: A New Diagnosis for Health Care's "Financial Flu" » Healthcare Financial Management. 1 mai 2002. https://link.galegroup.com/apps/doc/A86064162/AONE?sid=lms.

Kitchin, Rob. 2014. « Big Data, New Epistemologies and Paradigm Shifts ». *Big Data & Society* 1 (1): 2053951714528481. https://doi.org/10.1177/2053951714528481.

Kolko, Jon. 2012. « The Ethics of Disruptive Innovation in Wicked Problems – Austin Center for Design ». avril 2012. https://www.ac4d.com/2012/04/the-ethics-of-disruptive-innovation-in-wicked-problems/.

Kononenko, Igor. 2001. « Machine learning for medical diagnosis: history, state of the art and perspective ». *Artificial Intelligence in Medicine* 23 (1): 89-109. https://doi.org/10.1016/S0933-3657(01)00077-X.

Lahlou, Saadi. 2015. « Un monde numérique : le renversement du miroir ». Dans Vol. 53. Variances.

Lavrač, Nada, et Blaž Zupan. 2010. « Data Mining in Medicine ». Dans *Data Mining and Knowledge Discovery Handbook*, édité par Oded Maimon et Lior Rokach, 1111-36. Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-09823-4_58.

LeCun, Yann, Yoshua Bengio, et Geoffrey Hinton. 2015. « Deep Learning ». *Nature* 521 (7553): 436-44. https://doi.org/10.1038/nature14539.

Lee, Howard. 2014. « Paging Dr. Watson: IBM's Watson Supercomputer Now Being Used in Healthcare ». *Journal of AHIMA* 85 (5): 44-47.

Lipworth, Wendy, Paul H. Mason, Ian Kerridge, et John P. A. Ioannidis. 2017. « Ethics and Epistemology in Big Data Research ». *Journal of Bioethical Inquiry* 14 (4): 489-500. https://doi.org/10.1007/s11673-017-9771-3.

Liu, S., S. Liu, W. Cai, S. Pujol, R. Kikinis, et D. Feng. 2014. « Early diagnosis of Alzheimer's disease with deep learning ». Dans *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, 1015-18. https://doi.org/10.1109/ISBI.2014.6868045.

Lovejoy, Christopher A., Varun Buch, et Mahiben Maruthappu. 2019. « Technology and Mental Health: The Role of Artificial Intelligence ». *European Psychiatry* 55 (janvier): 1-3. https://doi.org/10.1016/j.eurpsy.2018.08.004.

Luo, Jake, Min Wu, Deepika Gopukumar, et Yiqing Zhao. 2016. « Big Data Application in Biomedical Research and Health Care: A Literature Review ». *Biomedical Informatics Insights* 8 (janvier): BII.S31559. https://doi.org/10.4137/BII.S31559.

Maimon, Oded, et Lior Rokach. 2010. « Introduction to Knowledge Discovery and Data Mining ». Dans *Data Mining and Knowledge Discovery Handbook*, édité par Oded Maimon et Lior Rokach, 1-15. Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-09823-4_1.

Manrai, Arjun K., Chirag J. Patel, et John P. A. Ioannidis. 2018. « In the Era of Precision Medicine and Big Data, Who Is Normal? » *JAMA* 319 (19): 1981-82. https://doi.org/10.1001/jama.2018.2009.

Máttyus, Gellért, et Raquel Urtasun. 2018. « Matching Adversarial Networks ». Dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8024–8032.

Merkow, Jameson, Robert Lufkin, Kim Nguyen, Stefano Soatto, Zhuowen Tu, et Andrea Vedaldi. 2017. « DeepRadiologyNet: Radiologist Level Pathology Detection in CT Head Images ». *arXiv:1711.09313 [cs]*, novembre. http://arxiv.org/abs/1711.09313.

Meunier, Jean-Guy. 2017. « Humanités numériques et modélisation scientifique ». *Questions de communication*, nº 31 (septembre): 19-48. https://doi.org/10.4000/questionsdecommunication.11040.

Miotto, Riccardo, Fei Wang, Shuang Wang, Xiaoqian Jiang, et Joel T. Dudley. 2018. « Deep Learning for Healthcare: Review, Opportunities and Challenges ». *Briefings in Bioinformatics* 19 (6): 1236-46. https://doi.org/10.1093/bib/bbx044.

Mirnezami, Reza, Jeremy Nicholson, et Ara Darzi. 2012. « Preparing for Precision Medicine ». *New England Journal of Medicine* 366 (6): 489-91. https://doi.org/10.1056/NEJMp1114866.

Mittelstadt, Brent Daniel, et Luciano Floridi. 2016. « The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts ». *Science and Engineering Ethics* 22 (2): 303-41. https://doi.org/10.1007/s11948-015-9652-2.

Moore, Robert J. 2011. « Eric Schmidt's "5 Exabytes" Quote Is a Load of Crap ». The Data Point. 2011. https://blog.rjmetrics.com/2011/02/07/eric-schmidts-5-exabytes-quote-is-a-load-of-crap/.

Müller, Vincent C., et Nick Bostrom. 2016. « Future progress in artificial intelligence : a survey of expert opinion ». Dans *Fundamental issues of artificial intelligence*, Springer International Publishing Switzerland, 555-68. V.C. Müller.

OMS. 2012. « National eHealth Strategy Toolkit ». World Health Organization.

———. 2016. « Atlas of eHealth country profiles : the use of eHealth in support of universal health coverage ». Global Observatory for eHealth. https://www.who.int/goe/publications/atlas_2015/en/.

Paillé, Pierre. 1994. « L'analyse par théorisation ancrée ». *Cahiers de recherche sociologique*, nº 23: 147-81. https://doi.org/10.7202/1002253ar.

Panch, Trishan, Heather Mattie, et Leo Anthony Celi. 2019. « The "Inconvenient Truth" about AI in Healthcare ». *Npj Digital Medicine* 2 (1): 1-3. https://doi.org/10.1038/s41746-019-0155-4.

Patel, Vimla L., Edward H. Shortliffe, Mario Stefanelli, Peter Szolovits, Michael R. Berthold, Riccardo Bellazzi, et Ameen Abu-Hanna. 2009. « The coming of age of artificial intelligence in medicine ». *Artificial Intelligence in Medicine*, Artificial Intelligence in Medicine AIME' 07, 46 (1): 5-17. https://doi.org/10.1016/j.artmed.2008.07.017.

Paul, Michael J., Abeed Sarker, John S. Brownstein, Azadeh Nikfarjam, Matthew Scotch, Karen L. Smith, et Graciela Gonzalez. 2016. « Social media mining for public health monitoring and

surveillance ». Dans *Biocomputing 2016: Proceedings of the Pacific Symposium*, 468–479. World Scientific.

Pavie, Xavier, et Julie Egal. 2014. « Innovation and Responsibility: A Managerial Approach to the Integration of Responsibility in a Disruptive Innovation Model ». Dans *Responsible Innovation 1: Innovative Solutions for Global Issues*, édité par Jeroen van den Hoven, Neelke Doorn, Tsjalling Swierstra, Bert-Jaap Koops, et Henny Romijn, 53-66. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-017-8956-1_4.

Peck, Richard W. 2018. « Precision Medicine Is Not Just Genomics: The Right Dose for Every Patient ». *Annual Review of Pharmacology and Toxicology* 58 (1): 105-22. https://doi.org/10.1146/annurev-pharmtox-010617-052446.

Peek, Niels, Carlo Combi, Roque Marin, et Riccardo Bellazzi. 2015. « Thirty years of artificial intelligence in medicine (AIME) conferences: A review of research themes ». *Artificial Intelligence in Medicine*, Artificial Intelligence in Medicine AIME 2013, 65 (1): 61-73. https://doi.org/10.1016/j.artmed.2015.07.003.

Quero, Giuseppe, Alfonso Lapergola, Luc Soler, Mouhamad Shabaz, Alexandre Hostettler, Toby Collins, Jacques Marescaux, Didier Mutter, Michele Diana, et Patrick Pessaux. 2019. « Virtual and Augmented Reality in Oncologic Liver Surgery ». *Surgical Oncology Clinics* 28 (1): 31-44. https://doi.org/10.1016/j.soc.2018.08.002.

Rial-Sebbag, Emmanuelle. 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): 39-50.

Ritchie, Marylyn D. 2012. « The success of pharmacogenomics in moving genetic association studies from bench to bedside: study design and implementation of precision medicine in the post-GWAS era ». *Human Genetics* 131 (10): 1615-26. https://doi.org/10.1007/s00439-012-1221-z.

Satava, Richard M. 2003. « Biomedical, Ethical, and Moral Issues Being Forced by Advanced Medical Technologies ». *Proceedings of the American Philosophical Society* 147 (3): 246-58.

Schwab, Klaus. 2016. *La quatrième révolution industrielle*. Dunod. Suisse: World Economic Forum.

Shafqat, Sarah, Saira Kishwer, Raihan Ur Rasool, Junaid Qadir, Tehmina Amjad, et Hafiz Farooq Ahmad. 2018. « Big Data Analytics Enhanced Healthcare Systems: A Review ». *The Journal of Supercomputing*, février. https://doi.org/10.1007/s11227-017-2222-4.

Shakhovska, Khrystyna. 2017. « Making answer algorithm for chat-bot ». Dans *Litteris et Artibus: матеріали*, 394–395. Видавництво Львівської політехніки.

Shameer, Khader, Kipp W. Johnson, Benjamin S. Glicksberg, Joel T. Dudley, et Partho P. Sengupta. 2018. « Machine Learning in Cardiovascular Medicine: Are We There Yet? » *Heart* 104 (14): 1156-64. https://doi.org/10.1136/heartjnl-2017-311198.

Sharon, Tamar. 2016. « The Googlization of Health Research: From Disruptive Innovation to Disruptive Ethics ». *Personalized Medicine* 13 (6): 563-74. https://doi.org/10.2217/pme-2016-0057.

Shmilovici, Armin. 2010. « Support Vector Machines ». Dans *Data Mining and Knowledge Discovery Handbook*, édité par Oded Maimon et Lior Rokach, 231-47. Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-09823-4_12.

Thompson, Reid F., Gilmer Valdes, Clifton D. Fuller, Colin M. Carpenter, Olivier Morin, Sanjay Aneja, William D. Lindsay, et al. 2018. « Artificial Intelligence in Radiation Oncology: A Specialty-Wide Disruptive Transformation? » *Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology* 129 (3): 421-26. https://doi.org/10.1016/j.radonc.2018.05.030.

Torkamani, Ali, Kristian G. Andersen, Steven R. Steinhubl, et Eric J. Topol. 2017. « High-Definition Medicine ». *Cell* 170 (5): 828-43. https://doi.org/10.1016/j.cell.2017.08.007.

Villani, Cédric. 2018. « Donner un sens à l'intelligence artificielle. Pour une stratégie nationale et européenne. » https://www.aiforhumanity.fr/pdfs/9782111457089_Rapport_Villani_accessible.pdf.

Wadhwa, Vivek. 2014. « Laws and Ethics Can't Keep Pace with Technology ». MIT Technology Review. 2014. https://www.technologyreview.com/s/526401/laws-and-ethics-cant-keep-pace-with-technology/.

Webster, Dan E., Christine Suver, Megan Doerr, Erin Mounts, Lisa Domenico, Tracy Petrie, Sancy A. Leachman, Andrew D. Trister, et Brian M. Bot. 2017. « The Mole Mapper Study, Mobile Phone

Skin Imaging and Melanoma Risk Data Collected Using ResearchKit ». *Scientific Data* 4 (février): 170005. https://doi.org/10.1038/sdata.2017.5.

Yi, Xin, Ekta Walia, et Paul Babyn. 2018. « Generative Adversarial Network in Medical Imaging: A Review ». *arXiv:1809.07294 [cs]*, septembre. http://arxiv.org/abs/1809.07294.

Zhang, G. Peter. 2010. « Neural Networks For Data Mining ». Dans *Data Mining and Knowledge Discovery Handbook*, édité par Oded Maimon et Lior Rokach, 419-44. Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-09823-4_21.

Chapitre 3 – Limites et enjeux éthiques de l'utilisation des systèmes d'intelligence artificielle en santé

L'analyse de données massives par le biais de systèmes d'intelligence artificielle (IA) permet ainsi de nombreuses applications prometteuses pour le système de santé. Ces technologies pourraient bien transformer la prévention, le diagnostic ou les soins, en favorisant l'efficacité et la précision des traitements, en rentabilisant les coûts ou en offrant de nouvelles perspectives de suivi pour les patients. Cependant, l'avènement des systèmes d'IA en santé n'est pas sans défi. Leur utilisation se heurte d'abord à trois principaux enjeux inhérents à la nature de leur fonctionnement : 1) des limites interprétatives et informationnelles ; 2) la nécessité de partage et 3) l'opacité des réseaux de neurones. Ces enjeux techniques sont susceptibles de conduire à différentes préoccupations éthiques relatives à la protection de la vie privée et de la confidentialité, au respect du consentement libre et éclairé ou de la justice sociale, à la déshumanisation des soins et du patient et à la protection de la sécurité. Ces différents défis sont à considérer pour assurer le développement d'une innovation numérique responsable en santé, soit d'en maximiser les bénéfices tout en limitant les risques.

1. Enjeux inhérents au fonctionnement des systèmes d'intelligence artificielle

1.1. Limites interprétatives et informationnelles de l'analyse des données massives

Comme mentionné dans le Chapitre 2, l'un des principaux défis dans le contexte du *Big Data* est de donner du sens à ces ensembles gigantesques de données plus ou moins structurées. L'analyse de données massives par des systèmes d'IA connait ainsi différentes limites interprétatives et informationnelles qu'il est nécessaire de prendre en considération. En santé, la recherche sur les données massives est reconnue pour être « observationnelle plus qu'expérimentale », et il existe un risque d'amplifier les lacunes de la recherche traditionnelle considérant le volume et la qualité souvent sous optimale de l'information caractéristique des données massives (Lipworth et al. 2017). Due à l'évolutivité des méthodes de *data mining* - qui fonctionnent dans un environnement avec un grand nombre de données, une forte dimensionnalité

et une grande hétérogénéité - une des difficultés réside dans le stockage efficient et la possibilité de traiter rapidement les volumes disponibles (Maimon et Rokach 2010b). Les ensembles de données massives qui font l'objet d'analyse étant issues de sources variées, leur qualité peut être compromise par leur trop grande hétérogénéité ou par différentes erreurs, biais ou observations manquantes - en particulier pour les données non-structurées (Zhang 2010; Lipworth et al. 2017).

Une autre des difficultés est également de limiter les erreurs liées à la fragmentation des données. Typiquement, les données vont être divisées en deux : un échantillon d'entrainement pour l'ajustement du modèle et un échantillon test pour évaluer ses capacités de prédiction (Maimon et Rokach 2010a). Il est critique que l'échantillon test soit indépendant de l'échantillon d'entrainement (Maimon et Rokach 2010a) afin d'éviter le risque de « surapprentissage », soit de garantir que le niveau de performance des prédictions sur de nouvelles données (échantillon test) ne soit pas affecté par le fait que celles-ci soient trop proches des données d'apprentissage (Devillers 2017). C'est l'écart de performance entre les données d'entrainement et les données test qui va permettre de mesurer ce que l'on appelle l'erreur de généralisation (Zhang 2010; Lipton 2016). L'objectif des modèles (et en particulier des modèles prédictifs) étant qu'ils soient généralisables, il est ainsi nécessaire qu'ils ne soient pas trop dépendants des données sur lesquelles ils apprennent pour être performants (Zhang 2010).

Également, les analyses de *data mining* sont exposées au risque de supériorité sélective : tous les algorithmes ne sont pas performants pour toutes les tâches ni tous les domaines, car ils contiennent tous des biais potentiels qui mènent à préférer certaines généralisations plutôt que d'autres (Maimon et Rokach 2010b). D'autres barrières techniques peuvent limiter le pouvoir des analyses sur les données massives. Pour n'en citer que quelques-unes : l'interopérabilité (soit le manque de comparabilité entre les ensembles de données, notamment à travers le temps); le risque élevé de « faux positifs »; ou encore la validité des liens entre les ensembles de données, en particulier considérant la fragmentation de leurs sources en santé (Lipworth et al. 2017).

Il n'est pas question ici de remettre en question la validité ou l'importance des analyses issues de systèmes d'IA. Il n'existe en effet aucune approche, méthode ou technique qui ne connaissent des limites. Cependant, il semble nécessaire d'évaluer la réelle portée des résultats obtenus par le biais de ce type d'analyses pour se prémunir d'un certain engouement qui pourrait

conduire à en surestimer les bénéfices. À l'heure où l'IA connait une grande popularité, il semble indispensable d'adapter le niveau de confiance et d'interprétation des décisions algorithmiques :

> Some decision-making tools overstate or obfuscate their usefulness or accuracy, inducing more trust than they deserve » (CDT, 2017).

La valeur potentielle des données doit être évaluée dès le départ et utilisée pour orienter la justification des efforts déployés pour le traitement et l'analyse (Brown et al. 2018).

Kitchin (2014) décrit que, pour certains auteurs, l'automatisation pourraient conduire à un certain appauvrissement de la qualité des analyses (dans le contexte des domaines qui relèvent des sciences humaines) :

> For many, then, the digital humanities is fostering weak, surface analysis, rather than deep, penetrating insight. It is overly reductionist and crude in its techniques, sacrificing complexity, specificity, context, depth and critique for scale, breadth, automation, descriptive patterns and the impression that interpretation does not require deep contextual knowledge (p. 8).

Pour Coutellec and Weil-Dubuc (2017), les analyses de données massives ne produisent pas tant de la connaissance qu'une information « auto-signifiante » :

> Une donnée ne nous est jamais donnée, elle est prise dans un dispositif de collecte dont le paramétrage et le calibrage (la métrologie) dépendent de certaines hypothèses scientifiques et d'intentions de recherche; une donnée n'est plus une data à partir du moment où l'on applique sur elle un dispositif technique et une série de filtres interprétatifs, ce dont nous avons à faire est donc plutôt de l'ordre du big ficta (p. 67).

La distinction entre données et information est ici essentielle. Par exemple, si l'ADN est toujours une donnée, elle nécessite une interprétation pour être transformée en information, une séquence de nucléotides brute n'étant que peu informative (Hallinan et De Hert 2016).

Il est ainsi nécessaire de ne pas surestimer l'utilité des systèmes d'IA. Ceci est d'autant plus important que les algorithmes sont réputés être plus performants, moins biaisés, et plus précis que les humains (CDT 2017). La mise en relation automatisée des données, puisqu'elle semble parer la subjectivité humaine, pourrait ainsi donner une impression « d'objectivité absolue » trompeuse (Rouvroy et Berns 2013). Cette perception d'objectivité absolue pourrait conduire à ignorer les limites mentionnées :

> There are also numerous problems with the view that big data is somehow objective, including that this obscures the fact that all research questions, methods, and interpretations are value-laden; makes it easier to ignore technical quality issues and biases; and, more generally, makes it easier to justify unbounded use of big data (Lipworth et al. 2017 p. 495).

Il n'y a, par exemple, aucune garantie qu'un système d'apprentissage automatique supervisé mette en évidence un lien de causalité, mais seulement des associations (Lipton 2016). Plus largement, pour Coutellec et Weil-Dubuc, la production de connaissances d'une science guidée par les données se fait sur la base de corrélations qui remplace la recherche de causalité, selon une rationalité statistique qui pourrait conduire à négliger la cause des phénomènes (Coutellec et Weil-Dubuc 2017). Cette impression d'objectivité absolue ne devient cependant problématique, selon Rouvroy et Berns (2013) que si les interprétations algorithmique ne sont pas remises en question, en particulier quant au poids qu'on pourrait leur donner dans la décision politique ou scientifique.

Une mauvaise interprétation des analyse issues de systèmes d'IA n'est pas sans conséquences en santé :

> Caruana et al. (2015) describe a model trained to predict probability of death from pneumonia that assigned less risk to patients if they also had asthma. In fact, asthma was predictive of lower risk of death. This owed to the more aggressive treatment these patients received. But if the model were deployed to aid in triage, these patients would then receive less aggressive treatment, invalidating the model (Lipton 2016, p. 3).

Il est donc important de tempérer la portée des analyses de données massives par des systèmes d'IA, et de nuancer les bénéfices potentiels des avenues prometteuses en santé. Par exemple, accorder trop d'importance aux données (en particulier, génétiques) dans le contexte de la médecine de précision pourrait faire en sorte que différents facteurs socio-économiques, qui ont un impact non-négligeable sur la qualité et l'espérance de vie des patients, ne soient pas suffisamment considérés (Bayer et Galea 2015). Ceci conduit différents auteurs à questionner, par exemple, la réelle plus-value du développement de la médecine de précision, qui risque d'augmenter les coûts de manière considérable sans pour autant réellement améliorer la qualité de vie (Bayer et Galea 2015; Jameson et Longo 2015). Également, si des modèles comme les jumeaux numériques se développent, Torkamani et al. (2017) précisent que, dû à leur nature incomplète, ils pourraient avoir tendance à surestimer les risques et pourraient conduire à l'initiation de thérapies non-nécessaires. Ces considérations peuvent également mener au risque de considérer les patients uniquement sur

la base de leurs données (omettant d'autres informations pertinentes) conduisant à des enjeux relatifs au *quantified self*[57]. Ainsi, les limites interprétatives et informationnelles des analyses de données massives par les systèmes d'IA doivent être prise en considération en vue d'une évaluation éthique pour un équilibre appropriée entre risques et bénéfices.

1.2. De la nécessité du partage pour l'optimisation de l'analyse des données massives par les systèmes d'intelligence artificielle

S'il n'existe pas de règles spécifiques concernant la taille nécessaire de l'échantillon pour qu'un système soit considéré comme valide, plus l'échantillon est grand, plus le réseau de neurones a des chances de fonctionner efficacement (Zhang 2010; LeCun, Bengio, et Hinton 2015; Chartrand et al. 2017). Le développement de l'innovation numérique en santé (et plus largement de toutes applications de systèmes d'IA) est donc dépendant du partage de données (ou *data sharing*). Cet aspect est extrêmement important pour assurer la validité et la portée des prédictions des systèmes qui seront développés, qu'il s'agisse des systèmes experts comme Watson (Lee 2014), de la mutualisation des données pour des perspectives de santé publique (Brouard 2017) ou du développement de la médecine de précision (Azencott 2018; Fiore et Goodman 2016; Iyengar, Kundu, et Pallis 2018).

Ce partage est généralement vu comme étant nécessaire lorsque l'on considère l'intérêt du plus grand nombre, en particulier en ce qui a trait à la recherche. Comme le reconnaît Rial-Sebbag (2017):

> Le principe est relativement simple : afin d'accélérer la production de connaissances dans les sciences de la vie il est nécessaire d'encourager les chercheurs à partager leurs données pour le « bien » de la science et afin d'éviter toute réplication inutile dans les recherches (p. 48).

[57] Brouard (2017) définit le « quantified-self » comme « l'auto-mesure de soi », concept apparu en Californie en 2007 qui « *consiste à mesurer des données relatives à notre organisme et à nos activités physiques. Il se situe à la croisée de la santé connectée et des services de bien-être, car il s'adresse aux patients, mais aussi aux individus en bonne santé* » (p. 27). Cependant, la traduction du *quantified-self* par un « soi quantifié » semble plus appropriée car elle élargit le phénomène à la connaissance de soi par les nombres (Ajana 2017), que ces nombres soient issue de mesures effectuées par l'individu lui-même ou par d'autres personnes.

L'essor de l'innovation numérique en santé s'inscrit alors dans la veine des discussions autour de la « science ouverte », une forme de nouvel impératif social, comme le défend la Recommandation révisée de l'UNESCO concernant la Science et les Chercheurs scientifiques (2018b). La Recommandation reconnait la science comme un bien commun, demande de faciliter les mécanismes pour une science ouverte et collaborative, de garantir un accès libre et équitable aux données et contenus scientifiques et d'assurer le partage du fruit de toutes recherches à tous (UNESCO 2018b). C'est bien dans une perspective de partage que s'inscrit également le phénomène du *quantified self*, qu'Ajana (2017) décrit comme une « biopolitique du soi » ou le partage de données personnelles permet d'informer la communauté médicale selon un phénomène qui incite les individus à partager, par exemple, leurs informations d'activités physiques ou leur données biologiques (Ajana 2017) donnant ainsi accès à un nouveau type de données de santé, de haute qualité et auparavant inaccessibles. La nécessité du partage et de la mutualisation des données se défend sur la base d'un principe de solidarité (Ajana 2017; Sharon 2017; Woods 2016), ou selon un certain impératif moral de « sacrifice » des patients qui prennent le risque de partager leur données afin d'assurer de potentialiser les bénéfices (collectifs) de l'innovation numérique en santé (Longo et Drazen 2016; Woods 2016).

Différentes initiatives nationales ou internationales de création de bases de données accessibles à tous voient le jour, visant la mutualisation en vue de l'avancée des connaissances en santé. Par exemple, le gouvernement canadien a mis en place en 2011 le Portail de données ouvertes du Canada[58] – afin de fournir aux canadiens les données qui sont produites, recueillies et utilisées par différents ministères et organismes. Le but est d'offrir des ensembles de données ouvertes[59] en vue de soutenir l'innovation, la recherche et les consommateurs. À partir de ces ensembles, différentes applications sont développées, notamment en santé pour les consommateurs comme pour les professionnels; telle que l'application de visualisation des données fondée sur la Base de

[58] Voir : https://ouvert.canada.ca/fr/donnees-ouvertes

[59] Le gouvernement canadien définit les données ouvertes comme « des données structurées, lisibles par machine, qui peuvent être librement partagées, utilisées et mises à profit par quiconque, sans restriction » (voir : https://ouvert.canada.ca/fr/donnees-ouvertes). Elles se caractérisent par leur accès et leur disponibilité, doivent permettre la réutilisation (à des fins commerciales ou non) et favoriser la participation universelle (Verdier et Murciano 2017).

données Canada Vigilance qui permet de visualiser les effets indésirables de médicaments[60]. Dans la même lignée, en France, la plateforme ouverte des données française *data.gouv.fr* rend accessible, entres autres, les données relatives à l'offre et la consommation de soins, à l'efficacité du système de santé, à la santé publique ou aux médicaments[61]. Cette plateforme représente aujourd'hui une « communauté vivante de 17 500 contributeurs » offrant à la société civile la possibilité de partager et d'améliorer les données (Verdier et Murciano 2017).

Dans la même lignée, le Open Data Institute, une compagnie indépendante basée à Londres et fondée en 2012, a pour objectif de favoriser l'utilisation de données ouvertes dans le but d'offrir des bénéfices à l'échelle mondiale. L'institut travaille, avec différentes compagnies et gouvernements, à la construction d'un écosystème de données qui pourrait permettre aux individus de prendre de meilleures décisions tout en participant à la gestion d'impacts négatifs potentiels. Différentes initiatives de chaînes de blocs (ou *Blockchain)*[62] offrent également de nombreuses opportunités pour le secteur de la santé, qu'il s'agisse de la gestion de la santé publique, de la recherche médicale basée sur les données personnelles des patients ou de la contrefaçon de médicaments (Mettler 2016).

Les avantages du partage ne sont pas uniquement relatifs aux données mais concernent également l'accès, en *open source*, au code des algorithmes. La création de l'Écosystème IBM Watson pour aider différents secteurs (incluant la santé) à tirer parti des compétences de Watson en est un exemple. Grâce à cet écosystème, IBM offre un accès ouvert à sa plateforme permettant de créer des applications personnalisées, notamment afin que leurs partenaires puissent développer une vaste gamme de produits, tout en rendant Watson accessible au plus grand nombre (Lee 2014). Ce partage démontre ainsi autant un intérêt scientifique qu'économique. Il est parfois même

[60] Voir : https://ouvert.canada.ca/fr/apps/base-donnees-canada-vigilance-effets-indesirables-medicaments-canadiens
[61] Voir : https://www.data.gouv.fr/fr/topics/sante-et-social/
[62] Le *Blockchain* est une technologie de stockage et de gestion des données décentralisée et distribuée qui se présente comme une séquence continue de « blocs » ou d'informations. Ces informations proviennent des utilisateurs (généralement via un pseudonyme), elles sont vérifiées et les différentes actions réalisées dans le temps sont transparentes. Ces technologies sont particulièrement intéressantes pour la création de bases de données publiquement accessibles à différentes parties qui ont besoin de la même information tout en respectant la protection de l'anonymat des utilisateurs et de la propriété des données (Mettler 2016).

considéré comme un mode de gouvernance à part entière : celui des (biens) communs numériques, définis comme l'ensemble des ressources et savoirs librement accessibles, à la fois partagés et cocréés selon un mode d'organisation coopératif qui assure l'horizontalité des échanges entre les pairs, lesquels décident eux-mêmes des formes de régulation de cette organisation (Le Crosnier 2018).

Ce partage, une des conditions essentielles à l'existence du *paradigme du Big data*, ouvre alors la porte à de nombreuses préoccupations éthiques relatives à la vie privée et au consentement des individus qui génèrent des données pouvant être infiniment réutilisées, de différentes manières et sans qu'il soit possible de prévoir avec précision pour quelles fins elles le sont (Christen et al. 2016; Jones, Kaufman, et Edenberg 2018; Mittelstadt et Floridi 2016b). L'ouverture des contenus numériques expose également au risque de *re-enclosure*[63] privée (qui s'observe lors de croisement de données ouvertes et propriétaires) ou par l'État (Verdier and Murciano 2017; Le Crosnier 2011). Certains s'inquiètent également que la grande majorité des données demeurent la propriété de grosses firmes internationales (Sharon 2016), situation qui s'avère particulièrement préoccupante dans le cadre de données de santé où une asymétrie de pouvoir entre système publique et privé serait potentiellement dommageable.

Le potentiel économique ainsi que le pouvoir associé à la propriété des données massives font de ces dernières le « nouveau pétrole » de nos sociétés numériques (Malik 2013; WEF 2012). Selon cette conception, les données (dont les données de santé) deviennent des marchandises à croissance rapide qui confèrent aux géants du traitement des données (les GAFAM) un caractère imparable (The Economist 2017). Cette mutualisation pose aussi des problèmes de contrôle et de règlementation, notamment parce-que le partage dépasse les frontières et doit alors répondre à des législations différentes (Danaher 2015), conférant aux enjeux relatifs à l'ouverture des données une portée internationale. Ainsi, les enjeux de stockage et de gestion de ces bases de données gigantesques et publiquement accessibles dépassent les seuls défis de moyens techniques et financiers.

[63] C'est-à-dire des données publiques qui deviennent ou redeviennent la propriété d'une partie.

1.3. Opacité des réseaux de neurones : la « boite noire » de l'intelligence artificielle

Le développement de l'innovation numérique en santé se heurte également aux difficultés associées à l'opacité des réseaux de neurones artificiels. L'enjeu sous-jacent est celui de l'interprétabilité des algorithmes, ou comment remédier à la fameuse « boite noire » de l'IA. Les appels à l'explication sont une réaction à deux propriétés des modèles d'apprentissage qui conduisent à une perception d'opacité : l'impénétrabilité (le fait qu'il soit difficile de comprendre complètement le modèle) et la non-intuitivité (le fait qu'on ne comprenne pas pourquoi les règles du modèle sont ce qu'elles sont) (Selbst et Barocas 2018). L'opacité des méthodes d'apprentissage automatique est soulignée par Lavrač et Zupan (2010) dans le contexte médical :

> For data analysis tasks, however, the most serious limitation is the lack of explanational capabilities: the induced weights together with the network's architecture do not usually have an obvious interpretation and it is usually difficult or even impossible to explain «why» a certain decision was reached (p. 1124).

La prise de décision algorithmique étant devenue aujourd'hui quasi synonyme de prise de décision inexplicable (Selbst et Barocas 2018), nombreux sont ceux qui appellent à plus de transparence des systèmes d'IA (Jobin, Ienca, et Vayena 2019; Floridi et al. 2018). Cependant, une ambiguïté persiste relativement à ce qu'impliquerait une telle interprétabilité ou la manière de la mettre en place (Selbst et Barocas 2018; Lipton 2016).

L'interprétabilité des modèles peut se définir en termes de transparence (soit, l'intelligibilité des systèmes) ou en termes d'interprétations post-hoc (soit, l'explication des prédictions sans élucider le mécanisme par lequel le modèle fonctionne) (Lipton 2016). L'opacité des réseaux de neurones est particulièrement signifiante pour les réseaux de neurones profonds considérant la complexité des systèmes (Chartrand et al. 2017) et augmente avec le nombre de couches « cachées ». En santé, l'interprétabilité des décisions (ex. dans le cas du diagnostic) revêt une importance particulière car elle peut placer patients et médecins face à des dilemmes sérieux quand il s'agit de décisions médicales qui pourraient mettre en jeu la vie des patients (Castelvecchi 2016). Toutefois, il existe de nombreuses limitations à la transparence. Rendre les systèmes d'IA transparents se heurte à des défis techniques, temporels, liés au maintien du secret professionnel et à la compétitivité économique (Ananny et Crawford 2018; Selbst et Barocas 2018). La transparence

dépend également des compétences de ceux qui cherchent à comprendre le système (Ananny et Crawford 2018) puisque elle implique la compréhension du mécanisme par lequel le modèle fonctionne (Lipton 2016). Ainsi, il ne s'agit pas seulement de rendre le code des algorithmes accessible mais aussi compréhensible, celui-ci pouvant être obscur pour beaucoup d'utilisateurs qui ne détiendraient pas les compétences techniques pour le comprendre. Les codes des algorithmes sont parfois indéchiffrables par leurs créateurs eux-mêmes considérant la complexité et la rapidité de leur évolution (Ananny et Crawford 2018). En découle certains arguments contre le développement de tels algorithmes « boîte noire », d'autant plus si ces derniers égalent ou surpassent les capacités humaines (Lipton 2016).

Cependant, selon la deuxième conception (l'interprétabilité post-hoc), faisant référence à la possibilité de donner des informations utiles aux utilisateurs concernant une décision algorithmique sans pour autant élucider comment le modèle fonctionne, l'opacité des réseaux n'est pas problématique en soi – si l'on considère que les préoccupations relatives à la transparence surviennent lorsque l'existence d'un processus décisionnel est connu alors que le véritable processus qui a mené à la décision ne l'est pas (Selbst et Barocas 2018). Ces informations peuvent par exemple être fournies sous la forme d'explications en langage naturel, de visualisations ou d'explications par exemple (Lipton 2016). Selon cette conception de l'interprétabilité, les modèles linéaires ne sont alors pas forcément plus interprétables que les modèles d'apprentissage profond, notamment du fait que ces derniers apprennent de représentations riches qui peuvent être visualisées ou verbalisées (Lipton 2016). Cette position s'inscrit dans la lignée de ceux qui défendent l'idée que les décisions humaines elles-mêmes ne sont interprétables qu'après coup car « *nous avons tous une boite noire dans la tête* » (Castelvecchi 2016, traduction libre). Pour Selbst et Barocas (2018), ce problème n'est pas propre à la technologie : être l'objet d'une décision alors que les éléments à la base de la décision sont inconnus est une situation qui arrive fréquemment sans algorithme. En d'autres termes, il n'y aurait pas plus de raison d'exiger davantage de transparence des algorithmes que des humains, les réseaux de neurones biologiques sollicités et le processus cérébral qui amène à une décision humaine n'étant pas plus intelligible.

Cette vision ne fait cependant pas consensus, notamment car il est plus facile d'interroger un individu qu'une machine sur les raisons motivant une décision :

> One could argue that an accurate opaque system is preferable to an inaccurate transparent one, and that a human expert's image analysis can similarly be relatively opaque to a nonexpert. Nevertheless, it is currently much easier to interrogate a human expert's thought process than to decipher the inner workings of a deep neural network with millions of weights. Furthermore, an automated system's ability to clearly justify its analysis would be highly desirable for it to become widely acceptable for making critical judgments regarding patients' health (Chartrand et al. 2017 p. 2129).

Ainsi, il ne s'agit pas uniquement d'une nécessité de comprendre les raisons pour lesquelles une décision algorithmique est prise mais également de l'acceptabilité sociale du recours à ces technologies en santé.

L'interprétabilité n'est pas, en effet, un concept monolithique, mais fait référence à plusieurs idées distinctes : elle peut être comprise comme un moyen d'engendrer la confiance en la technologie ou un droit à l'explication sur son fonctionnement (Lipton 2016). Cet impératif de transparence prend racine dans les cultures épistémologiques scientifiques et sociotechniques (Ananny et Crawford 2018). La transparence et l'explicabilité des décisions émanant de systèmes d'IA posent des problèmes relatifs à la confiance tant du public que des professionnels de santé qui utilisent ces systèmes (Castelvecchi 2016; Lipton 2016). L'absence de transparence peut avant tout créer une certaine méfiance à l'égard des institutions, notamment car il n'est pas possible de savoir sur quelle base contester une décision algorithmique (CDT 2017). Il en découle des préoccupations relatives au contrôle possible relativement aux décisions qui pourraient émaner de ces technologies, qui nous échapperaient en laissant peu de place à une remise en question du raisonnement qui en est à la source (CDT 2017).

La transparence algorithmique peut alors référer à un mode de gouvernance qui prédispose à la confiance, à l'imputabilité et à l'autonomie voire à la prise de décision juste et éthique (Lipton 2016; Ananny et Crawford 2018; Selbst et Barocas 2018). Certains voient cependant dans ces impératifs de transparence des attentes qui dépassent les capacités des modèles techniques (Lipton 2016), et qui remettent en question l'applicabilité réelle du principe de transparence des systèmes pour mener à leur compréhension et afin de les gouverner (Ananny et Crawford 2018). Répondre au défi de la transparence ne représente pas, en effet, une solution *per se*, l'obligation redditionnelle ne devant pas tant concerner le fonctionnement des modèles que leur adéquation aux usages pour

lesquels ils sont destinés (Selbst et Barocas 2018). L'importance résiderait alors plutôt dans le fait de justifier la pertinence ou la robustesse de leur utilisation. Dans cette perspective il est donc nécessaire de ne pas laisser cette visée de transparence limiter les bénéfices potentiels des méthodes analytiques, notamment en santé : « the short term goal of building trust with doctors by developing transparent models might clash with the longer-term goal of improving health care » (Lipton 2016 p. 7).

Ainsi, l'opacité des réseaux de neurones peut être considéré comme l'un des trois principaux enjeux techniques inhérents à l'utilisation de systèmes d'IA, qui demandent de tempérer les bénéfices potentiels que l'IA et les données massives pourraient apporter dans le contexte de la santé, invitant ainsi à une certaine prudence. Ces différentes barrières (valeur et fiabilité des modèles, partage des données et des codes ou opacité des réseaux de neurones) sont d'autant plus importantes à considérer qu'elles sont à la source de potentiels enjeux éthiques préoccupants.

2. Les principaux enjeux éthiques de l'utilisation des systèmes d'intelligence artificielle en santé

2.1. Protection de la vie privée et de la confidentialité

La collecte et le partage d'une grande quantité de données (personnelles, sinon sensibles) font du risque d'entrave à la confidentialité[64] et à la vie privée une des préoccupations éthiques majeures de l'innovation numérique en santé (Azencott 2018; Iyengar, Kundu, et Pallis 2018; Devillier 2017a; Hager et al. 2019; Christen et al. 2016). C'est en effet un des enjeux éthiques les plus discutés dans la littérature sur le sujet (voir par exemple : Mittelstadt and Floridi 2016b et Stahl and Wright 2018). Cet enjeu, tant relatif à l'avènement de l'IA qu'aux données massives, est souvent centré sur la protection des données personnelles (IEEE 2017; Stahl et Wright 2018; Mittelstadt et Floridi 2016b).

[64] Les enjeux de protection de la vie privée sont souvent discutés dans le contexte biomédical en termes de protection de la confidentialité, soit la non-divulgation d'informations privées dans le contexte d'une relation professionnelle ou contractuelle (B. D. Mittelstadt et Floridi 2016b).

Si, dans le domaine biomédical, cette préoccupation n'est pas nouvelle – tant du point de vue de l'éthique de la recherche que de l'éthique clinique – le risque d'atteinte à la confidentialité semble accru dans le contexte du développement des données massives en santé, comme le décrivent Lipworth et al. (2017) :

> The size, complexity, dispersion, linkage, long-term storage, automaticity, globalization, and commercialization of big data are all seen to both amplify the need to protect confidentiality and add complexity to efforts to do so. The heterogeneity and networking of big data and global moves towards open data blur conceptual distinctions such as those between health-related data and non-health-related data, between personal and nonpersonal data, between individual and group-level privacy, and between primary and secondary uses of data—distinctions that often form the basis of existing confidentiality (and consent) rules (p. 491- 492).

Plusieurs considérations spécifiques en ce qui a trait à la confidentialité et à la vie privée des patients, des participants à la recherche, voire des professionnels de la santé sont à considérer dans le cadre de l'utilisation de systèmes d'IA (Iyengar, Kundu, et Pallis 2018; Lipworth et al. 2017). Des données plus personnelles et reflétant des informations plus détaillées qu'auparavant peuvent aujourd'hui être collectées et analysées alors que les bases de données, conçues pour conserver ces données à perpétuité, rendent inopérantes les limitations temporelles traditionnelles imposées (Mittelstadt et Floridi 2016b).

L'usage croissant d'algorithmes et de données massives conduit plusieurs auteurs à reconnaître qu'aujourd'hui l'anonymisation, souvent considérée comme le minimum requis pour protéger la vie privée (Mittelstadt et Floridi 2016b), n'est plus suffisante voire n'est plus possible (Azencott 2018; Mittelstadt and Floridi 2016; Rumbold and Pierscionek 2017; Rial-Sebbag 2017). Si supprimer les informations identificatoires ou autres indicateurs uniques comme le nom ne suffit plus, c'est notamment parce-que le couplage avec des données « auxiliaires », rendu possible par le recours à certains algorithmes, permet la ré-identification. En d'autres mots, les bases de données jusqu'ici considérées comme anonymisées perdent leur caractère non-identificatoire car il est possible de les coupler à d'autres données (Azencott 2018; Rumbold and Pierscionek 2017; boyd and Crawford 2012; Lipworth et al. 2017; Rial-Sebbag 2017).

Cependant, alors que différentes études se penchent sur des techniques qui permettraient de minimiser les risques d'identification, elles se confrontent toujours à une perte d'informations essentielles (Rumbold and Pierscionek 2017; Azencott 2018; Lipworth et al. 2017). La protection de la vie privée et de la sécurité des données se heurte alors à l'utilité des systèmes et la pertinence de leurs analyses tant pour la recherche que pour les soins. En effet, plus on tend vers une limitation de la ré-identification, moins les données sont utiles (boyd et Crawford 2012; Rumbold et Pierscionek 2017; B. D. Mittelstadt et Floridi 2016a; Iyengar, Kundu, et Pallis 2018). Par exemple, dans le cadre de la recherche en génomique à large échelle utilisant des biobanques, le couplage des données génétiques avec des données ethnographiques et géographiques peut s'avérer particulièrement pertinent, mais augmente, par le fait même, le risque de ré-identification (Goodman 2016). Plusieurs mettent en évidence la nécessité de trouver des méthodes qui permettraient de préserver (voire maximiser) l'utilité des systèmes tout en garantissant la protection d'informations sur les individus qui risqueraient de porter atteinte à leur vie privée (Hager et al. 2019; Azencott 2018). Ceci est particulièrement important dans le domaine de la médecine, où l'avènement de l'IA et son efficacité sont hautement dépendants de la qualité des données et de la possibilité de suivre le patient dans son parcours de soins (Villani 2018).

Il en découle que certains perçoivent cet impératif de protection de la vie privée comme une barrière au progrès et à l'innovation (Hager et al. 2019; Spiekermann, Korunovska, et Langheinrich 2018) et remettent en question ce devoir de protection, qui n'aurait jamais été absolu, en particulier en santé :

> Breaching confidentiality is mandated for imminent threats, certain contagious diseases, and other public health purposes; privacy is regularly 'violated' in the name of health business activities, on behalf of those who lack legal capacity, and for certain research purposes (Fiore and Goodman 2016, p. 85).

Certains objectifs visés par la réutilisation de données ont justifié par le passé une telle entorse à la confidentialité, comme dans le cas de l'identification de victimes lors d'attentats ou de catastrophes naturelles (O'Doherty et al. 2016). Ainsi, selon ces perspectives, la protection de la vie privée serait à balancer avec les intérêts collectifs du plus grand nombre dans une visée de bien commun. C'est

par exemple ce que défendent Fiore et Goodman (2016) dans le contexte de la médecine de précision :

> In the research context, it has been argued that privacy and confidentiality as they have been understood should perhaps be replaced by some sort of 'responsible use' doctrine. Results from clinical sequencing have value for the patients for whom the sequencing was ordered; there are also related others for whom that data is also valuable, even vital, and, as part of a large data set, it can lead to better treatments and health for many others (p. 86).

Ou encore Goodman (2016) dans le cadre de la recherche utilisant des données génétiques, qui défend que les préoccupations liées à la protection de la confidentialité des données sont trop centrées sur de potentiels torts issus d'utilisation de toutes façons non-autorisées, au détriment des bénéfices que pourraient représenter les découvertes qui en découlent.

Spiekermann, Korunovska, and Langheinrich (2018) décrivent pour leur part une certaine ambivalence vécue par les ingénieurs[65] associée à la valeur de vie privée (*privacy*) : certains défendent que cette valeur est dépassée à une époque où l'informatique est omniprésente, où les individus partagent beaucoup de données sur les réseaux sociaux, et où plus de données semblent promettre plus de connaissance – d'autant que, d'un point de vue technique, la protection de la vie privée implique un coût en temps et en argent qui pourrait saper la profitabilité des modèles commerciaux qui reposent sur la vente de données personnelles (Spiekermann, Korunovska, and Langheinrich 2018).

D'un autre côté, certains défendent que le respect de la vie privée est un droit fondamental essentiel pour équilibrer les pouvoirs dans les démocraties fonctionnelles, en particulier face à celui des entreprises (Spiekermann, Korunovska, et Langheinrich 2018). C'est par exemple ce que défend Azencott (2018) dans le cadre de la médecine de précision :

> In an era where data is sometimes touted as the new oil, there is a growing need to keep personal data private without impeding the technological, scientific and societal advances that will come from analyzing large data sets (p. 2).

[65] Terme utilisé par les auteurs qui inclut (mais ne se limite pas) aux développeurs de systèmes d'IA.

Intégrer la protection des données dans la conception des systèmes présente pour d'autres auteurs des avantages commerciaux et réduit la responsabilité et les risques des entreprises sans nécessairement nuire à la sécurité ou à l'innovation (Spiekermann, Korunovska, et Langheinrich 2018). Dans un contexte de santé, ce serait d'autant plus le cas des données génétiques, qui bénéficieraient d'un certain « exceptionnalisme » de par leur nature (Goodman 2016) le génome pouvant être considéré comme « l'identifiant ultime » (Azencott 2018). Azencott (2018) présente trois principales raisons de faire particulièrement attention aux enjeux relatifs à la protection de la vie privée dans le cadre de l'utilisation de données génomiques : 1) notre compréhension de ces données ne cessant de croître, on ne sait toujours pas exactement ce que le génome peut révéler sur un individu; 2) le génome contient également des informations sur d'autres personnes que celle à qui il appartient (par exemple, la famille) et 3) ces données ne changent jamais. Trouver un équilibre entre confidentialité, vie privée et utilité semble ainsi essentiel afin de préserver la confiance du public en la médecine (Lipworth et al. 2017) et limiter la perception négative des industries qui pourrait en découler (IEEE 2017).

Également, la « portabilité »[66] des systèmes de collecte amène à reconsidérer la vie privée (Villani 2018) même pour des données qui pourraient paraitre de nature moins sensible. À l'heure de la santé connectée, la collecte de données originellement organisée en silos est remplacée par une collecte horizontale, ubiquitaire et relativement « invasive », sortant des espaces traditionnels pour intégrer des lieux (ex. le domicile) qui ont jusqu'ici bénéficié d'une protection forte (IEEE 2017; Villani 2018). Dans le contexte biomédical, les enjeux relatifs au respect de la vie privée sont parfois même discutés en terme d'intrusion (même lorsque les données sont anonymes), en particulier pour les données générées de manière « participative » en provenance d'internet et des médias sociaux (Mittelstadt et Floridi 2016b). Ces préoccupations sont également présentes lorsqu'il s'agit du développement de robots sociaux (de soins) capables d'acquisition autonome de connaissances (Devillers 2017). La collecte ubiquitaire et la mutualisation des données amènent alors à une potentielle atteinte à la vie privée car elle augmente le nombre et le types d'acteurs qui peuvent avoir accès aux données de santé. En effet, comme mentionné précédemment, la variabilité

[66] La « portabilité » fait ici références à la mobilité des systèmes de collecte de données de santé dans le contexte de l'innovation numérique, en particulier les téléphones intelligents et les capteurs, dont les caractéristiques physiques permettent une collecte en tout temps et depuis différents lieux (et pas seulement des lieux traditionnels de soins).

des sources en ce qui a trait aux données relatives à la santé introduit de nouveaux acteurs dans le parcours de soins, comme les GAFAM, qui soulève différentes préoccupations relatives à la façon dont la vie privée sera protégée considérant que ces acteurs ne répondent pas aux mêmes cadres déontologiques que ceux qui s'appliquent traditionnellement à la santé. C'est par exemple ce que soulève Sharon (2016) dans le contexte de la santé mobile qui, en offrant de nouveaux moyens de collecte de données multidimensionnelles, permet à de nouvelles compagnies d'entrer dans l'espace de la recherche en santé, s'accompagnant ainsi de nouveaux enjeux éthiques. Or, selon Lahlou, le véritable problème ne réside pas dans le partage de données qui revêtent un caractère privé par nature, mais dans qui y accède (Lahlou 2008). Par exemple, que des professionnels de santé aient accès à nos données biomédicales ne représente pas une atteinte à la vie privée, mais devient problématique seulement lorsque ces données sont accessibles à d'autres acteurs.

En conséquence de cette collecte de plus en plus ubiquitaire naît le développement d'une relation de soins de plus en plus intrusive qui demande de protéger l'intimité des patients – au-delà de la simple vie privée. Ce nouveau mode de collecte offre en effet la possibilité d'avoir accès à des données jusqu'alors inaccessibles et hautement personnelles, comme l'illustre ici Sharon (2016) :

> The ResearchKit mPower Parkinson's app, for example, uses sensors to track tremors, balance and gait, certain vocal characteristics and memory, combined with surveys and tasks to be completed before and after medication. Dr Ray Dorsey, the neurology expert for the study, has said, "We're seeing how people do on Saturdays and Sundays, how people do between 5 pm and 8 pm… Heretofore we'd had no good way of measuring that" [23]. "At the individual level," a Stanford enthusiast has explained, "we just haven't had any data like this at all… It's simply not possible to detect it until you have mobile devices gathered." (p. 565).

C'est également le cas dans le cadre du développement des robots sociaux, comme le discute Devillers (2017), qui stipule qu'il est nécessaire de réfléchir à des règles éthiques relatives à l'intimité qu'il serait possible de développer avec les machines :

> Peu importe que le robot soit une machine qui simule, si elle sait s'adapter à nous, nous comprendre d'une certaine manière, nous créerons une intimité avec elle (p. 149).

Le respect d'un principe d'intimité est d'ailleurs défendu dans plusieurs rapports portant sur l'éthique de l'IA (Déclaration de Montréal IA Responsable 2018; IEEE 2017) qui demande d'assurer la protection « d'espaces d'intimité dans lesquels les personnes ne sont pas soumises à

une surveillance, ou à une évaluation numérique » mais également protéger de l'intrusion « les pensées intimes, les émotions ou l'identité » (Déclaration de Montréal IA Responsable 2018).

Enfin, cette possibilité pour les individus de partager directement des données de santé (notamment *via* les applications mobiles et les capteurs) demande de veiller à leur capacitation face au partage et à la gestion de leurs données soit, par exemple, d'accompagner les patients dans le partage de celles-ci et d'évaluer l'impact d'un éventuel bris de confidentialité (Villani 2018). En effet, la collecte horizontale complique le contrôle et la détermination des rôles et responsabilités dans la gestion de la confidentialité (IEEE 2017).

2.2. Repenser le consentement des patients et des participants à la recherche

L'avènement de l'utilisation de systèmes d'IA et de l'analyse de données massives en santé pose également des problèmes relatifs au respect du consentement des patients et des participants à la recherche. Le consentement en santé est un des modèles de transparence formelle qui permet d'autoriser l'utilisation des données pour la recherche de manière appropriée, permettant notamment de préciser quelles sont les intentions cliniques ou quelles informations le patient souhaite partager et recevoir (ce qui est particulièrement pertinent dans le cas de découvertes fortuites, comme par exemple sur les données génomiques) (Fiore et Goodman 2016). L'obtention du consentement représente un des principaux modèles pour limiter les préjudices associés aux entraves à la vie privée. Selon la conception classique du consentement, les individus consentent à participer à une étude considérant l'équilibre entre les bénéfices et les risques, assistés par des professionnels de santé informés, en amont de la recherche ou de la prestation de soins et pour une intervention, un projet ou une juridiction bien définis (Mittelstadt and Floridi 2016a; ÉPTC2 2018; Mittelstadt and Floridi 2016b; CERNA 2018; Woolley 2016).

Cependant, l'avènement d'une médecine guidée par les données complique le respect des termes du consentement dans sa compréhension traditionnelle (Woolley 2016; Christen et al. 2016). En premier lieu, parce-que le contexte de la collecte, du stockage et de l'analyse des données massives rend souvent impossible l'obtention d'un consentement libre et éclairé relativement aux utilisations secondaires desdites données (Rumbold et Pierscionek 2017; Woolley 2016; Christen

et al. 2016). En effet, il devient impossible de prévoir à l'avance les usages qui seront fait des données collectées, comme le présente Mittelstadt et Floridi (2016b):

> Secondary effects of pharmaceuticals can be identified by comparing data not only from multiple clinical trials, but 'informal sources' as well, such as incidental self-reporting via social media and search engine queries. In this type of research the connections that can be revealed through linking multiple data sets cannot be accurately predicted prior to carrying out the research. As a result, 'consent' cannot be 'informed' in the sense that data subjects cannot be told about future uses and consequences of their data, which are unknowable at the time the data is collected or aggregated (p. 454).

Un défi en ce qui concerne les données massives en santé est en effet la réutilisation infinie que leur collecte et leur stockage permettent. Considérant les impératifs de partage, il devient presque impossible de prédire à l'avance les finalités de leur usage (Rial-Sebbag 2017). Cette possibilité infinie de réutilisation tend ainsi à transformer la collecte dans le cadre de la recherche biomédicale : les données ne sont plus recueillies dans le contexte spécifique des objectifs de recherche mais ce sont les ensembles de données massives déjà collectées qui offrent de multiples opportunités de recherche (Rumbold et Pierscionek 2017; Brouard 2017; Zwitter 2014).

Tel que mentionné précédemment, la portabilité associée à la santé connectée et la possibilité pour les individus de partager directement des données de santé conduisent à l'apparition d'une collecte horizontale qui complique le contrôle et les rôles de la gestion des données (IEEE 2017; Villani 2018). Même lorsque les individus consentent à la collecte des données *via* leurs objets connectés, ces services dépendent de paramètres évolutifs codés pour des finalités qui ne sont pas toujours explicites, les concepteurs eux-mêmes peuvent sous-estimer l'impact de leurs applications sur l'environnement digital global (CERNA 2018). Il est alors souvent difficile pour les utilisateurs de déterminer quelles informations sont collectées, de modifier ou de gérer ces données, voire de connaître l'étendue avec laquelle leurs données vont être publiquement diffusées et analysées, en dehors des espaces ou elles ont été générées (IEEE 2017; Mittelstadt et Floridi 2016b).

Si le modèle traditionnel de consentement ne fonctionne plus, c'est également en raison des opportunités qui existent de mettre en lien les ensembles de données médicales et non-médicales (Mittelstadt et Floridi 2016a). Les données ne sont plus forcément collectées dans un contexte biomédical (où le consentement reste un standard) mais proviennent de médias sociaux ou

d'applications mobiles où le consentement explicite est souvent absent (Mittelstadt et Floridi 2016b). Lorsqu'il est présent, le consentement digital (souvent un « click ») peut également mettre à mal l'autonomie des individus, considérant par exemple la surcharge d'informations des clauses de confidentialité qui ne sont finalement plus informatives – voire qui ne sont même pas lues (Jones, Kaufman, et Edenberg 2018). Le consentement dans ce contexte peut d'autant plus être questionné si l'on considère les algorithmes « invisibles » pour les utilisateurs qui peuvent aujourd'hui accéder aux données longtemps après qu'elles aient été fournies (IEEE 2017).

L'opacité des réseaux de neurones représente également une potentielle entrave à la conception traditionnelle du consentement et donc à leur autonomie, car il devient difficile d'expliquer pourquoi une décision a été prise ou d'informer les patients et les participants en amont de ce qui sera découvert par le modèle prédictif. En effet, les réseaux de neurones peuvent fonder leurs décisions ou recommandations sur la base de paramètres appris implicitement auxquels nous n'avons pas forcément accès (Castelvecchi 2016). Ceci peut avoir des conséquences face aux choix de santé des individus et des professionnels, par exemple dans le cadre de la mise en place de traitements préventifs lourds où la décision peut être encore plus difficile à prendre en l'absence de connaissance des facteurs de risques (Castelvecchi 2016). L'apparition d'une forme de paternalisme technologique est également à craindre considérant que les algorithmes peuvent conseiller, recommander et influencer les comportements de santé (CERNA 2018). Par extension, il est possible de questionner la réelle liberté de choix des individus à consentir à l'usage de l'IA et des données massives dans un système de santé où ces technologies tendent à devenir ubiquitaires.

Des préoccupations relatives à la nature véritablement éclairée du consentement apparaissent également. Ces préoccupations concernent la capacité des utilisateurs à comprendre le potentiel des données massives et de l'IA et de prendre des décisions informées, notamment sur la qualité publique des données qu'ils génèrent et des conséquences non-envisagées de leurs utilisations (Floridi et Taddeo 2016; Zwitter 2014), mais surtout leur capacité à comprendre le

fonctionnement des systèmes utilisés, dans un contexte où il est reconnu qu'améliorer le niveau de littératie numérique[67] est essentiel (Déclaration de Montréal IA Responsable 2018).

S'observe alors également, dans le contexte du respect du consentement, une tension entre la protection des intérêts individuels (ici, l'autonomie des patients et des participants) et des intérêts collectifs (ceux de la recherche en santé et par extension, des avancées de la médecine) :

> A basic tension currently exists between the regulation of biomedical information and the goals of research. On the one hand is the protection of basic rights of data providers and fears of abuse by data users. On the other is the promised greater societal good of research into the aetiology of disease (Woolley 2016 p.173).

L'apprentissage automatique tend en effet à détourner le consentement de l'utilisation individuelle de données personnelles vers un consentement collectif permettant d'utiliser ces systèmes (CERNA 2018). Une restriction trop importante relativement au consentement requis des utilisateurs dans ce contexte est parfois vue comme une barrière financière et bureaucratique au partage des données et à la recherche (Mittelstadt et Floridi 2016b; Wellcome Trust 2013). Il se pourrait, face à des restrictions trop importantes, que ces données ne soient pas partagées entre chercheurs et ce même si les individus auraient potentiellement consenti à le faire (considérant qu'il est difficile de savoir s'ils ont réellement consenti ou non dans les situations décrites) risquant ainsi de limiter l'avancée des connaissances (Mittelstadt et Floridi 2016b). Selon ces perspectives, éliminer ou tempérer le besoin de consentement se fait de manière pragmatique (relevant de l'altruisme) ou substantielle (relevant de la solidarité ou du bien commun) et peut se justifier si les bienfaits sociétaux dépassent les risques d'atteintes aux droits individuels (notamment, ici, l'autonomie) (Mittelstadt et Floridi 2016b; Wellcome Trust 2013).

[67] La littératie numérique correspond à la capacité de comprendre et d'utiliser l'information issue des outils numériques ou des technologies en réseaux (UNESCO 2018a). Assurer que l'ensemble de la population acquiert les compétences à la fois techniques et critiques est essentiel afin de garantir que « tout individu puisse agir de façon autonome, éclairée et responsable », par l'entremise du système formel de formation ou hors de ce dernier (Déclaration de Montréal IA Responsable 2018). Ces compétences sont incontournables lorsqu'il est question de consentement : « La littératie numérique ne se résume donc pas seulement au fait de savoir utiliser des outils technologiques, elle inclut également une dimension critique amenant à savoir prendre des décisions éclairées quant à cette utilisation. » (Vézy C. dans Déclaration de Montréal IA Responsable 2018 p. 272).

Néanmoins, un consentement valide (soit libre et éclairé) est considéré comme fondamental en ce qui a trait au respect de l'autonomie des patients et des participants à la recherche et ce depuis les débuts de la bioéthique (Jones, Kaufman, et Edenberg 2018). La notion de consentement informé est au centre des considérations éthiques depuis le procès de Nuremberg afin de protéger les participants à la recherche médicale contre d'éventuels préjudices, volontaires ou non (Christen et al. 2016). Pour certains, protéger l'avancée du progrès et le bien commun au détriment de l'autonomie individuelle ne peut être éthiquement justifié sans débat public, doit tenir compte de la manière dont la recherche sur les données massives risque de porter atteinte aux individus et ne peut se faire au nom d'un paternalisme médical qui limiterait les options de choix des individus qui ne comprendraient pas réellement la portée des projets (Mittelstadt et Floridi 2016b; Christen et al. 2016).

Ainsi, la nécessité de repenser le consentement – notamment dans le contexte de la recherche et de la médecine – face à l'IA et aux données massives est soutenue par plusieurs auteurs (Mittelstadt et Floridi 2016b; 2016a; Jones, Kaufman, et Edenberg 2018; Christen et al. 2016; Woolley 2016). En lieu et place d'un « consentement à usage unique » aujourd'hui pratiquement impossible à renouveler pour chaque utilisation (Mittelstadt et Floridi 2016b) la mise en place de consentements « ouverts » (ex. un patient atteint d'un Alzheimer précoce peut consentir à donner ses données pour toute recherche qui fait avancer les connaissances sur cette maladie) (Christen et al. 2016) ou « dynamiques » (ex. une interface interactive qui permettrait aux participants de choisir et de changer leur consentement en temps réel) (Woolley 2016; Villani 2018; Jones, Kaufman, et Edenberg 2018) sont actuellement en discussion. La nécessité d'accompagner les individus et de favoriser leur autonomie face au partage de leurs données devient alors une urgente nécessité (Villani 2018; Christen et al. 2016).

2.3. Différentes préoccupations relatives à la justice sociale

L'avènement des systèmes d'IA en santé soulève également des préoccupations relatives à la justice sociale, tant concernant l'accès aux technologies que les biais potentiels que pourraient contenir les données et les algorithmes. Ces préoccupations se manifestent surtout sous la forme de la crainte d'un risque de discrimination injustifiée qui pourrait en découler. Qu'il s'agisse d'accès aux technologies ou de leurs biais potentiels, l'innovation numérique en santé (incluant

mais ne se limitant pas à l'IA et aux données massives) est soit perçue comme une source d'iniquité, soit comme un moyen de remédier aux inégalités existantes, selon l'utilisation qui en est faite.

L'innovation numérique en santé peut être perçue comme une opportunité d'augmenter l'accès aux soins – notamment, par le biais de l'utilisation des systèmes d'IA. L'organisation mondiale de la santé (OMS) voit en effet dans les données massives, l'apprentissage automatique, la télémédecine ou la santé mobile, un moyen de garantir un accès aux soins universels : « eHealth is now an integral part of delivering improvements in health » (Global observatory for eHealth 2016 p. 7). Ces technologies permettraient de distribuer largement les services de santé dans les zones aux accès restreints, en utilisant par exemple les téléphones intelligents : « mHealth has emerged rapidly in developing countries as a result of the large penetration of mobile phones and the lack of other, modern health infrastructure » (OMS 2012 p.79).

Ayant réalisé un sondage auprès de ses pays membres, l'OMS a mis en évidence que plus de la moitié d'entre eux possèdent déjà une stratégie de e-santé, et 90% de ces stratégies réfèrent à une couverture de santé universelle. 80% des pays à faibles revenus ont rapporté avoir au moins un programme de santé mobile (contre 91% des pays aux revenus élevés). Cependant, seulement 14% des pays sondés possédaient en 2016 une évaluation gouvernementale de ces programmes, et différentes barrières seraient encore à surmonter, comme par exemple le besoin de former les professionnels de santé, de mettre en place des modes de gouvernance appropriés ou d'assurer un financement adéquat (Global observatory for eHealth 2016).

D'autres s'inquiètent au contraire que l'innovation numérique devienne une source d'iniquité relativement à la distribution des soins, et qu'une partie de la population demeure exclue d'un système de santé connecté (Déclaration de Montréal IA Responsable 2018). Le Global Network of Internet and Society Centers (NoC) ayant sondé les participants au symposium « AI and Inclusion » de 2017 sur les principaux défis de la création d'une société plus inclusive avec l'IA, a mis en évidence que les participants issus des les pays à faibles revenus ou à revenus

intermédiaires[68] sont particulièrement inquiets au regard de l'accès et du coût des technologies relatives à l'IA (NoC 2017). Ils sont également plus préoccupés par les inégalités préexistantes, le manque de connaissances et de compétences potentielles que les pays aux revenus élevés[69]. La distribution équitable parmi les régions du monde des technologies d'IA ne semble donc pas garantie, et le développement de la e-santé comme solution à l'accès aux soins risque de demander des efforts d'implémentation.

À cet effet, Ganascia s'inquiète de l'apparition d'une éventuelle fracture sociale au sein même des sociétés où l'accès à ces technologies serait restreint aux personnes ayant des moyens financiers élevés si leur coût ne permet pas une couverture par les assurances sociales de santé, ce qui conduirait à la nécessité d'un arbitrage (Ganascia 2018). Enfin, l'utilisation de systèmes d'IA en santé expose également au risque d'exclusion de certaines catégories de la population qui seraient moins enclines à la connectivité ou qui vivent « en marge du *Big data* » (ex. les personnes âgées ou certains groupes socio-économiquement désavantagés) (Kim 2016; Campolo et al. 2017; Déclaration de Montréal IA Responsable 2018; B. D. Mittelstadt et Floridi 2016b).

Les enjeux relatifs à l'équité sont également largement discutés en termes de biais[70] potentiels que les systèmes d'intelligence artificielle pourraient perpétuer (Villani 2018; Lipworth et al. 2017; Charlet 2018; Campolo et al. 2017; Barocas et Selbst 2016; Goodman 2016). D'un côté, le recours aux données massives et aux algorithmes peut être considéré comme plus efficace en termes de représentativité que les études randomisées traditionnelles – notamment, car ces analyses sont plus rapides, moins coûteuses, et en apparence plus neutres que celles découlant de décisions humaines qui peuvent être influencées par les biais implicites ou explicites inhérents à tout individu (Lipworth et al. 2017; Barocas et Selbst 2016; Kim 2016). D'un autre côté, le développement des décisions automatisées conduit également à l'apparition de préoccupations relatives à l'équité considérant que les machines ne sont pas dépourvues, elles-mêmes, de biais

[68] Désignés par le terme *Global South*.
[69] Désignés par le terme *Global North*. Ces derniers étaient quant à eux plus préoccupés par les défis relatifs aux biais et à la discrimination, à la confiance, à l'explicabilité, à la transparence ou à la responsabilité.
[70] Pour une typologie des différents types de biais des analyses de *data mining*, voir Kim (2016), bien qu'elle soit discutée dans le cadre de la discrimination à l'embauche.

éventuels (Friedler, Scheidegger, et Venkatasubramanian 2016; Russell, Dewey, et Tegmark 2015), pouvant mener à des disparités disproportionnées relatives à l'origine ethnique, au genre ou à la classe sociale (Russell, Dewey, et Tegmark 2015). Selon Barocas et Selbst (2016), deux situations peuvent être à l'origine d'un impact discriminatoire des décisions automatisées : 1) la présence d'erreurs dans les données ou leur exploration et 2) des données trop précises – soit des systèmes qui modélisent avec trop de précision les inégalités. Les limites interprétatives et informationnelles de l'analyse des données massives ne sont donc pas sans conséquence sur le respect de la justice sociale.

Relativement à la première situation (la présence d'erreurs dans les données ou leur exploration), les modèles d'apprentissage automatique repose en effet généralement sur l'hypothèse que les échantillons sont représentatifs de la population et que les caractéristiques de celle-ci vont demeurer les mêmes lors de l'application du modèle (Goodman 2016; Charlet 2018). Certains ensembles de données souffrent cependant de biais de sélection ou d'échantillonnage marqué (Lipworth et al. 2017; Villani 2018; Goodman 2016; Charlet 2018). Il se peut que certains groupes soient sous-représentés ou surreprésentés dans l'échantillon (Goodman 2016). Il est également possible pour les analystes de faire différentes erreurs qui pourraient conduire à un impact discriminatoire – par exemple, choisir une variable particulièrement corrélée à une catégorie protégée, ou omettre des variables qui pourraient éviter l'effet discriminatoire (Barocas et Selbst 2016; Kim 2016).

Les modèles n'ont cependant pas besoin d'inclure des erreurs pour être à la source de conséquences discriminatoires. Ceci renvoie à la deuxième situation (des données trop précises – soit des systèmes qui modélisent avec trop de précision les inégalités). Certains modèles s'appuient sur des statistiques significatives mais objectivement discriminantes qui sont contenues dans des ensembles de données où les groupes historiquement marginalisés peuvent être exclus ou sous-représentés, conduisant à un « impact disparate » (Barocas et Selbst 2016; CDT 2017). Les données existantes reflètent souvent un historique de discrimination et de biais implicites relatifs aux préjudices passés (CDT 2017; Villani 2018; Barocas et Selbst 2016). Par exemple, les données issues des essais cliniques traditionnels sont susceptibles de surreprésenter les personnes âgées ou

sous-représenter certains groupes en fonction de leurs origines ou de leur genre (Charlet 2018; Campolo et al. 2017) :

> Worryingly, data sets used to train health-related AI often rely on clinical trial data, which are historically skewed toward white men, even when the health conditions studied primarily affect people of color or women. Even without AI amplifying such biases, African Americans with sickle cell anemia are overdiagnosed and unnecessarily treated for diabetes based on insights from studies that excluded them (Campolo et al. 2017, p 19).

Les algorithmes non discriminatoires peuvent conduire à des déséquilibres dans la distribution des risques et des bénéfices considérant les conséquences des décisions qui en émanent, le résultat de leur analyse ne dépendant pas seulement de l'algorithme lui-même mais aussi d'une variété de facteurs empiriques tels que la distribution des groupes dans la population, le type de classification faite ou les coûts et bénéfices escomptés de chaque décision (Altman, Wood, et Vayena 2018).

Quelle que soit la situation, ces biais sont perpétués, voire exacerbés considérant l'échelle, la portée et la systématisation des décisions algorithmiques, par les algorithmes qui apprennent de ces ensembles de données et émettent des prédictions en fonction de celles-ci (Charlet 2018; CDT 2017; Goodman 2016; Villani 2018; Barocas et Selbst 2016). La perception d'injustice peut être particulièrement marquée considérant l'opacité des réseaux de neurones : une décision automatique et difficile à expliquer peut en effet être plus facilement considérée comme arbitraire (Zarsky 2016). L'impact disparate de ces décisions n'est pas seulement à considérer du point de vue des individus, mais peut également conduire à la discrimination ou la stigmatisation de groupes, par exemple lorsque les patients ou les participants sont regroupés en fonction de caractéristiques géographiques, ethniques ou socio-économiques (Mittelstadt et Floridi 2016b). Dans ces conditions, le préjudice peut affecter un groupe entier, même si les données sont anonymisées et si les participants ont donné leur consentement à titre individuel (Mittelstadt et Floridi 2016b).

Dans le contexte de la médecine de précision ou des biobanques, le risque de discriminations génétiques est emblématique (Azencott 2018; Fiore and Goodman 2016). La situation pourrait s'avérer préoccupante dans le cadre de discriminations d'assurances fondées sur la prédisposition (génétique) aux maladies (Mittelstadt et Floridi 2016b). Pour Charlet, c'est l'aspect personnalisation de ces technologies qui pourraient exacerber les discriminations

arbitraires, par exemple, si les assurances venaient à augmenter leurs franchises pour les personnes à risque (ex. les fumeurs) remettant ainsi en cause les fondements de la mutualisation (Charlet 2018).

Les critères précis qui définissent ce que doit faire un algorithme équitable n'ont toutefois pas été très explorés, et les mesures proposées pour mesurer l'équité semblent insuffisantes ou incompatibles (Friedler, Scheidegger, et Venkatasubramanian 2016; Binns 2018). Limiter l'impact disparate des analyses de données massives se heurte à différentes difficultés, qu'elles soient internes (ex. identifier les variables cibles, identifier les données biaisées ou étiqueter les exemples de manière appropriés) ou externes (ex. répondre adéquatement aux contraintes politiques et constitutionnelles) comme le reconnaissent Barocas and Selbst (2016) :

> Policies that compel institutions to correct tainted datasets or biased samples will make impossible demands of analysts. In most cases, they will not be able to determine what the objective determination should have been or independently observe the makeup of the entire population (p. 722).

Si la discrimination est presque toujours une propriété émergente non voulue de l'utilisation de l'algorithme plutôt qu'un choix conscient de la part de ses programmeurs (Barocas et Selbst 2016), il n'en demeure pas moins que leurs choix vont avoir un impact normatif – ils doivent par exemple choisir quels types d'erreurs sont les plus importants ou quels groupes doivent être classifiés plus précisément que d'autres pour atteindre une répartition équitable des bénéfices (Altman, Wood, et Vayena 2018), comme c'est d'ailleurs déjà le cas des professionnels de santé dans leur pratique. Dans certaines circonstances, il serait même plus judicieux d'intégrer les variables sensibles (ex. genre, âge, origine ethnique) dans l'analyse – même si le fait de les collecter augmente le risque de discrimination potentielle – afin de tempérer leur effet discriminatoire ou de connaitre leur influence (Kim 2016; Charlet 2018). Selon comment les données vont être utilisées, l'IA pourrait alors également permettre de diminuer les biais existants (en les détectant) et d'augmenter les opportunités pour les groupes traditionnellement désavantagés (Kim 2016). La tension entre intérêts individuels et collectifs se manifeste alors également lorsqu'il est question du respect de la justice sociale, à savoir que le partage des données semble une condition essentielle à

la mutualisation en vue de garantir un accès aux données et aux bénéfices équitables, exposant cependant les individus qui les partagent aux risques d'être discriminés de manière infondée.

2.4. Déshumanisation des soins et du patient

L'utilisation de systèmes d'IA en santé s'accompagne également de préoccupations relatives à une certaine déshumanisation. Celle-ci pourrait se manifester de deux principales manières : 1) une relation de soins déshumanisante liée à la diminution du contact humain et 2) un patient déshumanisé considérant les impacts du numérique sur l'identité et l'individualité.

Premièrement, le recours croissant aux systèmes d'IA et aux données massives pourrait exacerber la déshumanisation de la relation de soins, en augmentant la distance entre les professionnels de santé et les patients, notamment car ces technologies ne requièrent peu voire pas d'intervention humaine (Nakrem et al. 2018). Ceci pourrait ainsi conduire à une réduction du contact humain dans la relation de soins (Coeckelbergh 2015). Cette distance pourrait également être exacerbée par le temps que les professionnels de santé vont devoir consacrer à la compréhension et à l'utilisation des systèmes d'IA plutôt qu'au chevet du patient[71]. Cette déshumanisation de la relation de soins s'accompagne d'un appauvrissement des aspects émotionnels et psychologiques essentiels, en plus d'une expertise médicale plus formelle, à la qualité des soins (Coeckelbergh 2015) ainsi qu'à la confiance entre patient et professionnels de santé.

La distance à la base de la déshumanisation des soins pourrait également être favorisée par un isolement des patients. Cet isolement découlerait soit de l'exclusion de personnes (vulnérables) laissées aux mains des machines, soit des individus qui pourraient s'isoler eux-mêmes en préférant interagir avec des machines (Devillers 2017; Coeckelbergh 2012, 2015). Le risque d'exclure les patients qui nécessitent un certain niveau d'assistance clinique (comme les personnes âgées) est par exemple dénoncé par Coeckelbergh (2015) :

[71] Ce point est souligné par Robert Truog (2019) dans un essai produit pour le projet « AIship » (voir : https://aiship.org/fr/projets/les-nouveaux-etats-detre/), qui n'est pas encore publié.

> In discussions about care robots for elderly care, an important concern is that the care robots will reduce human contact. (…) Nightmare scenarios are sketched in which elderly people are abandoned and left in the hands of machines, shielded from the rest of the world (p. 268).

La crainte que les individus s'excluent eux-mêmes en se cantonnant uniquement à des relations avec les machines est également soulevée par Devillers (2017), notamment car ils pourraient préférer interagir avec des robots prévisibles et développés de manière à être toujours agréables à des humains imprévisibles. Ceci serait particulièrement le cas avec les robots « empathiques » ou lorsque les frontières entre humains et systèmes d'IA se brouillent (comme par exemple dans le cas des *chatbots*) (Devillers 2017). Préférer les interactions avec les systèmes d'IA plutôt qu'avec d'autres humains serait favorisé par le potentiel engagement émotionnel vis-à-vis des machines, qui se manifeste selon différents mécanismes de projection d'intention ou d'anthropomorphisation (Devillers 2017; Iphofen et Kritikos 2019). Le recours grandissant au numérique n'étant pas sans impact sur les capacités sociales des individus, certains s'inquiètent que cela nuise à l'exercice de « capacités humaines fondamentales » - en particulier, les compétences sociales comme l'empathie, l'introspection et la compassion (Schwab 2016).

Deuxièmement, l'avènement de l'IA et des données massives contribuerait à une perception déshumanisée du patient considérant les impacts du numérique sur l'identité et l'individualité. Ce risque de déshumanisation s'inscrit dans le phénomène que Lahlou (2015) nomme le « retournement du miroir » : avec la numérisation grandissante, nos sociétés et les individus qui la composent deviennent « icodynamiques », c'est-à-dire qu'ils se définissent par leur image (numérique) qui finit par ne plus correspondre à la réalité (physique) (Lahlou 2015). Le phénomène du *soi quantifié* pourrait conduire à percevoir les patients comme des ensembles de données plutôt que des individus à part entière et dans toute leur complexité (Rouvroy 2014; Coutellec et Weil-Dubuc 2017; Ajana 2017). Ce phénomène serait favorisé par la création de profils de santé en vue d'une médecine de précision. C'est ce que dénoncent Coutellec et Weil-Dubuc, discutant du profilage des patients qui pourrait conduire à la perception du « moi » comme un « agrégat de données » : « Le profil est une représentation appauvrie de l'humain, un 'homme disloqué' puis agrégé » (Coutellec et Weil-Dubuc 2017 p. 71).

Les conséquences du profilage algorithmique – quel qu'en soit les applications – conduisent à un certain paradoxe de l'individualisation de la statistique (Ibekwe-Sanjuan 2014) qui pourrait amener à une « désubjectivation » (Rouvroy et Berns 2013) en négligeant la personne (Rouvroy 2014) par une industrialisation la personnalisation (Ibekwe-Sanjuan 2014). L'établissement de profils contournerait les sujets humains se basant sur des données insignifiantes sans jamais en appeler au sujet lui-même. Ces profils se substitueraient parfois à l'identité des individus (Rouvroy 2014), ou en tout cas participeraient à la création d'une « identité digitale » qui contribuerait à la dépersonnalisation des soins, ne reflétant pas un assemblage qu'il est possible d'assimiler au soi (Rouvroy et Berns 2013; Coutellec et Weil-Dubuc 2017). Le profilage amène alors à des préoccupations relatives au contrôle et à la compréhension qu'aurait les individus de leur identité digitale (IEEE 2017), à une nouvelle objectivation du corps des patients (CCNE 2019) qui impacterait l'image de l'humain, la place de l'individu, la dignité et le sentiment d'être unique (CNIL 2017; CCNE 2019).

Le recourt croissant aux technologies pourraient également conduire à une aliénation du personnel médical et du patient, issue d'un potentiel glissement de la perception des humains comme des objets et du soin comme d'un produit (Coeckelbergh 2015). L'automatisation des tâches pourrait même transformer la perception que les professionnels de santé ont d'eux-mêmes en automatisant leur travail (qui seraient alors réduit à celui des machines); et conduirait à des patients « *managed and processed* » de manière impersonnelle (Coeckelbergh 2015). Un glissement vers une atteinte à l'authenticité est alors envisageable, soit à ce qui constitue l'identité propre des individus (patients).

Cette transformation potentielle du rapport à soi s'inscrit cependant dans la lignée des préoccupations relatives à la biomédicalisation de la société, déjà décriée par certains auteurs avant le développement de systèmes d'IA en santé. Rose définit par exemple ce qu'il nomme les *neurochemical selves*, caractéristiques de nos sociétés psychopharmacologiques (Rose 2003). La compréhension du « soi » se fait en termes de cerveau et de corps, et a pour conséquence une

profonde transformation de l'identité individuelle (Rose 2003). Ceci fait échos aux concepts de *biosocialité* et de *biocitoyen*, qui mettent en lumière une redéfinition des identités collectives dans les sociétés occidentales contemporaines au travers des attributs génétiques, somatiques, ou physiques que les individus partagent et autour desquels ils sont mobilisés (Collin 2016). Le *soi quantifié* et les préoccupations qui s'y rattachent semble alors s'inscrire dans la même lignée, exacerbant potentiellement ces phénomènes.

La déshumanisation des soins n'est en effet pas le propre de l'IA, déjà dénoncée par exemple avec l'avènement de la médecine fondée sur les données probantes (ou *l'evidence-based medicine*), où la standardisation de la pratique médicale conduit à une déshumanisation de la relation de soin par le biais d'une médecine « procédurale » qui rationnalise la pratique médicale (Azria 2012). Dans cette situation, les modèles (statistiques) conduisant à la production de « preuves » viennent « nier l'incertitude », ce qui revient à nier la dimension humaine des soins considérant qu'il est impossible de contrôler tous les paramètres des patients (Azria 2012). L'objectivation des patients est ici à la base d'une tension dans la pratique médicale :

> La pratique médicale impose en effet la mise en relation de deux univers ; l'un est scientifique, il est celui du général et du multiple, des études sur populations, des probabilités et autres modélisations du risque. L'autre univers, restreint à l'individu, domaine d'expression de sa singularité et de sa variabilité, est celui des affects et de l'inquantifiable. La médecine se fait dans un va-et-vient constant entre l'un de l'individu et le multiple de la connaissance scientifique (Azria 2012).

L'utilisation de systèmes d'IA et l'analyse des données massives en santé ne feraient alors qu'exacerber une déshumanisation des soins qui s'observait déjà. Pour certains, le recours à des systèmes d'IA représenterait même, au contraire, une opportunité d'améliorer la relation de soins dans le contexte d'une médecine déshumanisée ou les professionnels de santé n'ont que peu de temps à consacrer aux patients (Topol 2019). En offrant la possibilité d'automatiser la plupart des tâches répétitives aujourd'hui réalisées par les professionnels de santé, l'IA dégagerait l'espace et le temps nécessaires pour créer et entretenir une connexion humaine et des échanges avec le patient (Topol 2019).

2.5. Sécurité des systèmes d'intelligence artificielle

L'utilisation de systèmes d'IA en santé soulève également des enjeux relatifs à la sécurité. La robustesse des systèmes d'IA et du stockage des données massives est essentielle afin de limiter les différentes conséquences éthiques et sociales présentées précédemment – soit, afin de favoriser la protection de la vie privée et de la confidentialité, le respect du consentement libre et éclairé, ou la protection contre les biais et la discrimination potentiels. Assurer la sécurité se heurte également aux principaux enjeux techniques de l'utilisation des systèmes d'IA, et ce de différentes manières.

Les risques concernant la sécurité des systèmes d'IA peuvent être issus de failles techniques inhérentes à ces technologies (IEEE 2017) ou d'une l'IA qui ne remplirait pas ses fonctions de manière sécuritaire ayant été mal évaluée (Mai V. dans Déclaration de Montréal IA Responsable 2018). Qu'il s'agisse de conséquences non-intentionnelles ou d'acteurs malveillants qui exploiteraient les vulnérabilités des systèmes (Brundage et al. 2018; Villani 2018), différentes failles peuvent être à l'origine d'entraves à la sécurité. Lipton (2016) alerte par exemple au fait que les interprétations *post-hoc* formulées par les algorithmes d'apprentissage profond (pour remédier à l'opacité des réseaux de neurones) sont facilement manipulables, de manière volontaire ou non. Ceci serait d'autant plus préoccupant si les explications créées sont humainement plausibles – conditions dans lesquelles il serait difficile d'identifier quand les explications sont fausses (Lipton 2016). Il existe également des risques associés à des comportements non-anticipés ou inattendus des systèmes, qui peuvent provenir de difficultés dans les choix d'architecture des réseaux de neurones, des échecs lors de l'entrainement, ou des erreurs dans l'implémentation (IEEE 2017). Il est également nécessaire de contrôler l'accessibilité aux données, tant pour les protéger de la corruption que pour empêcher de perdre l'accès à des informations à cause d'interruptions de système imprévues (Floridi et Taddeo 2016), de veiller à ce que les systèmes adaptatifs (comme les robots de soins) soit régulièrement contrôlés (traçabilité) et que les systèmes connectés ne soient pas piratables (Devillers 2017).

De nombreuses vulnérabilités non-résolues à ce jour laissent place à la possibilité de différents types « d'attaques » des systèmes d'IA comme, par exemple, l'inclusion de données faussées dans l'échantillon d'entrainement afin que le modèle apprenne des erreurs, l'exploitation

de défauts dans la conception des objectifs du système autonome ou encore les attaques antagonistes (*adversarial attacks*) (Brundage et al. 2018; Villani 2018). Les attaques par exemples antagonistes (ou *adversarial examples attacks*) correspondent à l'utilisation de données d'entrées (*inputs*) choisies pour causer une modification du résultat de l'analyse (*output*) du réseau de neurones sans que ce changement ne soit perceptible par un humain (Brown et al. 2017; Finlayson et al. 2018). Généralement, ce genre d'attaque est réalisé sur des images pour duper les réseaux de classification et les forcer à faire des erreurs (Brown et al. 2017; Finlayson et al. 2018), en modifiant un peu chaque pixel, ou un nombre défini de pixels – modification communément appelé « *patch* » (Brown et al. 2017). Ce genre d'attaques peut cependant également être dirigé contre des modèles qui utilisent d'autres types de données que des images, comme le NLP (Finlayson et al. 2018). Ces attaques ne fonctionnent pas seulement dans des conditions de laboratoire mais également dans le monde réel (Brown et al. 2017; Finlayson et al. 2018), et peuvent être partagées sur internet (Brown et al. 2017).

Finlayson et al. (2018) mettent en évidence que le domaine médical est particulièrement vulnérable aux attaques antagonistes, pour différentes raisons : la présence d'incitatifs financiers inhérents au système de santé (ex. les données diagnostiques représentent de l'argent car elles sont associées à une valeur monétaire de remboursement), la vulnérabilité technique des systèmes d'IA en santé (ex. le manque de diversité dans les architectures des réseaux de neurones utilisés) voire du système de santé lui-même (ex. l'infrastructure médicale est difficile à mettre à jour et de nombreux professionnels de santé sont peu, voire pas, formés aux méthodes et techniques d'IA).

Si l'intérêt des recherches sur les exemples antagonistes est surtout axé sur la mise en évidence des limites des systèmes actuels, ces recherches ont aussi retenu l'attention étant donné les menaces inhérentes que les attaques antagonistes représentent en termes de cyber-sécurité (Finlayson et al. 2018). C'est plus généralement le cas de nombreuses recherches sur l'IA, l'IA étant considérée comme une technologie à double-usage (Brundage et al. 2018) soit, qui présente des usages à la fois potentiellement bénéfiques et potentiellement néfastes (Selgelid 2013). Un usage problématique ou « mauvais » est issu du fait que les connaissances scientifiques partagées (publiquement ou au sein du monde académique) pourraient être utilisées par des acteurs

« malveillants » à des fins allant à l'encontre des intérêts de sécurité nationale et de la santé publique, portant notamment atteinte aux droits des individus (Miller et Selgelid 2007; Revill et Dando 2008; Selgelid 2009b; 2009a; 2013). La nécessité de partage inhérente à l'innovation numérique en santé ainsi que la mutualisation des données augmentent inexorablement les possibilités de double-usage. Cependant, le partage peut aussi être en faveur d'une meilleure sécurité, comme le soutient le rapport Villani (2018), traitant ici du cas des voitures autonomes : le partage de données pourrait permettre aux développeurs d'envisager un maximum de possibilités en vue d'assurer la fiabilité des systèmes qu'ils développent.

L'étude de Brown et al. (2017) est un exemple de recherche qui relève de l'IA où le double-usage potentiel est évident. Les chercheurs ont démontré qu'ils pouvaient générer un *patch* universel, robuste et ciblé qui permet de duper les algorithmes de classification peu importe la localisation ou la taille du *patch*, et ce sans connaitre préalablement l'image à attaquer. Si les motivations de leur travail étaient de mettre en évidence des failles potentielles dans la robustesse des systèmes d'IA afin d'en améliorer la sécurité, la mise en évidence de ces failles peut également inquiéter si ces informations venaient à être utilisées par des acteurs « malveillants ».

Les recherches à double-usage font face à un dilemme, qui réfère au conflit entre les valeurs défendables de protection de la santé et la sécurité publique versus la promotion du progrès scientifique (Selgelid 2009a; 2009b; Miller et Selgelid 2007; Resnik 2009). Il s'agit en effet, d'une part, de protéger la liberté académique afin d'assurer l'avancée des connaissances pour le bien commun et, d'autre part, d'empêcher un mésusage potentiel et prévenir ou gérer les risques issus de son développement (Selgelid 2009a; 2009b; Miller et Selgelid 2007; Resnik 2009). Selon Brundage et al. (2018), il est impossible pour les chercheurs en IA de simplement éviter la possibilité de mésusage de leur recherche. Cette notion de mésusage fait écho à trois sources possibles d'atteintes à la sécurité identifiées lors de la consultation de la Déclaration de Montréal (Mai V. dans Déclaration de Montréal IA Responsable 2018, p. 185) : une IA conçue dans le but de menacer la sécurité publique (soit, avec une intention de nuire); l'utilisation des données collectées pour des fins autres que celles envisagées (que l'intention de nuire soit volontaire ou non); un détournement volontaire des systèmes d'IA (ex. piratage). L'utilisation des

données de santé pour des fins autres qu'envisagées est en effet préoccupant, comme le cas d'une personne, en 2013, qui s'est vue refuser l'entrée aux États-Unis sur la base de son passé dépressif, suite à l'accès à son dossier médical (O'Doherty et al. 2016).

Considérant l'intention de nuire, les performances des systèmes d'IA peuvent également se retrouver à la source d'atteintes à la sécurité, si utilisées à des fins problématiques - ou « mal utilisées » (IEEE 2017). Il existerait en effet un risque d'expansion des menaces existantes liées à différents facteurs : considérant l'évolutivité et la puissance croissante des systèmes d'IA, les attaques pourraient devenir plus faciles à réaliser, d'autant plus que la portée de ces systèmes est de plus en plus large, leur accès de plus en plus facile et leur diffusion rapide (Brundage et al. 2018). Apparaît alors également un dilemme entre la protection de la sécurité et l'efficacité des systèmes. Lors de la consultation réalisée dans le cadre de la Déclaration de Montréal, ce dilemme entre la protection de la sécurité et de l'efficacité des systèmes a d'ailleurs été souligné. Certains ont exprimé la crainte que des règles de sécurité trop restrictives pourraient nuire à l'efficacité des modèles et ont fait ressortir les enjeux liés à la recherche de compromis entre un système fiable et un système inopérant (Déclaration de Montréal IA Responsable 2018).

Enfin, certains craignent que les capabilités croissantes des systèmes d'IA conduisent à une perte de contrôle. Loin du catastrophisme des scénarios dystopiques associés au développement d'une *superintelligence* ou d'une *explosion intelligente*, qui conduiraient les systèmes à prendre le contrôle de l'humanité, une préoccupations majeures est relative au comment garantir un contrôle humain sur les systèmes d'IA de plus en plus autonomes et garantir la fiabilité de ce contrôle : en assurant par exemple qu'ils fonctionnent selon des valeurs similaires aux nôtres ou en assurant leur incorruptibilité (soit une protection contre la manipulation et le sabotage) (Russell, Dewey, et Tegmark 2015; Davis 2015; Shulman, Jonsson, et Tarleton 2009; Bostrom et Yudkowsky 2011). Une étude menée par Müller et Bostrom (2016) sur un groupe d'experts en IA a mis en évidence que selon ces experts, il existe une forte probabilité que des systèmes *superintelligents* soient développés dans les 30 années à venir (pour 75% d'entre eux), et les chances que ces

développements soient « mauvais » ou « extrêmement mauvais » s'élèveraient, selon leur perception, à 31%[72] (Müller et Bostrom 2016).

Plusieurs défendent alors qu'il est nécessaire de réfléchir à l'éthique de ces dispositifs - et notamment aux solutions face aux enjeux de sécurité - dès la conception, en amont des conséquences problématiques, afin de garantir la robustesse des systèmes (IEEE 2018; Brundage et al. 2018; Villani 2018; CNIL 2017). Ceci rejoint la nécessité de répondre à un principe de « loyauté » des plateformes qui fonctionneraient sur la base de systèmes d'IA ou de données massives, « visant à ce que l'outil algorithmique ne puisse trahir sa communauté d'appartenance (consumériste ou citoyenne), qu'il traite ou non des données personnelles » (CNIL 2017, p. 6). Son application consiste à assurer le fonctionnement des systèmes « de bonne foi » et « sans chercher à l'altérer ou à le détourner à des fins étrangères à l'intérêt des utilisateurs » (CNIL 2017, p. 48). Il semble en effet impossible que les aspects relatifs à la sécurité soient intégrés dans les systèmes *a posteriori*, notamment parce-que : « il n'est pas possible d'intégrer la dimension sécurité après coup sans détruire une grande partie de ce qui a été construit » (Villani 2018, p. 50). Quel que soit le type de système en cause, Devillers (2017) (citant Gérard Berry) rappelle cependant qu'un système n'est jamais inattaquable mais devient sûr seulement quand il devient trop cher de l'attaquer.

3. Conclusion

Le développement des systèmes d'IA en santé s'accompagne ainsi de différents enjeux techniques inhérents à l'utilisation des technologies en question. D'abord, il existe différentes limites interprétatives et informationnelles de l'analyse systématisée des données massives qu'il est nécessaire de considérer pour ne pas surestimer leur portée. Le fonctionnement optimal de ces technologies nécessite également que données et algorithmes soient ouverts et partagés, ouvrant la porte à différentes préoccupations relatives au contrôle ou à l'accès aux données comme aux algorithmes. Enfin, il n'est pas toujours possible d'expliquer les raisons qui amènent des systèmes

[72] Dans ce sondage, les répondants ont assigné une probabilité à différentes propositions concernant l'impact positif ou négatif de machines à haut niveau d'intelligence. Le résultat cité est la moyenne des probabilités proposées par l'ensemble des participants.

d'IA à formuler des recommandations ou prendre des décisions, amenant aux considérations relatives au manque de transparence – soit, à la boite noire de l'IA.

Ces différentes considérations techniques ne sont pas sans conséquences éthiques et sociales, conduisant à cinq principaux enjeux éthiques relatifs à l'utilisation des systèmes d'IA en santé. Des préoccupations relatives à la protection de la vie privée, de la confidentialité voire de l'intimité apparaissent. L'usage de l'IA et des données massives en santé vient également défier le respect du consentement libre et éclairé des patients et des participants à la recherche. Différentes préoccupations relatives à la justice sociale ressortent également, qu'il s'agisse d'assurer l'accès aux technologies de l'innovation numérique en santé ou de limiter la discrimination issue des biais que les algorithmes pourraient perpétuer. Le recours grandissant aux systèmes d'IA vient exacerber la déshumanisation des soins, en augmentant la distance entre professionnels de santé et patients ou en conduisant à une perception déshumanisée des individus, cantonnés à leurs ensembles de données. Afin de répondre à ces préoccupations éthiques et de limiter les conséquences des enjeux techniques présentés, il est alors nécessaire de renforcer, autant que possible, la sécurité et la robustesse des systèmes d'IA.

Deux principaux éléments ressortent de la description des principaux défis de l'utilisation des systèmes d'IA en santé. Premièrement, il semble exister un lien inextricable entre les enjeux techniques et les enjeux éthiques associés à l'avènement de l'IA et des données massives. Le Tableau 4 illustre l'influence des trois principaux enjeux techniques mentionnées sur 4 des 5 enjeux éthiques soulevés[73].

[73] Il s'agit des conséquences sur la vie privée et la confidentialité, sur le consentement libre et éclairé, la justice sociale, et la déshumanisation des soins et du patient. Assurer la sécurité des systèmes d'IA (le 5ème enjeu) demande de répondre aux 4 enjeux précédemment identifiés et entremêle caractéristiques techniques et éthiques des systèmes d'IA. Il n'a donc pas été inclus dans le tableau mais doit tenir compte de l'ensemble des éléments qui y figurent.

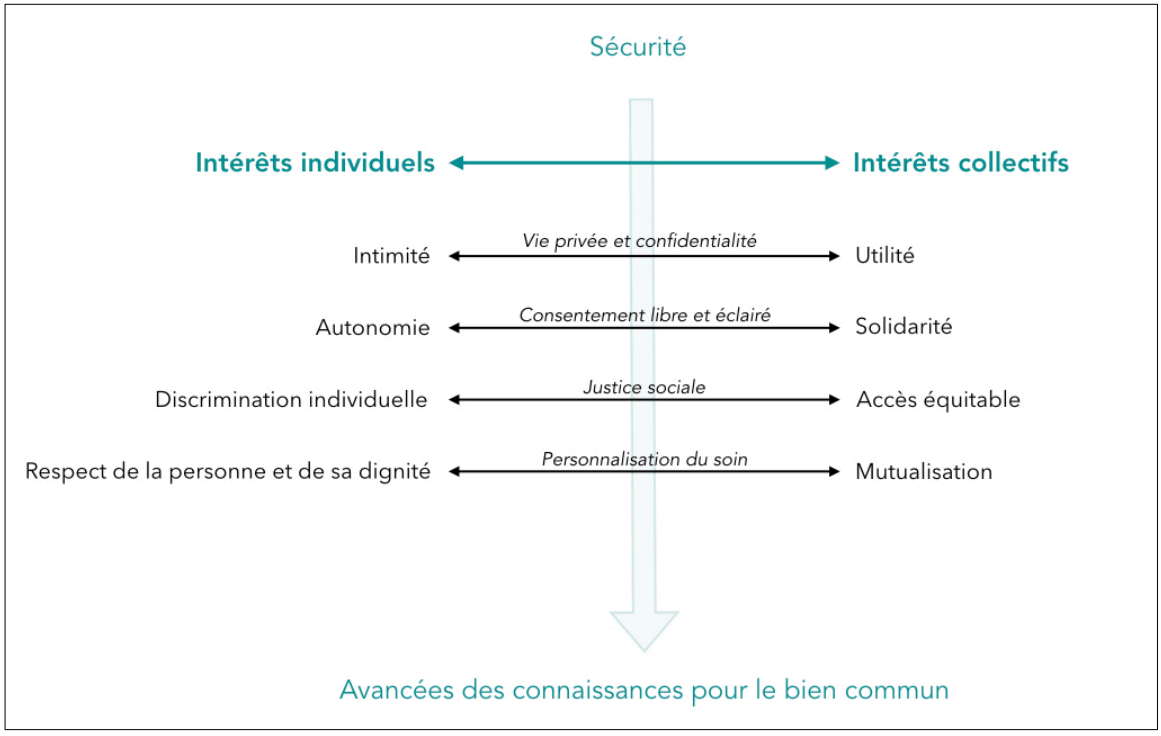
Tableau 5. – Les principaux défis associés aux enjeux éthiques relatifs au développement des systèmes d'IA et leurs principales influences techniques.

Principaux Enjeux éthiques	Vie Privée et Confidentialité	Consentement Libre et Éclairée	Justice Sociale	Déshumanisation
Principaux défis associés	• Préserver l'anonymisation et limiter la réidentification. • Limiter l'intrusion et protéger l'intimité des patients.	• Prévoir pour quelles fins les données sont collectées ou informer sur les utilisations secondaires potentielles. • Garantir l'obtention d'un consentement valide pour les données collectées en dehors du contexte médical ou de la recherche en santé. • Assurer un certain niveau de littératie numérique et de transparence des décisions algorithmiques pour l'obtention d'un consentement véritablement éclairé.	• Assurer l'accès aux technologies. • Protéger de la discrimination sous toutes ses formes.	• Préserver le contact humain, prévenir l'isolement et limiter la distance entre patients et professionnels de santé. • Préserver le patient et les professionnels de santé de la désubjectivation.
Principales influences techniques	• **Partage** (favorise l'accès aux données personnelles à un plus grand nombre d'acteurs).	• **Limites interprétatives et informationnelles** (impactent la compréhension) • **Partage** (augmente les possibilités de réutilisation) • **Opacité** (impacte l'interprétabilité)	• **Limites interprétatives et informationnelles** (risques de perpétuer les biais ou de conduire à une généralisation excessive) • **Partage** (les conditions du partage déterminent partiellement l'accès)	• **Limites interprétatives et informationnelles** (impactent la façon de percevoir les individus et les professionnels de santé). • **Partage** (le contact direct avec un professionnel de santé n'est plus nécessaire pour obtenir des données).

Deuxièmement, il existe une tension marquée entre la protection des intérêts individuels et celle des intérêts collectifs, à différents niveaux. La protection des intérêts individuels (en particulier, concernant la protection de l'intimité, de l'individualité de la discrimination

individuelle ou de l'autonomie des patients) peut s'opposer à la protection d'intérêts et de bénéfices collectifs (relativement à l'utilité des systèmes, l'accès équitable aux données ou aux bénéfices, la solidarité ou la mutualisation). Le respect de l'ensemble de ces intérêts demande d'assurer à la fois la sécurité des systèmes d'IA et de protéger l'avancées des connaissances en vue d'un certain bien commun. Cette tension est illustrée dans le Schéma 3. Il est à noter cependant que ce dilemme, qui oppose une « éthique individualiste » ancrée dans les traditions d'autonomie et de droits individuels avec une éthique de la santé publique basée sur le bien commun et la solidarité n'est pas nouveau - mais persistant - dans les préoccupations de santé publique et de santé des populations (Kenny, Sherwin, et Baylis 2010).

Schéma 3 - Tension entre le respect des intérêts individuels et collectifs relativement au développement éthique des systèmes d'IA.



Ainsi, les qualités intrinsèques des technologies apparentées aux systèmes d'IA sont à la source de conséquences éthiques et sociales préoccupantes qu'il est nécessaire de considérer dans l'optique d'une innovation responsable. Répondre à ces défis fait aujourd'hui l'objet de

nombreuses initiatives et lignes directrices relatives au développement responsable de l'IA et des données massives, qu'il est nécessaire d'explorer en vue de dégager les éléments essentiels à un encadrement de l'innovation numérique en santé éthique et pertinent.

Références bibliographiques

Ajana, Btihaj. 2017. « Digital Health and the Biopolitics of the Quantified Self ». *DIGITAL HEALTH* 3 (janvier): 2055207616689509. https://doi.org/10.1177/2055207616689509.

Altman, M., A. Wood, et E. Vayena. 2018. « A Harm-Reduction Framework for Algorithmic Fairness ». *IEEE Security Privacy* 16 (3): 34-45. https://doi.org/10.1109/MSP.2018.2701149.

Ananny, Mike, et Kate Crawford. 2018. « Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability ». *New Media & Society* 20 (3): 973-89. https://doi.org/10.1177/1461444816676645.

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): 20170350. https://doi.org/10.1098/rsta.2017.0350.

Azria, Élie. 2012. « L'humain face à la standardisation du soin médical ». *La Vie des idées*, juin. http://www.laviedesidees.fr/L-humain-face-a-la-standardisation-du-soin-medical.html.

Barocas, Solon, et Andrew D. Selbst. 2016. « Big Data's Disparate Impact Essay ». *California Law Review* 104: 671-732.

Bayer, Ronald, et Sandro Galea. 2015. « Public Health in the Precision-Medicine Era ». *The New England Journal of Medicine* 373 (6): 499-501. https://doi.org/10.1056/NEJMp1506241.

Binns, R. 2018. « What Can Political Philosophy Teach Us about Algorithmic Fairness? » *IEEE Security Privacy* 16 (3): 73-80. https://doi.org/10.1109/MSP.2018.2701147.

Bostrom, Nick, et Eliezer Yudkowsky. 2011. « The Ethics of Artifical Intelligence ». Dans *The Cambridge Handbook of Artificial Intelligence*, 316-35. Cambridge University Press.

boyd, Danah, et Kate Crawford. 2012. « Critical Questions For Big Data Provocations for a Cultural, Technological, and Scholarly Phenomenon ». *Information Communication & Society* 15 (5): 662-79. https://doi.org/10.1080/1369118X.2012.678878.

Brouard, Benoît. 2017. « Chapitre 2. Utilisation des Big Data en santé : le cas des objets connectés ». *Journal international de bioethique et d'ethique des sciences* Vol. 28 (3): 27-30.

Brown, Nathan, Jean Cambruzzi, Peter J. Cox, Mark Davies, James Dunbar, Dean Plumbley, Matthew A. Sellwood, et al. 2018. « Chapter Five - Big Data in Drug Discovery ». Dans

Progress in Medicinal Chemistry, édité par David R. Witty et Brian Cox, 57:277-356. Elsevier. https://doi.org/10.1016/bs.pmch.2017.12.003.

Brown, Tom B., Dandelion Mané, Aurko Roy, Martín Abadi, et Justin Gilmer. 2017. « Adversarial Patch ». *arXiv:1712.09665 [cs]*, décembre. http://arxiv.org/abs/1712.09665.

Brundage, Miles, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, et Bobby Filar. 2018. « The malicious use of artificial intelligence: Forecasting, prevention, and mitigation ». *arXiv preprint arXiv:1802.07228*.

Campolo, Alex, Madelyn Sanfilippo, Meredith Whittaker, et Kate Crawford. 2017. « AI Now 2017 Report ». https://ainowinstitute.org/AI_Now_2017_Report.pdf.

Castelvecchi, Davide. 2016. « Can We Open the Black Box of AI? » *Nature News* 538 (7623): 20. https://doi.org/10.1038/538020a.

CCNE. 2019. « Données massives (big data) et santé : une nouvelle approche des enjeux éthiques ». Avis 130. Comité Consultatif National d'Éthique français. https://www.ccne-ethique.fr/sites/default/files/avis_130.pdf.

CDT, (Center for Democracy and Technology). 2017. « Digital Decisions ». *Center for Democracy & Technology* (blog). 2017. https://cdt.org/issue/privacy-data/digital-decisions/.

CERNA. 2018. « Research Ethics in Machine Learning ». Research Ethics Board of Allistene, the Digital Sciences and Technologies Alliance. cerna-ethics-allistene.org/digitalAssets/54/54730_cerna_2017_machine_learning.pdf.

Charlet, Jean. 2018. *Intelligence artificielle et algorithmes en santé*. ERES. https://www.cairn.info/traite-de-bioethique-iv--9782749260839-page-541.htm?contenu=resume.

Chartrand, Gabriel, Phillip M. Cheng, Eugene Vorontsov, Michal Drozdzal, Simon Turcotte, Christopher J. Pal, Samuel Kadoury, et An Tang. 2017. « Deep Learning: A Primer for Radiologists ». *RadioGraphics* 37 (7): 2113-31. https://doi.org/10.1148/rg.2017170077.

Christen, Markus, Josep Domingo-Ferrer, Bogdan Draganski, Tade Spranger, et Henrik Walter. 2016. « On the Compatibility of Big Data Driven Research and Informed Consent: The Example of the Human Brain Project ». Dans *The Ethics of Biomedical Big Data*, édité par Brent Daniel Mittelstadt et Luciano Floridi, 199-218. Law, Governance and Technology

Series. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-33525-4_9.

CNIL (Commission nationale informatique et libertés). 2017. « Comment permettre à l'homme de garder la main ? Les enjeux éthiques des algorithmes et de l'intelligence artificielle ».

Coeckelbergh, Mark. 2015. « Artificial Agents, Good Care, and Modernity ». *Theoretical Medicine and Bioethics* 36 (4): 265-77. https://doi.org/10.1007/s11017-015-9331-y.

———. 2012. « "How I Learned to Love the Robot": Capabilities, Information Technologies, and Elderly Care ». Dans *The Capability Approach, Technology and Design*, édité par Ilse Oosterlaken et Jeroen van den Hoven, 77-86. Philosophy of Engineering and Technology. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-3879-9_5.

Collin, Johanne. 2016. « On Social Plasticity: The Transformative Power of Pharmaceuticals on Health, Nature and Identity ». *Sociology of Health & Illness* 38 (1): 73-89. https://doi.org/10.1111/1467-9566.12342.

ÉPTC2 : Conseil de recherches en sciences humaines du Canada, Conseil de recherches en sciences naturelles et en génie du Canada, et Instituts de recherche en santé du Canada. 2018. « Énoncé de politique des trois Conseils : Éthique de la recherche avec des être humains ». http://www.ger.ethique.gc.ca/fra/policy-politique_tcps2-eptc2_2018.html.

Coutellec, Léo, et Paul-Loup Weil-Dubuc. 2017. « Chapitre 7. Big data ou l'illusion d'une synthèse par agrégation. Une critique épistémologique, éthique et politique ». *Journal international de bioethique et d'ethique des sciences* Vol. 28 (3): 63-79.

Crosnier, Hervé Le. 2011. « Une bonne nouvelle pour la théorie des biens communs ». *Vacarme*, nº 56: 92-94. https://doi.org/10.3917/vaca.056.0092.

Crosnier, Hervé LE. 2018. « Communs numériques et communs de la connaissance. Introduction ». *tic&société*, nº Vol. 12, N° 1 (mai): 1-12.

Danaher, John. 2015. « Philosophical Disquisitions: Is effective regulation of AI possible? Eight potential regulatory problems ». *Philosophical Disquisitions* (blog). 7 juillet 2015. http://philosophicaldisquisitions.blogspot.com/2015/07/is-effective-regulation-of-ai-possible.html.

Davis, Ernest. 2015. « Ethical Guidelines for a Superintelligence ». *Artificial Intelligence* 220: 121-24. https://doi.org/10.1016/j.artint.2014.12.003.

Déclaration de Montréal IA Responsable. 2018. « Rapport de la Déclaration de Montréal pour un Développement Responsable de l'Intelligence Artificielle ». https://docs.wixstatic.com/ugd/ebc3a3_d806f109c4104c91a2e719a7bef77ce6.pdf.

Devillers, Laurence. 2017. *Des robots et des hommes: Mythes, fantasmes et réalité*. Plon.

Devillier, Nathalie. 2017. « Chapitre 5. Santé et Big data : l'émergence d'un droit d'infrastructure dans l'espace numérique ». *Journal international de bioethique et d'ethique des sciences* Vol. 28 (3): 51-56.

Finlayson, Samuel G., Hyung Won Chung, Isaac S. Kohane, et Andrew L. Beam. 2018. « Adversarial Attacks Against Medical Deep Learning Systems ». *arXiv:1804.05296 [cs, stat]*, avril. http://arxiv.org/abs/1804.05296.

Fiore, Robin N., et Kenneth W. Goodman. 2016. « Precision Medicine Ethics: Selected Issues and Developments in next-Generation Sequencing, Clinical Oncology, and Ethics ». *Current Opinion in Oncology* 28 (1): 83-87. https://doi.org/10.1097/CCO.0000000000000247.

Floridi, Luciano, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, et al. 2018. « AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations ». *Minds and Machines* 28 (4): 689-707. https://doi.org/10.1007/s11023-018-9482-5.

Floridi, Luciano, et Mariarosaria Taddeo. 2016. « What Is Data Ethics? » *Philosophical Transactions of the Royal Society A-Mathematical Physical and Engineering Sciences* 374 (2083): 20160360. https://doi.org/10.1098/rsta.2016.0360.

Friedler, Sorelle A., Carlos Scheidegger, et Suresh Venkatasubramanian. 2016. « On the (im)possibility of fairness ». *arXiv:1609.07236 [cs, stat]*, septembre. http://arxiv.org/abs/1609.07236.

Ganascia, Jean-Gabriel. 2018. « Éthique, intelligence artificielle et santé ». Dans *Traité de bioéthique*, 527–540. ERES.

Global observatory for eHealth. 2016. « Global diffusion of eHealth : Making universal health coverage achievable ». World Health Organization.

Goodman, Bryce. 2016. « What's Wrong with the Right to Genetic Privacy: Beyond Exceptionalism, Parochialism and Adventitious Ethics ». Dans *The Ethics of Biomedical Big Data*, édité par Brent Daniel Mittelstadt et Luciano Floridi, 139-67. Law, Governance

and Technology Series. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-33525-4_7.

Goodman, Bryce W. 2016. « Economic Models of (Algorithmic) Discrimination. » Dans . Vol. 6. 29th Conference on Neural Information Processing Systems.

Hager, Gregory D., Ann Drobnis, Fei Fang, Rayid Ghani, Amy Greenwald, Terah Lyons, David C. Parkes, et al. 2019. « Artificial Intelligence for Social Good ». *arXiv:1901.05406 [cs]*, janvier. http://arxiv.org/abs/1901.05406.

Hallinan, Dara, et Paul De Hert. 2016. « Many have it wrong–samples do contain personal data: the data protection regulation as a superior framework to protect donor interests in biobanking and genomic research ». Dans *The ethics of biomedical big data*, 119–137. Springer.

Ibekwe-Sanjuan, Fidelia. 2014. « Big Data, Big machines, Big Science : vers une société sans sujet et sans causalité ? » Dans *XIXème Congrès de la Sfsic. Penser les techniques et les technologies : Apports des Sciences de l'Information et de la Communication et perspectives de recherches.*, 1-10. Toulon, France. https://hal.archives-ouvertes.fr/hal-01066202.

IEEE, Institute of Electrical and Electronics Engineers. 2017. « Ethically aligned design - Version 2 - For Public Discussion ». https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf.

Iphofen, Ron, et Mihalis Kritikos. 2019. « Regulating artificial intelligence and robotics: ethics by design in a digital society ». *Contemporary Social Science* 0 (0): 1-15. https://doi.org/10.1080/21582041.2018.1563803.

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

Jameson, J. Larry, et Dan L. Longo. 2015. « Precision Medicine--Personalized, Problematic, and Promising ». *The New England Journal of Medicine* 372 (23): 2229-34. https://doi.org/10.1056/NEJMsb1503104.

Jobin, Anna, Marcello Ienca, et Effy Vayena. 2019. « Artificial Intelligence: the global landscape of ethics guidelines ». *arXiv:1906.11668 [cs]*, juin. http://arxiv.org/abs/1906.11668.

Jones, M. L., E. Kaufman, et E. Edenberg. 2018. « AI and the Ethics of Automating Consent ». *IEEE Security Privacy* 16 (3): 64-72. https://doi.org/10.1109/MSP.2018.2701155.

Kenny, .Nuala P., Susan B. Sherwin, et Françoise E. Baylis. 2010. « Re-Visioning Public Health Ethics: A Relational Perspective ». *Canadian Journal of Public Health* 101 (1): 9-11. https://doi.org/10.1007/BF03405552.

Kim, Pauline T. 2016. « Data-Driven Discrimination at Work ». *William & Mary Law Review* 58: 857-936.

Kitchin, Rob. 2014. « Big Data, New Epistemologies and Paradigm Shifts ». *Big Data & Society* 1 (1): 2053951714528481. https://doi.org/10.1177/2053951714528481.

Lahlou, Saadi. 2008. « Identity, Social Status, Privacy and Face-Keeping in Digital Society ». *Social Science Information* 47 (3): 299-330. https://doi.org/10.1177/0539018408092575.

———. 2015. « Un monde numérique : le renversement du miroir ». Dans . Vol. 53. Variances.

Lavrač, Nada, et Blaž Zupan. 2010. « Data Mining in Medicine ». Dans *Data Mining and Knowledge Discovery Handbook*, édité par Oded Maimon et Lior Rokach, 1111-36. Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-09823-4_58.

LeCun, Yann, Yoshua Bengio, et Geoffrey Hinton. 2015. « Deep Learning ». *Nature* 521 (7553): 436-44. https://doi.org/10.1038/nature14539.

Lee, Howard. 2014. « Paging Dr. Watson: IBM's Watson Supercomputer Now Being Used in Healthcare ». *Journal of AHIMA* 85 (5): 44-47.

Lipton, Zachary C. 2016. « The Mythos of Model Interpretability ». *arXiv:1606.03490 [cs, stat]*, juin. http://arxiv.org/abs/1606.03490.

Lipworth, Wendy, Paul H. Mason, Ian Kerridge, et John P. A. Ioannidis. 2017. « Ethics and Epistemology in Big Data Research ». *Journal of Bioethical Inquiry* 14 (4): 489-500. https://doi.org/10.1007/s11673-017-9771-3.

Longo, Dan L., et Jeffrey M. Drazen. 2016. « Data Sharing ». *New England Journal of Medicine* 374 (3): 276-77. https://doi.org/10.1056/NEJMe1516564.

Maimon, Oded, et Lior Rokach, éd. 2010a. *Data Mining and Knowledge Discovery Handbook*. 2e éd. Springer US. //www.springer.com/gp/book/9780387098227.

———. 2010b. « Introduction to Knowledge Discovery and Data Mining ». Dans *Data Mining and Knowledge Discovery Handbook*, édité par Oded Maimon et Lior Rokach, 1-15. Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-09823-4_1.

Malik, P. 2013. « Governing Big Data: Principles and practices ». *IBM Journal of Research and Development* 57 (3/4): 1:1-1:13. https://doi.org/10.1147/JRD.2013.2241359.

Mettler, M. 2016. « Blockchain technology in healthcare: The revolution starts here ». Dans *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)*, 1-3. https://doi.org/10.1109/HealthCom.2016.7749510.

Miller, Seumas, et Michael J. Selgelid. 2007. « Ethical and Philosophical Consideration of the Dual-Use Dilemma in the Biological Sciences ». *Science and Engineering Ethics* 13 (4): 523-80. https://doi.org/10.1007/s11948-007-9043-4.

Mittelstadt, Brent Daniel, et Luciano Floridi. 2016a. « Introduction ». Dans *The Ethics of Biomedical Big Data*, Springer, 1-16. Brent Daniel Mittelstadt; Luciano Floridi.

———. 2016b. « The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts ». *Science and Engineering Ethics* 22 (2): 303-41. https://doi.org/10.1007/s11948-015-9652-2.

Müller, Vincent C., et Nick Bostrom. 2016. « Future progress in artificial intelligence : a survey of expert opinion ». Dans *Fundamental issues of artificial intelligence*, Springer International Publishing Switzerland, 555-68. V.C. Müller.

Nakrem, Sigrid, Marit Solbjør, Ida Nilstad Pettersen, et Hanne Hestvik Kleiven. 2018. « Care relationships at stake? Home healthcare professionals' experiences with digital medicine dispensers – a qualitative study ». *BMC Health Services Research* 18 (janvier). https://doi.org/10.1186/s12913-018-2835-1.

NoC (Global Network of Internet and Society Centers). 2017. « AI and Inclusion Global Symposium - Pre-event survey responses ». https://drive.google.com/file/d/1xXynk73DPxlcw7iD5f2ZsuNH2I1izRfe/view?usp=sharing.

O'Doherty, Kieran C., Emily Christofides, Jeffery Yen, Heidi Beate Bentzen, Wylie Burke, Nina Hallowell, Barbara A. Koenig, et Donald J. Willison. 2016. « If You Build It, They Will Come: Unintended Future Uses of Organised Health Data Collections ». *Bmc Medical Ethics* 17 (septembre): 54. https://doi.org/10.1186/s12910-016-0137-x.

OMS. 2012. « National eHealth Strategy Toolkit ». World Health Organization.

Resnik, David B. 2009. « What is "dual use" research? A response to Miller and Selgelid ». *Science and engineering ethics* 15 (1): 3–5.

Revill, James, et Malcolm Dando. 2008. « Life Scientists and the Need for a Culture of Responsibility: After Education … What? » *Science and Public Policy* 35 (1): 29-35. https://doi.org/10.3152/030234208X270469.

Rial-Sebbag, Emmanuelle. 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): 39-50.

Rose, Nikolas. 2003. « Neurochemical Selves ». *Society* 41 (1): 46-59. https://doi.org/10.1007/BF02688204.

Rouvroy, Antoinette. 2014. « Des données sans personne: le fétichisme de la donnée à caractère personnel à l'épreuve de l'idéologie des Big Data ». *Contribution en marge de l'Étude annuelle du Conseil d'État. Le numérique et les droits et libertés fondamentaux*.

Rouvroy, Antoinette, et Thomas Berns. 2013. « Gouvernementalité algorithmique et perspectives d'émancipation ». *Réseaux*, nº 177 (mai): 163-96. https://doi.org/10.3917/res.177.0163.

Rumbold, John M. M., et Barbara K. Pierscionek. 2017. « A Critique of the Regulation of Data Science in Healthcare Research in the European Union ». *Bmc Medical Ethics* 18 (avril): 27. https://doi.org/10.1186/s12910-017-0184-y.

Russell, Stuart, Daniel Dewey, et Max Tegmark. 2015. « Research Priorities for Robust and Beneficial Artificial Intelligence ». *AI Magazine* 36 (4): 105-14.

Schwab, Klaus. 2016. *La quatrième révolution industrielle*. Dunod. Suisse: World Economic Forum.

Selbst, Andrew D., et Solon Barocas. 2018. « The Intuitive Appeal of Explainable Machines ». SSRN Scholarly Paper ID 3126971. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=3126971.

Selgelid, Michael J. 2009a. « Governance of dual-use research: an ethical dilemma ». *Bulletin of the World Health Organization* 87 (9): 720-23. https://doi.org/10.1590/S0042-96862009000900017.

———. 2009b. « Dual-Use Research Codes of Conduct: Lessons from the Life Sciences ». *NanoEthics* 3 (3): 175-83. https://doi.org/10.1007/s11569-009-0074-y.

———. 2013. « Dual-Use Research ». *The International Encyclopedia of Ethics*. http://onlinelibrary.wiley.com/doi/10.1002/9781444367072.wbiee607/full.

Sharon, Tamar. 2016. « The Googlization of Health Research: From Disruptive Innovation to Disruptive Ethics ». *Personalized Medicine* 13 (6): 563-74. https://doi.org/10.2217/pme-2016-0057.

———. 2017. « Self-Tracking for Health and the Quantified Self: Re-Articulating Autonomy, Solidarity, and Authenticity in an Age of Personalized Healthcare ». *Philosophy & Technology* 30 (1): 93-121. https://doi.org/10.1007/s13347-016-0215-5.

Shulman, Carl, Henrik Jonsson, et Nick Tarleton. 2009. « Machine ethics and superintelligence ». Dans , 95–97. Tokyo, Japan: Carson Reynolds and Alvaro Cassinelli.

Spiekermann, S., J. Korunovska, et M. Langheinrich. 2018. « Inside the Organization: Why Privacy and Security Engineering Is a Challenge for Engineers[40pt] ». *Proceedings of the IEEE*, 1-16. https://doi.org/10.1109/JPROC.2018.2866769.

Stahl, B. C., et D. Wright. 2018. « Ethics and Privacy in AI and Big Data: Implementing Responsible Research and Innovation ». *IEEE Security Privacy* 16 (3): 26-33. https://doi.org/10.1109/MSP.2018.2701164.

The Economist. 2017. « The world's most valuable resource is no longer oil, but data ». *The Economist*, 6 mai 2017. https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data.

Topol, Eric J. 2019. *Deep Medicine : How Artificial Intelligence Can Make Healthcare Human Again*. Basic Books. https://www.basicbooks.com/titles/eric-topol/deep-medicine/9781541644649/.

Torkamani, Ali, Kristian G. Andersen, Steven R. Steinhubl, et Eric J. Topol. 2017. « High-Definition Medicine ». *Cell* 170 (5): 828-43. https://doi.org/10.1016/j.cell.2017.08.007.

UNESCO. 2018a. « A draft report on a global framework on digital literacy skills for indicator ». http://gaml.cite.hku.hk/wp-content/uploads/2018/03/DLGF-draft-report-for-online-consultation-all-gaml.pdf.

———. 2018b. « Recommandation concernant la science et les chercheurs scientifiques ». UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000263618_fre.

Verdier, Henri, et Charles Murciano. 2017. « Les communs numériques, socle d'une nouvelle économie politique ». *Esprit* Mai (5): 132-45.

Villani, Cédric. 2018. « Donner un sens à l'intelligence artificielle. Pour une stratégie nationale et européenne. » https://www.aiforhumanity.fr/pdfs/9782111457089_Rapport_Villani_accessible.pdf.

WEF, (World Economic Forum). 2012. « Big data, Big Impact : New possibilities for Development ». Genève.

Wellcome Trust. 2013. « Impact of the draft European data protection regulation and proposed amendments from the rapporteur of the LIBE committee on scientific research. » http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_ communications/documents/web_document/WTP055584.pdf.

Woods, Simon. 2016. « Big Data Governance: Solidarity and the Patient Voice ». Dans *The Ethics of Biomedical Big Data*, édité par Brent Daniel Mittelstadt et Luciano Floridi, 221-38. Law, Governance and Technology Series. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-33525-4_10.

Woolley, J. Patrick. 2016. « How Data Are Transforming the Landscape of Biomedical Ethics: The Need for ELSI Metadata on Consent ». Dans *The Ethics of Biomedical Big Data*, édité par Brent Daniel Mittelstadt et Luciano Floridi, 171-97. Law, Governance and Technology Series. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-33525-4_8.

Zarsky, Tal. 2016. « The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making ». *Science, Technology, & Human Values* 41 (1): 118-32. https://doi.org/10.1177/0162243915605575.

Zhang, G. Peter. 2010. « Neural Networks For Data Mining ». Dans *Data Mining and Knowledge Discovery Handbook*, édité par Oded Maimon et Lior Rokach, 419-44. Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-09823-4_21.

Zwitter, Andrej. 2014. « Big Data Ethics ». *Big Data & Society* 1 (2): 2053951714559253. https://doi.org/10.1177/2053951714559253.

Chapitre 4 – Gouvernance éthique des systèmes d'intelligence artificielle

Considérant les différents risques et enjeux éthiques auxquels expose l'utilisation des systèmes d'intelligence artificielle (IA), différentes institutions, gouvernements et organisations se penchent actuellement sur l'encadrement de leur utilisation. De nombreuses initiatives ont vu le jour, de portée nationale ou internationale, aboutissant à la proposition de déclarations de principes et de lignes directrices éthiques nécessaires au développement responsable de l'IA, quels que soient les secteurs d'application. Ces principes ont pour vocation d'orienter l'encadrement ou l'intendance des systèmes d'IA afin d'assurer que leur développement se fasse de manière éthique et responsable, soit de guider les mécanismes de la gouvernance algorithmique. Comme dans toute démarche d'éthique appliquée, l'opérationnalisation des principes de l'éthique de l'IA se heurte à différentes difficultés qu'il est nécessaire de prendre en considération en vue d'une innovation numérique responsable – en santé comme dans d'autres secteurs.

1. De l'ambiguïté de la gouvernance algorithmique

Les réflexions qui portent sur la mise en place d'un encadrement éthique de l'utilisation des systèmes d'IA renvoient à la notion de gouvernance algorithmique, soit les « procédures qui permettent d'encadrer les dispositifs relatifs à la prise de décision autonome (à des degrés variables) par un système automatisé » (Déclaration de Montréal IA Responsable 2018 p. 263). Or, cette notion soulève une ambiguïté notable. Cette ambiguïté, comme le souligne Musiani (2013), n'est pas l'effet d'un terme mal choisi mais révèle un état de fait essentiel :

> By naming a conference held at New York University last May "Governing Algorithms", its organisers were making a deliberate choice of ambiguity – hinting at both the governance of algorithms, the extent to which political regulation can affect the functioning of the instructions and procedures subtending technology, and the governing power of algorithms themselves (p. 2).

La gouvernance algorithmique peut ainsi référer soit aux différentes manières de gouverner les algorithmes, soit à la manière dont ils nous gouvernent (Déclaration de Montréal IA Responsable 2018).

La seconde conception de la gouvernance algorithmique (la manière dont les algorithmes nous gouvernent) est sous-tendue par l'idée que les algorithmes deviennent une figure de pouvoir dans nos sociétés actuelles. Leur (omni)présence, parfois invisible pour l'utilisateur, vient influencer différents aspects de la vie quotidienne comme les décisions des plus importantes. On accorde aujourd'hui facilement aux algorithmes un certain pouvoir de « modifier les règles de l'économie, les choix politiques des électeurs ou la vie quotidienne des individus » (Cardon 2018). Ils peuvent devenir des *Weapons of maths destruction* (O'Neil 2016), contribuent à la mise en forme des informations auxquelles nous avons accès et leur organisation (Musiani 2013) et exercent leur influence par différents effets de « bulle filtrante »[74], quand il ne s'agit pas directement d'orienter les individus vers des comportements jugés appropriés et « bons » pour eux par le biais de *nudges*[75] (Hausman et Welch 2010; Sonntag 2016). Rouvroy et Berns décrivent trois temps de ce qu'ils nomment la « gouvernementalité algorithmique ». Le premier est celui de la collecte de données massives ; le second, leur traitement, dont découle la production de connaissances ; le troisième réfère quant à lui à l'usage de ces savoirs probabilistes et statistiques (notamment à des fins de profilage) (Rouvroy et Berns 2013). C'est, selon les auteurs, lors du troisième temps que les algorithmes « nous gouvernent » (Rouvroy et Berns 2013). Par le biais du profilage qu'ils opèrent (quel que soit le secteur d'application) les algorithmes exercent une certaine normativité relativement à nos traces numériques (Ibekwe-Sanjuan 2014). Cette normativité suppose une « exposition permanente de soi », qui pourrait rendre suspect tout individu qui ne s'y conforme pas (Ibekwe-Sanjuan 2014).

L'échelle de grandeur, tant concernant la quantité des données sur lesquelles les algorithmes apprennent que la puissance de leur analyse, élargie la portée des décisions qui en émanent. Outre l'impression d'indomptabilité que cette échelle confère aux systèmes d'IA et aux bases de données

[74] L'effet de « bulle filtrante » est issu du fait que les algorithmes, en s'appuyant sur les caractéristiques des profils des utilisateurs, augmentent leur tendance à rechercher des contenus conformes à leurs propres goûts (Déclaration de Montréal IA Responsable 2018). « Un individu se retrouve donc enfermé dans une 'bulle filtrante', c'est-à-dire dans un espace de recommandations toujours conforme au profil qu'il alimente par son comportement numérique. » (Dilhac M, dans Déclaration de Montréal IA Responsable 2018, p. 290).

[75] Les *nudges* réfèrent à des architectures de choix qui vont encourager ou orienter les individus à faire certains choix (relatifs – ou non – à la santé), considérés comme « bons », plutôt que d'autres (Hausman et Welch 2010; Sonntag 2016). Les *nudges* sont parfois critiqués pour le paternalisme sous-jacent à leur développement et parfois considérés comme de la persuasion rationnelle, les individus conservant tout de même la liberté de faire le mauvais choix (Hausman et Welch 2010; Sonntag 2016).

massives, qui peuvent facilement être perçus comme hors de contrôle, cette échelle de grandeur favorise l'émergence de la perception des algorithmes comme figure de pouvoir :

> Plus que la simple collecte des données numériques, souvent figurée comme le principal enjeu du *Big Data*, c'est donc la force et la précision des calculs (notamment leur capacité à effectuer des traitements massifs en temps réel) qui expliquent l'émergence des algorithmes comme une nouvelle figure du pouvoir (Cardon 2018 p. 63).

La gouvernance par les algorithmes devient alors associée à une « double automatisation » : celle de la collecte et de l'analyse des données (tâches qui deviennent impossible à réaliser « manuellement ») et celle de l'automatisation de la prise de décision qui émane des résultats de ces analyses (Musiani 2013). Cette double automatisation soulève alors des enjeux relatifs à l'agentivité et au contrôle des systèmes d'IA (Musiani 2013). D'autres réduisent ce pouvoir aux intérêts (économiques) de leurs concepteurs (Cardon 2018) quand il ne s'agit pas de considérer explicitement les algorithmes comme de véritables agents moraux (Shulman, Jonsson, et Tarleton 2009; Moor 2006). Plusieurs auteurs proposent en effet d'intégrer des représentations explicites de normes, principes ou valeurs morales dans les algorithmes (Bostrom et Yudkowsky 2011; E. Davis 2015; Moor 2006; Scheutz 2016; Shulman, Jonsson, et Tarleton 2009). Certains auteurs questionnent même la possibilité d'atteindre une morale ultime *via* des algorithmes créés de façon à être incorruptibles (Bostrom et Yudkowsky 2011; E. Davis 2015; Moor 2006).

Pour ne pas que les algorithmes nous gouvernent, plusieurs proposent de gouverner les algorithmes en limitant le pouvoir qu'on leur accorde, en favorisant le contrôle des humains sur les machines (ex. CNIL 2017; Déclaration de Montréal IA Responsable 2018) et en développant également différents mécanismes de gouvernance. L'ensemble de ces réponses renvoient ainsi à la première conception de la gouvernance algorithmique, soit la manière de gouverner les algorithmes. La gouvernance algorithmique ne saurait se réduire à la gouvernance éthique de l'IA, mais la mise en place de principes éthiques pour guider cette gouvernance semble être le point de départ de nombreuses initiatives à travers le monde.

2. Des principes éthiques pour guider la gouvernance de l'intelligence artificielle

Différentes organisations gouvernementales, académiques ou entreprises privées ont produit des déclarations de principes ou lignes directrices éthiques en vue du développement responsable de l'IA. Le Tableau 6 en présente quelques exemples, ainsi que les principaux principes éthiques mobilisés dans ces documents. Si cette liste ne saurait être exhaustive[76], elle tente de démontrer l'ampleur du travail réalisé et la diversité des parties prenantes qui se sont penchées sur l'établissement de principes directeurs relevant de la gouvernance éthique de l'IA.

Tableau 6. – Exemples de différentes initiatives éthiques en vue du développement responsable de l'IA et les principaux principes mobilisés.

[76] Pour plus d'informations sur l'ensemble des principes et lignes directrices développés à travers le monde, voir la cartographie de Jobin, Ienca, et Vayena (2019). D'autres travaux similaires ont également été réalisés, comme la cartographie de Fjeld et collaborateurs (2019) https://ai-hr.cyber.harvard.edu/primp-viz.html; la Global AI Policy Database de Sixt (https://www.charlottestix.com/ai-policy-resources) ou le Global AI Policy du *Future of Life institute* (https://futureoflife.org/ai-policy/).

Titre du document	Date	Auteurs	Principes mobilisés
Asilomar AI Principles	Janvier 2017	Future of life Institute	1. Safety 2. Failure Transparency 3. Judicial Transparency 4. Responsibility 5. Value Alignment 6. Human Values 7. Personal Privacy 8. Liberty and Privacy 9. Shared Benefit 10. Shared Prosperity 11. Human Control 12. Non-subversion 13. AI Arms Race.
Japanese Society for Artificial Intelligence Ethical Guidelines	Février 2017	Japanese Society for Artificial Intelligence (JSAI)	1. Contribution to humanity 2. Abidance of laws and regulations 3. Respect for the privacy of others 4. Fairness 5. Security 6. Act with integrity 7. Accountability and social responsibility 8. Communication with society and self-development 9. Abidance of ethics guidelines by AI
ITI AI policy principles	Octobre 2017	Information Technology Industry Council (ITI)	• Responsible Design and Deployment • Safety and Controllability • Robust and Representative Data • Interpretability • Liability of AI Systems Due to Autonomy • Cybersecurity and Privacy • Democratizing Access and Creating Equality of Opportunity
Comment permettre à l'homme de garder la main ?	Décembre 2017	Commission internationale de l'informatique et des libertés française (CNIL)	• Loyauté • Vigilance et réflexivité
Déclaration de Toronto	Mai 2018	Amnesty International et Access Now	• Equality and Non-discrimination • Diversity and Inclusion • Transparency and Accountability
Google AI Principles[77]	Juin 2018	Google	1. Be socially beneficial 2. Avoid creating or reinforcing unfair bias 3. Be built and tested for safety 4. Be accountable to people 5. Incorporate privacy design principles 6. Uphold high standards of scientific excellence 7. Be made available for uses that accord with these principles
Code of Conduct for Data-driven Health and Care Technology	Septembre 2018	National Health Service (UK)	1. Understand users, their needs and the context 2. Define the outcome and how the technology will contribute to it 3. Use data that is in line with appropriate guidelines for the purpose for which it is being used 4. Be fair, transparent and accountable about what data is being used 5. Make use of open standards

[77] Disponible ici : https://ai.google/principles/

			6. Be transparent about the limitations of the data used and algorithms deployed 7. Show what type of algorithm is being developed or deployed, the ethical examination of how the data is used, how its performance will be validated and how it will be integrated into health and care provision 8. Generate evidence of effectiveness for the intended use and value for money 9. Make security integral to the design 10. Define the commercial strategy
Déclaration de Montréal	Décembre 2018	Université de Montréal	1. Bien-être 2. Respect de l'autonomie 3. Protection de l'intimité et de la vie privée 4. Solidarité 5. Participation démocratique 6. Équité 7. Inclusion de la Diversité 8. Prudence 9. Responsabilité 10. Développement soutenable
Microsoft AI Principles[78]	2018	Microsoft	1. Fairness 2. Reliability and Safety 3. Privacy and Security 4. Inclusiveness 5. Transparency 11. Accountability
Ethics Guidelines for Trustworthy AI	Avril 2019	High-Level Expert Group on Artificial Intelligence (Commission Européenne)	1. Respect for human autonomy 2. Prevention of harm 3. Fairness 4. Explicability
Artificial Intelligence : Australian Ethics Framework.	Avril 2019	CSIRO's Data61 et Australian Government – Department of Industry, Innovation and Science	1. Generates net-benefits 2. Do no harm 3. Regulatory and legal compliance 4. Privacy protection 5. Fairness 6. Transparency & Explainability 7. Contestability 8. Accountability
OECD AI Principles	Mai 2019	Organisation de coopération et de développement économiques (OCDE)	1. Inclusive growth, sustainable development and well-being 2. Human-centered values and fairness 3. Transparency and explainability 4. Robustness, security and safety 5. Accountability
Beijing AI Principles	Mai 2019	Beijing Academy of Artificial Intelligence (BAAI), Pekin University, Tsinghua University, Institute of	Research and development • Do Good For Humanity • Be Responsible • Control Risks • Be Ethical • Be Diverse and Inclusive • Open and Share Use

[78] Disponible sur https://www.microsoft.com/en-us/ai/our-approach-to-ai

		Automation and Institute of Computing Technology in Chinese Academy of Sciences, and an AI industrial league involving firms like Baidu, Alibaba and Tencent.	• Use Wisely and Properly • Informed-consent • Education and Training Governance • Optimizing Employment • Harmony and Cooperation • Adaptation and Moderation • Subdivision and Implementation • Long-term planning
Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition	Août 2019	Institute of Electrical and Electronics Engineers (IEEE)	1. Human rights 2. Well-Being 3. Data Agency 4. Effectiveness 5. Transparency 6. Accountability 7. Awareness of misuse 8. Competence

Les documents sont présentés dans l'ordre chronologique de leur parution. Les principes sont présentés en français ou en anglais selon les traductions officielles disponibles. Les principes sont numérotés tel que dans les documents d'origine, lorsque disponible.

2.1. Organisations internationales

Plusieurs organisations internationales et intergouvernementales se sont positionnées sur l'éthique du développement de l'IA, de la robotique ou des données massives. Le 16 mai 2018, Amnesty International et Access Now lance la Déclaration de Toronto, qui propose de protéger les droits à l'égalité et à la non-discrimination relativement à l'utilisation des systèmes d'apprentissage automatique (Amnesty International 2018). Utilisant la Déclaration universelle des droits humains[79] comme cadre de référence pour défendre l'égalité et promouvoir la diversité et l'inclusion, cette Déclaration est centrée sur l'accès aux technologies, la non-discrimination et le rôle des États et des compagnies privées dans la gouvernance de l'IA. Elle a été endossée, entre autres, par Wikimédia Foundation et Human Rights Watch (Amnesty International 2018).

De son côté, l'Organisation Mondiale de la Santé (OMS) publie en 2018 un rapport émanant d'une consultation internationale sur l'éthique des données massives et de l'IA en santé. L'objectif de l'initiative est d'identifier le spectre des enjeux éthiques associés à leur utilisation et leur

[79] Réfère ici à la traduction de « Universal Declaration of Human Rights », qui a été préférée à la traduction plus commune de « Déclaration universelle des droits de l'Homme » pour sa nature plus inclusive.

développement, en vue d'informer la création de lignes directrices et de principes éthiques pour guider différentes parties prenantes (WHO 2018). La Commission Européenne a quant à elle mis sur pied un groupe d'experts de haut niveau sur l'IA (High-level Expert Group on Artificial Intelligence) conduisant au lancement, en avril 2019, de lignes directrices éthiques pour une IA « digne de confiance » (AI HLEG 2019). Quatre principes éthiques sont à la base des recommandations du rapport associé : le respect de l'autonomie humaine, la prévention du préjudice, l'équité et l'explicabilité (AI HLEG 2019). Ces principes sont eux-mêmes déclinés en sept exigences clés.

Le 22 mai 2019, l'ensemble des pays membres de l'OCDE (et d'autres pays d'Amérique latine – soit 42 pays) adopte la Déclaration de principes promulguée par cette organisation, qui se décline en cinq grands principes éthiques. Ces principes constituent, selon l'OCDE, les premiers standards internationaux agréés par les gouvernements en vue de l'intendance responsable de l'IA[80]. L'UNESCO, qui souhaite également prendre part au façonnement de l'avenir de l'IA, a récemment pris position sur la question de son développement responsable en défendant une approche humaniste[81]. Selon cette approche, le développement de l'IA doit être « centré sur l'humain », assurer une égalité d'accès au savoir, la diversité des expressions culturelles, ou ne doit pas creuser la fracture technologique entre les pays et au sein d'entre eux. L'organisation a organisé deux conférences mondiales sur le sujet et publié différents rapports, notamment sur la promotion de l'égalité des genres dans le développement de l'IA (UNESCO et EQUALS Skills Coalition 2019).

L'Institute of Electrical and Electronics Engineers (IEEE)[82] a lancé, en 2019, le rapport *Ethicaly aligned design* première édition, dans sa troisième version, dans la lignée de leur initiative sur l'éthique des systèmes autonomes et intelligents. Cette dernière a pour but de réunir les voix des communautés technologiques et scientifiques concernées afin d'identifier et de trouver, entres autres, des consensus quant au développement des systèmes d'IA (IEEE 2019). Cette version a été

[80] Voir : https://www.oecd.org/going-digital/ai/
[81] Voir : https://fr.unesco.org/artificial-intelligence
[82] L'IEEE se définit comme la plus grande organisation professionnelle technique au monde pour le progrès de la technologie, regroupant plus de 420 000 membres relevant de plus de 160 pays.

développée à partir de plus de 200 pages de contributions, recueillies sur la base de la première version (publiée en Décembre 2016) et plus de 300 pages de contributions sur la deuxième version (publiée en Décembre 2017). Le rapport présente cinq principes éthiques dont découlent différentes recommandations de politiques publiques.

2.2. Initiatives nationales

De nombreux gouvernements et organisations nationales se sont également penchés sur la question du développement éthique et responsable de l'IA, proposant des déclarations de principes définissant l'orientation de potentielles politiques publiques. Par exemple, en octobre 2016, le UK House of Commons Science and Technology Committee sort un rapport à l'attention du gouvernement britannique, qui se penche sur les enjeux éthiques et légaux de vérification et de validation des systèmes d'IA, le manque de transparence de la prise de décision algorithmique, la minimisation des biais, la protection du consentement des utilisateurs et la vie privée, ou encore l'imputabilité en cas de conséquences néfastes de l'utilisation des systèmes d'IA (House of Commons Science and Technology Committee 2016). La Commission nationale de l'informatique et des libertés française (CNIL) s'appuie quant à elle dans son rapport *Comment permettre à l'homme de garder la main ?*, parut en décembre 2017, sur deux principes fondateurs pour le développement de l'IA : loyauté et vigilance. Ces principes sont déclinés en six recommandations de politiques publiques qui visent, entre autres, à protéger l'autonomie et l'identité humaine, la confidentialité des données ou encore protéger de la discrimination (CNIL 2017).

Au Japon, le Comité d'éthique de la société japonaise pour l'IA lance en février 2017 les *Japanese Society for Artificial Intelligence Ethical Guidelines*, qui visent à définir une direction éthique pour les professionnels (JSAI 2017). L'organisation mixte Beneficial AI Japan a organisé, en octobre 2017, l'atelier Beneficial AI Tokyo qui visait à regrouper différentes parties prenantes de l'IA afin d'explorer les défis de la construction d'une communauté mondiale efficace et d'assurer un développement sécuritaire et bénéfique. Les participants ont exprimé un engagement à travailler sur le développement d'une « AI for good » et ont souscrit à la Déclaration de Tokyo (*The Tokyo Statement : Cooperation for beneficial AI*)[83]. Si les principes éthiques ne sont pas

[83] Voir : http://bai-japan.org/en/tokyo-statement/

explicites, la Déclaration défend qu'il est nécessaire de s'emparer du défi du développement d'une IA bénéfique dans un esprit de coopération et non de compétition et que ce développement doit se faire en accord avec les valeurs de la communauté dans laquelle les technologies seront déployées.

Certaines des initiatives sont spécifiques à la santé, comme le National Health Service du Royaume-Uni qui publie, en septembre 2018, le *Code of Conduct for data-driven health and care technology* (Departement of Health and Social Care 2018). Ce code de conduite promeut 10 principes pour capitaliser de manière responsable sur les opportunités qu'offrent les technologies de santé guidées par les données, notamment en vue des bénéfices potentiels pour les patients. Au Canada, l'équipe de la Déclaration de Montréal[84] lance, en décembre 2018, une déclaration de 10 principes éthiques pour guider le développement responsable de l'IA. Ces principes ont été développés sur la base d'une large consultation citoyenne (principalement au Québec) et ont été adoptés par plus de mille signataires – des citoyens, diverses organisations (ex. l'Ordre des ingénieurs du Québec) ou des compagnies privées (ex. AiFred Health, Imagia)[85].

En février 2018, la Commission de réflexion sur l'Éthique de la Recherche en sciences et technologies du Numérique d'Alllistene (CERNA) française publie le rapport *Éthique de la recherche en apprentissage machine*. Le rapport s'adresse aux chercheurs en technologie de l'information, développeurs et concepteurs, et recommande notamment que les chercheurs assurent la qualité des données sur lesquelles les systèmes apprennent ou veillent à ce qu'elles reflètent la diversité des groupes d'utilisateurs. Selon ce rapport, les parties concernées doivent rester vigilantes dans la communication au public, prêter attention au manque d'explicabilité des systèmes, être transparents face à ce manque d'explicabilité ou face aux intentions d'utilisation, ou encore respecter le consentement des utilisateurs et des participants à la recherche. Elle recommande également que les résultats issus des analyses de systèmes d'IA soient toujours interprétables par un humain (CERNA 2018b).

[84] Voir le Chapitre 1 : Méthodologie où le travail de l'équipe de la Déclaration de Montréal est présenté en détail.
[85] Voir : https://www.declarationmontreal-iaresponsable.com/signataires

Le Comité Consultatif National d'Éthique (CCNE) français emboite le pas en publiant un avis le 29 mai 2019 *Données massives et santé : une nouvelle approche des enjeux éthiques*, qui met de l'avant le respect de la personne, la justice, la bienfaisance et la non-malfaisance (CCNE 2019). Le CCNE propose également différents « principes d'action » relatifs au consentement et ses nouvelles formes ainsi que 12 recommandations qui expriment « la certitude que de nouveaux équilibres doivent être trouvés dans l'exercice d'une démarche éthique qui doit accompagner les conséquences qu'introduisent les sciences et technologies du numérique, sans freiner les bénéfices attendus, mais sans affaiblir les principes qui fondent la qualité d'être humain et la relation humaine » (p. 85).

Récemment, le gouvernement australien a pris part à la discussion sur l'éthique de l'IA en lançant le *Australia's Ethics Framework.* Le document, publié sous la forme d'un « *discussion paper* » en avril 2019, vise à stimuler les discussions et appelle à contribuer à son contenu (Dawson et al. 2019). Il regroupe différentes recommandations clés pour le développement de politiques publiques ainsi que huit principes fondamentaux en vue du développement de l'IA, afin de protéger notamment la confiance du public et la réalisation des bénéfices de ces technologies. Le travail a été guidé par un comité directeur composé de différents experts de l'industrie, du gouvernement et de différentes organisations communautaires.

2.3. Initiatives des acteurs de la sphère privée

Différentes firmes privées se sont également penchées sur le développement de principes éthiques. Google a par exemple déployé, en juin 2018, sept principes pour guider le développement de ses technologies d'IA[86]. La compagnie a également formé en mars 2019, en complément de leur déclaration de principes, un comité d'éthique (le Advanced Technology External Advisory Council), qui avait pour but de réfléchir et d'aider le développement responsable de l'IA[87], bien que démantelé à peine une semaine après sa création suite à différentes protestations, notamment des employés de la compagnie[88].

[86] Voir : https://ai.google/principles/
[87] Voir : https://www.blog.google/technology/ai/external-advisory-council-help-advance-responsible-development-ai/
[88] Les employés de Google ont dénoncé la vision conservatrice d'une des membres du comité, qui n'était pas alignée avec les valeurs de la compagnie (Knight 2019). De plus, une lettre ouverte signée par environ un millier de personne (notamment des chercheurs) dénonçait le développement d'un conseil prétendument indépendant pour guider l'éthique des projets d'IA de Google (Knight 2019).

IBM[89], qui défend que l'éthique « fait partie intégrante » de la mission des entreprises, a de son côté développé le code de conduite : *Everyday Ethics for Artificial Intelligence* (la dernière version datant de 2019) qui se penche sur cinq sujets d'éthique de l'IA : l'imputabilité, l'alignement des valeurs, l'explicabilité, l'équité et les droits d'utilisateurs. La compagnie Microsoft a quant à elle adopté six principes éthiques pour guider le développement et l'utilisation de l'IA « with people at the center of everything ». Ces principes ont encore une fois pour vocation de garantir le développement d'une IA digne de confiance, qui prend racine dans différentes valeurs[90]. Microsoft a également publié, en novembre 2018, dix lignes directrices pour la conception de *bots* responsables, qui demandent, en plus de respecter leurs principes éthiques, d'articuler l'objectif du *bot*, être transparent sur son utilisation (et notamment assurer qu'il soit clair qu'il ne s'agit pas d'un humain) ou d'assurer des échanges *humains-bots* fiables (Microsoft Corporation 2018).

Les principes, définis par le biais de ces différentes initiatives, sont généralement développés de manière à être suffisamment abstraits afin d'être flexibles et de permettre d'assurer leur pérennité (considérant les avancées rapides du domaine) mais également de permettre aux différentes parties prenantes du développement de l'IA de s'approprier les principes et de les adapter à leur réalité pratique. C'est par exemple ce que mentionne l'OCDE relativement à sa Déclaration, précisant que les principes ont été développés de manière à être « practical and flexible enough to stand the test of time in a rapidly evolving field »[91]. Ces déclarations sont établies sur la base d'une vision large et globale des défis du développement de l'IA, et ne sont pas toujours spécifiques à un secteur particulier. L'adoption de principes éthiques ne représente pas une solution *per se* aux différents défis du développement responsable de l'IA mais vient uniquement guider la mise en place de solutions pratiques.

[89] Voir : https://www.ibm.com/watson/ai-ethics/
[90] Voir : https://www.microsoft.com/en-us/ai/our-approach-to-ai
[91] Voir : https://www.oecd.org/going-digital/ai/principles/

3. Opérationnalisation des principes de l'éthique de l'intelligence artificielle

3.1. Traduire les principes éthiques en mesures concrètes

Il existe différentes manières de mettre en application les principes de l'éthique de l'IA. Le rapport de la Commission Européenne les regroupe selon deux principales méthodes : les méthodes techniques et les méthodes non-techniques, qui font référence à l'ensemble du cycle de développement des systèmes d'IA (AI HLEG 2019).

Les méthodes techniques demandent de traduire, dès la conception, les principes en procédures pour les inclure dans les architectures des algorithmes d'IA (AI HLEG 2019). Elles peuvent être incorporées à différentes étapes du processus de développement, de la conception à l'utilisation (AI HLEG 2019). Parmi ces mesures techniques, celles qui relèvent d'une *ethics by design* font couler beaucoup d'encre. Il s'agit de développer des valeurs dès la conception et expliciter les liens entre les principes éthiques abstraits que le système doit respecter et les décisions d'implémentation (AI HLEG 2019; Dignum et al. 2018; Borrett, Sampson, et Cavoukian 2017; d'Aquin et al. 2018; Iphofen et Kritikos 2019). De nombreux rapports sur l'éthique de l'IA défendent la nécessité d'une éthique dès la conception, en amont de la chaine du développement de l'IA comme l'IEEE qui défend la nécessité de développer des systèmes « safe by design » (IEEE 2017). Dans le secteur de la santé comme ailleurs, différents auteurs proposent le développement de mesures d'éthique *by design*[92], qu'il s'agisse de la protection de la vie privée (Iyengar, Kundu, et Pallis 2018; Azencott 2018; Cavoukian 2016); de la sécurité (Cavoukian 2016) ou encore de la transparence (Mascharka et al. 2018). Il semble qu'aucun de ces outils d'éthique d'appliquée ne soit réellement implémenté à l'heure actuelle, mais demeure « in the academic research stage » et demande encore beaucoup de travail - notamment pour la communauté de l'IA (Morley et al. 2019).

[92] Azencott (2018) présente par exemple une revue des techniques qui existent à l'heure actuelle pour partager les données et maximiser l'utilité scientifique tout en minimisant l'impact sur la vie privée des patients – soit offrir une protection appropriée.

Concernant les méthodes « non-techniques », elles peuvent référer à la traduction des principes en lois ou autres mécanismes juridiques. Par exemple, le Règlement Général de Protection des Données européen (RGPD[93]) a été adopté en 2016 en ce qui a trait à la gestion des données personnelles, qui constituent bon nombre de données massives – sans pour autant s'y restreindre. Ce règlement défend notamment un droit à l'oubli (qui implique de pouvoir supprimer les liens comportant des données à caractère personnel auprès d'un responsable de traitement) (Tambou 2016) et un droit à la portabilité (soit la possibilité pour l'utilisateur de pouvoir récupérer ses données dans un format lisible) (Maurel 2019). Cependant, un encadrement juridique de l'IA et des données massives se heurte à différents défis pratiques et conceptuels[94] (quelle que soit la juridiction) qu'il est encore nécessaire de résoudre (Danaher 2015; Scherer 2015).

S'il n'est pas possible de toutes les nommer, il existe différentes méthodes non-techniques (plus ou moins contraignantes) pour opérationnaliser les principes. Le White House Office of Science and Technology Policy (OSTP) américain, a par exemple formulé différentes recommandations en vue de guider les futurs développements de l'IA. L'OSTP recommande notamment aux institutions publiques et privées d'examiner de quelle manière elle pourraient tirer profit de l'IA et de l'apprentissage automatique tout en assurant que cela profite à la société ; que les agences fédérales priorisent les données ouvertes et créent des standards pour guider leur gestion ; d'assurer la sécurité et l'efficacité de leur utilisation ; ou d'intégrer dans la formation initiale des développeurs les sujets relatifs à l'éthique de l'IA tels que la vie privée, l'équité et la sécurité (OSTP 2016). L'OSTP fait également de l'éthique une de ses sept stratégies de recherche et développement, soit par la demande de comprendre les implications éthiques, légales et sociales du développement de l'IA et d'évaluer les systèmes d'IA par l'entremise de standards.

[93] Le RGPD est disponible ici : https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1532348683434&uri=CELEX:02016R0679-20160504

[94] Pour ne citer qu'eux, une règlementation de l'IA se heurte au problème de la discrétion (la recherche et le développement en IA pourraient avoir lieu à l'aide d'infrastructures difficilement visibles par les régulateurs) ; de la diffusion (les systèmes peuvent être développés par des équipes distantes tant sur le plan organisationnel que géographique) ou de l'opacité des réseaux de neurones (Danaher 2015). Elle se heurte également à un défi conceptuel, soit comment définir l'IA afin de l'encadrer et comment lui assigner une éventuelle responsabilité lorsqu'elle cause des dommages (Scherer 2015).

Cédric Villani, mandaté par le gouvernement français de « conduire une mission sur la mise en œuvre d'une stratégie française et européenne en intelligence artificielle »[95], présente également dans son rapport de nombreuses recommandations de politiques publiques. Il défend notamment la nécessité d'ouvrir la boite noire de l'IA et de soutenir la recherche sur l'explicabilité, de développer une évaluation citoyenne et des audits des systèmes d'IA, d'intégrer l'éthique dans formation des ingénieurs, ou encore d'instaurer des études d'impact sur les discriminations. Le rapport recommande également la mise en place d'un « Comité consultatif national d'éthique pour les technologies numériques et l'intelligence artificielle » afin de développer une gouvernance « spécifique à l'IA » (Villani 2018). La CERNA publie également le rapport : *La souveraineté à l'heure du numérique : rester maître de nos choix et de nos valeurs* en octobre 2018 qui synthétise les débats ayant eu lieu lors de plusieurs journées d'étude, et formule huit recommandations et huit suggestions de nature politique, comme par exemple de mettre en place des moyens de souveraineté scientifique ou de partage équitable des données entre disciplines académiques (CERNA 2018a). Le rapport de la Déclaration de Montréal recommande quant à lui la mise en place d'un organisme indépendant de veille et de consultation citoyenne, une politique d'audit et de certification des systèmes d'IA, la mise en place de formations en éthique pour tous ou encore de s'engager vers la mise en place de politiques publiques qui répondent à l'urgence environnementale en minimisant notamment l'impact énergétique du numérique (Déclaration de Montréal IA Responsable 2018).

L'opérationnalisation des principes de l'éthique de l'IA peut ainsi se faire par l'entremise de la création de différents types de comités ou d'organismes de surveillance, la formation et la sensibilisation de la population ou des professionnels, ou la mise en place de mécanismes de normalisation et de certification – notamment sur la base du développement de métriques communs d'évaluation (CERNA 2018b; Villani 2018; Déclaration de Montréal IA Responsable 2018; Floridi et al. 2018). Plusieurs des recommandations des rapports cités précédemment en appellent également à une autorégulation, qui se traduirait par une culture de la réflexivité pour les développeurs, les entreprises et autres acteurs du développement des systèmes d'IA; laquelle serait

[95] Voir : https://www.ladocumentationfrancaise.fr/rapports-publics/184000159/index.shtml

encouragée par différents incitatifs, lignes directrices et codes de conduites d'ordre déontologique (Floridi et al. 2018; IEEE 2017; CERNA 2018b)[96].

L'opérationnalisation des principes n'est cependant pas sans défi. Relativement à l'éthique de l'IA, différents auteurs dénoncent en effet un décalage entre principes et pratique qui semble difficile à dépasser (Mittelstadt 2019; Morley et al. 2019). Un des principaux arguments est que la multiplicité des interprétations possibles fait des principes de l'éthique de l'IA des concepts ambigus, difficiles à traduire en normes, en particulier lorsqu'il s'agit de les intégrer dans les architectures des systèmes d'IA (Morley et al. 2019). C'est notamment ce que défend Mittelstadt (2019) :

> AI Ethics initiatives aim to address this gap by defining broadly acceptable principles to guide the people and processes responsible for the development, deployment, and governance of AI across radically different contexts of use. This may be an impossible task. The great diversity of stakeholders and interests involved necessarily pushes the search for common values and norms towards a high level of abstraction. The results are statements of principles or values based on abstract and vague concepts, for example commitments to ensure AI is 'fair', or respects 'human dignity', or enables 'human flourishing', which are not specific enough to be action-guiding (p. 5).

Ainsi, si 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), ils en deviendraient par la même occasion des outils ambigus et difficilement utilisables et traductibles en normes ou règles. C'est cependant bien ce niveau d'abstraction qui différencie les principes des règles ou des normes (Massé 2003).

Ces considérations demandent alors de se pencher sur deux polarisations de l'éthique appliquée[97] (toutefois relativement simplificatrices), auxquelles s'est notamment confrontée la

[96] Plusieurs initiatives relatives au développement de lignes directrices éthiques pour un développement responsable de l'IA sont actuellement en cours au Québec. Pour ne citer qu'elles, la Commission de l'éthique en science et en technologie du Québec travaille actuellement sur l'éthique du développement de l'IA en santé dans le cadre d'une collaboration avec la France et les Fonds de recherche du Québec ont mis sur pied un Comité de travail sur les aspects d'éthique de la recherche dans les domaines du numérique, de l'intelligence artificielle et des données massives.

[97] Le terme « éthique appliquée » a cependant soulevé suffisamment de polémiques, notamment relativement aux polarisations présentées, que le Conseil de recherche en sciences humaines du Canada lui a préféré le terme d'« éthique sectorielle » (Durand 2005; Doucet 1999).

pratique de la bioéthique, et dans lesquelles se retrouvent les difficultés de l'opérationnalisation des principes « abstraits » de l'éthique de l'IA. La première concerne la portée des principes et réfère à une apparente opposition entre universalisme et relativisme éthique, qui demande de se pencher plus spécifiquement sur les enjeux du respect du pluralisme éthique qui s'observe au sein des sociétés et entre celles-ci. Cette polarisation est liée à une deuxième opposition qui concerne la manière de faire de l'éthique, et oppose une application déductive de principes (potentiellement considérés comme universels) à une approche inductive sensible au contexte.

3.2. Difficultés associées à l'identification de la portée des principes

La gouvernance éthique de l'IA, qui tend à se définir en prenant comme point de départ des principes directeurs généraux, ravive ainsi les débats relatifs à une opposition schématique entre les défenseurs de l'existence de valeurs morales universellement partagées (ex. Kluckhohn 1955; Shaw 2000) et ceux d'un certain relativisme éthique qui promeut que la notion de « bien » ou de « mal » est dépendante d'un système moral présent dans un groupe ou société donnés – aucun système ne pouvant prévaloir sur un autre (Harman 1975; Massé 2003). Si le relativisme éthique s'est cependant très souvent vu critiqué pour l'inconsistance de ses arguments, Harman (1975) en présente une conception plus précise. Selon sa vision, la moralité découle d'un accord implicite au sein d'un groupe de personnes sur la base de critères issus de leurs obligations réciproques (Harman 1975). Puisqu'il n'est pas question de discuter des fondements théoriques d'une telle opposition, seront plutôt présentées les convergences (qui renvoient à une certaine idée de l'universalisme éthique) et les divergences (qui renvoient, elles, plutôt au pluralisme[98] qu'au relativisme) de l'éthique de l'IA. S'il est essentiel de se pencher sur ces convergences et divergences éthiques, c'est qu'une coordination internationale semble inévitable relativement à la gouvernance éthique de l'IA.

3.2.1. De la nécessité d'une coordination internationale

Que la gouvernance éthique de l'IA se définisse sur la base de principes universels ou non, la nécessité d'une coopération internationale semble de mise, considérant, comme mentionné dans les précédents chapitres, que la numérisation de la société ne connaît pas de frontières. L'urgence de cette coopération est d'ailleurs soulignée par Jobin, Ienca, et Vayena (2019) :

[98] La notion de pluralisme éthique renvoie plutôt à l'existence d'une diversité de valeurs ou de différentes manières de hiérarchiser les principes, lesquelles sont observables entre différents groupes ou cultures (Massé 2003).

> At the policy level, they urge increased cooperative efforts among governmental organisations to harmonize and prioritize their AI agendas, an effort that can be mediated and facilitated by inter-governmental organisations (p. 16)

Il peut en effet paraître vain, de prime abord, de mettre en place une gouvernance régionalisée de la gestion des données massives alors qu'il semble aujourd'hui difficile de circonscrire les frontières de leur collecte[99]. L'analyse qui en est faite par les systèmes d'IA ne peut pas non plus se limiter à un usage prédéfini ni à une région du monde, considérant les innombrables réutilisations que les bases de données permettent, étant de plus en plus accessibles[100]. Également, l'influence des compagnies qui détiennent un quasi-monopole sur les développements de l'IA connaît une portée mondiale non-négligeable et les conséquences de la mise en place de mesures restrictives, même géographiquement délimitées (ex. RGPD), ont des répercussions à l'international. Une harmonisation de la gestion des risques éthiques de l'utilisation des systèmes d'IA semble alors incontournable.

L'organisation diffuse du développement de l'IA est en effet un enjeu de taille pour toute mesure de gouvernance et défie – quand elle ne rend pas caduque – la mise en place de mécanismes de gouvernance locaux. C'est par exemple ce que souligne Danaher (2015) dans le cas de la mise en place de dispositions juridiques, discutant de ce qu'il nomme le « *diffuseness problem* » :

> It is the problem that arises when AI systems are developed using teams of researchers that are organisationally, geographically, and perhaps more importantly, jurisdictionally separate. Thus, for example, I could compile an AI program using researchers located in America, Europe, Asia and Africa. We need not form any coherent, legally recognisable organisation, and we could take advantage of our jurisdictional diffusion to evade regulation.

Si la gouvernance de l'IA semble incontestablement un enjeu de politique internationale, il est cependant légitime de questionner la manière selon laquelle ces principes éthiques pourront être opérationnalisés en concordance, tout en respectant l'autonomie décisionnelle des pays du monde. D'un côté, une certaine convergence des principes éthiques développés à l'international vient soutenir la possibilité d'une gouvernance mondiale basée sur des principes universels et partagés, laquelle impliquerait néanmoins une collaboration étroite entre les différents pays. D'un autre, le

[99] Voir le Chapitre 2 et le Chapitre 3 sur l'ubiquité et la portabilité des systèmes d'IA.
[100] Voir le Chapitre 3, Section 1.2. sur la nécessité de partage et Section 2.2. sur les enjeux de l'utilisation secondaire des données massives.

respect d'un pluralisme éthique et de son expression au sein des sociétés et entre celles-ci demande de prêter attention aux conditions de la portée universelle des principes directeurs de l'éthique de l'IA, notamment dans leur opérationnalisation.

3.2.2. Les convergences de l'éthique de l'intelligence artificielle

Relativement à la recherche en IA ou à la gestion des données massives, la nécessité d'une approche globale a été mise de l'avant à plusieurs reprises. Pour différents auteurs, la littérature actuelle concerne essentiellement des applications spécifiques, problématisant la responsabilité à un aspect particulier et de façon *ad hoc,* n'offrant ainsi que peu de directives générales tant relatives au rôle que l'IA doit jouer dans nos sociétés que pour les chercheurs aux domaines d'applications différents (Brundage 2016; Floridi et Taddeo 2016; Cath et al. 2016). Face au besoin d'une approche générale et d'une vision globale et proactive des risques (Brundage 2016; Cath et al. 2016; Floridi et Taddeo 2016), plusieurs soutiennent qu'il est nécessaire de mettre en place un code d'éthique universel – ou d'autres solutions aux répercussions mondiales – en ce qui a trait à l'éthique de l'IA. C'est par exemple le cas de l'UNESCO, qui promeut le développement d'un « code d'éthique mondial » pour la recherche en IA[101]; du IEEE qui vise à devenir « a leader in global ethics » en mettant en place différents mécanismes d'inclusion interdisciplinaires et multiculturels (Mattingly-Jordan 2017) ; mais également de différents auteurs qui proposent la création d'un système universel de l'éthique de l'IA (Waser 2008) ou d'un cadre réglementaire international uniformisé sur la base d'une collaboration entre les différents pays du monde (Erdélyi et Goldsmith 2018).

Une certaine forme d'universalisme (ou tout du moins de portée universelle) est également défendue par certains des acteurs du domaine de l'éthique des machines, qui vise à introduire des valeurs morales dans les architectures des réseaux de neurones afin que les systèmes d'IA puissent agir et prendre des décisions sur la base de valeurs humainement partagées. Afin de construire une IA sécuritaire quelles que soient les conséquences (incluant celles que les concepteurs n'ont pas explicitement envisagées au départ), il est en effet nécessaire de définir en amont ce qu'est un « bon » comportement, de manière à s'assurer que ces conséquences ne soient pas néfastes pour

[101] Voir : https://en.unesco.org/courier/2018-3/towards-global-code-ethics-artificial-intelligence-research

les humains (Bostrom et Yudkowsky 2011). Dans l'idéal, ces valeurs morales reflèteraient donc celles de « l'humanité » (Shulman, Jonsson, et Tarleton 2009). Dans le même ordre d'idée, l'IEEE défend également la nécessité que les systèmes d'IA soient développés selon des impératifs moraux partagés. Citant le Future of Life Institute, l'IEEE mentionne que les systèmes d'IA « should only be developed in the service of widely shared ethical ideals, and for the benefit of all humanity rather than one state or organization » (Futur of Life Institute 2017 dans IEEE 2017 p. 75).

Or, la mise en évidence d'une convergence éthique quant aux déclarations de principes relatives à l'IA semble soutenir l'idée de la mise en place de principes ou de solutions universelles[102]. Par exemple, dans leur cartographie du corpus des principes et lignes directrices éthiques relatives à l'IA, Jobin, Ienca, et Vayena (2019), qui ont analysé 84 des documents produits à travers le monde, révèlent une convergence de ces documents vers cinq principes éthiques : transparence, justice et équité, non-malfaisance, responsabilité et vie privée[103] (Jobin, Ienca, et Vayena 2019). Les auteurs rapportent également que, de manière générale, la communauté actuelle de l'éthique de l'IA priorise une certaine obligation morale de prévenir les dommages plutôt que la promotion du « bien » (Jobin, Ienca, et Vayena 2019). De manière similaire, les auteurs du projet *AI4People* soulèvent une convergence de l'éthique de l'IA vers quatre principes de référence en bioéthique : bienfaisance, non-malfaisance, autonomie et justice (Floridi et al. 2018). Leur analyse comparative a cependant mis en évidence la nécessité de créer un autre principe, celui de l'explicabilité, compris comme intégrant à la fois l'intelligibilité et l'imputabilité (Floridi et al. 2018). Selon Mittelstadt (2019), cette convergence vers les principes bioéthiques est également soulevée par l'OCDE et par la Commission Européenne :

> This position was subsequently endorsed by the OECD and the European Commission's High Level Expert Group on Artificial Intelligence (HLEG), which proposed four principles to guide the development of 'trustworthy' AI: respect for human autonomy, prevention of harm, fairness, and explicability (p. 2).

[102] Cette convergence n'est cependant pas surprenante car c'est parfois l'objectif annoncé de certaines des initiatives présentées.

[103] Il s'agit des catégories de convergences créées pas les auteurs. Relativement à la fréquence d'apparition dans les documents analysés, les principaux principes sont : transparence, justice et équité, non-malfaisance, responsabilité, vie privée, bienfaisance, liberté et autonomie, confiance, dignité, soutenabilité et solidarité (Jobin, Ienca, et Vayena 2019).

Les auteurs du *Moral Machine Experiment*[104] ont également mis en évidence des convergences éthiques en faveur du développement d'une éthique des machines universelle, ayant identifié des préférences éthiques partagées à travers l'ensemble des millions de personnes issues de 233 pays ayant participé à leur étude (Awad et al. 2018). Ils soutiennent également que les variations individuelles ne sont que peu utiles pour les décideurs politiques, n'ayant pas d'impact mesurable, et que les désaccords observés lors de l'expérience ne sont pas inéluctables.

La convergence des principes éthiques ne saurait évidemment pas à elle seule justifier l'universalisme de leur portée. En bioéthique, les conditions qui autorisent l'universalisme de l'éthique appliquée ont déjà été critiquées (Metz 2010b; 2010a). Différents auteurs défendent des approches qui demandent de tenir compte des caractéristiques culturelles, historiques et personnelles relatives aux différentes situations dans lesquelles des enjeux éthiques émergent, posant des défis supplémentaires à l'application de principes éthiques généraux (Cleret de Langavant 2001; Durand 2005; Foy et al. 2011; Toulmin 1981; Massé 2003). Bien qu'il soit possible de condamner certaines pratiques même si elles sont issues de traditions culturelles (Macpherson et Macklin 2010), plusieurs ont dénoncé le risque que la généralisation des principes conduise à une certaine « myopie culturelle » (De Vries 2011) voire constitue une nouvelle forme d'impérialisme occidental qui risque de détourner la bioéthique des conditions initiales du problème qu'elle tente de résoudre, notamment les causes culturelles et structurelles qui en sont à l'origine, et la conduise à « abuser de la théorie morale » (De Vries 2011). Il est ainsi nécessaire, en vue de l'opérationnalisation des principes, de porter une attention particulière au divergences potentielles de l'éthique de l'IA.

3.2.3. Les divergences de l'éthique de l'intelligence artificielle

Il existe des divergences potentielles non-négligeables dans l'interprétation des principes de l'éthique de l'IA (Jobin, Ienca, et Vayena 2019; Mittelstadt 2019). D'abord, il est possible de questionner la représentativité (mondiale) des principes éthiques qui émanent des initiatives

[104] Le *Moral Machine Experiment* est une expérience réalisée en 2016-2017, visant à jauger les attentes normatives et la moralité publique par l'entremise d'une plateforme en ligne permettant d'explorer les dilemmes moraux des véhicules autonomes. L'expérience a mis en évidence des préférences morales généralisées (ex. sauver des humains plutôt que des animaux ou sauver des vies plus jeunes) et trois regroupements moraux distincts (Regroupement Occidental, Oriental et du Sud) qui présentaient des convergences morales d'autant plus significatives, corrélées aux variations économiques et culturelles entre pays.

présentées, en particulier en ce qui concerne la portée de la convergence mentionnée. En effet, une sous-représentation de certains pays et cultures s'observe apparemment relativement aux différentes initiatives présentées précédemment. Par exemple, la revue des 84 documents de Jobin, Ienca, et Vayena (2019) montre une forte représentation des pays aux revenus élevés : ils proviennent pour plus du tiers d'entre eux des États-Unis et du Royaume Uni, suivi par le Japon et différents pays d'Europe. Les pays d'Afrique et d'Amérique du Sud ne sont aucunement représentés dans cet échantillon, sauf par le biais des organisations internationales (ex. OCDE). Il se pourrait ainsi que la convergence éthique en ce qui a trait à l'IA soit essentiellement occidentale. Cependant, l'existence d'initiatives similaires dans d'autres pays, comme en Chine[105] ou au Mexique[106] ainsi que les préférences morales généralisées mises en évidence dans le *Moral Machine Experiment* demande d'explorer plus en profondeur une possible convergence mondiale.

La diversité culturelle pourrait en effet conduire à des divergences non-négligeables dans l'interprétation des principes. Goodman (2016) rappelle par exemple, dans le contexte des biobanques et des données génétiques, que la valeur associée à la nature privée des données (*privacy*) n'est pas universelle. Ce qui est privé se définit selon l'auteur en fonction de préférences individuelles et de normes culturelles. Concernant le respect de la vie privée et de la confidentialité en santé, l'approche africaine diffère par exemple de l'approche occidentale (Metz 2010b; 2010a) relativement au « consentement de groupe », qui implique l'inclusion des proches et de la famille dans les prises de décisions médicales en opposition avec les valeurs individualistes occidentales (Metz 2010a; Graboyes 2010).

Relativement aux enjeux de l'opérationnalisation des principes de l'éthique de l'IA, le pluralisme n'est cependant pas seulement relatif à la culture. Il est également présent au sein même du champ de l'éthique et réfère aux dissensus théoriques existants entre les experts et approches du domaine. Gordon (2019) souligne qu'il s'agit d'un défi majeur pour les développeurs qui visent

[105] Comme les *Beijing AI Principles, cf.* Tableau 6.

[106] Commandité par l'Ambassade britannique au Mexique et développé par C Mind, Oxford Inshights et le gouvernement mexicain, le rapport « *Towards an AI Strategy in Mexico : Harnessing the AI revolution* » est sorti en juin 2018. Il présente des recommandations similaires à celles des rapports susmentionnés, et propose notamment de créer un Conseil Mexicain de l'Éthique de l'IA. Ces recommandations sont basées sur l'analyse qualitative d'entrevues réalisées avec plus de 60 experts de différents domaines.

la création de systèmes d'IA éthiques. Le pluralisme théorique les confronte en effet à des enjeux méthodologiques de taille, en particulier face aux dilemmes moraux qui ne trouvent pas les mêmes solutions en fonction des approches (ex. déontologie, casuistique, utilitarisme, éthique de la vertu) (Gordon 2019), ce qui complique les tentatives d'une éthique *by design* qui demanderait des critères précis en vue d'une certaine systématisation des conditions de l'éthique au sein des architectures de réseaux de neurones.

Les divergences de l'éthique de l'IA réfèrent alors à la fois à un pluralisme culturel et théorique. C'est également un des constats de Morley et al. (2019), qui défendent que le pluralisme au niveau de l'interprétation des principes est un problème tant pour leur mise en application que pour l'évaluation d'une opérationnalisation réussie :

> Producing tools to fill in the white space on the typology is likely to be challenging. There is a distinct lack of agreement on what the aims of such tools would be. Key terms such as 'fairness', 'accountability', 'transparency', and 'interpretability' have myriad definitions, and sometimes (e.g. in the case of 'fairness') many statistical implementations that are not compatible and require informed decisions about trade-offs (p. 11).

Rien qu'en ce qui a trait à la non-malfaisance, les interprétations qui en sont faites dans les différents rapports réfèrent à différentes conceptions du préjudice (Jobin, Ienca, et Vayena 2019). Le choix des valeurs éthiques à inscrire dans les algorithmes peut alors sembler difficile relativement à l'absence de consensus dans leur interprétation.

Bien qu'elle démontre l'existence d'une certaine convergence, l'analyse comparative de Jobin, Ienca, et Vayena (2019) révèle également qu'il existe différents niveaux de divergences tant dans l'interprétation des principes que dans les moyens à mettre en place pour les respecter :

> Our thematic analysis reveals substantive divergences in relation to four major factors: (i) how ethical principles are interpreted, (ii) why they are deemed important, (iii) what issue, domain or actors they pertain to, and (iv) how they should be implemented. Furthermore, unclarity remains as to which ethical principles should be prioritized, how conflicts between ethical principles should be resolved, who should enforce ethical oversight on AI and how researchers and institutions can comply with the resulting guidelines (p. 13-14).

C'est alors dans l'opérationnalisation des principes que pourrait s'exprimer le pluralisme éthique, ce qui est encouragé par la grande diversité des propositions relativement à leur opérationnalisation

(*cf.* section 3.1. du présent chapitre). C'est également dans l'opérationnalisation des principes qu'il est possible de mettre à mal le respect du pluralisme éthique, ce que reconnaît d'ailleurs Massé (2003), discutant de la polarisation entre portée universelle des principes et relativisme éthique relativement à leur mise en application dans le contexte de l'éthique de la recherche internationale :

> Ce sont moins les droits et principes à prétention universelle, en eux-mêmes, qui posent problème que le dogmatisme et l'absolutisme qui accompagnent leur application. Bref, même si les principes sont définis localement ou universellement, s'ils font l'objet d'une application sans sensibilité au contexte, de façon mécaniste, les dérapages au niveau de leur application les rendront non éthiques. […] Le débat sur le principisme doit départager le questionnement sur la pertinence des principes du procès à l'encontre des usages sociaux, politiques et éthiques qui en sont faits (p. 23)

Ainsi, c'est plutôt dans la pratique de l'éthique que vont se manifester les risques que soulèvent les critiques de l'universalisme (notamment, le nouvel impérialisme ou la myopie culturelle) ou celles d'une approche basée sur les principes. Ceci amène à se pencher sur la deuxième polarisation de l'éthique appliquée : celle qui relève de la manière de faire de l'éthique et qui oppose, entre autres, éthique inductive et déductive.

3.3. Considérations relatives à la manière de faire de l'éthique

L'opérationnalisation des principes de l'éthique de l'IA ne semble pas sans difficulté lorsqu'il est question de traduire ces principes « abstraits » en différents mécanismes de gouvernance ou d'éthique appliquée. Ces difficultés prennent en partie racine dans une opposition classique entre opérationnalisation inductive et déductive des principes et théories qui relèvent de l'éthique. Il est possible de distinguer deux conceptions de « l'éthique appliquée » : elle peut se comprendre soit comme une application des principes généraux aux cas concrets, de manière mécanique et déductive; soit comme une « éthique en pratique », concrète ou sectorielle qui démontre que l'éthique s'intéresse (indépendamment de la méthode de réflexion) aux situations singulières ou à la prise de décision (Durand 2005).

Comme mentionné par Massé, cette opposition peut s'illustrer dans les débats bioéthiques relatifs au principisme et à sa mise en application. En effet, pendant de nombreuses années, la voix dominante de la bioéthique « privilégiait une éthique normative appliquée qui 'appliquait' mécaniquement des principes philosophiques déterminés d'avance à certains domaines

biomédicaux » (Cleret de Langavant 2001 p. 22). Ces principes sont, en bioéthique, généralement ceux du principisme de Beauchamp et Childress (2001) qui définissent quatre principes élémentaires en vue de la résolution de problème en bioéthique : l'autonomie, la non-malfaisance, la bienfaisance et la justice[107]. Le principisme, parfois considéré comme le fondement incontournable de l'éthique appliquée en bioéthique (Doucet 1999) a fait l'objet de plusieurs critiques, notamment considérant la possibilité de conflits entre les principes; le « mutisme » de l'approche relativement à la manière de justifier la primauté d'un principe sur un autre; le risque d'en faire une utilisation « mécaniste » dans son application à la prise de décision ou encore risquant d'exagérer les ressemblances entre les problèmes éthiques, sans tenir compte de la spécificité de chacune des situations et des processus qui conduisent à résoudre lesdits problèmes (Cleret de Langavant 2001; Toulmin 1981; De Vries 2011; R. B. Davis 1995; Clouser et Gert 1990; Lacroix et Létourneau 2000).

En bioéthique également, les principes ont été critiqués pour leur niveau d'abstraction qui rend difficile leur mise en pratique ou le risque d'absolutisme dans leur application (Massé 2003). Toulmin soulignait déjà, en 1981, l'opposition entre une « tyrannie des principes » associée à un certain universalisme éthique et un relativisme sans substance :

> These days, public debates about ethical issues oscillate between, on the one hand, a narrow dogmatism that confines itself to unqualified general assertions dressed up as "matters of principle" and, on the other, a shallow relativism that evades all firm stands by suggesting that we choose our "value systems" as freely as we choose our clothes (p. 31).

Face à cette opposition, l'auteur défend une approche casuiste pour une éthique opérationnelle et adaptée au contexte plutôt qu'une éthique basée sur des principes qui, en pratique, contraint les débats et est aveugle aux spécificités contextuelles (Toulmin 1981).

Relativement à l'éthique de l'IA, une certaine convergence entre les différents principes éthiques définis par les initiatives présentées précédemment (Jobin, Ienca, et Vayena 2019; Floridi

[107] Mittelstadt (2019) présente cependant cinq principes de références en éthique biomédicale : les quatre principes mentionnés et celui de la confidentialité. L'ouvrage de Beauchamp et Childress ne présente cependant pas explicitement un principe de « confidentialité » mais un ensemble de principes associés à la relation entre professionnels de santé et patients qui impliquent, entre autres, le respect de la véracité, de la fidélité, de la vie privée et de la confidentialité (Beauchamp et Childress 2001).

et al. 2018) ne devrait effectivement pas pour autant encourager une tendance à universaliser les réponses à apporter aux problèmes éthiques, ni conduire à la traduction des principes en formules d'application directe. Il est en effet nécessaire de se prémunir d'une éthique comme quête de justification, ce que pourrait encourager une éthique *by design* principiée (si elle est déductive) ou ce que Mittelstadt présente comme un « solutionnisme technologique ». Les préoccupations relatives à la gouvernance algorithmique – celle qui se fait « par » les algorithmes – demande de prêter une attention particulière au pluralisme. En effet, considérant l'impact des algorithmes et l'ampleur des conséquences de leurs décisions sur l'ensemble de la société, il serait potentiellement préoccupant qu'ils intègrent et valorisent une conception de l'éthique plutôt qu'une autre[108].

Une éthique déductive risquerait également de freiner les tentatives de réflexion sur les conséquences éthiques de la mise en application de dispositifs car il serait de toutes façons « éthiquement validés ». Dans la même veine que les critiques relatives à la quantification de l'étude des moralités qui entraîne la validation de projets « bioéthiquetés » (Néron 2017), l'éthique appliquée ne doit pas devenir un moyen de légitimer les dispositifs eux-mêmes mais devrait plutôt conduire à stimuler l'autocritique et laisser place au doute relativement aux conséquences éthiques de leur développement et de leur utilisation. Si cet aspect est important, c'est que l'éthique de l'IA se retrouve en effet en partie dans les actes quotidiens de ses acteurs, comme le défend Vézy (2018) discutant de la portée des gestes et décisions des « architectes de l'IA » :

> La production du sens de l'IA se réalise ainsi non seulement dans les rapports et recommandations à des fins stratégiques d'encadrement mais aussi dans les pratiques elles-mêmes de ces architectes de l'IA qui façonnent au quotidien des outils artificiellement intelligents.

Ceci demande alors de s'intéresser à la pratique de ces acteurs, qui naviguent selon Vézy dans une certaine ambiguïté, notamment s'ils ont à répondre à des impératifs d'application déductive.

L'éthique inductive ne réfère pas seulement à des motivations et attitudes individuelles (ici, celles des architectes de l'IA), mais peut également faire référence à différentes méthodes et manières de faire en vue d'appréhender les enjeux soulevés. Afin de respecter le pluralisme éthique

[108] Ce qui ne signifie pas qu'il n'est pas souhaitable que les algorithmes intègrent certaines considérations éthiques plutôt que d'autres, dans une certaine mesure.

et la diversité des valeurs mobilisées, plusieurs auteurs défendent en effet la nécessité d'une approche inductive (*bottom-up*) et contextuelle pour un développement responsable de l'IA (Jobin, Ienca, et Vayena 2019; Mittelstadt 2019; Morley et al. 2019), notamment parce-que l'IA réfère à de nombreux contextes d'implémentation différents et de nombreuses technologies différentes (Mittelstadt 2019). En bioéthique, ce souci du contexte et de la singularité de chaque cas se retrouve dans différentes approches qui se sont développées en parallèle du principisme[109] (Durand 2005; Cleret de Langavant 2001). Elles ne sont pas, par définition, incompatibles avec l'existence de principes éthiques généraux comme repères pour guider la réflexion, notamment si l'on considère que le rôle de l'éthique relève plus de l'aide à l'action que de l'intervention (Lacroix et Létourneau 2000).

Pour certains auteurs en bioéthique, même le principisme n'est pas en contradiction avec une approche inductive[110]. C'est par exemple, le cas du « principisme spécifié » proposé par Massé (2003) qui se base à la fois sur des principes directeurs et une application contextuelle :

> Au-delà d'une polarisation entre un principisme intégriste défendant l'imposition mécanique de valeurs universelles et un relativisme radical, nous défendrons la pertinence d'une approche fondée sur les valeurs phares et la discussion éthique, approche qui retient certaines composantes constructives d'un principisme spécifié, sensible aux contextes socioculturels et arrimée à une éthique de la discussion (p. 21).

C'est dans la même veine que St-Arnaud défend que les principes ne représentent pas une fin *per se* mais plutôt « la manifestation de la diversité des valeurs dont il faut tenir compte pour reconnaître la complexité du réel » (Saint-Arnaud 1999). Dans ce contexte, l'éthique appliquée ne réfère pas à une méthode déductive mais bien « au caractère concret et pratique de l'entreprise » (Saint-Arnaud 1999).

[109] Il s'agit par exemple d'approche qui relève de l'éthique de la vertu, qui demandent de s'intéresser aux individus plutôt qu'au savoir technique et principes objectifs; de la casuistique, qui demande de se pencher sur l'expérience vécue des malades et les conditions individuelles des patients; ou de l'éthique narrative, qui demande d'intégrer l'histoire du patient et le sens qu'il donne à son vécu (Durand 2005).

[110] Bien que Lacroix et Létourneau (2000) reconnaissent que l'élément déductif en constitue, à la base, l'élément clé, ce qui n'invalide pas forcément les critiques formulées à l'égard du principisme.

Ainsi, l'existence de principes universellement partagés n'est pas systématiquement associée à une application absolutisante. Des principes éthiques généraux pourraient servir de références universelles et leurs applications varieraient en fonction du contexte d'application des systèmes d'IA, comme le soutien Gordon dans le contexte de l'éthique de l'IA (ce qui renvoie aux considérations précédemment présentées relativement à la bioéthique appliquée) :

> When it comes to pluralism in ethics, it seems reasonable to avoid extreme moral relativism and to accept a firm core of universal moral norms that all human beings acknowledge (e.g. one must not commit murder or rape, insult other people, or violate the human rights of others). Beyond this core of moral norms, people should be free to act according to their particular moralities in their given community, as long as those particular norms do not conflict with the core of universal norms or with the lingua franca of international human rights (Gordon 2019 p.10).

Le pluralisme pourrait s'exprimer dans les différentes hiérarchisations possibles entre principes mais également dans les différentes interprétations que leur niveau d'abstraction autorise, ce qui a été défendu en bioéthique (Lacroix et Létourneau 2000). Les deux visions se réunissent alors autour de l'idée de l'existence de principes universels sur lesquels un certain consensus s'observe, qui pourrait guider les décisions politiques – en particulier relativement aux utilisations les plus risquées (ex. armes autonomes, piratage politique, véhicule autonome) comme cela a pu déjà s'observer pour d'autres situations à haut risque (ex. armes biologiques) (Erdélyi et Goldsmith 2018). Les déclarations de principes serviraient ainsi de guide international définissant des valeurs inaliénables, à l'image d'autres déclarations d'influence (ex. La Déclaration des droits humains ou la Déclaration d'Helsinki). S'il semble possible de trouver un équilibre entre le respect de la convergence des principes avec la pluralité de leurs interprétations dans la manière de faire de l'éthique, il semble alors qu'il est nécessaire de se pencher sur les défis du respect des mécanismes de mise en pratique des principes éthiques de l'IA – quelle que soit la méthode (inductive) privilégiée – notamment dans le contexte de la santé.

4. Pistes de réflexion au regard de la gouvernance éthique de l'intelligence artificielle en santé

Considérant le travail majeur déjà réalisé tant sur l'identification et la définition des principes de l'IA que ceux de la bioéthique, les défis de la gouvernance éthique de l'utilisation des systèmes d'IA en santé demandent de se pencher sur les difficultés de leur mise en pratique. Relativement à ce point, il est intéressant de se pencher sur le travail de Mittelstadt, qui décrit une éthique de l'IA qui serait,

selon l'auteur, potentiellement trop principiée pour réussir. Mittelstadt souligne quatre faiblesses potentielles de l'éthique de l'IA comparativement au domaine médical qui limiteraient l'impact d'une approche basée sur les principes (Mittelstadt 2019). Selon lui, comparativement au domaine médical, le domaine de l'IA manque :

> 1) **d'objectifs communs** (en médecine, il s'agit de promouvoir la santé et le bien-être des patients) **et d'obligations fiduciaires** (en médecine, celles-ci dérivent de la relation médecin-patient qui suppose la confiance et que les professionnels vont agir dans le meilleur intérêt des patients) clairement définis;
> 2) **d'histoire et de normes professionnelles ancrées** (en médecine, il s'agit des différents codes, standards et normes professionnelles qui existent depuis le Serment d'Hippocrate alors qu'il n'existe pas de normes actuellement bien définies relativement à ce qu'est être un « bon » développeur);
> 3) **de méthodes éprouvées pour traduire les principes en pratique** (ex. comités d'éthique, codes de conduites);
> 4) **de solides mécanismes de responsabilité juridique et professionnelle** qui autorisent des sanctions en cas de fautes (Mittelstadt 2019).

Pour Mittelstadt, ces différences entre le secteur de l'IA et celui de la médecine suggèrent que nous ne devrions pas encore célébrer un consensus autour de principes de haut niveau masquant un profond désaccord politique et normatif. Cependant, les quatre faiblesses soulignées amènent à deux principales considérations importantes lorsqu'il est question de la gouvernance éthique de l'IA en santé :

> 1) les quatre faiblesses définies comparativement aux forces d'une éthique appliquée au domaine médical ne semblent pas tenir compte des aspects disruptifs de l'innovation numérique en santé (décrits précédemment dans le Chapitre 2) qui impactent (et dépassent) le simple domaine médical;
> 2) La critique de Mittelstadt semble basée sur un fait contestable : celui que les principes et lignes directrices de l'éthique de l'IA s'adressent uniquement aux développeurs, ingénieurs ou chercheurs du domaine de l'IA, sans égard aux autres parties prenantes de son développement responsable.

La critique de Mittelstadt pourrait d'abord conduire à penser que l'opérationnalisation des principes de l'éthique de l'IA devrait être particulièrement prometteuse dans le cadre de la pratique clinique, car il existe déjà de nombreux mécanismes de gouvernance pour assurer l'opérationnalisation de principes éthiques dans le domaine médical. La convergence des principes de l'éthique de l'IA vers ceux de la bioéthique encourage aussi l'idée d'une application prometteuse, ne venant pas bouleverser les fondements théoriques du champ de la bioéthique ni de sa mise en pratique. Cependant, c'est sans compter les différents éléments disruptifs qui accompagnent l'innovation numérique en santé ou les points de rupture qui accompagnent l'utilisation des systèmes d'IA dans le domaine médical. Ces points de ruptures, comme décrit dans le Chapitre 2, concernent l'échelle de grandeur (tant concernant la taille des bases de données massives à gérer que la portée des décisions algorithmiques), l'entrée de nouveaux acteurs dans le système de santé, une exacerbation de l'automatisation des soins ou encore l'apparition de nouveaux lieux de soins et de collecte de données relatives à la santé. Tous ces éléments viennent ainsi défier le respect des normes en vigueur et des balises éthiques dans leur conception traditionnelle, notamment relativement au respect de la vie privée et de la confidentialité, du consentement, de la déshumanisation du soin et du patient ou encore de la sécurité, comme cela a été présenté en détail dans le Chapitre 3.

Cela étant dit, la similitude dans les principes directeurs de l'éthique de l'IA et de la bioéthique et les différents éléments soulevés par Mittelstadt permettent de suggérer qu'il n'est pas forcément nécessaire de créer de nouveaux mécanismes de gouvernance pour la gestion de l'utilisation éthique de l'IA en santé. Également, la bioéthique, qui s'est penchée sur l'application de principes au domaine biomédical, a pu développer différentes méthodes pour répondre aux tensions de la portée de l'application des principes ou celle qui apparaît entre éthique inductive et déductive; et son institutionnalisation et sa force normative varie grandement entre les pays (Durand 2005). Cependant, il semble essentiel d'adapter les mécanismes existants, considérant que l'utilisation des systèmes d'IA en santé vient exacerber différents risques et enjeux éthiques et revêt des spécificités qu'il est nécessaire d'explorer en vue d'une innovation responsable. Également, les difficultés de la mise en pratique de l'éthique ou de la bioéthique dépassent la simple dichotomie entre éthique appliquée inductive et déductive ou portée des principes universelle ou particulière. Comme le reconnaît Cleret de Langavant (2001), la complexité des problèmes éthiques

contemporains demande de se pencher sur des méthodes intégratives sensibles au contexte et à la contradiction en vue de leur compréhension, de leur définition et de leur résolution. Ces considérations soulèvent la nécessité de développer de nouveaux outils conceptuels et de nouvelles approches méthodologiques qui tiennent compte de la complexité des situations et des paradoxes qui y sont associés, notamment relativement à l'émergence potentielle de valeurs et de différentes dimensions relatives à l'écologie organisationnelle et celle de l'action d'acteurs – lesquels sont de plus en plus difficile à identifier (Cleret De Langavant 2001).

Ce point amène ainsi à la seconde considération. S'il est vrai que certaines des déclarations présentées s'adressent explicitement aux ingénieurs, développeurs et autres parties prenantes du domaine de l'IA (ex. CERNA 2018; IEEE 2018), les principes de l'éthique de l'IA ne sont pas toujours destinés exclusivement à ces acteurs. Pour plusieurs auteurs, la responsabilité du développement responsable de l'IA incombe en effet aux concepteurs et informaticiens créateurs des algorithmes en question (Brundage 2016; Alexiou, Psixa, et Vlamos 2011; Cath et al. 2016; Russell, Dewey, et Tegmark 2015; Sharkey 2008). Par contre, que les seuls développeurs portent le poids du développement responsable de l'IA ou celui du choix des valeurs éthiques qui influenceraient les individus et façonneraient notre société dans son ensemble pose différents problèmes – non seulement ceux de l'éthique appliquée présentés précédemment, mais aussi celui de la « surcharge morale » des membres de la communauté de l'IA que cette attribution de la responsabilité aux seuls concepteurs seuls pourrait engendrer (IEEE 2017).

L'innovation responsable en ce qui a trait au développement des systèmes d'IA ou des données massives relève en effet tant du secteur privé et des gouvernements que du milieu académique (Brundage 2016). Institutions publiques et entreprises privées ont ainsi un rôle à jouer dans le développement éthique de l'IA. Si les discussions autour de la gouvernance numérique ont souvent opposé ces deux types d'institution, certains auteurs ont identifié l'existence d'une troisième voie concernant la gouvernance de l'IA : celles des communs numériques, comme le discute Verdier et Murciano (2017) :

> Longtemps, les débats sur la révolution numérique ont opposé les États, présumés rigides et conservateurs, et les géants de la Silicon Valley, supposés ouverts et innovants. Mais cette aporie occulte la troisième voie ouverte par les communs numériques : Wikipédia,

> OpenStreetMap, Open Food Facts, les logiciels libres ou en *open source* comme Linux, Apache ou MySql existent de fait. Ni privés ni publics, produits et utilisés par des communautés actives de contributeurs qui en garantissent la pérennité et l'accessibilité, ces ressources constituent un pan majeur de l'économie numérique (p. 132).

À l'heure de la collecte ubiquitaire des données et des codes en *open source* c'est, dans une certaine mesure, la société dans son ensemble qui devient partie prenante du développement responsable de l'IA. Cette idée est d'ailleurs soutenue par les nombreuses démarches participatives qui ont conduit aux déclarations de principes présentées (ex. CNIL 2017; Déclaration de Montréal IA Responsable 2018; IEEE 2017; Dawson et al. 2019). Une volonté d'impliquer un large panel d'acteurs clés est manifeste dans la plupart des initiatives décrites, voire l'inclusion des citoyens par le biais de consultations ou d'appels à contributions dans plusieurs d'entre elles. Différents rapports et différents auteurs en appellent à la nécessité d'impliquer le public dans les discussions relatives à l'éthique ou à la gouvernance de l'IA, à la mise en place de mécanismes délibératifs ou de « saisines citoyennes » dans différentes régions du monde afin de faire vivre le débat mais également de s'entendre sur les désaccords relatifs à l'interprétation des principes (Jobin, Ienca, et Vayena 2019; Villani 2018; Déclaration de Montréal IA Responsable 2018; House of Commons Science and Technology Committee. 2016). L'implication du public se fait en vue d'explorer l'acceptabilité sociale des technologies en jeu mais également de favoriser une utilisation responsable des systèmes d'IA.

S'ajoutent à ces parties prenantes, dans le contexte de l'utilisation des systèmes d'IA en santé, les professionnels de santé eux-mêmes, qui n'évolueraient pas dans un milieu qui souffrirait des quatre faiblesses décrites par Mittelstadt; tout du moins dans le contexte de la pratique clinique, mais qui devraient tout de même faire face à l'utilisation de technologies de plus en plus complexes dont il devient difficile de saisir l'ensemble des conséquences (positives et négatives) associées à leur usage.

Enfin, l'apparition des praticiens du domaine de l'IA ou des données dans le parcours de soins demeurant un des éléments disruptifs précédemment mentionnés, les faiblesses mises en évidence par Mittelstadt restent pertinentes en vue d'une innovation numérique en santé. En plus des scientifiques des données et de nouveaux acteurs privés qui entrent de manière inédite dans le secteur de la santé (ici, principalement les GAFAM), il faut mentionner que des développeurs et

d'autres ingénieurs informatiques prennent maintenant part, bien que relativement indirectement, à la relation de soin. C'est ce que soulève le CCNE dans son avis Données massives et santé :

> S'ajoutant aux traditionnels acteurs publics et privés de la santé, de nouveaux acteurs interviennent dans la relation de soin, ainsi que sur le marché de la santé et du bien-être. Il s'agit d'abord des '*data scientists'*, qui interviennent quel que soit le domaine d'application. Ils occupent une place centrale puisqu'ils sont responsables de la gestion des données et de leur exploitation en vue de produire de nouvelles informations. Beaucoup d'initiatives viennent des patients eux-mêmes, qui recherchent et partagent une information médicale, à l'image de la plate-forme *PatientsLikeMe*; mais ce sont principalement des entreprises privées qui sont le moteur de l'innovation en matière de technologies numériques y compris dans le domaine de la santé (p.18).

Or ces acteurs ne répondent effectivement pas aux mêmes normes déontologiques que les professionnels de santé, en clinique comme en recherche. Par exemple, au Canada, les recherches du domaine de l'IA sont parfois conduites en dehors de l'évaluation des comités d'éthique de la recherche ou de la surveillance éthique des organismes responsables de la distribution et de la gestion des fonds de recherche (ex. bailleurs de fonds) car celles-ci se résume à la recherche sur des sujets humains ou sur des animaux. Ainsi, la gouvernance éthique de l'utilisation des systèmes d'IA en santé demande d'explorer les défis du respect des mécanismes d'opérationnalisation des principes déjà en place, afin de les adapter en vue d'une gouvernance éthique des systèmes d'IA en santé effective.

5. Conclusion

De très nombreuses initiatives à travers le monde ont permis de définir les principes directeurs d'une gouvernance éthique de l'IA, ayant entretenu le débat public sur les questions de l'éthique en général et de l'éthique de l'IA en particulier. Le travail relatif à l'identification des principes directeurs de l'éthique de l'IA étant conséquent, il est nécessaire que les initiatives à venir se tournent vers les défis de leur mise en application.

Lorsqu'il est question d'appliquer les principes, la gouvernance éthique des systèmes d'IA ravive des tensions de l'éthique appliquée, ici présentées comme la nécessité de respecter les convergences et les divergences potentielles de l'éthique de l'IA. Si une gouvernance mondiale semble de mise considérant que la collecte et l'analyse des données ne connaît pas de frontières,

mais également que nombreuses des parties prenantes du développement de l'IA sont des acteurs internationaux, elle se heurte au défi de donner préséance aux valeurs morales qui feraient globalement consensus tout en respectant la pluralité qui pourrait s'observer dans la mise en application et l'interprétation des principes. La diversité des modes d'opérationnalisation des principes semble encourager l'idée d'une mise en application plurielle, compatible avec l'existence de principes éthiques généraux qui permettraient d'entretenir la réflexivité des parties prenantes du développement responsable de l'IA, et devrait favoriser des approches inductives qui respectent la singularité des situations.

La gouvernance éthique de l'IA apparaît être l'affaire de tous - bien qu'il faille tenir compte des asymétries de pouvoir et de savoir non négligeables entre les différents acteurs - et ne saurait se limiter à un ordre professionnel en particulier considérant ses applications dans tous les secteurs, mais aussi considérant l'implication d'acteurs aux expertises variées pour un seul domaine d'application (notamment, ici, la santé). Si, contrairement à d'autres secteurs, le domaine médical connaît un ensemble de mécanisme ancrés relativement à la mise en application des principes éthiques, le caractère disruptif de l'avènement des systèmes d'IA en santé demande cependant de se pencher sur les éléments qui pourraient défier le respect des normes et cadres existants relativement à leur utilisation, soit ce qui pourrait nuire à l'exercice de la responsabilité des différents acteurs en jeu. S'il n'est pas possible de se pencher sur l'ensemble des aspects complexes de la gouvernance éthique de l'IA en vue d'une innovation responsable, il semble pertinent de se pencher sur l'un d'entre eux, soit ce qui pourrait défier la manière de faire de l'éthique. Relativement à ce point, les discussions ayant eu lieu lors de la coconstruction de la Déclaration de Montréal sont particulièrement intéressantes, permettant de mettre en évidence différents défis de l'exercice de la responsabilité face à l'utilisation des systèmes d'IA, qui risqueraient ainsi d'entraver la pratique de l'éthique, soit notamment l'application des principes identifiés.

Références bibliographiques

AI HLEG, (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.

Alexiou, Athanasios, Maria Psixa, et Panagiotis Vlamos. 2011. « Ethical Issues of Artificial Biomedical Applications ». Dans *Artificial Intelligence Applications and Innovations*, 297-302. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-23960-1_36.

Amnesty International, Access Now. 2018. « The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems ». RightsCon Toronto. https://www.accessnow.org/toronto-declaration.

Aquin, Mathieu d', Pinelopi Troullinou, Noel E. O'Connor, Aindrias Cullen, Gráinne Faller, et Louise Holden. 2018. « Towards an "Ethics by Design" Methodology for AI Research Projects ». Dans *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 54–59. AIES '18. New York, NY, USA: ACM. https://doi.org/10.1145/3278721.3278765.

Awad, Edmond, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, et Iyad Rahwan. 2018. « The Moral Machine Experiment ». *Nature* 563 (7729): 59. https://doi.org/10.1038/s41586-018-0637-6.

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): 20170350. https://doi.org/10.1098/rsta.2017.0350.

Beauchamp, Tom L., et James F. Childress. 2001. *Principles of Biomedical Ethics*. Oxford University Press. Fifth Edition.

Borrett, Donald S, Heather Sampson, et Ann Cavoukian. 2017. « Research Ethics by Design: A Collaborative Research Design Proposal ». *Research Ethics* 13 (2): 84-91. https://doi.org/10.1177/1747016116673135.

Bostrom, Nick, et Eliezer Yudkowsky. 2011. « The Ethics of Artifical Intelligence ». Dans *The Cambridge Handbook of Artificial Intelligence*, 316-35. Cambridge University Press.

Brundage, Miles. 2016. « Artifical intelligence and responsible innovation ». Dans *Fundamental issues of artificial intelligence*, Springer International Publishing Switzerland, 543-53. V.C. Müller.

Cardon, Dominique. 2018. « Le pouvoir des algorithmes ». *Pouvoirs* N° 164 (1): 63-73.

Cath, Corinne J. N., Sandra Wachter, Brent Mittelstadt, Mariarosaria Taddeo, et Luciano Floridi. 2016. « Artificial Intelligence and the "Good Society": The US, EU, and UK Approach ». SSRN Scholarly Paper ID 2906249. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=2906249.

Cavoukian, A. 2016. « International Council on Global Privacy and Security, By Design ». *IEEE Potentials* 35 (5): 43-46. https://doi.org/10.1109/MPOT.2016.2569741.

CCNE. 2019. « Données massives (big data) et santé : une nouvelle approche des enjeux éthiques ». Avis 130. Comité Consultatif National d'Éthique français. https://www.ccne-ethique.fr/sites/default/files/avis_130.pdf.

CERNA. 2018a. « La souveraineté à l'ère du numérique. Rester maîtres de nos choix et de nos valeurs ». http://cerna-ethics-allistene.org/digitalAssets/55/55708_AvisSouverainete-CERNA-2018.pdf.

———. 2018b. « Research Ethics in Machine Learning ». Research Ethics Board of Allistene, the Digital Sciences and Technologies Alliance. cerna-ethics-allistene.org/digitalAssets/54/54730_cerna_2017_machine_learning.pdf.

Cleret de Langavant, Ghislaine. 2001. *Bioéthique : Méthode et complexité*. Presses de l'Université du Québec. Québec.

Clouser, K. Danner, et Bernard Gert. 1990. « A Critique of Principlism ». *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine* 15 (2): 219-36. https://doi.org/10.1093/jmp/15.2.219.

CNIL (Commission nationale informatique et libertés). 2017. « Comment permettre à l'homme de garder la main ? Les enjeux éthiques des algorithmes et de l'intelligence artificielle ».

Danaher, John. 2015. « Philosophical Disquisitions: Is effective regulation of AI possible? Eight potential regulatory problems ». *Philosophical Disquisitions* (blog). 7 juillet 2015. http://philosophicaldisquisitions.blogspot.com/2015/07/is-effective-regulation-of-ai-possible.html.

Davis, Ernest. 2015. « Ethical Guidelines for a Superintelligence ». *Artificial Intelligence* 220: 121-24. https://doi.org/10.1016/j.artint.2014.12.003.

Davis, Richard B. 1995. « The Principlism Debate: A Critical Overview ». *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine* 20 (1): 85-105. https://doi.org/10.1093/jmp/20.1.85.

Dawson, D, E Schleiger, J Horton, J McLaughlin, C Robinson, G Quezada, J Scowcroft, et S Hajkowicz. 2019. « Artificial Intelligence : Australia's Ethics Framework. A Discussion Paper. » Australia: Data61 CSIRO. https://consult.industry.gov.au/strategic-policy/artificial-intelligence-ethics-framework/supporting_documents/ArtificialIntelligenceethicsframeworkdiscussionpaper.pdf.

De Vries, Raymond. 2011. « The Uses and Abuses of Moral Theory in Bioethics ». *Ethical Theory and Moral Practice* 14 (4): 419-30. https://doi.org/10.1007/s10677-011-9290-y.

Déclaration de Montréal IA Responsable. 2018. « Rapport de la Déclaration de Montréal pour un Développement Responsable de l'Intelligence Artificielle ». https://docs.wixstatic.com/ugd/ebc3a3_d806f109c4104c91a2e719a7bef77ce6.pdf.

Departement of Health and Social Care. 2018. « Code of Conduct for data-driven health and care technology ». Royaume-Uni. https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology.

Dignum, Virginia, Matteo Baldoni, Cristina Baroglio, Maurizio Caon, Raja Chatila, Louise Dennis, Gonzalo Génova, et al. 2018. « Ethics by Design: Necessity or Curse? » Dans *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 60–66. AIES '18. New York, NY, USA: ACM. https://doi.org/10.1145/3278721.3278745.

Doucet, Hubert. 1999. « Preface ». Dans *Enjeux éthiques et technologies biomédicales : Contribution à la recherche en bioéthique.*, 7-11. Montréal: PUM (Les presses de l'Université de Montréal).

Durand, Guy. 2005. *Introduction générale à la bioéthique : histoire, concepts et outils.* Éditions Fides.

Erdélyi, Olivia J., et Judy Goldsmith. 2018. « Regulating Artificial Intelligence: Proposal for a Global Solution ». Dans *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and*

Society, 95–101. AIES '18. New York, NY, USA: ACM. https://doi.org/10.1145/3278721.3278731.

Floridi, Luciano, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, et al. 2018. « AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations ». *Minds and Machines* 28 (4): 689-707. https://doi.org/10.1007/s11023-018-9482-5.

Floridi, Luciano, et Mariarosaria Taddeo. 2016. « What Is Data Ethics? » *Philosophical Transactions of the Royal Society A-Mathematical Physical and Engineering Sciences* 374 (2083): 20160360. https://doi.org/10.1098/rsta.2016.0360.

Foy, Robbie, John Ovretveit, Paul G. Shekelle, Peter J. Pronovost, Stephanie L. Taylor, Sydney Dy, Susanne Hempel, Kathryn M. McDonald, Lisa V. Rubenstein, et Robert M. Wachter. 2011. « The Role of Theory in Research to Develop and Evaluate the Implementation of Patient Safety Practices ». *BMJ Quality & Safety* 20 (5): 453-59. https://doi.org/10.1136/bmjqs.2010.047993.

Goodman, Bryce. 2016. « What's Wrong with the Right to Genetic Privacy: Beyond Exceptionalism, Parochialism and Adventitious Ethics ». Dans *The Ethics of Biomedical Big Data*, édité par Brent Daniel Mittelstadt et Luciano Floridi, 139-67. Law, Governance and Technology Series. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-33525-4_7.

Gordon, John-Stewart. 2019. « Building Moral Robots: Ethical Pitfalls and Challenges ». *Science and Engineering Ethics*, janvier. https://doi.org/10.1007/s11948-019-00084-5.

Graboyes, Melissa. 2010. « Fines, Orders, Fear . . . And Consent? Medical Research In East Africa, C. 1950s ». *Developing World Bioethics* 10 (1): 34-41. https://doi.org/10.1111/j.1471-8847.2009.00274.x.

Harman, Gilbert. 1975. « Moral Relativism Defended ». *The Philosophical Review* 84 (1): 3-22.

Hausman, Daniel M., et Brynn Welch. 2010. « Debate: To Nudge or Not to Nudge ». *Journal of Political Philosophy* 18 (1): 123-36. https://doi.org/10.1111/j.1467-9760.2009.00351.x.

House of Commons Science and Technology Committee. 2016. « Robotics and artificial intelligence ». London, UK. http://www.publications. parliament.uk/pa/cm201617/cmselect/cmsctech/145/145.pdf.

Ibekwe-Sanjuan, Fidelia. 2014. « Big Data, Big machines, Big Science : vers une société sans sujet et sans causalité ? » Dans *XIXème Congrès de la Sfsic. Penser les techniques et les technologies : Apports des Sciences de l'Information et de la Communication et perspectives de recherches.*, 1-10. Toulon, France. https://hal.archives-ouvertes.fr/hal-01066202.

IEEE, Institute of Electrical and Electronics Engineers. 2017. « Ethically aligned design - Version 2 - For Public Discussion ». https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf.

———. 2019. « Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems ». First Edition. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf.

Iphofen, Ron, et Mihalis Kritikos. 2019. « Regulating artificial intelligence and robotics: ethics by design in a digital society ». *Contemporary Social Science* 0 (0): 1-15. https://doi.org/10.1080/21582041.2018.1563803.

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

Jobin, Anna, Marcello Ienca, et Effy Vayena. 2019. « Artificial Intelligence: the global landscape of ethics guidelines ». *arXiv:1906.11668 [cs]*, juin. http://arxiv.org/abs/1906.11668.

JSAI. 2017. « The Japanese Society for Artificial Intelligence Ethical Guidelines ». Japon.

Kluckhohn, Clyde. 1955. « Ethical Relativity: Sic et Non ». *The Journal of Philosophy* 52 (23): 663-77. https://doi.org/10.2307/2022567.

Knight, Will. 2019. « Google employees are lining up to trash Google's AI ethics council ». *MIT Technology Review* (blog). 2019. https://www.technologyreview.com/s/613253/googles-ai-council-faces-blowback-over-a-conservative-member/.

Lacroix, André, et Alain Létourneau, éd. 2000. *Méthodes et interventions en éthique appliquée*. FIDES. Québec.

Macpherson, Cheryl, et Ruth Macklin. 2010. « Standards and Practices in a Diverse World: An Investigation into Shared Values ». *Developing World Bioethics* 10 (1): 30-33. https://doi.org/10.1111/j.1471-8847.2010.00278.x.

Mascharka, David, Philip Tran, Ryan Soklaski, et Arjun Majumdar. 2018. « Transparency by Design: Closing the Gap Between Performance and Interpretability in Visual Reasoning ». *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, juin, 4942-50. https://doi.org/10.1109/CVPR.2018.00519.

Massé, Raymond. 2003. « Valeurs universelles et relativisme culturel en recherche internationale: les contributions d'un principisme sensible aux contextes socioculturels ». *Autrepart* n° 28 (4): 21-35.

Mattingly-Jordan. 2017. « Becoming a Leader in Global Ethics Creating a Collaborative, Inclusive Path for Establishing Ethical Principles for Artificial Intelligence and Autonomous Systems ».

Maurel, Lionel. 2019. « Contre le pouvoir des plateformes, établir une portabilité sociale des données ? » Dans *Rapport annuel OPTIC "ETHICS & TECH 2019" - (Re)construire la confiance dans les technologies.* https://hal.archives-ouvertes.fr/hal-02144473.

Metz, Thaddeus. 2010a. « African and Western Moral Theories in a Bioethical Context ». *Developing World Bioethics* 10 (1): 49-58. https://doi.org/10.1111/j.1471-8847.2009.00273.x.

———. 2010b. « An African Theory of Bioethics: Reply to Macpherson and Macklin ». *Developing World Bioethics* 10 (3): 158-63. https://doi.org/10.1111/j.1471-8847.2010.00289.x.

Microsoft Corporation. 2018. « Responsible bots: 10 guidelines for developers of conversational AI ». https://www.microsoft.com/en-us/research/uploads/prod/2018/11/Bot_Guidelines_Nov_2018.pdf.

Mittelstadt, Brent. 2019. « AI Ethics – Too Principled to Fail? » SSRN Scholarly Paper ID 3391293. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=3391293.

Moor, J. H. 2006. « The Nature, Importance, and Difficulty of Machine Ethics ». *IEEE Intelligent Systems* 21 (4): 18-21. https://doi.org/10.1109/MIS.2006.80.

Morley, Jessica, Luciano Floridi, Libby Kinsey, et Anat Elhalal. 2019. « From What to How. An Overview of AI Ethics Tools, Methods and Research to Translate Principles into Practices ». *arXiv:1905.06876 [cs]*, mai. http://arxiv.org/abs/1905.06876.

Musiani, Francesca. 2013. « Governance by algorithms ». *Internet Policy Review*, août. https://policyreview.info/articles/analysis/governance-algorithms.

Néron, Adeline. 2017. « La Bioéthique, Science d'État. La fabrique du gouvernement de la morale des corps humains biomédicaux ». Paris, École des hautes études en sciences sociales.

O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

OSTP, (White House Office of Science and Technology Policy). 2016. « Preparing for the Future of Artificial Intelligence ».

Rouvroy, Antoinette, et Thomas Berns. 2013. « Gouvernementalité algorithmique et perspectives d'émancipation, Faced with algorithmic governmentality ». *Réseaux*, nº 177 (mai): 163-96. https://doi.org/10.3917/res.177.0163.

Russell, Stuart, Daniel Dewey, et Max Tegmark. 2015. « Research Priorities for Robust and Beneficial Artificial Intelligence ». *AI Magazine* 36 (4): 105-14.

Saint-Arnaud, Jocelyne. 1999. *Enjeux éthiques et technologies biomédicales : Contribution à la recherche en bioéthique.* PUM. Montréal.

Scherer, Matthew U. 2015. « Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies ». SSRN Scholarly Paper ID 2609777. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=2609777.

Scheutz, Matthias. 2016. « The need for moral competency in autonomous agent architectures ». Dans *Fundamental issues of artificial intelligence*, Springer International Publishing Switzerland, 517-27. V.C. Müller.

Sharkey, Noel. 2008. « The Ethical Frontiers of Robotics ». *Science* 322 (5909): 1800-1801. https://doi.org/10.1126/science.1164582.

Shaw, W.H. 2000. « Relativism and objectivity in ethics ». Dans *Ethical theory: a concise anthology.*, Broadview Press. Giersson, H., & Holmgren, M.

Shulman, Carl, Henrik Jonsson, et Nick Tarleton. 2009. « Machine ethics and superintelligence ». Dans , 95–97. Tokyo, Japan: Carson Reynolds and Alvaro Cassinelli.

Sonntag, Daniel. 2016. « Persuasive AI Technologies for Healthcare Systems ». Dans *2016 AAAI Fall Symposium Series*. https://www.aaai.org/ocs/index.php/FSS/FSS16/paper/view/14087.

Tambou, Olivia. 2016. « Protection des données personnelles: les difficultés de la mise en oeuvre du droit européen au déréférencement ». *RTDeur. Revue trimestrielle de droit européen* 2016 (2). https://hal.archives-ouvertes.fr/hal-01408535.

Toulmin, Stephen. 1981. « The Tyranny of Principles ». *The Hastings Center Report* 11 (6): 31-39. https://doi.org/10.2307/3560542.

UNESCO, et EQUALS Skills Coalition. 2019. « I'd blush if I could: closing gender divides in digital skills through education. » https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=1.

Verdier, Henri, et Charles Murciano. 2017. « Les communs numériques, socle d'une nouvelle économie politique ». *Esprit* Mai (5): 132-45.

Vézy, Camille. 2018. « L'éthique de l'intelligence artificielle : la faire au quotidien ». *Cahier d'Écoles* (blogue). 2018. http://cahier-ecole.com/l-etique-de-l-intelligence-artificielle.html.

Villani, Cédric. 2018. « Donner un sens à l'intelligence artificielle. Pour une stratégie nationale et européenne. » https://www.aiforhumanity.fr/pdfs/9782111457089_Rapport_Villani_accessible.pdf.

Waser, Mark R. 2008. « Discovering the Foundations of a Universal System of Ethics as a Road to Safe Artificial Intelligence. » Dans *AAAI Fall Symposium: Biologically Inspired Cognitive Architectures*, 195–200.

WHO, (World Health Organization). 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.

Chapitre 5 – Craintes et attentes citoyennes relatives à trois grands défis de l'exercice de la responsabilité face à l'utilisation des systèmes d'intelligence artificielle en santé

Les citoyens participants à la coconstruction de la Déclaration de Montréal ont exprimé différentes craintes et attentes relatives à l'exercice de la responsabilité (morale) face à l'utilisation des systèmes d'IA en santé. Ces craintes et ces attentes sont relatives à trois grands défis qui accompagnent cette utilisation : 1) la préservation des capabilités humaines; 2) le partage des responsabilités face à la multiplication des acteurs qui sont parties prenantes du parcours de soins et 3) l'agentivité attribuée aux systèmes d'IA, au détriment de celle des humains qui les utilisent. L'ensemble de ces craintes et attentes sont présentées dans le Tableau 7. Ces défis sont essentiels à considérer du point de vue de l'innovation numérique en santé responsable. Afin de rester au plus proche de la parole citoyenne, les discussions et les catégories qui s'y rattachent sont présentées dans leurs mots bien qu'elles renvoient à des notions et concepts de bioéthique ou de philosophie morale qui peuvent connaître d'autres appellations.

Tableau 7. – Craintes et attentes citoyennes face aux trois grands défis de l'exercice de la responsabilité

DÉFIS	CRAINTES			ATTENTES	
PRÉSERVER LES CAPABALITÉS HUMAINES	Incapacitation des professionnels de santé	• Perte de compétences • Dépendance à la technologie • Perte de confiance en soi		• Préserver l'autonomie décisionnelle • Éducation et formation des professionnels de santé et des patients	
	Incapacitation des patients	• Perte de la capacité à consentir de manière éclairée • Atteinte à la liberté de choix • Perte de l'esprit critique			
LE PROBLÈME DES MAINS MULTIPLES	Un grand nombre d'acteurs impliqués			Un contrat social	Responsabilité partagée : • De la recherche et des chercheurs • Des développeurs • Des utilisateurs • Des professionnels de santé • Des patients • Des entreprises • Des institutions publiques
				Identification des mécanismes existants	
	Craintes relatives aux conséquences des mains multiples	Propriété et protection des données	• Propriété des données • Marchandisation des données • Sécurité • Traçabilité • Asymétrie de pouvoir • Rétroaction	Attentes normatives relatives au partage des responsabilité	• Transparence des institutions • Précaution • Propriété et gestion des données

		• Vie privée et confidentialité	
		Perte de lien naturel entre professionnels de santé et patients Potentiels conflits d'intérêts	
AGENTIVITÉ ARTIFICIELLE	Les systèmes d'IA sont des agents Risques associés à la reconnaissance d'une agentivité artificielle	Les biais algorithmiques	• Les systèmes d'IA sont des outils • L'humain garde la main • Des systèmes d'IA transparents
	Une transformation du rapport à la technologie		
ENTRE CAPABILITÉS HUMAINES ET AGENTIVITÉ ARTIFICIELLE	Remplacement des humains par les machines La déshumanisation du soin	• Perte de contact humain • Portrait incomplet des patients • Perte de l'individualisation	Coopération humain-machine

1. Préserver les capabilités humaines

1.1. Technologies capacitantes et incapacitantes

Un des défis de l'exercice de la responsabilité devant l'avènement des systèmes d'IA en santé est relatif à la préservation de ce que les individus sont effectivement capables de faire et d'être - soit, leurs capabilités. Plus précisément, l'approche des capabilités est issue des travaux d'Amartya Sen (ex. Sen 2005; 1992), qui s'est penché sur la question de l'évaluation du développement et des inégalités, repris plus tard par Martha Nussbaum (ex. Nussbaum 2011). Comme le reconnait Oosterlaken (2015), qui reprend ladite approche en vue de l'évaluation technologique, les théories de Sen et Nussbaum permettent d'évaluer le progrès d'une manière nouvelle :

> These two thinkers both argue that assessment of development progress should not be made in terms of income or resource possession, but in terms of valuable individual human capabilities – or what people are effectively able to do and be (p. 2)

L'agentivité est un des concepts clés de l'approche des capabilités (Oosterlaken 2015). Ce concept est cependant polysémique. Il réfère généralement à l'exercice ou à la manifestation de la capacité d'agir (Schlosser 2015). La conception « standard » de l'action, bien que parfois controversée, offre selon Schlosser (2015) une théorie de l'agentivité :

> The standard theory of action provides us with a theory of agency, according to which a being has the capacity to act intentionally just in case it has the right functional organization: just in case the instantiation of certain mental states and events (such as desires, beliefs, and intentions) would cause the right events (such as certain movements) in the right way. According to this standard theory of agency, the exercise of agency consists in the instantiation of the right causal relations between agent-involving states and events.

Ainsi, cette conception de l'agentivité demande de se pencher sur les impacts de l'utilisation des systèmes d'IA sur la capacité (intentionnelle[111]) d'agir des individus qui les utilisent, en vue de préserver cette capacité.

[111] Il est à noter cependant qu'il existe d'autres conceptions de l'action d'un agent que celle liée à la notion d'intentionnalité (Schlosser 2015). Pour les fins de ce chapitre, cette conception est cependant suffisante, donnant les éléments essentiels à la compréhension des craintes et attentes citoyennes en vue de la préservation des capabilités.

Plus précisément, se basant sur les travaux de Sen, Oosterlaken (2015) définit ainsi l'agentivité : "agency refers to the ability that humans have to reflect on what they value, to set goals and to pursue the realization of those goals" (p. 5). Selon cette conception, il ne s'agit pas seulement de préserver la capacité d'agir de manière intentionnelle, mais de préserver la capacité à valoriser certains objectifs plutôt que d'autres et celle de les poursuivre. Dans la perspective de préserver les capabilités humaines, il est essentiel de se pencher sur les craintes et attentes citoyennes afin de dépeindre quels objectifs ceux-ci valorisent.

Selon le mouvement de la technologie appropriée - ou *appropriate technology movement* (ATM) qui se base également sur l'approche des capabilités, plus que de préserver les capabilités humaines, un développement technologique approprié devrait assurer l'expansion de celles-ci. Il n'est possible d'atteindre cet objectif, selon Oosterlaken, qu'en préservant l'agentivité des individus et en portant une attention spécifique aux particularités contextuelles (Oosterlaken 2015). Cette approche revêt une dimension éthique, car elle demande de tenir compte des préférences des individus, relativement à leur idée de la vie bonne[112] et leurs capacités de faire des choix en conséquence (Oosterlaken 2015). Ainsi, si l'on suit la vision de l'ATM, l'avènement des systèmes d'IA en santé doit se faire dans le sens d'une capacitation des individus, soit d'accroitre leurs capabilités, et demande de se pencher sur leurs préférences – ce qui inclut leurs craintes et leurs attentes – face au développement des technologies en jeu.

En ce qui a trait aux initiatives technologiques, l'ATM se base sur la question clé suivante : "Do such initiatives truly empower people – in all their human diversity- to lead the lives they have reason to value?" (Oosterlaken 2015, p. 41). Le progrès technologique n'est ainsi pas un progrès en lui-même, mais demande d'évaluer l'effet – positif ou négatif – sur les capabilités des individus.

[112] Si la vie bonne connaît de nombreuses interprétations, elle renvoie dans ce cadre plus particulièrement à la notion d'épanouissement. Selon van den Hoven (2012), la vie bonne est le point de convergence entre l'éthique et la conception (ou l'ingénierie) technologique. À cet effet, citant le philosophe Ortega et l'historien Basalla, l'auteur mentionne que l'objectif du développement technologique est la vie bonne, dans le sens qu'il doit permettre « d'améliorer » les façons de vivre des humains, afin de faire du monde un endroit plus facile à vivre en fonction de ceux que les individus valorisent, soit, il existe autant de technologies différentes que de façon de valoriser la vie (van den Hoven 2012).

Selon les acteurs et selon les usages, les systèmes d'IA peuvent ainsi soit augmenter les capabilités (ou « capaciter » les individus), soit les réduire (ou « incapaciter » les individus)[113]. Dans la littérature, différentes préoccupations relèvent effectivement des conséquences de l'automatisation des soins sur la perception qu'ont les professionnels de santé de leur expertise et de la perte de compétences associées (ex. Coeckelbergh 2015, 2010). D'autres s'inquiètent également de préserver la capacité des utilisateurs à comprendre les technologies en question (un des éléments essentiels afin de garder un contrôle sur les conséquences de leur utilisation) notamment considérant la « boite noire » de l'IA ou leur complexité croissante (cf. Chapitre 2 et section 2.2. du Chapitre 3 sur le consentement).

Or, comme le reconnaît Noorman (2016), il existe un lien tangible entre l'exercice de la responsabilité et l'agentivité. L'auteur soutient que la responsabilité, et plus particulièrement la responsabilité morale, concerne les actions humaines, les intentions à l'origine de ces actions et leurs conséquences, la personne ou le groupe à l'origine de l'action étant généralement identifiés à un « agent » :

> Generally speaking a person or a group of people is morally responsible when their voluntary actions have morally significant outcomes that would make it appropriate to blame or praise them.

Selon cette conception, l'utilisation de systèmes d'IA en santé pourrait défier l'exercice de la responsabilité en incapacitant les humains qui les utilisent, notamment car elle implique une relation causale entre les états de l'agent et les évènements. Une incapacitation des individus pourrait nuire au contrôle qu'ils ont sur les technologies en question, sur leur agir libre et, par extension, sur leur capacité à être responsables face aux conséquences de l'utilisation des systèmes d'IA[114].

Face à l'impact des systèmes d'IA sur les capabilités, les citoyens ont, sans surprise, soulevé des **craintes** relatives à l'incapacitation des acteurs du système de santé lors des discussions de la

[113] Il est également possible d'envisager que les technologies permettent de « préserver » les capacités existantes des individus, ce qui revient indirectement à assurer une capacitation.
[114] Les conditions et les conséquences d'une incapacitation sur la responsabilité sont cependant discutées plus en détails dans le Chapitre 6.

coconstruction de la Déclaration de Montréal. Ces craintes concernent soit l'incapacitation des professionnels de santé soit celle des patients. Concernant les professionnels de santé, cette incapacitation pourrait se manifester selon l'apparition d'une perte de compétence, d'une dépendance à la technologie ou d'une perte de confiance en soi face à l'avènement de systèmes de plus en plus performants. Concernant les patients, les citoyens craignent l'apparition d'une perte de la capacité à consentir de manière éclairée, une atteinte à la liberté de choix ou une perte de l'esprit critique. Concernant les **attentes** citoyennes, elles tendent vers une capacitation des professionnels de santé et des patients, en préservant l'autonomie décisionnelle, notamment par le biais de l'éducation et de la formation.

1.2. Craintes citoyennes : incapacitation des professionnels de santé et des patients

1.2.1. Incapacitation des professionnels de santé

Face au recours croissant aux systèmes d'IA, les citoyens craignent l'apparition d'une incapacitation des professionnels de santé, issue d'une perte de compétences, d'une certaine dépendance à la technologie voire d'une perte de confiance en soi.

Relativement à la ***perte de compétence*** potentielle des professionnels de santé, la crainte concerne une certaine perte de « dextérité », de leur « capacité de jugement » ou de la « finesse » de leurs analyses. Certains craignent que l'usage de systèmes d'IA entraine une perte de « connaissances » relatives au développement scientifique voire de « l'expertise » des professionnels de santé. Cette perte serait issue du fait que les professionnels de santé finiraient par « trop se fier » à l'IA, que l'IA finirait par faire « tout le travail », voire que les professionnels de santé en deviennent dépendants.

La crainte d'une perte de compétences est ainsi intrinsèquement liée à celle de la ***dépendance à la technologie***. Les citoyens participants ont défendu la nécessité de maintenir l'indépendance du médecin face aux recommandations de systèmes d'IA afin que ces derniers ne

les influencent pas; et de maintenir le niveau d'analyse du médecin au lieu de se fier uniquement à la machine. L'efficacité de l'IA risquerait ici de menacer l'autonomie des professionnels de santé et d'exercer une pression concurrentielle.

> P42 : Le point de vue médecin, je veux dire, quand tu prends le système d'aide à la décision. Les médecins vont se servir d'un outil qui leur fournit des bonnes réponses. À un moment donné, ils vont faire trop confiance à la machine. La machine, des fois, elle se trompe. Là, c'est toujours qu'ils soient vigilants, la machine est là juste en support. Elle va donner une réponse, mais on doit la questionner, parce qu'elle n'est pas fiable à 100%.

La dépendance à la technologie accompagnée de la perte potentielle de compétences amène alors des préoccupations relatives à une ***perte de confiance*** en soi, ou en sa propre expertise et intuition, notamment s'il advenait que les professionnels de santé aient plus confiance en celle de l'IA, d'autant plus dans les situations où les recommandations humaines et algorithmiques seraient contradictoires. Les citoyens ont parfois décrit un « effet pervers » de l'outil qui ne viendrait plus appuyer les recommandations du médecin, mais, au contraire, insécuriser le médecin.

> P28 : Le médecin vient qu'à se censurer lui-même pour pas avoir l'air fou.

À trois reprises, il a été mentionné que ces enjeux n'étaient cependant ***pas nouveaux*** : par exemple, les médecins peuvent déjà faire face à des dilemmes où les données statistiques ne vont pas dans le sens des recommandations basées sur leur expertise et qu'il s'agit dans le fond plus d'une « question de société que d'IA » en tant que telle, soit d'un effet de l'incertitude scientifique.

1.2.2. Incapacitation des patients

Les citoyens ont également manifesté des craintes relatives à l'incapacitation des patients, que celle-ci concerne la perte de la capacité à consentir de manière éclairée, l'atteinte à la liberté de choix, ou la perte de « pensée critique »; les trois étant intimement liées.

D'abord, une crainte de l'apparition d'une ***perte de la capacité à consentir de manière éclairée*** a été soulevée. Cette perte de capacité à consentir peut être issue de la complexité croissante des technologies en jeu (qui deviennent difficile à comprendre) ; de la complexité et de la longueur des clauses de confidentialité (en particulier pour les données collectées en dehors du système de santé) ou du manque de transparence des systèmes d'IA. Selon les participants,

l'accessibilité de l'information ne semble pas toujours garantie dans un monde de plus en plus numérique.

> P15 : Des fois, on a des applications […] et on a une politique d'utilisation, une politique de vie privée d'une centaine de pages. Je pense qu'il y a 0,001 % des gens qui vont la lire.

Également, le manque de traçabilité des données pourrait nuire à la capacité à consentir, à savoir qu'il est difficile de comprendre ou d'identifier à quelles fins les données sont collectées, comment et quand elles vont être utilisées et réutilisées.

> P42 : Parce que quand tu donnes ton consentement, ça va où dans 30 ans, ta donnée ? Comme patient, tu dois savoir où elle est ta donnée, qui l'a touchée, qui y a accès.

Cette perte de la capacité à consentir est parfois décrite comme une « naïveté » face à la collecte de données :

> P56 : On a une naïveté par rapport à qu'est ce qui est fait de nos données sans savoir qu'en utilisant des logiciels et tout ça, que tu donnes accès à tes données. Une naïveté. Là on est là : « Il faut de la confidentialité », mais tu utilises ton téléphone, tu es géolocalisé, tu as ta carte métro d'épicerie, ils savent ce que tu consommes… tout ça c'est vendu. Tu vas sur Facebook …

Pour certains, le consentement réfère à des décisions morcelées dans le temps et dans l'espace, qu'il semble impossible de généraliser à d'autres usages ou à d'autres périodes. L'obtention d'un véritable consentement explicite et la possibilité de se rétracter deviennent alors difficiles.

> P63 : Il faudra s'exprimer chaque jour de ta vie pour que l'algorithme ait les bonnes données à traiter.

> P36 : La difficulté qu'on a, c'est de consentir aujourd'hui pour demain, après-demain, dans cinquante ans, dans trente ans. Et on ne consent pas pour les mêmes choses. Ce qu'on met sur les médias sociaux versus ce qu'on donne à notre médecin versus ce qu'on met dans notre rapport d'impôts versus ce qu'on envoie à l'OMS par philanthropie des données… Toutes ces décisions-là, c'est des décisions morcelées qui sont prises à un moment où on ne sait pas à quoi on consent. Et on consent, et est-ce qu'on est capable de concevoir pour cinquante ans ? Non. Ce à quoi je dis oui aujourd'hui, dans trente ans, je vais peut-être dire non.

Les citoyens ont également soulevé des craintes relatives à une ***atteinte à la liberté de choix***, qui pourrait nuire à la capacité d'agir des patients. Différents choix seraient affectés dans le cadre de l'utilisation de système d'IA en santé : le choix de partager ou non ses données de santé; de

respecter ou non les recommandations algorithmiques; d'être mis au courant ou non des recommandations ou diagnostics issus de systèmes d'IA; ou enfin d'être placé en institution plutôt que d'avoir recours à un robot de soin à domicile.

> P68 : Est-ce qu'on donne le choix aux gens ou pas d'être informés de prédispositions en matière de santé ? Est-ce qu'éthiquement on doit les forcer à ça ?

> P4 : Si je dis 'Ben non, moi je ne veux pas que vous preniez ces données là pour faire du diagnostic.' Ben ça devrait être mon droit mais je ne devrais pas être pénalisé parce que je décide de faire ça. Ou si les données n'existent pas parce que je n'utilise pas Facebook, je n'utilise pas Instagram, je n'utilise rien de ça... donc est-ce que je vais être pénalisé parce que je n'utilise pas un certain service ? Non, je ne devrais pas l'être.

En effet, selon les citoyens, le choix ne doit pas se limiter au partage des données mais également au type de diagnostic. Par exemple, il a été mentionné que la population doit être libre de refuser certains diagnostics, notamment ceux qui seraient reçus par l'entremise de notifications.

> P41 : Il faut aussi donner à la population la liberté de ne pas vouloir partager des données, et à la limite, en disant que ma vie privée est maintenant un livre ouvert, est-ce qu'au moins, je peux décider : 'Textez-moi pas mon diagnostic'.

Concernant le choix d'être informé ou non des recommandations algorithmiques, il a été souligné que mettre au courant les patients ne représente pas en soi une manière de favoriser la capacité d'agir :

> P54 : Par exemple, il y a un ensemble de données qui, admettons qu'elles soient déjà analysées et qu'il y ait déjà un certain travail qui est fait pour faire ressortir des prédispositions. J'ai beau vouloir le savoir ou ne pas le savoir, encore faut-il que ces analyses-là me donnent des éléments qui soient significatifs, des éléments qui aient... ou bien qui me permettent d'agir sur cette base-là.

Plus largement, certains ont questionné la possibilité de choisir d'utiliser ou non les technologies en question, de pouvoir les « éteindre » en tout temps :

> P46 : Si le robot me dit de me laver les dents 3 fois par jour, à un moment donné, je veux avoir la possibilité de l'éteindre et de dire 'Fiche-moi la paix!'

Cependant, certains questionnent s'il est possible de choisir d'avoir recours à des systèmes d'IA dans un contexte où le choix des dispositifs de santé n'est pas forcément laissé aux patients, qu'il s'agisse ou non de systèmes d'IA :

> P36 : Les prothèses de hanche, quand ils font ça, il y a combien de modèles sur le marché ? Les gens ne les choisissent pas.

Cette atteinte à la liberté de choix pourrait se manifester par une forme de coercition relativement aux conséquences associées au fait de refuser de partager ses données de santé ou de suivre les recommandations de santé issues des algorithmes. Concernant le partage, les citoyens ont en effet mentionné à plusieurs reprises un « sentiment d'être forcé », de « se sentir obligé » ou que « l'incitatif est trop fort » à « tout communiquer ». Cette coercition se manifesterait notamment par un risque d'être pénalisé (ex. en ne se faisant pas ou moins rembourser) ou de se faire exclure du système de santé et de ses bénéfices en refusant de prendre part au partage de données.

> P56 : J'ajouterais un enjeu de contrôle, parce que, moi, c'est la première chose qui m'est venu. Je veux dire un enjeu de contrôle : on est en train de contrôler mon comportement en créant, en me forçant ou en m'obligeant à partager des données.

Certains ont mentionné que cette situation était particulièrement problématique lorsqu'il s'agissait de données issues des réseaux sociaux.

Les participants ont également soulevé qu'une forme de coercition pourrait apparaitre en pénalisant les individus qui refuseraient de suivre les recommandations algorithmiques et par extension, ils questionnent la liberté de choisir d'être en santé ou non. Quelles seraient les répercussions (en termes d'assurance par exemple) si un patient refuse de suivre les recommandations d'un algorithme ?

> P37 : Mon point, j'ai écrit juste une chose parce que c'est un peu pointilleux. Le respect de l'autonomie de la personne. Moi, ce qui m'a frappé, c'est un portrait très juridique, c'est à quel point tu peux garder l'autonomie de ce que tu veux comme traitement, de ce que tu veux comme vie dans un contexte où, à un moment donné, on te dit : « Voici les options pour toi ». Je trouve carrément ça étouffant. Tu vas avoir des conséquences. Non, tu vas garder ton autonomie, mais tu vas avoir des conséquences, tu ne seras plus assurable. On ne te donne pas de traitement, tu vas être plus loin sur la liste.

Certains ont souligné que ce type de coercition était déjà présent en santé et que les systèmes d'IA viendraient surtout exacerber le phénomène. Ils ont également mentionné un risque de « stigmatisation » des individus qui ne prendraient pas part au système de santé numérique et soulignent le besoin d'une « grande ouverture d'esprit » relativement aux choix individuels considérant l'aspect systématique des analyses issues de systèmes d'IA :

> P36 : La frontière est mince à dire : 'Oui, mais ceux qui font exprès, qui ne se tiennent pas en santé pis qui sont gros, on ne fait rien pour eux.' Je trouve qu'il y a un glissement facile. Dans la notion d'autonomie, j'aimerais qu'on garde cette idée que si on veut vraiment l'autonomie, ça veut dire qu'on accepte dans notre société toutes sortes de comportements, qu'on accepte toutes sortes d'attitudes...

Cette liberté de choix a été mise en relation avec la notion de consentement. Les citoyens ont soutenu qu'il est nécessaire de garantir que le consentement soit libre (pas seulement éclairé) et continu – certains ayant par exemple questionné s'il était possible de changer d'avis après avoir donné son consentement relativement au partage, et s'il était possible de « se retirer » ou de respecter la liberté des patients considérant, comme mentionné précédemment, qu'il s'agit de décisions morcelées dans le temps.

Des citoyens ont souligné un « retrait de l'autonomie de la personne », le choix des utilisateurs n'étant pas suffisamment pris en considération, ce qui risquerait d'éluder le patient :

> P31 : Moi, je pense que le principe d'autonomie n'est pas respecté dans ce cas-là, parce que tu as le droit à l'autodétermination comme patient. On est allé voir son petit fichier, on a mis ça dans la machine, ce que je comprends, et ça a fait des calculs magiques et il est arrivé à trois types de diagnostics, mais la communication patient-médecin doit toujours être présente. Moi, je pense que l'intervention de l'IA doit venir aider le médecin à prendre une décision, mais ne pas éluder le patient.

Certains ont cependant souligné que cette liberté de choix n'est « pas fondamentale » et que l'autonomie est toujours relative :

> P36 : Qu'est-ce qu'on n'a pas capté dans autonomie ? Pour moi, l'autonomie est toujours relative à une société, on est autonome dans les limites de ce que notre culture, nos institutions nous permettent de faire.

Et d'autres questionnent les conditions qui autorisent à remettre en question le respect de cette liberté – notamment lors d'un bénéfice collectif :

> P52 : Il y a un enjeu de ... si en obligeant les gens à fournir au moins certaines données de santé, on atteint une très bonne augmentation du bien-être collectif, c'est un véritable enjeu. Peut-être qu'on va décider effectivement d'obliger les gens à les donner.
> P54 : Justement, si tout le monde a intérêt à partager ses données pour avoir en contrepartie un bénéfice.
> P52 : Mais, des fois, on n'a pas un intérêt personnel à le faire, mais il y a un intérêt collectif. Là-dedans, il y a un enjeu.

Les citoyens ont ainsi remis en question la réelle capacité des individus à faire un choix dans le contexte de l'utilisation des systèmes d'IA en santé, et ont parfois même directement fait le lien avec l'exercice de la responsabilité (mais cette fois, en discutant également de la liberté de choix des professionnels de santé) :

> P52 : En fait, c'est qu'on restreint d'autant plus son autonomie, fait qu'on a une asymétrie de poids, on a des contraintes importantes qui font en sorte que la personne n'a pas vraiment le choix ...
> P56 : ... L'asymétrie des pouvoirs, je trouve que c'est bon.
> P52 : C'est une limitation de la véritable capacité de choix. On est encore dans un enjeu d'autonomie, à ce moment-là, pas mal plus que de responsabilité. Et c'est parce qu'effectivement, on parle d'un individu, ça ne serait pas la même chose du tout si on parlait d'un médecin ou d'un professionnel de la santé qui a d'autres obligations … Parce qu'en ce moment, on regarde dans des secteurs comme en oncologie par exemple, ils appliquent des protocoles extrêmement définis, et le médecin qui décide de déroger du protocole, il faut qu'il ait une maudite bonne raison parce qu'il peut se faire poursuivre. S'il n'a pas suivi le protocole et qu'il arrive une *bad luck*, il est responsable et il va le payer cher.

Enfin, la crainte de l'apparition d'une ***perte de « l'esprit critique »***, de la « capacité de jugement » est apparue, notamment issue d'un surplus de confiance en la technologie.

> P41 : Alors, comment tu humanises ça, pis tu sécurises une madame de 80 ans qui pense que... ou que le conjoint à côté la voit en arythmie par terre pendant 30 secondes en syncope disons et qu'elle se relève et dit : 'Non, non. Mon application me dit que c'est correct.' Et c'est vraiment ça qui se passe en ce moment : 'Mon application va me dire que c'est correct, je reste chez moi.'

Les discussions ont porté sur les inquiétudes relatives à un manque de connaissances techniques et critiques nécessaires à la bonne compréhension du fonctionnement des algorithmes

en santé, soit sur les inquiétudes relatives au niveau de littératie numérique. Les connaissances techniques ont été reconnues comme « complexes » limitant la pertinence de la transparence des algorithmes ou des codes « ouverts ». Un niveau de littératie numérique insuffisant pourrait également compromettre la compréhension qu'ont les patients de ce qu'il adviendrait des données de santé collectées.

> P47 : Pour les individus de comprendre qu'est-ce qui arrive, et qu'est-ce qui advient de leurs données, ça on n'est pas rendu là. On n'a pas encore trouvé comment on va faire pour faire comprendre à la personne qu'est-ce qui arrive.

> P56 : Comment je peux demeurer critique comme citoyen par rapport au pronostic qu'on va faire ?

> P62 : Une des choses qui me vient en tête, c'est : c'est quoi l'âge de cette personne et c'est quoi leur familiarité avec la technologie ? Est-ce que c'est une autre contribution à la stigmatisation d'une population qui n'est pas alphabète numérique et qui peut-être n'a pas cette capacité de se questionner sur les limites de cette technologie ? Donc, prendre ça trop comme... au lieu d'une recommandation, comme quasiment un diagnostic parce que c'est mon téléphone qui me le dit et forcément la technologie de mon téléphone est correcte !

Les participants ont également soulevé un risque d'une « perte de pensée critique par rapport à soi-même » ou « d'instinct » intimement lié aux risques de « suggestion » décrit comme un risque « de prophétie auto-réalisatrice » :

> P23 : Parce que à un moment donné elle exprime : tu as ton robot, tu ne te fies plus jamais à 'Ah ! Là j'ai mal à la tête ou là je me sens pas bien'. C'est comme si tu peux plus te fier à toi-même, ton corps, tes sens… à moins que ce soit un robot qui te dise 'Tu te sens bien, tu ne te sens pas bien'.

Ceci pourrait favoriser l'autodiagnostic et les différents risques associés (ex. créer de l'angoisse et un coût pour le système de santé), et s'accompagne d'une responsabilisation accrue du patient (qui n'a peut-être pas les capacités de répondre de cette responsabilisation) :

> P52 : Puis l'idée de la prophétie auto-réalisatrice c'est ça. Dire à quelqu'un : 'Vous pouvez développer une dépression', alors que la personne n'est pas déprimée, il va se dire 'Shit ! Est-ce que ça va bien dans ma vie ?'

Voire entraîner une perte de contrôle sur « son corps » et « sa vie » ; notamment avec une obligation de vie saine – et défier donc les capabilités des patients, considérant l'impact sur ce qu'ils sont effectivement capables « d'être » :

> P68 : Est-ce qu'on peut forcer les gens à vivre sainement ?
> P7 : Ouais, ça c'est juste fou en fait. Avec l'objectif de vivre longtemps, longtemps, longtemps puis on s'en fout dans quel état. Si on perd le contrôle sur notre corps …qui décide de jusqu'où faut qu'on aille ?
> P68 : On est en train de bouger dans le transhumanisme à ce moment-là.

1.3. Attentes citoyennes : capacitation des professionnels de santé et des patients

1.3.1. Préserver l'autonomie décisionnelle

Pour beaucoup des citoyens participants, la « population » doit être « libre de ne pas partager ses données » ou de « refuser ce genre de diagnostic » voire de refuser de connaitre son état de santé face à des technologies de plus en plus intrusives (ex. notifications sur téléphone intelligents). Ils ont mentionné à plusieurs reprises qu'il est nécessaire de respecter « l'autonomie du patient » et de préserver la « liberté de la personne », son « autonomie de choix », son « autodétermination », son « autonomie décisionnelle », sa « liberté d'action » ou son « libre arbitre », afin de « garder un contrôle sur sa vie, sur ses choix de vie ». Ils ont soulevé l'importance de mettre en place des balises « sinon, ça devient intrusif et coercitif ».

> P54 : L'autonomie joue à plusieurs niveaux. Comment je comprenais l'intervention c'était : une fois que je reçois cette information-là, je garde, quand même, est-ce que je garde l'autonomie de faire bien ce que je veux avec, mais il y a un niveau où la pression est tellement forte que finalement, justement, je perds mon libre arbitre.

La pertinence de la mise en place de consentement numérique ou digital a également été discutée, qui permettrait de préserver la capacité à consentir pour certains considérant la possibilité de rendre le consentement plus « dynamique »; mais ce consentement numérique ne règle pas le problème pour d'autres considérant la quantité d'informations à fournir, et que la présence d'une personne (ex. un professionnel de santé) est nécessaire (ex. pour que le patient puisse poser des questions ou pour rendre le consentement plus convivial). Considérant que le consentement est un processus, la nécessité d'éduquer les patients pour préserver cette capacité à consentir a été

mentionnée à plusieurs reprises, notamment car « on ne peut pas avoir une autonomie sans éducation » ou que les patients risqueraient de ne pas donner leur consentement simplement parce qu'ils ont « peur ».

Concernant plus spécifiquement les professionnels de santé, il a été mentionné qu'éviter « trop de confiance » envers les systèmes d'IA pourrait également être un moyen de préserver l'expertise :

> P30 : Ça va forcer le médecin à garder ses connaissances top niveau. Au lieu de se fier à la machine.

1.3.2. Éducation et formation des professionnels de santé et des patients

Face à ces préoccupations, les citoyens ont mis de l'avant la nécessité de mettre en place des formations (éthiques et techniques) pour les professionnels de santé comme pour les patients. Pour les professionnels de santé, ces formations devraient être continues afin de tenir compte de l'évolution rapide des technologies en question. Il a également été reconnu comme nécessaire de former et informer le public, en tant qu'utilisateurs des systèmes d'IA. Ces formations auraient pour objectifs d'assurer une utilisation « optimale et consciente » des systèmes d'IA en santé. Elles doivent répondre à la crainte d'une évolution rapide des connaissances du domaine de l'IA :

> P41 : Ma peur dans tout ça, c'est que ça va tellement vite. On va former et la semaine d'après, va(-tu) falloir reformer ?

L'éducation des professionnels de santé pourrait prendre la forme d'un guide de bonnes pratiques élaboré par le Collège des médecins, notamment pour garantir de toujours considérer la place de l'humain dans les soins face à l'avènement de l'IA en santé. Pour ce qui est des patients, la nécessité de les former sur leurs droits et sur les différentes possibilités qui s'offrent à eux face à l'innovation numérique en santé s'avère nécessaire pour les citoyens. Cette formation se ferait par le biais de la vulgarisation et permettrait de développer le sens critique et éclairer les décisions. Elle pourrait prendre la forme de campagnes de sensibilisation pour se « défendre de l'IA ».

> « L'algorithme de l'application santé se trompe : pas les Québécois ! Grâce à l'approche patient-partenaire, le Québec a évité une crise. Les Québécois qui ont tous suivi leur cours d'auto-défense intellectuel sur l'IA en santé, ont refusé de suivre les recommandations de leur application de santé évitant une crise d'opiacés. D'autres pays du monde ont eu plus de

mal à éviter que leurs citoyens suivent les recommandations de l'algorithme qui a en fait bogué. » La Une de la Table 6, rédigée collectivement.

P51 : Dès le primaire, d'instaurer autant le côté de la littératie numérique mais celle de santé également. Que ce soit un sujet intégré à tous les cycles de formations. De comprendre ce que c'est, ce que ce n'est pas ... Je pense que les gens ont vraiment... la technologie fait partie de la vie, mais [les gens] n'ont pas développé en même temps le sens critique que des personnes un peu plus vieilles ont développé, parce qu'on a vu ça arriver.

Ainsi, la nécessité de préserver le sens critique et éthique des « utilisateurs » (soit préserver la littératie numérique), de les éduquer sur les limites des technologies en question a été soulevée à plusieurs reprises (par exemple, à l'aide du développement d'un guide d'auto-défense intellectuelle[115]). Les participants ont souligné l'importance de préserver les savoirs et savoir-faire en santé, notamment pour conserver le « pouvoir des humains », l'indépendance de leurs « opinions » et éviter la « désinformation ». Pour certains, il n'est pas absolument nécessaire de comprendre tout le fonctionnement à la base des décisions algorithmiques mais seulement que le médecin connaisse les limites de la technologie utilisée.

Assurer un niveau minimum de littératie numérique en santé permettrait alors de favoriser l'autonomie de tous, en particulier les personnes qui seraient moins enclines que d'autres à comprendre les enjeux de l'innovation numérique en santé et impliquerait de vulgariser et de clarifier les différents aspects relatifs à l'utilisation des systèmes d'IA en santé.

2. Le partage de la responsabilité face à la multiplication des acteurs

2.1. Le problème des mains multiples (*many hands*)

Comme mentionné dans les précédents chapitres, l'utilisation de systèmes d'IA – soit l'usage d'algorithmes et de données massives - implique pour un seul usage une multitude d'acteurs depuis la génération des données à l'application médicale (ex. scientifiques des données, développeurs, chercheurs en santé, médecins et patients). Cette multiplicité des acteurs conduit, en ce qui a trait à la détermination de la responsabilité, au problème des mains multiples (ou « *many*

[115] La participante fait ici référence au livre *Petit cours d'autodéfense intellectuelle* de Normand Baillargeon (2005).

hands ») : la responsabilité ne pouvant être attribuée à un seul individu, organisation ou groupe, il est difficile de déterminer qui est responsable des conséquences (positives ou négatives) d'une action et dans quelle mesure (Noorman 2016).

Le problème des mains multiples est un problème initialement identifié dans le domaine de l'administration publique : Thompson (2004) soulève différentes difficultés relatives à l'application de principes moraux dans le développement de politiques publiques à cause du problème des *mains multiples* :

> Because many different officials contribute in many ways to decisions and policies of government, it is difficult even in principle to identify who is morally responsible for political outcomes. This is what I call the problem of many hands (p. 11).

Le problème survient ainsi dans des contextes où de multiples acteurs contribuent chacun à des conséquences qui s'observent au niveau d'un système mais où il est difficile de tenir un seul acteur comme responsable des conséquences en question (Dixon-Woods et Pronovost 2016). Pour Dixon-Woods et Pronovost (2016), le système de santé est un cas paradigmatique du problème des *mains multiples*, étant caractérisé par des acteurs autonomes et hétérogènes bien qu'interdépendants. Selon les auteurs, les acteurs du système de santé ne fonctionnent pas comme un collectif mais agissent plutôt comme un ensemble d'individus atomisés, principalement responsables d'eux-mêmes et non du système comme un tout. Dans ce contexte, il est alors difficile de coordonner leurs interactions, d'autant plus que ces acteurs peuvent être rivaux ou manquer d'engagements partagés vers des objectifs communs, créant de potentiels conflits sur la nature des problèmes auxquels ils font face et sur qui sera imputable de ces problèmes (Dixon-Woods et Pronovost 2016). Cette observation semble d'autant plus vraie aujourd'hui, l'avènement de l'utilisation des systèmes d'IA faisant intervenir, comme précédemment mentionné, de nouveaux acteurs dans le système de santé – notamment, les développeurs et scientifiques des données (CCNE 2019). Cette multitude d'acteurs défie le partage des rôles des différentes parties prenantes d'un système de santé où le numérique occupe une place de plus en plus importante et défie, par ce fait même, la distribution de la responsabilité.

Le problème des *mains multiples*, parce-qu'il rend difficile l'identification d'un responsable, peut ainsi favoriser une certaine déresponsabilisation des individus qui prennent part au développement des systèmes d'IA en santé. D'abord, parce que la multiplication des acteurs augmente la distance entre un individu et les conséquences de ses actions (Noorman 2016). Ensuite, parce que les acteurs impliqués dans la chaine qui conduit à une conséquence, aux expertises et aux intérêts variés (voire opposés), vient diluer la capacité d'agir (ou agentivité), comme le reconnaissent Dixon-Woods et Pronovost (2016): "*The profusion of agents obscures the location of agency*" (p. 1). Il est en effet difficile pour un seul acteur, contraint dans un rôle bien spécifique, d'identifier l'ensemble des conséquences de son travail.

Considérant la nature du problème des *mains multiples*, une tension importante entre la responsabilité individuelle et la responsabilité collective peut apparaître, en particulier lorsqu'il est question d'impacts négatifs (Dixon-Woods et Pronovost 2016; Thompson 2004). Selon Thompson (2004), cette tension se résume en trois étapes :

> The argument underlying the collective model begins by posing a version of the problem of many hands: many political outcomes are the product of the actions of many different people whose individual contributions may not be identifiable at all, and certainly cannot be distinguished significantly from other people's contributions. The second step is the claim that no one individual, therefore, can be morally blamed for these outcomes. At the final stage of the argument, its proponents reach two seemingly contradictory conclusions: one stating that every individual associated with the collectivity should be charged with moral responsibility, the other holding that only the collectivity can be so charged. But the conclusions are not so different since neither ascribes responsibility to persons on the basis of their specific and distinct connections to the outcome in question (p. 15).

Différentes « excuses » peuvent ainsi servir une certaine déresponsabilisation des individus. Pour ne citer qu'elles, il peut s'agir de blâmer la collectivité en général plutôt qu'un de ses membres en particulier, se dédouaner en renvoyant la responsabilité à d'autres acteurs de la chaine causale ou, considérant la séparation des rôles, que les conséquences ne relèvent pas de la fonction qu'occupe un seul individu mais de celles des autres (Thompson 2004). Systématiquement renvoyer la responsabilité à quelqu'un d'autre est en filigrane du problème des *mains multiples* et risque de conduire à une irresponsabilité partagée :

> To relieve one person of responsibility, the excuse asserts that other people (the alternative causal agents) would be responsible for the action. But if the excuse is valid, each of the other people would be exonerated, seriatim, just as the first person was. In other words, if the excuse is valid, no one is responsible (Thompson 2004 p. 20).

Lors des discussions ayant eu lieu au cours de la coconstruction de la Déclaration de Montréal, des craintes relatives aux problèmes des *mains multiples* ont été soulevées. D'abord, les citoyens ont mentionné à plusieurs reprises le fait qu'un grand nombre d'acteurs soit impliqué dans le parcours de soin complique l'attribution de la responsabilité. Ils ont également soulevé trois principales craintes relatives aux conséquences de la multiplication des acteurs qui seraient favorisées par le phénomène des *mains multiples*. Ces craintes sont relatives 1) à la propriété et à la protection des données; 2) à la perte de lien naturel et 3) aux conflits d'intérêts potentiels. Face à ces craintes, les citoyens ont manifesté des attentes relatives à un contrat social défini en fonction d'une certaine responsabilité partagée. Ils ont également identifié les mécanismes de gestion déjà en place, soit les acteurs existants qui ont une potentielle responsabilité face aux enjeux du développement des systèmes d'IA en santé. Les citoyens participants ont également discuté de normes et de principes à la base dudit contrat social : la transparence des institutions et la précaution. Ils ont formulé des attentes relatives aux partages de la gestion et de la propriété des données et ont finalement identifié les mécanismes de gestion existants pour répondre aux enjeux éthiques associés au développement des systèmes d'IA et qui permettent d'identifier les acteurs responsables.

2.2. Craintes citoyennes : un grand nombre d'acteurs impliqués et des conséquences sur la gestion des données de santé et sur les soins

2.2.1. Un grand nombre d'acteurs impliqués

Les citoyens participants ont soulevé qu'un grand nombre d'acteurs aux rôles et aux responsabilité différents est impliqué dans l'utilisation des systèmes d'IA en santé : les développeurs ou les ingénieurs, les chercheurs, les médecins, les gestionnaires, les fournisseurs de services, les entreprises, les patients, mais aussi les algorithmes (notamment les systèmes d'IA qui représentent « plusieurs expert*s* » - voir Section 3. du présent chapitre).

Face à ce grand nombre d'acteurs, les citoyens ont soulevé la difficulté de définir qui est responsable du développement des systèmes d'IA en santé, notamment en cas d'erreurs ou de dysfonctionnement :

> P65 : Il y a la question de la responsabilité aussi en cas de dysfonctionnement. Je veux dire… Bon, c'est des systèmes numériques, à IA, peu importe comment on les appelle. Mais en cas de dysfonctionnement, qui est tenu pour responsable ? Le propriétaire de la machine ? Le propriétaire des données ? Ou la machine elle-même ?

> P60 : Ce que je me demandais dans tout ça, la place du technicien, qui est ni... parce qu'au final, il y a quand même de l'humain qui rentre, (issu) du secteur médical et qui doit être en charge un peu de régler ces possibilités d'erreur et de voir en quoi l'algorithme a pu un peu mal fonctionner et, au final, cette personne-là est absolument pas en lien avec l'algorithme et elle est amenée à être en lien avec les données qui sont insérées dans l'algorithme.
> P62 : Et la personne est probablement distante dans une autre ville, dans un autre pays...
> P64 : En Inde !

Les citoyens ont mentionné que l'implication de nombreux acteurs (notamment distribués géographiquement) aux expertises variées complique l'attribution de la responsabilité. Ils ont alors à plusieurs reprises soulevé des questions relatives au partage des responsabilités, notamment considérant qu'il y a « plusieurs systèmes inter-imbriqués là-dedans » : Qui est imputable ? Les « personnes qui ont entré les données » ? Celles qui « créent les algorithmes » ? Le personnel médical ou le « médecin prescripteur du robot » ? Ceux qui développent les algorithmes ou ceux qui les exploitent ? Qui serait responsable si le système venait à être « hacké » ?

L'attribution de la responsabilité se complique également selon les citoyens lorsqu'il est question de savoir qui aura accès aux données de santé : les professionnels de santé ? La compagnie qui les collecte ? Le patient qui les génère ?

> P67 : … Si elle vient donner des soins la préposée, enfin préposée ou infirmière, elle doit aussi avoir un *feedback* pour savoir quoi faire. Est-ce qu'elle a accès ou pas ? Ils vont dire : « Non, puisque c'est une machine et c'est un robot, ça appartient à la compagnie et la personne qui s'occupe... »
> P65 : C'est ce qu'on disait tout à l'heure sur la responsabilité en cas de dysfonctionnement : Qui est tenu pour responsable ?
> P67 : Même si ça marche, même si ça fonctionne. […]
> P62 : Il y a un transfert d'informations.

P67 : Il y a un transfert d'informations. Est-ce qu'elle a accès ou pas ? Ou est-ce qu'elle doit passer par les parents ou par le …
P64 : Ou par le technicien lui-même ?

Les citoyens ont ainsi soulevé des craintes qui relèvent directement du problème des *mains multiples*, à savoir que les acteurs risquent de se renvoyer la responsabilité mutuellement :

P37: L'enjeu, c'est que si tout le monde pense que c'est quelqu'un d'autre qui est responsable, il me semble que...
P42 : C'est la machine !

2.2.2. Craintes relatives aux conséquences des mains multiples sur le soin et sur la santé

Les citoyens participants ont exprimé trois principales craintes qui peuvent être considérées comme des conséquences du problème des *mains multiples*[116]. Premièrement, la multiplication des acteurs conduit à des préoccupations relatives à la propriété et à la protection des données. Il devient en effet difficile de déterminer qui a accès aux données et qui en est propriétaire. Ceci peut conduire à une certaine asymétrie de pouvoir entre les acteurs et soulève des préoccupations relatives à la protection de la confidentialité et de l'intimité, si l'on considère que le respect de la vie privée n'est pas tant défié par la qualité des données que par les acteurs qui y ont accès (Lahlou 2008)[117]. Deuxièmement, l'implication d'un grand nombre d'acteurs soulève des préoccupations relatives à une perte de lien naturel entre professionnels de santé et patients, notamment car la multiplication des acteurs dans le parcours de soin fait qu'il devient difficile d'avoir une vision d'ensemble, ce qui pourrait nuire à la mission de délivrer des soins adaptés. Troisièmement, des craintes relatives aux potentiels conflits d'intérêts ont été soulevées, favorisées par la multiplication des acteurs – et donc des intérêts en jeu.

Propriété et protection des données

En premier lieu, les participants ont soulevé des craintes relatives à la ***propriété des données***. La multiplicité des acteurs impliqués dans le parcours de soin complique la détermination de qui est propriétaire des données et de qui devrait y avoir accès.

P40 : Si t'es l'auteur des données, est-ce que tu en es le propriétaire ou pas ?

[116] Soit de la multiplication des acteurs. Le lien avec la responsabilité est discuté dans le Chapitre 6.
[117] Cet aspect est discuté dans le Chapitre 3, Section 2.1. sur la protection de la vie privée et de la confidentialité.

P38 : Habituellement non.

P62 : Qui a le droit d'accès à ces informations ? Ça appartient à qui ? Dépendant de qui a acheté la machine, est-ce que la machine est louée ou est un remplacement à un soignant ? Parce que dépendant comment cette machine est insérée dans la vie de la personne, il y aura différentes notions de droits d'accès et de contrôle de ces informations.
P65 : Ça rejoint la propriété intellectuelle.
P62 : Mais au-delà, parce que la propriété intellectuelle va être autour de qui contrôle la machine, l'algorithme. C'est à qui appartient l'information. Parce que dans une relation de soin, c'est partagé.

Les citoyens ont questionné comment différencier les données publiques des données privées, en particulier quand celles-ci sont issues des réseaux sociaux : À partir de quel moment ces données ne nous appartiennent plus ? Faut-il que « *toutes traces de trucs sociaux* » soit « *la propriété de l'usager* » ? (P38).

Ils ont également discuté de la propriété des données médicales plus classiques :

P36 : Disons que c'est des données génomiques, elles appartiennent à qui ? À vos parents, à vos enfants ou à vous ? Qui vous les a données vos données ? La notion de « Ça nous appartient », c'est une notion... Oui, mon numéro de téléphone. Mes données génomiques, on me les a léguées, je vais peut-être les léguer, mais est-ce que c'est à moi?

Ils se sont également inquiétés que les professionnels de santé ou les patients aient toujours accès aux données de santé qui seraient la propriété de compagnies privées.

P34 : Si on revient en santé, c'est le débat actuel avec les données qui sont dans les dossiers médicaux électroniques. Combien tout le monde tombe de sa chaise comme quoi d'un seul coup, il y a des gens qui les vendent. Ben oui ! Pourquoi ? On les a laissé aller sur des serveurs on ne sait pas où, propriété de Telus et compagnie, et ils les vendent, et on se dit : « Ah, c'est incroyable ! » Mais à qui elles appartiennent ces données-là actuellement ? Pour se rapprocher du sujet, c'est plus nos données qu'on génère en se consultant. Je comprends que tu as le choix ou pas de mettre des choses sur Facebook, est-ce que tu as le choix ou pas d'aller voir ton médecin de famille ? Si tu as un problème de santé, pas vraiment. Les données se génèrent, du coup elles appartiennent au médecin, elles appartiennent à toi, elles appartiennent à la compagnie ?

La propriété des données de santé est ainsi également liée à des enjeux de ***marchandisation des données***, soit notamment la possibilité de les revendre, peu importe l'acteur qui les détient.

P6 : Avec l'IA ils peuvent cibler une personne mieux que n'importe qui. C'est de la marchandisation des données, les assureurs qui veulent savoir, les employeurs, tout le monde. Pour moi la marchandisation des données c'est un gros problème du point de vue de l'IA.

Les citoyens ont souligné dans ce contexte une multiplication du risque en fonction des propriétaires. Ils ont souligné que la valeur monétaire de la donnée se retrouve à la source de cet enjeu de marchandisation :

> P36 : Je suis d'accord avec « Protection » mais j'utiliserais le mot « achat ». Parce que, comme on dit souvent que c'est le nouveau pétrole, que les données c'est le nouveau pétrole. Je mettrais le mot « achat » parce qu'il y a une valeur économique.

En santé, la marchandisation des données peut prendre différentes formes : il peut également s'agir d'échanger des données pour des bénéfices en santé et non juste pour de l'argent. Les citoyens ont par exemple souligné qu'il serait problématique d'échanger des données contre le fait d'avoir accès de manière prioritaire aux soins :

> P50 : C'est très discutable de donner une valeur à la donnée, d'échanger ta donnée contre un accès aux soins de santé, contre un accès à du financement, contre un accès à un service.

> P51 : ... C'est l'accès... C'est malgré ce que je suis prête à donner comme données, je ne devrais pas en subir les conséquences de manière négative, si je ne veux pas fournir mes données. C'est lié avec la valeur de la donnée qu'on monnaye en échange de service. Ça va avec l'universalité.

Les citoyens se sont ainsi inquiétés que les données de santé deviennent une « monnaie d'échange » ce qui risquerait de « forcer la main »[118] et pourrait défier l'universalité de l'accès au soin. Pour d'autres, le problème est associé au fait de faire de l'argent sur des données produites gratuitement par les patients, comme P36 interrogé sur les enjeux que soulève le scénario :

> P36 : C'est facile, c'est l'achat de données pour des données qui ont déjà été données fait que gratuitement. Parce qu'il y a la partie où ça serait le gouvernement qui achète les données, mais ces données-là viennent des individus qui les ont partagées délibérément, pleinement.

La multiplication des acteurs dans le parcours de soin s'accompagne également de préoccupation relative à la ***traçabilité des données***. Leur partage dépassant les frontières, les citoyens s'inquiètent qu'il ne soit plus possible de savoir de quelle juridiction relèvent les données ni de déterminer qui serait responsable de leur gestion.

> P52 : Les serveurs dans différentes juridictions, savoir quel cadre légal s'applique. On parle de téléphones privés. Là, tous les téléphones des individus, avec tout ce qu'on connait

[118] Cf. les discussions relatives à l'atteinte à la liberté de choix, Section 1.2.2. sur l'incapacitation des patients du présent chapitre.

justement de…des accès qu'on peut avoir [avec ces téléphones]. Là, ça se met à se promener toutes ces informations-là…

P15 : Aussi, la complexité, c'est comme Amazon par exemple. Ils ont des centres de données ici au Canada...
P8 : Mais ça passe 10 fois autour du monde.
P15 : Exactement. Et la problématique, c'est que les données doivent être sauvegardées, parce qu'eux doivent assurer à leurs clients que les données sont disponibles en tout temps. S'ils ont un problème électrique au Canada, toi en tant que client, tu veux avoir accès, tu veux garder accès à tes données et les données, elles sont sauvegardées aux États-Unis, mais tu ne le sais même pas.

P15 : Il y a beaucoup de choses virtuelles comme l'application, ok, tu l'installes à Londres ou ailleurs, tes utilisateurs comment tu les... Ça va être impossible de contrôler ça.

La propriété des données pourrait créer une **asymétrie de pouvoir** entre les différents acteurs du développement des systèmes d'IA en santé, notamment entre les assureurs de santé et les patients :

P8 : Il y a un peu une asymétrie de pouvoir entre le client de l'assurance et l'assurance qui va avoir trop d'information.

P2 : Ben le fait qu'on a une entreprise d'assurance privée qui a accès aux données confidentielles et à la vie privée d'Olivier pour déterminer … C'est-à-dire qu'on leur donne un pouvoir quand même assez important. Moi ça, ça me dérange beaucoup. On dit qu'elles sont collectées pour le gouvernement mais tout à coup elles se retrouvent dans une compagnie privée à but lucratif qui va faire de l'argent en ne payant pas Olivier quand il va tomber malade.

La multiplication des acteurs dans le parcours de soin amène également à de potentielles conséquences relatives à ce que les citoyens participants à la coconstruction ont nommé la « ***rétroaction***[119]», soit à quelles données il est possible d'avoir accès ou pas, rétroactivement par rapport à la collecte. Dans cette perspective, les citoyens ont soulevé que c'est « le croisement qui est problématique » concernant la gestion et la protection des données.

P7 : J'ai un super bon exemple : la maudite carte de la SAQ : la carte à points. La SAQ c'est un organisme gouvernemental. Mettons que demain matin, finalement Santé Canada dit 'Bon, nous autres on va aller chercher des données là, de la carte Inspire'. 'Oh M. Caron, hey vous achetez une bouteille de rhum par mois (par jour !) depuis 6 ans, mon dieu mais c'est dégueulasse !' Fait que là, si je prends … exactement… c'est de la rétroaction dont je parle ! Ce serait dégueulasse que la SAQ transmette ces données là tout à coup et que là tu

[119] Qui fait écho, en éthique de la recherche, à la notion d'utilisation secondaire (ÉPTC2, 2018) voire au double-usage (Selgelid, 2013).

fasses ‘Ah ben finalement j'suis une mauvaise citoyenne mais je ne savais pas que c'était commencé votre affaire’.

L’enjeu est associé au fait que les données deviennent finalement accessibles à d’autres acteurs que ceux initialement prévus lors de la collecte :

> P1 : Ce que je verrais moi c'est par exemple quand tu donnes accès à ton dossier médical à tout le monde médical. [Tu ne veux pas] que n'importe quel médecin accède à tes résultats d’analyse de sang, etc. Et aujourd’hui je donne accès… parce que… Je ne veux pas que ces données soient partagées en dehors du monde médical.

> P54 : Là où il va y avoir des désaccords c’est, mettons, une question de confidentialité : à quel point on devrait ouvrir à d’avantages d’acteurs ces données-là ? Pour quelles raisons ?

Les inquiétudes relatives à la rétroaction concernent également les fins pour lesquelles les données vont être réutilisées :

> P42 : Ça revient sur le consentement éclairé d'un patient. Si je demande à un patient qui a le cancer : « Est-ce que tu serais d'accord que tes données personnelles servent à découvrir des traitements qui peut-être ne te soigneront pas toi parce que tu es en phase terminale, mais qui serviront à guérir le prochain ? » Tout le monde dit oui. Je dis : « Mais aussi, ça pourrait servir pour les compagnies d'assurances, ceux qui vont faire les assurances-vie », là c'est non. Et c'est vraiment à quelle fin la donnée sert.

Les citoyens ont souligné les enjeux associés à l’accessibilité aux données, qui potentialise le risque de rétroaction :

> P52 : On a des données qui viennent de plusieurs sources qui sont centralisées donc là on a un risque accru. Quelqu’un qui réussit à rentrer dans la base de données centrale, il a accès à énormément de choses auxquelles il n’aurait pas accès autrement. Ça, c’est un risque important. On parle des nuages, les nuages ce n’est pas évident à quel point c’est… on est encore en train de sécuriser toutes ces choses-là.

> P55 : Il peut y avoir des dérives autoritaires. On parle de contrôle de données, si on décide du jour au lendemain que... Ça peut facilement déraper...
> P52 : […] Si on change de dynamique sociale, et que tout ça est accessible, c'est extrêmement dangereux.

Certains citoyens ont cependant reconnu que la rétroaction peut également apporter des bénéfices, notamment pour les patients, et que l’évaluation du bon usage des données se fait selon un équilibre :

> P52 : Fait que là tu as d’emblée une tension avec… Plus il y a de données et plus il y a de gens qui les partagent, généralement plus on peut s'attendre à ce qu'on soit capable d'aller créer de la valeur, afin de valoriser ces données-là. En même temps, plus peuvent augmenter

> les risques qui viennent avec ça. Il y a une grosse balance … Mais ça c'est un choix social, jusqu'où on est prêts à aller ?

Le flou entourant les acteurs qui peuvent être propriétaires ou avoir accès aux données potentialise également le risque ***d'atteinte à la vie privée et à la confidentialité***. En ce qui a trait aux données de santé, les citoyens ont parfois mentionné qu'il serait problématique qu'elles deviennent accessibles ou qu'elles soient vendues à « d'autres instances que les instances de santé ». Cette entrave potentielle s'accompagne d'inquiétudes relatives aux limites à poser quant à la collecte des données, notamment en vue de cibler les patients.

> P19 : Comment s'assurer que la Vie Privée n'est pas marchandée ?

Le risque de ne pas respecter la « confidentialité », le « secret professionnel » ou la « confidentialité de l'opinion du médecin » a également été souligné.

Certains ont soulevé que la vie privée dans un contexte de santé n'est cependant pas absolue – ou pas la priorité - et qu'une entrave à la vie privée est parfois acceptable lorsqu'il s'agit de la vie ou de la santé des patients, notamment considérant que collecter plus de données (ou des données de différentes natures) pourrait potentialiser l'efficacité de l'IA (ex. donner aux algorithmes des données sur le style de vie pourrait permettre d'adapter et de personnaliser les recommandations de santé).

> P16 : Je vous dis dans le domaine de la santé, surtout dans la santé, la vie privée, elle ne vient pas en premier pour moi. […] C'est un enjeu dans d'autres, peut-être, domaines et tout ça, mais quand ta santé elle est en danger, ta vie privée, elle sera moins pire pour toi par rapport à ta santé.

Certains participants ont soulevé que protéger l'information est un enjeu purement technique et pas tant éthique (par exemple, relativement à la dé-nominalisation ou à la protection de la sécurité). Une crainte que la protection de la vie privée se fasse au détriment des bénéfices de l'IA a également été exprimée dans une perspective de bien commun ou de bénéfices collectifs.

> P52 : Mais là, effectivement les valeurs qui sont en jeu là-dedans, c'est vraiment le côté vie privée et de l'autre côté ... d'un côté vie privé et autonomie, et de l'autre côté le bien-être collectif. Parce que la question qui peut venir après ça c'est, si on a besoin des données de tout le monde pour générer le plus de bénéfice collectif, ils ne sont pas obligés les gens de donner leurs données.

Un citoyen participant a cependant souligné qu'il était peine perdue de se pencher sur la protection des données et qu'il n'avait pas confiance dans les compagnies qui les collectent :

P41 : Je vais régler tout de suite quelque chose : la protection de données en IA, je n'y crois pas du tout.
P35 : Et pourquoi ?
P41 : Ça n'arrivera pas. Il y a des hackers qui vont juste faire ça. Je vais être cru, mais on pourrait en parler toute la journée, mais je n'y crois pas. […]
P35 : Mais il n'y aura pas de plus en plus de protection ? Maintenant, qu'on est conscients de...
P41 : … J'y crois pas.

Perte de lien naturel

Les répondants s'inquiètent que l'avènement des systèmes d'IA en santé s'accompagne d'une ***perte de « lien naturel »*** ou une « déconnexion » entre patients et professionnels de santé, considérant l'absence d'approche « holistique » ou « exhaustive » des soins mais également le contexte de la collecte (massive) de données. Le grand nombre d'acteurs impliqués pourrait en effet conduire à une perte de la vision d'ensemble :

P5 : Pour la partie qui contrôle "machine/humains", y'a l'idée de déconnexion. Alors on a des systèmes complexes qu'on met en place avec des données et toutes sortes d'affaires, puis a un moment donné qui voit si ça a du bon sens par rapport à des facteurs globaux de la société et c'est ça qui est difficile à comprendre parce qu'il y a tellement d'acteurs impliqués …

L'utilisation des systèmes d'IA en santé s'accompagnerait d'une différence d'échelle notable en ce qui a trait à la gestion du système de santé, impliquant des systèmes d'IA de plus en plus complexes :

P5 : Donc on a ce truc-là complexe qui sort du contexte national et à un moment donné on perd le pouvoir de comprendre que ça à un lien avec ce qui est tolérable par la société et puis aussi on perd avec la particularité de cette société-là.

Notamment, un partage de données qui dépasserait l'échelle nationale – et impliquerait donc des acteurs de différents pays - poserait des problèmes relativement au contrôle que l'on a sur le système de santé :

P5 : Ben on perd le lien naturel : on a une problématique puis on a des moyens qu'on peut investir parce que moi je ne vais pas aller chercher les moyens de la Chine pour soigner les gens d'ici là. À date on a une organisation qui est relativement nationale. Donc si on devient dépendant de, par exemple d'un organisme qui collecte des données en Chine ben on perd un contrôle puis un lien de ...

L'implication de nombreux acteurs divise les rôles et par là-même pourrait influencer le travail du médecin qui n'aurait pas accès à tous les éléments :

> P61 : Deuxièmement, une affaire qu'il faudrait inclure dans la réflexion, c'est qu'un technicien, le technicien n'a même pas besoin de... Ce n'est pas la même personne qui joue dans l'algorithme et qui utilise les données. Dans le sens que ces analyses-là, statistiques, ce qu'ils appellent le *black box*, ça ne sert à rien de voir les données parce que tu ne sais même pas comment il a réussi à arriver au *output*. Donc, tu joues sur l'algorithme et la personne en fait, c'est ça qui est dangereux, ce n'est pas d'avoir les données, mais c'est de les passer dans d'autres algorithmes qui permettent de traduire d'autres choses. La donnée en elle-même, même le médecin n'aura jamais ces données-là. Il va avoir le *output* du robot parce qu'en fait, le robot va compiler.

Cette échelle trop grande participerait à une certaine perte d'individualisation :

> P62 : 'Oui, on a des données basées sur des populations, mais la personne devant nous, ce n'est pas une population, c'est un individu.' Donc, où est-ce que cette personne se situe sur le spectre peut être aidé par une analyse statistique, mais ça reste une analyse statistique hyper bien foutue, mais il y a quand même un aspect d'un individu dans une population. C'est quand même un individu.

Potentiels conflits d'intérêts

Les citoyens ont finalement soulevé des craintes relatives aux conflits d'intérêts des différentes parties prenantes de l'utilisation des systèmes d'IA en santé.

> P38 : Le problème, c'est que justement, il va y avoir des acteurs qui vont proposer des diagnostics qui ne sont pas tout à fait appropriés à cause de conflits d'intérêts cachés.

Ils ont soulevé que les différents acteurs impliqués ont différents rôles et intérêts :

> P13 : Les professionnels privés et publiques ont des intérêts différents …

Lesquels pourraient potentiellement devenir conflictuels, ce qui pourrait impacter la « crédibilité », la « fiabilité de l'information » et la « confiance ». Les citoyens ont par exemple mentionné que ces conflits d'intérêts risqueraient de nuire à « l'impartialité » ou à « l'indépendance » des recommandations issues des systèmes d'IA, notamment lorsqu'issues de lobbysme ou d'autres facteurs d'influence :

> P38 : Moi, je vais rebondir sur un truc que tu as dit avec un autre enjeu qui est la crédibilité. Parce qu'on parle d'efficacité, mais de crédibilité. Par exemple, une compagnie pharmaceutique qui dit : 'Voici les pilules que vous devez prendre.' À quel point c'est crédible ? Et à quel point, parce qu'il y a des intérêts financiers très forts derrière tout ça, et toutes ces données-là. À quel point, maintenant, les diagnostics qu'ils ont faits, par exemple,

> justement ils disent : 'Prenez cette pilule.' Oui, est-ce que ça a vraiment été contrôlé de manière séparée ?

Les discussions concernant les conflits d'intérêts ont majoritairement porté sur ceux des acteurs privés. Par exemple, les compagnies n'ont peut-être pas intérêt à répondre à des exigences de transparence vis-à-vis de leurs algorithmes considérant leur intérêt à rester compétitives. Elles ont également des intérêts financiers (notamment, considérant la valeur monétaire associée à la donnée) qui pourraient entrer en conflit avec les intérêts du patient qui les génère, « l'intérêt public » et « les intérêts du public ». Enfin, si les algorithmes venaient à recommander certains médicaments plutôt que d'autres, les systèmes d'IA risqueraient d'être touchés également par les conflits d'intérêts issus de l'influence des compagnies pharmaceutiques.

> P37 : On a découvert finalement que ce que font les compagnies, c'est qu'elles offrent aux pharmacies le logiciel qui leur permet de gérer les prescriptions gratuitement, en échange de quoi ? Eux récupèrent les données. Donc, quand vous allez à la pharmacie, les données de prescription sont comptabilisées. Tout ce qu'ils font, c'est qu'ils vendent ça à d'autres compagnies qui veulent mieux connaître les habitudes de prescriptions d'un médecin. Donc, vos informations sont colligées, après elles sont vendues et on les utilise pour essayer d'influencer le jugement des médecins sur tel médicament et ça a des coûts pour la santé publique.

Les citoyens ont toutefois également soulevé des conflits d'intérêts qui pourraient concerner les acteurs publics, comme ce citoyen qui discute du choix des experts à inclure dans un potentiel comité d'éthique de l'IA indépendant :

> P67 : Justement, moi, instance gouvernementale, dans le meilleur des cas, il n'y aurait aucun conflit d'intérêts, rien. Mais le problème, c'est que si c'est vraiment gouvernemental, on rentre dans des problèmes de conflits d'intérêts potentiels énormes. Et si c'est indépendant, la même chose. Qui c'est qui va nommer les experts ?

2.3. Attentes citoyennes : un contrat social selon une responsabilité partagée

2.3.1. Un contrat social

Face au problème des *mains multiples*, les citoyens ont formulé des attentes qui prennent la forme d'un « **contrat social** » entre les différents acteurs de l'utilisation et du développement des systèmes d'IA en santé. Ce contrat social devrait assurer la confiance de la « population » envers ces acteurs. Les citoyens ont défendu la nécessité d'une responsabilité « collective » ou

« partagée » des différentes parties prenantes « en amont et en aval », soit des concepteurs aux patients et en vue du « bien-être collectif » :

> P43 : Ce n'est pas juste ce que fait le robot, c'est toute la chaîne qu'il faut impliquer.
>
> P36 : L'autre volet, responsabilité, on peut le mettre en Responsabilité partagée pour reprendre les termes du patient-partenaire. Parce que le soin, ça reste une rencontre entre un malade et un soignant, fait qu'il y a deux volets à la médaille, c'est une responsabilité des deux côtés.
>
> P34 : Les différents acteurs doivent assumer leurs responsabilités contre les risques des innovations.

Cette responsabilité partagée ne devrait pas pour autant limiter la responsabilité individuelle :

> P35 : Je suis vraiment désolée, parce qu'alors, vous voulez mettre 'responsabilité individuelle et partagée' ? Parce que si on se dit : 'On partage d'un côté », il y a moins même l'accent sur, aussi, la responsabilité de chacun, non ?'

Afin que l'IA remplisse sa fonction d'assurer la santé des individus et des communautés, les citoyens ont parfois recommandé que les normes à la base dudit contrat social soient développées selon une gouvernance collective :

> P52 : Parce qu'encore là, il y a des valeurs qui sont en jeu, on va vouloir l'orienter d'une certaine manière. Si on ne veut pas ... parce qu'en ce moment, quand on fait de la priorisation, il y a un travail qui est fait sur quelles valeurs on met en œuvre. C'est extrêmement important.
> P51 : On commence par ça ! Quand on priorise, il faut dire : c'est quoi nos critères de base. Nos critères, c'est nos valeurs.
> P52 : Il faut trouver une manière d'intégrer ça, à une étape ou à une autre, dans le processus décisionnel de l'IA.
> P51 : C'est là que vous mentionniez que ce soit fait démocratiquement, que les critères de priorisation qui vont être considérés dans l'IA, ça soit décidé collectivement.

Laquelle permettrait de déterminer démocratiquement les normes à la base dudit contrat social :

> P52 : Deuxièmement, là-dessus, c'est : il faut consulter, il faut avoir un débat public sur cette question-là. Une information de qualité, mais, aussi, un forum qui permette un peu à chacun de ne pas juste être confronté individuellement, à une politique de confidentialité : « Je clique, j'accepte de partager mes données », mais d'avoir une délibération démocratique sur ces notions-là.

Ce qui n'est pas sans risques pour d'autres :

> P55 : Mais, est-ce que le commun des mortels est capable d'aborder ces question-là ? Il y a des gens qui sont spécialistes en éthique qui on réfléchit depuis de nombreuses années à ces

> [notions]-là, est-ce que ramener ça, demander l'avis de tout le monde par rapport à ça, ce n'est pas un danger de … ?
> P51 : ... Danger de rendre ça individualiste aussi.
> P55 : Ou réservé à une classe bien spécifique.

Face au fait qu'il devient de plus en plus difficile de départager les responsabilités entre les différents acteurs, notamment considérant le nombre croissant de collaborations et l'imbrication à la fois du secteur public et du secteur privé, certains citoyens ont mis en avant la nécessité de bien définir les rôles de chacun :

> P61 : Se départager bien le rôle et la fonction, la mission de l'académique, de l'université, et de l'industrie de la technologie.

Notamment relativement aux responsabilités du secteur privé :

> P37 : C'est un contrat social. Je pense que ça oblige à... Ce que je perçois, c'est que le secteur privé a besoin du secteur public pour avoir une législation robuste. Sinon, les gens n'auront pas confiance.

> P36 : Plus on voit des applications qui sont à des fins commerciales, plus les gens sont méfiants. Et je pense que pour que l'IA continue à rouler et à avancer, il va falloir que le secteur privé ait confiance et que les investisseurs aient confiance. Et ça, c'est public. D'une certaine façon, le contrat social implique le public du privé.

Dans le cadre de ce contrat social, les citoyens ont mentionné que la responsabilité des différentes parties prenantes se déclinait autour de leurs rôles respectifs et ont discuté de chacun d'eux.

2.3.2. Identification des mécanismes existants

A plusieurs reprise, les citoyens ont mentionné qu'il n'était pas nécessaire de mettre en place de nouveaux mécanismes de gouvernance mais plutôt de renforcer, d'adapter ou de spécialiser les mécanismes existants.

Par exemple, en spécialisant les ombudsmans des hôpitaux sur l'IA en santé, en renforçant la Charte des Droits et Libertés québécoise face à l'IA ou en incluant sous la « Loi d'accès universel » tous les nouveaux développements de l'IA en santé.

> P20 : Mais de la même manière, à un sens plus large, comme on a une Charte des Droits et Libertés québécoise je pense ou des droits de la personne. Dans ça on est responsables de nos choix, on est responsables de nos actes, on est autonomes. Fait que peut-être renforcer la Charte des Droits et Libertés des personnes face à l'IA d'une manière générale.

Ils ont également parfois mentionné que les enjeux soulevés relevaient de la responsabilité d'institutions déjà en place comme Santé Canada, la Food and Drug Administration (FDA) américaine, la Commission d'accès à l'information québécoise, le Commissariat à la protection de la vie privée canadien, ou encore l'INESSS.

> P52 : Il y a une espèce de double fonction ici. D'un côté, il y a un aspect de s'assurer des questions de vie privée, mais ce n'est pas juste ça. Il y a une portion qui, à la limite, pourrait relever de l'INESSS, qui est davantage de faire le … Il faut qu'il y ait quelqu'un qui soit capable de statuer dans une pondération entre le risque pour la vie privée et les bénéfices, et ça, l'IA, elle ne peut pas... Elle juge [seulement] sur les bénéfices de santé. C'est un jugement sur les crises de santé.

> P42 : Par exemple, la FDA, ils ont engagé douze experts en IA, fait qu'ils peuvent se référer à eux. Normalement, eux sont indépendants et représentent la FDA, et ne représentent pas la compagnie pharmaceutique. Je pense que ça prend des acteurs qui comprennent.

Il existe déjà, pour certains participants, de nombreux mécanismes pour protéger les patients et leurs données (ex. RGPD) :

> P42 : Ce que je peux dire, c'est que les mécanismes pour protéger les patients et les données patient sont déjà existants dans différents pays. Santé Canada, la FDA. L'IA est juste un outil qui arrive en plus, mais il ne faut pas se poser de questions sur qu'est-ce qu'une donnée personnelle. Les données personnelles sont déjà bien énumérées, bien comprises. Qu'est-ce qu'une donnée personnelle *versus* une donnée non-personnelle. Donc, l'IA ne va pas changer ça, mais va plutôt apporter des pressions sur les hôpitaux peut-être, à partager plus facilement les données, mais partager à qui ? Il faut que ce soit partagé avec des compagnies qui sont réglementées, qui sont habilitées à le faire, qui vont suivre l'ordre des choses. Il y a des comités d'éthique qui vont approuver les projets, il y a déjà tous ces mécanismes-là. Qu'est-ce qu'on a vu dans les médias, c'est que les compagnies au nom de l'IA ont détourné ces règles-là et ont été punies d'avoir fait ça. Il faut juste sensibiliser les gens...

Certains citoyens ont considéré les systèmes d'IA en santé comme des dispositifs médicaux qui devaient faire l'objet d'une approbation par Santé Canada (mais qui doivent être réapprouvés régulièrement considérant que les systèmes d'IA sont évolutifs).

> P54 : Une homologation que Santé Canada doit approuver. C'est homologué le truc comme un *medical device*.
> P52 : Ça vient avec un ensemble de responsabilité et des conditions d'homologation.

> P42 : Moi, dans ma perspective, l'IA, c'est un *medical device*, il est déjà couvert par le code de déontologie des médecins, il est déjà couvert par le comité d'éthique des hôpitaux, il est déjà couvert par les agences réglementaires qui vont approuver les produits d'IA comme ils le font depuis 98. L'IA est introduite depuis 98 dans les hôpitaux, pour la détection de cancer [de l'ovaire].
> P37 : Ça, c'est comme un postulat qu'il n'y a rien de différent avec l'IA.

P36 : Peut-être que c'est vrai.

Le problème serait alors d'assurer que les parties prenantes respectent les mécanismes en place plutôt que d'en créer de nouveaux, et remplissent leur mandat. Dans cette optique, il ne serait pas nécessaire de créer de nouveaux dispositifs de gestion des systèmes d'IA en santé complexes mais plutôt de mettre en place des mécanismes simples qui pourraient évoluer avec les avancées technologiques :

> P61 : Je vais surfer là-dessus. Je pense qu'une des recommandations qui pourrait être bien c'est de commencer petit, sur quelque chose qu'on comprend et faire emballer le truc en même temps que l'IA s'emballe. Dans le sens que ça ne sert à rien d'essayer de trouver un mécanisme politique immensément complexe, parce qu'on ne réussira même pas à le concevoir avant de le mettre en application. Commencer par une application simple, et le faire emballer tranquillement pas vite.

Enfin, les participants ont soulevé l'existence d'une « bulle spéculative » ou un certain « fatalisme » qui entoure « l'IA et le Big Data ». De cette idée découle une impression qu'il n'est pas possible de garder le contrôle sur le développement des systèmes d'IA, et favorise l'idée qu'il est nécessaire de mettre en place de nouveaux mécanismes.

> P60 : Et de toute façon, on a l'impression que la technologie va tellement vite qu'on n'a pas le temps de réfléchir, or que si, on a le temps de réfléchir.

P60 a ainsi précisé qu'il était en fait possible de garder le contrôle :

> Je pense aussi que le logo big data bénéficie comme le cyber de cette idée de « C'est hors de contrôle. » alors qu'en fait si. Et qu'on a tellement cette idée que c'est hors de contrôle que on le voit comme quelque chose de pas séquençable, et que ce n'est pas de petites choses qu'on peut gérer mais c'est un ensemble pas gérable, alors qu'en fait si.

2.3.3. Rôles et responsabilités des parties prenantes du développement responsable des systèmes d'intelligence artificielle

Sept catégories d'acteurs, parties prenantes du développement responsable de l'IA en santé, ont été identifiées lors des discussions de la coconstruction. Ces catégories ont été déterminées en fonction du rôle des individus dans le développement des systèmes d'IA dont découle leur responsabilité en termes de conséquences potentielles. Celles-ci ne sont pas mutuellement exclusives, un même individu pouvant se voir attribuer plusieurs rôles, et donc endosser différentes responsabilités (ex. tout individu peut, à un moment donné, endosser le rôle d'utilisateur et donc endosser la responsabilité qui en découle). Ayant discuté les rôles et responsabilités tantôt de

manière descriptive, tantôt de manière prescriptive, se dégagent en filigrane des propos des citoyens différentes recommandations concrètes relatives aux mécanismes de gouvernance à mettre en place[120].

Chercheurs

Les citoyens ont, à plusieurs reprises, discuté d'éléments relatifs au rôle de la recherche et des chercheurs. Ils ont discuté du poids ou de la place des chercheurs dans les décisions et ont mentionné que des derniers avaient une « autorité symbolique », notamment car ils pouvaient « faire des recommandations de gouvernance » mais « sans le pouvoir » (P36).

Ils ont discuté également de la responsabilité des chercheurs face aux découvertes fortuites, à savoir s'il est nécessaire de partager les informations découvertes par le biais des systèmes d'IA aux patients ou rester seulement « spectateur » de ces découvertes, comme discuté ici :

> P39 : J'étais à Santé Canada récemment. Les données ont été prises pour cette année, les patients ont consenti : « Vous pouvez les utiliser ». Après ça, eux les ré-utilisent pour la recherche, et découvrent de nouvelles choses et se demandent : « C'est quoi ma responsabilité moi comme personne qui ait ces données-là ? Est-ce que je dois informer que j'ai découvert quelque chose de nouveau à la personne ? » Donc, il y a ça, mais il y a l'autre côté qui se demande si c'est quoi sa responsabilité comme quelqu'un qui découvre quelque chose. C'est super embêtant.

Selon les citoyens, les chercheurs sont responsables de vulgariser les résultats de recherche, notamment dans l'optique de favoriser l'autonomie de « la population » :

> P41 : En recherche, tout ce qu'ils disent, dans n'importe quelle affaire de recherche, leur gros *outcome* à la fin, c'est *knowledge transfert*. Tu pourras vouloir dire tout ce que tu as voulu dire, mais c'est comment tu peux le …
> P42 : … le vulgariser...
> P41 : … le vulgariser à la fin pour que la population ou les utilisateurs s'en servent. Et si on manque cette coche-là, ça reste dans la boite noire.

Les citoyens ont également soulevé que les attentes en pratique relativement à la responsabilité pouvaient varier selon que les chercheurs exercent dans une institution publique ou privée, notamment car ils ne répondent pas aux mêmes normes de conduite responsable en recherche et d'intégrité scientifique.

[120] Comme la catégorisation de ces mécanismes a déjà été réalisée dans le cadre de l'analyse de la Déclaration de Montréal, ces propositions n'ont pas fait l'objet ici d'une analyse plus ciblée. Elles sont cependant reprises brièvement dans le Chapitre 6.

Les citoyens ont mis de l'avant l'importance de la recherche dans le développement responsable des systèmes d'IA :

> P55 : C'est important la recherche ! C'est qui va faire grandir le reste.
> P56 : Ça permet de faire développer les connaissances dans ce domaine-là.

> P35 : Je suis tout à fait d'accord, il faut de la recherche.
> P41 : Non, mais voir qu'il y a une lacune sur la recherche en IA et il faut promouvoir ça.

Ils ont à plusieurs reprises manifesté des attentes relatives au soutien de la recherche, notamment par le biais de différents incitatifs et subventions. Selon plusieurs d'entre eux, il est nécessaire d'encourager et de soutenir des programmes de recherche interdisciplinaires et intersectoriels (incluant la philosophie, les sciences sociales ou la bioéthique). Ces recherches pourraient être réalisées en vue de favoriser le développement des algorithmes en *open source* et la mutualisation, porteraient sur les biais, les conséquences de l'IA et ses impacts sur la santé des individus, l'accessibilité des soins ou encore l'amélioration des algorithmes.

Les citoyens ont particulièrement mis de l'avant qu'il était nécessaire de ne pas uniquement soutenir la recherche sur les aspects « techniques » du domaine de l'IA, car cette recherche ne permettrait pas d'anticiper aussi bien tous les enjeux attachés au développement responsable des systèmes d'IA, notamment parce que « *ça touche des aspects humains* » :

> P52 : Et pas uniquement de la recherche fondamentale sur : envoyer des algorithmes dans des bases de données, mais aussi la recherche non seulement sur ... en sciences sociales, sur les conséquences et tout l'environnement autour de l'IA.

Ils ont également mentionné qu'il était nécessaire de « tempérer la neutralité scientifique » car de là découle « le risque sociétal ».

Enfin, les citoyens ont mentionné la nécessité de développer des instruments similaires à l'Énoncé de Politique des Trois Conseils (version 2), développés collectivement par la communauté scientifique – notamment par le biais de consultations - dont le but serait d'accompagner les chercheurs :

> P62 : L'ÉPTC 2, c'est intéressant parce que ce que tu es en train de nommer, c'est le développement des normes qui sont ensuite encadrées dans un règlement, mais le développement est fait par la communauté scientifique pour la communauté scientifique.

C'est un acte collectif. Même s'il y a énormément de débats et de chicanes, mais c'est quand même un acte collectif. Ce n'est pas quelque chose d'imposé d'en haut, juste parce qu'il y a quelques décideurs qui ont décidé. C'est suite à des consultations à large échelle.
P61 : Et par définition, c'est...
P60 : ...C'est évolutif.

Développeurs

Les citoyens ont également reconnu une responsabilité aux développeurs, car ils sont en amont du développement des systèmes d'IA :

P28 : Oui, mais est-ce qu'on veut que ce soit seulement l'utilisation qui soit bien faite ou est-ce qu'on veut que le logiciel ait bien été conçu ? Dans un tel cas, c'est pas une question de formation, c'est une question de contrôle plus à la source.
[…]
P32 : En fait, t'as deux composantes, le corps médical et la conception du logiciel lui-même. Qui peuvent être complémentaires.
P31 : Donc, il a fallu qu'il y ait sa contribution... sa collaboration.
P28 : Moi, je pense qu'il faudrait prévenir à la source et imposer à la conception du logiciel, que certains mécanismes obligatoires existent. »

P52 : Soyons honnêtes un instant ! Gravité de la maladie, probabilité de traitement efficace, comportement de santé des patients : on a quand même des critères qui sont déterminés d'avance. Après ça, il y a plein d'éléments qui vont être balancés, mais fondamentalement il y a une programmation de base dans laquelle il y a des valeurs qui vont être intégrées et là-dessus, il y a un gros enjeu de responsabilité sociale. Comment est-ce qu'on va, au départ au moins, programmer nos algorithmes ?

Les développeurs seraient imputables des conséquences de l'utilisation de systèmes d'IA, notamment car ils sont responsables des « intentions qu'ils mettent dans le modèle » :

P43 : Non, je veux dire, ce n'est pas parce que c'est le robot qui va faire quelque chose que c'est lui qui a posé l'acte en tant que tel. Si l'algo est orienté de telle façon qu'on veut que la personne soit la plus calme possible, ce n'est pas parce que le robot il va lui donner plus de telle pilule que de telle autre que c'est lui qui pose l'acte. C'est la façon dont l'algo a été développé et les intentions qui ont été mises dans le modèle.

Les citoyens ont parfois soulevé la nécessité de créer des codes de déontologie pour les développeurs – notamment au sein des entreprises d'IA - afin de les rendre « responsables de ce qu'ils font », notamment concernant les enjeux de sécurité et de transparence :

P21 : Parce qu'en ce moment, comment ça se passe dans les entreprises c'est que n'importe quel codeur ou développeur à aucune responsabilité d'une certaine façon sur ce qu'il fait. C'est l'entreprise qui est responsable, il n'est pas possible de poursuivre le développeur s'il a fait une erreur. S'il a fait exprès, [on est] peut-être capable d'aller un peu plus loin, mais

> ultimement le code appartient à l'entreprise. Tandis que là ça serait de les rendre responsables de ce qu'ils font.

P20 réagit cependant à cette proposition en mentionnant que dans ces conditions, elle aurait « peur de coder ».

Les citoyens ont formulé plusieurs recommandations d'ordre technique, qui relayent ainsi la responsabilité aux développeurs, comme par exemple la mise en place de « garde-fous » dans les architectures des algorithmes qui empêcheraient de « dévier le diagnostic » et responsabiliseraient ainsi le développeur en cas d'erreur.

> P21 : Il a mis son... Ok, mettons que le code de déontologie mettrait, imposerait aux développeurs qu'il y ait toujours des garde-fous à l'intérieur des algorithmes pour pas qu'ils partent d'un bord ou de l'autre.

Les citoyens ont parfois soulevé que cette responsabilité est complémentaire de celle des utilisateurs :

> P28 : Moi, je pense qu'il faudrait prévenir à la source et imposer à la conception du logiciel, que certains mécanismes obligatoires existent. C'est sûr qu'après ça, il faut s'assurer que les gens l'utilisent comme il faut mais si le mécanisme n'existe pas, ils vont juste faire leur gros possible avec ce qu'ils vont avoir.

Utilisateurs

Les participants ont également reconnu une forme de responsabilité aux utilisateurs des systèmes d'IA face à l'usage de ces dispositifs. Cette responsabilité est notamment issue du fait que les algorithmes apprennent sur la base de données générées par les utilisateurs eux-mêmes. Les sources ayant « un grand impact sur le développement », les utilisateurs devraient utiliser les systèmes d'IA de manière responsable, puisqu'ils risquent par exemple d'y introduire des biais.

> P21 : Comme l'IA de Microsoft qui est devenu Nazi. C'est pas la responsabilité de Microsoft ou du développeur, c'est tous ceux qui ont nourrit...

Ont généralement été considérés comme utilisateurs des systèmes d'IA soit les patients, soit les professionnels de santé.

> P41 : À vrai dire, il y a l'utilisateur. Là, je dis ça de même, oui ou non. […] Tu en mets un du côté « médical/chercheur », et côté « patient ». On se dit : « Ok, on en fait un d'un bord et un de l'autre. » […]. Toujours penser qu'il n'y pas juste le patient là-dedans. Il n'y a pas juste l'utilisateur. L'utilisateur, c'est celui qui s'en sert pour les bienfaits et l'autre,

> l'utilisateur qui va avoir supposément une meilleure qualité de vie à la fin. Comme deux grosses familles d'utilisateurs, théoriquement.

Il est à noter cependant que tous les autres acteurs sont susceptibles d'être considérés comme des utilisateurs des systèmes d'IA, et ainsi voir leur responsabilité se définir sur la base de ce rôle.

Les citoyens ont, là aussi, souligné que cette responsabilité des utilisateurs est complémentaire à celle des développeurs :

> P41 : Finalement, peu importe, ça peut être n'importe quoi. […] Ça peut être du macramé. Mais c'est l'utilisateur du macramé qui va faire la différence. Fait que l'éthique, on devrait plutôt se protéger et viser les utilisateurs plutôt que de dire, c'est toujours l'IA qui est le …
> P36 : Là, on est dans la responsabilité...
> P41 : Responsabiliser les utilisateurs.
> […]
> P42 : Je peux te vendre un couteau et je dis : « Ce couteau-là, ça sert juste à mettre le beurre sur ton pain. » Mais, tu peux prendre un couteau et te faire mal. […]
> P36 : C'est pour ça que, moi, ça m'apparaît en ligne avec la responsabilité. Parce que la responsabilité de l'usage attendu, c'est la responsabilité de la personne qui le met en marché : manufacture, qui fait les étiquettes, qui explique comment l'utiliser.

Ils ont également reconnu qu'elle pouvait entrer en tension avec celle des entreprises au regard de la collecte des données, comme mentionné lors des discussions concernant les clauses de confidentialité (qui ne sont pas lues) desdites compagnies (qui collectent les données en dehors du système de santé) :

> P15 : C'est comme une responsabilité corporative. C'est pas l'utilisateur, mais beaucoup plus les corporations, non ? Parce qu'ils ont comme une responsabilité face à leurs utilisateurs. Eux ils ont leur...
> P19 : Oui, oui je sais bien, mais c'est notre responsabilité aussi. Quand on sait, c'est notre responsabilité.
> P15 : Oui, on est responsable.
> P17 : Il faut vraiment s'engager un peu mieux envers nous-mêmes. Il faut se prendre au sérieux. Il faut arrêter de faire les enfants.

Pour certains, une véritable « démocratie » impliquerait que la gestion du développement responsable des systèmes d'IA en santé ne soit pas dans les mains des institutions mais que ce développement soit collectivement géré par les « usagers eux-mêmes », à l'image de la gestion de « bases de données collectives ». Pour garantir une utilisation responsable, les citoyens ont recommandé que les utilisateurs soient formés à l'utilisation des systèmes d'IA afin qu'ils puissent développer leur sens critique, ce qui renvoie aux considérations relatives à la capacitation.

Professionnels de santé

Concernant les professionnels de santé, les citoyens ont souligné leur responsabilité face à l'usage des systèmes d'IA, notamment face au risque d'erreur. Les professionnels de santé ont la responsabilité de la bonne communication de l'information aux patients (notamment l'information probabiliste) ; de comprendre l'IA (bien que cela nécessite alors de mettre en place des formations) ; de faire un suivi auprès des patients et de ne pas les laisser « seuls aux mains des IA » ; et « d'utiliser la donnée à bon escient ».

Les participants ont défini le rôle des professionnels de santé comme une responsabilité professionnelle et déontologique. Ils considèrent que le code de déontologie des médecins devrait être modifié pour y intégrer les aspects relatifs à l'utilisation des systèmes d'IA, notamment en ce qui a trait à la « relation médecin-patient-IA » :

> P23 : Est-ce que c'est possible de juste le modifier ? Il faudrait voir à l'intégration dans le code de déontologie des médecins de la responsabilité face à l'IA.

Face au fait que la responsabilité des médecins peut être mise au défi par les systèmes d'IA[121], les citoyens ont mentionné que le médecin doit conserver le rôle de « constater la maladie ».

> P56 : Le médecin a une responsabilité par rapport à son patient. S'il dit : 'L'algorithme me dit ça, donc fais ça. Et si ça ne marche pas, c'est la faute de l'algorithme, ce n'est pas ma responsabilité professionnelle qui est en jeu.' On est dans une autre *game* là.

Les citoyens ont également soulevé une responsabilité des professionnels de santé de partager leurs expériences avec les membres de leur communauté :

> P29 : Réunion des professionnels concernés, évidemment. Si personne ne se parle et que chaque médecin vit les cas avec son robot à côté de lui, il ne partagera pas avec personne.

Et d'éviter le paternalisme médical :

> P41 : Tantôt on parlait de 'un mauvais patient' … 'un mauvais médecin'. On est en train de mesurer le médecin par rapport à l'approche qu'il a et là en revenant au numérique, c'est comme si on retournait en arrière, dans le paternalisme : 'Voici ton diagnostic, voici ce qui va se passer avec toi mon grand.' Et il n'y a aucune consultation ou [équipe] approche avec le patient et il y a un équilibre à y avoir dans tout ça.

[121] *Cf.* Section 3 du présent chapitre sur l'agentivité artificielle.

Patients

Les citoyens ont également reconnu le patient comme « partie prenante » du développement responsable des systèmes d'IA en santé, comme le mentionnent ici les citoyens discutant de la responsabilité de la patiente mise en scène dans le scénario :

> P47 : Elle est partie prenante.
> P44 : C'est vrai, elle a sa responsabilité. On a tendance à exclure la personne...

Avec l'utilisation des systèmes d'IA, la responsabilisation des patients serait accentuée :

> P46 : Je trouve que par rapport à la santé, c'est comme par rapport à l'autonomie et à la liberté de la personne. Il y a un effet de déresponsabilisation, parce que le robot dit : 'Le pilulier, tiens voilà ta pilule'. Mais il y a aussi un effet responsabilisation : 'Maintenant, vous avez un robot, ne venez pas nous dire que vous ne pouvez pas vous rendre en santé, parce que vous avez tous les éléments pour le faire'.

Il a été reconnu que le patient « a le dernier mot », notamment face à son traitement :

> P23 : Mais aussi par rapport à la responsabilité du patient ? Parce qu'aujourd'hui, moi je suis patiente, qu'il y ait un robot qu'il n'y ait pas de robot, mon médecin va me dire 'Je vais vous mettre un *PaceMaker*'. Ok mais pourquoi, c'est quoi les autres possibilités qu'est-ce que je peux avoir ? Pis je peux décider de me faire traiter ou de ne pas me faire traiter. Fait que la responsabilité de la personne reste quand même, devrait rester au centre.

Les patients pourraient également répondre à un certain devoir d'avoir recours à des systèmes d'IA, comme le discute ici un citoyen se référant à l'utilisation de robot de soin pour assurer le maintien à domicile des personnes âgées :

> P1 : Dans ce cas-là, finalement, c'est un contrat entre une personne qui est en perte d'autonomie et la société complètement. […] on lui dit : « Ok, tu peux rester chez toi, mais tu es en perte d'autonomie, à la condition que tu acceptes que ce robot t'accompagne. » Au départ, c'est un contrat. Sans ça, la personne n'aura pas le choix. On va institutionnaliser une personne parce qu'elle est un danger à elle-même. Tandis que là, on lui propose une alternative. 'Tu peux rester chez toi, mais à condition que ...' Aujourd'hui ce qu'on fait, on leur dit : 'Tu peux rester chez toi, à condition que tu es une garde-malade'.

D'autres ont cependant tempéré le rôle des patients ici en mentionnant qu'il était nécessaire de s'assurer que cette obligation d'utilisation ne devienne pas « dictatoriale ».

Entreprises

Concernant les entreprises, les citoyens leur ont reconnu une responsabilité par rapport au bon fonctionnement des systèmes d'IA, notamment au regard des conséquences de l'usage des technologies qu'ils développent, à l'image d'autres types d'entreprises :

> P24 : C'est comme un peu la responsabilité des compagnies qui créent la cigarette.

Cette responsabilité pourrait prendre la forme d'une obligation pour les compagnies qui créent des systèmes d'IA d'établir des normes, des comités d'éthiques, des codes de conduite ou de déontologie :

> P24 : Pour moi le code de déontologie c'est plus créer un comité d'éthique au sein des entreprises qui créent l'IA. Ce n'est pas juste les médecins, les entreprises privées qui créent l'IA.

Tandis que certains des citoyens participants restent sceptiques à l'idée que les comités d'éthique soient de la responsabilité des compagnies :

> P52 : Permettez-moi, honnêtement, d'être sceptique des comités d'éthiques dans les entreprises et des bonnes pratiques ...
> P55 : On ne va pas dire que c'est mal, mais il faut y croire.
> P52 : Ben moi, je ne le crois pas, je ne pense pas que c'est utile.

Les compagnies pourraient également avoir comme devoir de déterminer une personne responsable :

> P52 : C'est parce que l'idée, c'est qu'il faut obliger toute entreprise ou organisation qui a ce genre de pratiques-là, à avoir au moins en son sein quelqu'un qui est légalement responsable.

Par exemple, un membre de la communauté médicale pourrait être nommé dans chacune des entreprises et responsable du respect du bien-être des patients.

Pour d'autres, nommer une personne responsable n'est pas la solution :

> P52 : Parce que, ultimement ce que tu veux, ce n'est pas tant savoir... C'est un peu ça que je disais tantôt. Ce n'est pas tant de savoir qui poursuivre quand ça va mal, c'est avoir les meilleures manières de t'assurer que ça n'aille pas mal. Dans une entreprise, si tu as quelqu'un de responsable, au pire, c'est un siège éjectable on passe à d'autres choses. C'est la manière dont tu organises tout ça… est extrêmement sensible.

Les compagnies seraient garantes de la privatisation et de la sécurité des données qu'elles collectent, car il s'agit d'une question d'« intégrité de la compagnie » mais aussi de « responsabilité

corporative ». Concernant les données collectées, les citoyens ont discuté du partage de la responsabilité entre utilisateurs et compagnies qui les collectent par le biais de systèmes d'IA :

> P62 : Ça c'est un bon point lié à la responsabilité : Est-ce que la compagnie reste toujours responsable ? Ou, si j'achète ça, après 30 jours, c'est toute ma responsabilité ? *Right* ? S'il y a un problème, c'est à moi à gérer. Est-ce que ça reste toujours, probablement dans un contexte de robot hyper compliqué, est-ce que ça reste toujours la propriété d'Alphabet ou d'une compagnie et c'est loué à des utilisateurs ?

Les citoyens ont recommandé la mise en place d'approbations ou de certifications qui attesteraient du bon fonctionnement des dispositifs et rendraient les entreprises imputables en cas de problème. D'autres ont mentionné qu'il fallait inciter les compagnies à « travailler avec le public ». Pour que les systèmes d'IA d'une compagnie privée soient certifiés, certains ont proposé de les forcer à rendre les algorithmes publics, et de créer des lois pour limiter leur influence.

Institutions publiques

Les citoyens participants ont reconnu une responsabilité des institutions publiques de protéger l'intérêt du public face au développement des systèmes d'IA, qui prendraient différentes formes.

Les citoyens ont parfois dénoncé le « vide juridique » qui entoure l'IA que le gouvernement se doit de combler, ou d'adapter le **cadre réglementaire et législatif**, notamment en faisant intervenir des personnes « indépendantes et formées à tous les niveaux » :

> P36 : Ce que P42 a fait ressortir, c'est que pour l'IA la réglementation actuelle ne nous permet pas de capter ce qui serait les systèmes intelligents autonomes.
> P42 : C'est ça. C'est un vide.
> P36 : C'est un vide réglementaire.

À de nombreuses reprises, les citoyens participants ont mentionné qu'il était nécessaire de mettre en place des « lois et règlements » afin de garantir une certaine « surveillance ». Ces dispositions devraient favoriser la protection des renseignements personnels, la transparence des algorithmes utilisés en santé, l'interdiction d'être pénalisé si on ne veut pas partager ses données de santé ou celles issues des réseaux sociaux, l'encadrement de la provenance des sources qui servent à l'algorithme pour garantir qu'il n'y ait pas de biais ou encore la garantie d'accès à tous les soins quelle qu'en soit la forme (incluant les systèmes d'IA).

P52 : Mais [j'ai] deux idées moi en fait. D'une part, comme à peu près tout ce qui est collecte d'informations personnelles, avoir un encadrement qui permette de fournir une information transparente et claire aux utilisateurs, donc ça c'est une condition fondamentale de l'exercice de l'autonomie et du consentement.

P24 : C'est ça, c'est écrit ici : une loi pourrait porter sur la provenance, la transparence et l'utilisation des ensembles de données utilisées par l'algorithme.

Certains ont proposé que le gouvernement crée des politiques standardisées sur la vie privée et s'assure de les faire respecter :

P15 : Que ça soit dans un code d'éthique, ça, c'est bon. Mais sinon, avec des politiques de vie privée aussi qui soient vraiment très, très claires aussi et qui doivent être respectées par les organismes. Parce qu'on va avoir une combinaison d'organismes qui vont y avoir accès, des cliniques, et ça va être très compliqué. Sinon, des règles, comment ça s'appelle, des politiques de vie privée qui sont standardisées.
P17 : Des règlementations.
P15 : Exact. Parce que si tu donnes à chaque organisme de créer sa politique de vie privée...
P19 : … Ses règlements …
P8 : C'est pas bon, ça serait trop cafouiller.

Ils ont également soulevé la nécessité de créer des **acteurs institutionnels** responsables du développement de l'IA en santé comme par exemple : une plateforme d'audit publique tenue par des experts indépendants ; une « police d'experts » collective ; une « autorité-médico-sociale » ou « institutionnelle soumise à déontologie » qui puisse « valider que ça reste dans un cadre éthique »; une « instance indépendante » qui évalue le rapport coût-efficacité des algorithmes :

P8 : Comme la cour des comptes : qui évalue la manière dont on va dépenser l'argent, qu'est-ce qu'on gagne en retour et si l'argent est bien dépensé ?
P5 : C'est ça, par exemple combien d'acteurs on a, qui a été rétribué et comment, finalement est-ce que ça a eu un impact si on a rajouté 100 millions dans la machine.

Certains ont soulevé qu'un « organisme pour faire appel » à l'attention des personnes qui considèrent être injustement traitées par un algorithme serait mieux que le développement de lois :

P3 : Moi je suis plus pessimiste parce que les algorithmes, c'est comme les lois. Ce ne sera jamais parfait. Ce sera toujours à interprétation. La même chose que les lois. Par exemple, quand les gens ne sont pas d'accord, ils vont en cour, ils passent devant un juge, des jurés. Les juges peuvent faire pencher la loi d'un côté ou de l'autre mais avec les algorithmes ça ne se fera pas comme ça. Ce qui serait intéressant ce serait d'avoir un acteur institutionnel, d'avoir un organisme où les patients qui ne sont pas d'accord avec l'algorithme, qui pensent qu'il y a quelque chose, qu'ils ont été traités injustement, qui ait… qu'ils puissent avoir des

> droits […]. Qu'ils puissent aller voir et vérifier qu'effectivement l'algorithme est dans le champ, que ce soit ajusté ...

Les citoyens ont également proposé la mise en place d'un « chien de garde » un peu comme un « protecteur du citoyen » ; un « endroit où les patients peuvent faire valoir leurs droits » ; un « ombudsman » ou un « gestionnaire des plaintes » pour régler les litiges.

> P29 : C'est ça, régler les litiges entre les trois. Ça serait un Ombudsman entre le patient, le médecin et l'IA. Le patient n'a pas à être pris à se battre contre ces deux-là.

Les citoyens ont également proposé que le gouvernement s'assure de conserver les données sur le territoire québécois ou qu'un code d'éthique global soit mis en place et destiné à tous les professionnels et usagers qui touchent aux données dans le contexte de la santé dans le territoire canadien.

Ils ont également recommandé que la mise en place d'une **certification** relève de la responsabilité de « comités multipartites indépendants », de Santé Canada ou du Ministère de la santé et qu'elle certifie le stockage et la collecte des données, les algorithmes ou les robots de soin et leur trousse d'outils, notamment afin de garantir la sécurité ou la non-discrimination.

> P51 : Pour l'IA, une certification pourrait s'assurer de la non-discrimination, de la loyauté des algorithmes, de la réalisation d'essais rigoureux et indépendants.

Les citoyens ont plus particulièrement reconnu un rôle **à « l'État » ou au « gouvernement »**, qui devrait assurer l'« accès universels à ces machines, sans négliger les options alternatives », par exemple par l'entremise de subventions. D'autres ont proposé que l'État s'assure de contrôler le marché – qui ne devrait pas rester aux mains de compagnies privées - en s'assurant plus particulièrement que les systèmes d'IA demeurent « *made in Canada* » ou répondent à certaines exigences de transparence :

> P14 : Par rapport à ce que tu viens de dire, il peut y avoir des barrières à l'entrée. Par exemple, moi je suis le premier ministre du Québec ou du Canada maintenant dans mon système de santé, les entreprises qui veulent entrer dans mon écosystème de santé doivent rendre leurs algorithmes publics. Il n'y a pas de problème après pour vendre le modèle de robot à tel ou tel prix. Par contre l'algorithme...

Ils ont soulevé la nécessité que le gouvernement canadien conserve sa souveraineté au regard de l'éthique de l'IA :

> P42 : L'autre question aussi, c'est d'un point de vue autonomie, si on regarde juste le bateau passer, les solutions de l'IA vont venir de la Chine. Si on veut rester souverains, si on veut rester autonomes, nous, on doit prendre le leadership pis influencer nos solutions d'IA.

Ils ont enfin mentionné que les acteurs publics ont la responsabilité de rester « neutres » face à différentes formes d'influence (cf. également la section sur les conflits d'intérêts potentiels du présent chapitre) :

> P51 : Une autre affaire qui me chicotte beaucoup c'est toute la question du lobbysme auprès du gouvernement, de Santé Canada… On le voit avec le lobby alimentaire ces temps-ci là, est si qu'ils vont vraiment réussir à être complètement neutre par rapports aux différents lobbys ?

2.3.4. Attentes normatives relatives au partage des responsabilités

Les citoyens participants ont formulé des attentes normatives relatives au partage des responsabilités entre les différents acteurs du développement responsable des systèmes d'IA. Ils ont notamment formulé des attentes relatives à la transparence des institutions, à la précaution et au partage de la gestion des données.

Les citoyens ont manifesté des attentes relatives à la **transparence des institutions**, qu'il s'agisse des entreprises ou du système de santé, afin de permettre la traçabilité des données et de qui les consulte :

> P1 : Ce serait une obligation pour le système de santé de rendre transparent le chemin que prennent nos données. Ou bien un peu comme, moi par exemple, je vais sur Linkedin et aujourd'hui je reçois beaucoup 'Un tel, un tel a consulté votre profil' ... Ben j'aimerais bien que pour mon profil de santé, savoir qui a consulté mon profil de santé, au moins que je sois vigilante.

Certains des citoyens participants souhaitaient en faire une obligation pour le système de santé « de documenter et de rendre transparent pour le patient l'accès à ses données par des tierces parties ».

> P5 : Informer le producteur initial de la donnée de l'utilisation de sa donnée.

Ils ont également proposé que les systèmes d'IA utilisés par le système de santé fassent l'objet d'une évaluation périodique – notamment concernant la validité des algorithmes - et que cette évaluation soit rendue accessible pour les usagers du système de santé ; avec une « clause de déclaration de toute modification » :

> P5 : Mais moi je rajouterais une clause d'obligation de déclaration de toutes modifications dans le temps. Faut qu'on ait un historique de ce qui s'est passé parce que si tu fais ça et que à la date ... tu veux comprendre ce qu'il s'est passé en 2016 et que tu regardes l'algorithme de 2017, t'es mal pris. Il faut que tu saches ce qu'il s'est passé en 2016, avec quel algorithme on travaillait en 2016. Donc, il faut un principe de conservation de l'historique. De tout ce qui agit sur le public.

Les citoyens ont également recommandé que les institutions soient transparentes au regard de leur respect des règles.

Les citoyens ont parfois présenté la transparence comme un moyen de favoriser l'intelligence collective. Par exemple, si tout le monde ne comprend pas le code à la base des algorithmes, le fait de permettre à ceux qui le comprennent d'y avoir accès a été présenté comme un moyen de « garantir l'intégrité ». Suivant cette idée, développer des logiciels en *open source* pourrait permettre de favoriser la « mutualisation », la « transparence » et la « responsabilisation ».

> P10 : Moi, je dirais privilégier le développement des logiciels d'IA en *open source* de façon à assurer un coût minimal et une réutilisation maximale.

La transparence permettrait également de démontrer à l'avance (et non « après-coup ») que les institutions en jeu sont capables d'utiliser les systèmes d'IA « correctement ». En effet, selon les citoyens participants, le partage des responsabilités devrait se faire dans la perspective d'une certaine ***précaution*** :

> P3 : Mais, pour moi ce n'est pas nécessairement de la modération après-coup, c'est plus la modération dans la conception.

Ils ont mentionné qu'il était nécessaire d'étudier les impacts de l'IA « en amont », afin de ne pas se laisser dépasser par l'avancée rapide des technologies en question :

> P53 : Je pense qu'il faut augmenter les connaissances par rapport aux impacts, parce qu'on est toujours en projection. On se projette en 2025 : est-ce que ça va améliorer ou nuire au système de santé ? On ne le sait pas.

Ou de faire de la sécurité un principe opérationnel dès la conception du logiciel selon une « *privacy by design* » :

> P62 : Dès le design, on est en train de trouver des moyens pour échanger de l'information qui est pertinente et effacer toute l'information qui peut être confidentielle et qui assure que dans toute la logique de développement, on coupe des ponts pour protéger la sécurité. Donc la sécurité est un principe opérationnel dès le début de la construction du logiciel.

Ils ont également mentionné un « devoir éthique de prudence » face aux utilisations des données qui conduisent à faire des prédictions de santé. Cette précaution pourrait être garantie par la mise en place de « mesures non contraignantes », dont le but ne serait pas de brimer ou de décourager, mais plus de de prévenir les dommages.

Pour certains, une approche préventive doit tenir compte du fait que l'utilisation de systèmes d'IA **n'est pas une obligation** et qu'il ne faut pas oublier les « options alternatives » ou « traditionnelles » :

> P5 : Ok on a une problématique et puis on va aller chercher des moyens pour résoudre cette problématique dans les nouveaux systèmes d'IA. Mais bon, est-ce qu'on peut mettre en regard avec des moyens plus simple de prévention de base ?
>
> P11 : Dans ce que tu proposais, c'était que plutôt que d'investir pour développer une IA plutôt l'investir sur d'autres manières de répondre aux mêmes besoins. Pas nécessaire en utilisant une IA. Sauf que si on part dans cette idée-là l'exercice n'a plus vraiment de sens...
> P14 : Oui mais c'est un bon enjeu, parce que pour toutes les solutions qui vont être développées ou même imaginées en IA… Il y a une nouvelle technologie qui arrive et tout le monde se jette dessus. Ce n'est peut-être pas toujours la meilleure.
>
> P13 : C'est juste qu'en technologie je pense que, surtout pour les gros débats de société, c'est important de se rappeler, il y a un dicton en anglais qui dit : "*Because you can doesn't mean you should.*"
>
> P52 : On n'a pas besoin d'avoir tout cet appareillage-là, technologique, pour donner ce genre de recommandations-là de santé publique. On peut juste le faire comme une belle campagne de santé publique tout à fait traditionnelle.

Selon cette conception, certains participants ont mentionné qu'ils souhaitent pouvoir se dire que l'on peut encore « rejeter tout ça » et que notre choix ne soit pas limité au comment mettre en place les systèmes d'IA, ce qui nous met « devant le fait accompli » - bien que cela ne fasse pas consensus :

> P40 : J'aimerais croire qu'on a l'option de dire on ne veut pas d'IA, on peut faire ça. On peut dire : « C'est trop risqué. C'est trop risqué, ce n'est pas le genre de société qu'on veut avoir ». Je pense.
> P35 : Mais on l'a déjà avec nous. On vit avec.
> P40 : On peut l'arrêter. On peut reculer. On peut faire plein de choses.
> P35 : C'est assez difficile.

> P40 : Oui, mais si on voit des conséquences, si on voit des impacts et on décide... Ce n'est pas essentiel à la vie.

Pour d'autres, il n'est pas possible d'aller contre le développement des systèmes d'IA. P28 a par exemple mentionné qu'il était « peut-être utopique » d'imaginer qu'on pourrait maintenant ne pas utiliser l'IA.

> P38 : Ce n'est pas l'IA versus le non-IA, c'est comme si par exemple, est-ce qu'on dit : « On arrête de faire de la science » ? C'est du même niveau.

> P1 : On ne peut pas lutter actuellement contre l'implantation de systèmes comme Soline[122], etc. Et il y a beaucoup d'aspects positifs (plutôt que de laisser une personne seule à la maison et responsable), mais en même temps, quelle recommandation on peut faire pour qu'il n'y ait pas un dérapage ? Et, c'est vers là qu'on doit s'en aller dans les discussions. C'est ça que je veux amener. On parle déjà que c'est quelque chose de positif, sans ça, personne ne va l'implanter. Rappelez-vous que c'est la société qui veut ça.

D'autres ont également mentionné qu'il ne fallait pas encadrer le développement des systèmes d'IA de manière trop restrictive afin de ne pas nuire à l'avancée des connaissances :

> P1 : Parce qu'il ne faut pas ... en mettant trop de contraintes on va aussi empêcher la recherche. Il est important de ne pas empêcher la progression de la science.

Les citoyens ont exprimé différentes attentes relatives à la **propriété et à la gestion des données**. Plusieurs ont défendu que les données doivent rester la propriété des usagers afin qu'ils gardent le contrôle sur leurs données de santé :

> P38 : Si les gouvernements mettaient leur culotte et disaient : « Les données sont les propriétés des usagers » plutôt que de la compagnie, ça serait différent.

Ils ont également recommandé de restreindre l'accès aux données, notamment aux compagnies privées (bien que ce soit eux qui les collectent) ou d'interdire de revendre les données (peu importe l'organisme qui les détient ou qui les collecte).

Pour d'autres, il n'est pas possible d'exercer une propriété sur les données de santé, qui devraient être accessibles à la demande :

> P5 : Quand on touche à la santé, je pense que ça devient publique par définition. Donc, là on est comme dans une sphère particulière de l'entreprise qui est ... moi je mettrais des clauses de données, données ouvertes obligatoires. Dès lors que ton terrain de jeu financier

[122] Le citoyen fait ici en fait référence au robot de soin mis en scène dans le scénario.

> c'est la santé des autres personnes ben tu t'engages à rendre accessible peut-être avec l'instance là. Donc à la demande, mais tu t'engages à rendre accessible ton historique, tes affaires...

Cette accessibilité va de pair avec une certaine responsabilité sociale des personnes qui utilisent les données :

> P43 : C'est facile. [Le résultat] de cette accessibilité, c'est une responsabilité sociale. Donc, dans le cadre de la responsabilité, on a la responsabilité en direct avec le patient : Qui est responsable de quoi ? Et puis, dans le cadre de l'accessibilité, on a la responsabilité sociale de donner, de fabriquer une société plus juste, où on est responsable de qui va avoir accès à quoi.

Les citoyens ont également mentionné qu'il était nécessaire d'assurer sécurité et protection dans le contrôle des données, et ont formulé des attentes relatives au partage des données entre les différents acteurs :

> P61 : Oui et... En tout cas, une idée à réfléchir : Il y a aussi le fait que la donnée peut être transportée, transférée peut-être potentiellement entre des acteurs proximaux, mais après si tu veux te transposer sur un autre acteur, tu ne devrais peut-être pas transférer tes données brutes initiales du *output*, mais transférer toujours le résultat et non pas la donnée brute initiale. Fait que finalement le transfert va juste se faire entre des acteurs proximaux qui comprennent le bien éthique, pis la donnée brute ne va jamais être extrapolée au niveau global, parce que ça va être juste le résultat qui va monter de niveau.

En mentionnant que les institutions bancaires gèrent déjà bien ce genre de partage avec les systèmes de chaines de blocs.

Enfin, les citoyens ont souligné un potentiel « focus individualiste » et ont questionné s'il était possible d'obliger le partage des données en vue du bien commun :

> P52 : Moi, je trouve qu'on a un focus extrêmement individualiste, mais ça peut être ce qui est choisi. C'est beaucoup l'autonomie de la personne qui est mise de l'avant ...
> P58 : Le bien commun n'apparaît pas beaucoup...

3. L'agentivité humaine au défi de l'agentivité artificielle

3.1. De nouveaux agents (moraux) ?

Le troisième grand défi de l'exercice de la responsabilité soulevé par les citoyens est celui issu de la reconnaissance d'une certaine agentivité aux systèmes d'IA[123]. L'agentivité présupposant généralement l'autonomie (bien qu'il existe de nombreuses conceptions de ces deux concepts) (Noorman 2008), l'autonomie croissante des technologies numériques, en particulier des systèmes d'IA, posent aujourd'hui des préoccupations et défis cruciaux en termes de responsabilité (Floridi et Taddeo 2016). En effet, les avancées en IA – et particulièrement celles de l'apprentissage profond - compliquent les conditions de l'attribution de la responsabilité et amènent certains philosophes à reconsidérer l'agentivité morale, jusqu'ici attribuée seulement aux êtres humains (Noorman 2016). L'utilisation de systèmes d'IA augmente la complexité et la distance entre le concepteur de l'algorithme et les conséquences de ses applications, en laissant naître un risque de déresponsabilisation des humains face aux conséquences des technologies qu'ils ont eux-mêmes créées (Zwitter 2014; Noorman 2016). Une conception possible de la responsabilité est alors de reconnaître que l'agentivité morale n'est pas l'apanage des seuls êtres humains. Comme mentionné dans le Chapitre 4, plusieurs questionnent alors dans quelle mesure les agents artificiels peuvent ou doivent être considérés comme des agents moraux (Noorman 2016; Bostrom et Yudkowsky 2011; Moor 2006; Bucher 2016), notamment en inscrivant des valeurs similaires aux nôtres dans l'architecture des algorithmes (Scheutz 2016).

Certains chercheurs en éthique des machines s'accordent pour considérer les agents artificiels comme des agents moraux implicites, c'est-à-dire des agents qui s'occupent de leur mission attendue de façon sûre et responsable sans pour autant qu'ils soient capables d'étendre ces fonctions à d'autres situations (Moor 2006; Allen, Wallach, et Smit 2006). S'il n'existe actuellement pas d'exemples clairs qu'ils puissent être des agents moraux explicites, soit capables de décrire précisément une situation morale et poser un jugement (Moor 2006), les développements du domaine montrent pour certains qu'ils tendent à le devenir (Shulman, Jonsson, et Tarleton

[123] Comprise ici strictement comme la capacité d'agir ou de prendre des décisions et, le plus souvent, d'agir ou de prendre des décisions tel qu'un professionnel de santé. Elle est notamment ici entendue comme une condition à l'imputabilité. La pertinence de cette conception est discutée dans le Chapitre 6.

2009). Qu'ils puissent devenir des agents éthiques à part entière, soit des agents capables de poser des jugements éthiques explicites et de les justifier raisonnablement (Moor 2006), relève cependant toujours de la prospective, bien que pour certains ces limites seront bientôt dépassées (Shulman, Jonsson, et Tarleton 2009; Gordon 2019).

Or, jusqu'à récemment, la responsabilité face au développement technologique ne s'était articulée qu'autour du développement de technologies non-autonomes, relayant la responsabilité des effets de leurs usages aux humains qui les conçoivent ou qui les utilisent. Des inquiétudes relatives aux conséquences de l'automatisation du système de soins apparaissent alors, notamment relativement aux erreurs issues d'une prise de décision algorithmique qui échapperait au contrôle des humains (particulièrement problématique s'il s'agit, par exemple, d'une erreur diagnostique) (Kononenko 2001; Sharkey 2008; Alexiou, Psixa, et Vlamos 2011). Qui peut être tenu responsable d'une « mauvaise » prise de décision : l'analyste qui a développé le système, l'utilisateur qui ne comprend pas les limites du logiciel, ou l'algorithme lui-même ?

Sous-jacente à ces considérations, l'analogie implicite entre l'autonomie humaine et artificielle (Noorman 2008) vient alimenter la possible déresponsabilisation humaine, étant basée sur la supposition que les machines vont devenir des entités animées qui vont définir ou accomplir leurs objectifs de manière indépendante (Noorman 2008). Cette perception vient brouiller la distinction entre l'humain et les machines, notamment parce que les systèmes d'IA deviennent plus proches des humains à mesure que leurs capacités augmentent (Noorman 2008) ou sont perçus comme tels selon différents mécanismes d'anthropomorphisation et de projection d'intention (Devillers 2017; Proudfoot 2011). Une tension marquée entre l'agentivité humaine et l'agentivité des machines apparaît alors. En assurant une certaine responsabilité des agents artificiels par le biais de l'inscription de valeurs morales dans leurs algorithmes, on ne fait d'une certaine manière que déplacer le lieu et le moment de la responsabilité des personnes qui les conçoivent, soit en responsabilisant l'agent artificiel à la place des individus. En filigrane de ces préoccupations, notre rapport à la technologie pourrait ainsi se retrouver bouleversé, comme le présente Noorman (2008) :

> A change in human/technology relationships that would make the analogy between human autonomy and autonomy of computers less contested would require a major shift in how

> we think about technologies as well as in our more fundamental beliefs about moral responsibility (p. 67)

Concernant l'agentivité (ou la perception d'agentivité) des systèmes d'IA, le discours des citoyens participants est assez consensuel : leur principale crainte est de reconnaitre une agentivité aux systèmes d'IA, et leur principale attente est que les systèmes d'IA soient considérés comme des outils. S'ils craignent que les systèmes d'IA deviennent des agents de soin, c'est relativement au biais qu'ils pourraient contenir, ou aux conséquences sur la prestation des soins du fait de leur accorder une trop grande importance – notamment, une perte de l'individualisation et une déshumanisation. Reconnaitre une agentivité aux systèmes d'IA est intrinsèquement lié à la confiance que l'humain place en la technologie et relève d'une transformation du rapport à la machine, thèmes qui ont été largement discutés lors de la coconstruction.

3.2. Craintes relatives à l'agentivité des systèmes d'intelligence artificielle

Lors des discussions, les citoyens ont souligné à plusieurs reprises une crainte que les systèmes d'IA soient considérés comme des agents. Cette crainte est associée au fait que ce sont des algorithmes qui « prennent la décision » ou « gèrent la priorité d'accès aux soins » avec peu, voire aucune, modération humaine. Ils ont soulevé à plusieurs reprises que cette agentivité est associée à la relative autonomie des systèmes d'IA, notamment dans leurs rapports aux patients. Les citoyens ont questionné ce niveau d'autonomie. Par exemple, les systèmes d'IA transmettent-ils seulement de l'information ou est-ce qu'ils peuvent également « intervenir » ?

> P9 : J'ai une question sur l'autonomie, c'est-à-dire jusqu'où peut aller Vigilo ? C'est-à-dire que si, comme ta mère, elle décide qu'elle ne prend pas ses pilules, mais qu'elle les cache ? Est-ce que Vigilo s'en rend compte ? Est-ce qu'il fait juste transmettre l'information ou est-ce qu'il peut intervenir ?

Cette autonomie s'accompagne de questionnements relatifs à la responsabilité qu'il est possible d'attribuer aux systèmes d'IA – à la différence des technologies médicales traditionnelles :

> P21 : Au moment où quelqu'un se mettrait à rédiger un code de déontologie, admettons pour les développeurs, il y va avoir une réflexion qui va devoir se faire. C'est à partir de quel moment que l'algorithme devient responsable de lui-même ?

> P43 : Quand vous allez passer une échographie ou que vous faites une radiographie pulmonaire, qui est imputable ?

> P47 : Généralement, il y a plusieurs personnes qui opèrent. Les niveaux de responsabilité sont partagés. Sauf que le robot, il est seul dans le domicile.
> P46 : Par exemple, un *pacemaker*, est-ce qu'il y a une responsabilité attachée au *pacemaker* ?

En agissant de manière autonome, des questionnements relatifs à quel type d'informations le robot peut révéler au patient ou transmettre à l'équipe médicale ont été soulevés :

> P9 : Est-ce qu'elle peut demander à Vigilo : Est-ce que je me dégrade vraiment très vite ? À quelle rapidité ça va ? À ce rythme-là, j'en ai pour combien de temps ?

Reconnaitre une certaine autonomie aux agents artificiels viendrait, selon les citoyens, défier la responsabilité du médecin. L'opposition entre agentivité humaine et agentivité artificielle s'illustre dans les dilemmes auxquels les professionnels de santé pourraient être confrontés face à des prises de décisions médicales, en particulier lors de diagnostics. L'utilisation de systèmes d'IA exposent les médecins à de nouveaux dilemmes dans leurs prises de décisions (ex. si l'opinion du médecin va à l'encontre de la recommandation algorithmique, comme mis en scène dans le scénario *Décision thérapeutique à l'hôpital*), ou exacerbe les difficultés déjà rencontrées avec d'autres technologies médicales. C'est bien l'autonomie qui fait la distinction avec les autres outils médicaux selon les citoyens. Les systèmes d'IA se distingueraient en effet des technologies traditionnelles car les systèmes « apprennent tout seuls » :

> P47 : La différence du robot, c'est qu'il est apprenant, c'est qu'il y en a une partie qui vient de lui-même. De la façon dont les données sont traitées dans son système apprenant. C'est ça qui le différencie de la machine d'imagerie, ou le *pacemaker*. C'est qu'il est apprenant.

Également, attribuer une agentivité aux systèmes d'IA serait issu des avancées majeures du domaine, qui conduisent à donner plus de crédit aux technologies qu'auparavant. L'IA serait perçue comme « plus fiable » ou « plus efficace » que les technologies médicales traditionnelles. Les citoyens ont d'ailleurs mentionné qu'attribuer une certaine agentivité aux systèmes d'IA serait « plus sécuritaire » et permettrait de « réduire l'incertitude ». Cependant, ceci risquerait d'influencer le médecin, de rendre la décision d'autant plus difficile à renverser et viendrait limiter le fait que l'on s'autorise à remettre en question les recommandations algorithmiques, présumant que les systèmes sont « parfaits ».

> P27 : Ça ne sera pas juste des médecins ou des groupes de médecins, là ça va être une intelligence qui a été prouvée avec 25 millions versus un docteur.

> P26 : L'algorithme quand on peut l'utiliser comme ça, c'est comme si c'était 10 000 experts. C'est une statistique de plusieurs personnes.

Cette situation pourrait conduire à donner une grande importance aux systèmes d'IA et par là-même leur reconnaitre une certaine agentivité. S'en suivrait potentiellement une déresponsabilisation (par exemple, des développeurs) ou une perte de jugement :

> P23 : Ok mais je voudrais juste soulever un enjeu. Parce que oui au départ il y a des gens qui vont coder, mais après ça c'est l'algorithme qui apprend de lui-même. Fait que là c'est l'algorithme qui va devenir responsable de la propre création de l'algorithme qui a créé pour... c'est pas super évident qu'ils vont avoir une responsabilité...

> P46 : L'enjeu 2, c'est : la perte de jugement. Je trouve que les décisions sont prises par Vigilo, ou on se repose beaucoup sur les décisions prises par Vigilo et trop d'attentes dans cette situation-là envers le robot.

La décision devient également difficile à contester car il n'est pas possible de « dialoguer » avec l'IA; peu importe sa réelle capacité à décider. Certains se demandent quelles seront les conséquences sur le pouvoir décisionnel du médecin :

> P62 : Ça, c'est un autre élément que tu as soulevé. La peur. Moi, je pose la question : si l'algorithme est devenu le *goal center*, le nord, est-ce que le médecin a encore sa liberté professionnelle de dire : ' Non, je ne suis pas convaincu par l'algorithme'. Ou est-ce qu'il va dire : 'Oui, je vais le faire quand même. Même si je me sens mal à l'aise parce que j'ai la peur de poursuite si je vais contre l'algorithme qui est maintenant le standard'. Parce que l'algorithme a une puissance légale et même si c'est jamais démontré dans une cour, quand même il y a un sentiment de peur de poursuite. Et ça, surtout aux États-Unis, mais quand même au Canada.

Reconnaitre une agentivité aux machines au détriment de celle des humains soulève plusieurs questions. Quelle décision prévaut : celle du médecin ou celle de l'algorithme ? Serait-il possible de reprocher à un médecin de ne pas avoir suivi la recommandation algorithmique ?

> P54 : Si l'IA faisait une recommandation erronée qui avait des conséquences irréversibles sur la santé, qui serait responsable ?

> P52 : Le médecin a une responsabilité par rapport à son patient. S'il dit : L'algorithme me dit ça, donc fais ça. Et si ça ne marche pas, c'est la faute de l'algorithme, ce n'est pas ma responsabilité professionnelle qui est en jeu.

La tension entre agentivité humaine et artificielle vient ainsi questionner qui, de l'humain ou de la machine, est responsable de différentes décisions de santé, face à une erreur de diagnostic, ou du choix des données collectées.

P23 : Mettons, tu es responsable des données que tu donnes à ton algorithme, après quand il apprend lui il devient responsable d'où il va chercher ses données ?

P44 : Je me suis demandé aussi qui était imputable pour les erreurs du robot. Il y a les pilules et tout ça... c'est comme... ok. Donc, est-ce que c'est le robot qui remplit le rapport d'accident-incident si jamais y'a une erreur ?

Également, considérant l'autonomie des systèmes apprenants, certains ont questionné qui, de l'humain ou de l'algorithme, détermine les « règles » du fonctionnement des systèmes. Qui décide quels tests sont administrés par les robots de soin ? Est-ce que l'algorithme peut les déterminer seul, par apprentissage ?

Reconnaitre une agentivité aux SIA s'accompagne de questionnements relatifs à leur encadrement; à savoir s'ils sont « soumis aux mêmes lois » et « aux mêmes règlements » et s'il faut les considérer comme « l'extension d'une équipe professionnelle » :

P11 : Si c'était une aide à domicile, elle a un code déontologique etc. Est-ce que Vigilo est soumis aux mêmes lois et au même code d'éthique ?

L'agentivité artificielle risquerait ainsi de déresponsabiliser les professionnels de santé. Le médecin pourrait se « déculpabiliser s'il fait une erreur », notamment car le recours à des systèmes d'IA favorise le « détachement » :

P20 : Si on imagine que plus ça va plus pis plus il se fie à un algorithme plutôt qu'à lui-même, plus il y a comme un détachement par rapport à ça. Fait qu'il y a la responsabilité, mais il la ressent pas. Dans un sens, pas du tout.

Avoir recourt à des systèmes d'IA pourrait également déresponsabiliser l'entourage du patient :

P1 : L'autre problème que j'ai vu, c'est la déresponsabilisation de l'entourage par rapport aux personnes âgées. On se dit : « Le robot est là, pas besoin d'être présent. » À mon avis, c'est quelque chose qui me paraît important.

Voire toute l'équipe de soins :

P67 : Il y a aussi la responsabilité puisque l'introduction d'un robot, est-ce que ce n'est pas aussi une déresponsabilisation de la famille et de la société ?
P62 : Et le médecin et le professionnel soignant.
P64 : Toute l'équipe soignante.

Ainsi, selon cette crainte, le déplacement de la responsabilité (ou la déresponsabilisation) s'opère des différents humains impliqués vers les systèmes d'IA.

3.3. Risques associés à la reconnaissance d'une agentivité artificielle : les biais algorithmiques

Selon les citoyens, reconnaitre une certaine agentivité aux systèmes d'IA serait problématique considérant les biais qui pourraient être présents dans les données initiales ou dans la programmation des algorithmes ; qu'il s'agisse des biais des humains qui programment ou des biais « qui ne sont pas vus » dans les systèmes. Les citoyens ont soulevé différents types de biais (« de mesure » ou « d'échantillonnage ») qu'il serait nécessaire de considérer en santé, et qui pourrait nuire à l'efficacité des algorithmes, ce qu'ils ont reconnu comme particulièrement problématique si on leur laisse le « contrôle » de la décision de santé.

> P1 : C'est ce qu'on appelle un biais d'échantillonnage dans le sens où les données vont être basées juste sur une population, et la population qui a accès aux médias sociaux etc. Tous ceux qui s'éliminent (même les jeunes) de Facebook (donc pas seulement des personnes âgées) on n'a pas accès à leurs données. Donc il peut y avoir ici un biais d'échantillonnage.

Certains soulignent que les données seront forcément biaisées si les individus non « éduqués en santé » qui consentent à donner leurs données étaient « exclus ». Comment s'assurer de la neutralité des algorithmes tout en considérant les données pertinentes à la décision de santé ? Comment garantir la qualité des données, comprendre les échantillons sélectionnés ?

Les participants ont souligné que laisser la décision aux algorithmes est problématique considérant le niveau de confiance que l'on peut accorder au diagnostic, la qualité des données, ou le risque de faux positif :

> P61 : Il y a un facteur [prendre en compte] aussi là-dessus, c'est un peu la même chose que ce dont on parle avec l'éthique de la recherche et les femmes qui sont sous-représentées. Ces mêmes gens-là, c'est aussi les gens qui ont le moins de données sur Alphabet, fait que la précision de l'analyse, la précision de l'algorithme, elle est suggestible. Et c'est eux autres qui prennent ça pour du *cash*. Alors que ceux-là qui comprennent la technologie, qui l'utilisent en malade, et qui ont *full* de données sur internet, c'est eux autres qui sont le mieux orientés par l'algorithme.

> P56 : Comment à partir de l'ensemble des données qui ont été collectées on est en mesure de cibler un diagnostic aussi précis et pointu qu'un risque de dépression ? Puis moi ça m'a tout de suite allumé sur un biais potentiel de sélection des sujets, pour arriver à circonscrire un groupe de personnes susceptible d'avoir de la dépression.

D’autant que les données sont parfois mesurées par l’individu lui-même, ce qui est potentiellement problématique au regard de leur validité (biais de l’auto-mesure de soi).

Pour certains des citoyens participants, c’est une erreur de mettre la responsabilité des biais sur le dos des systèmes d’IA, puisqu’il s’agit avant tout d’un problème lié aux humains et un besoin de confirmation objective :

> P61 : Moi, une des choses que j'haïs en ce moment, c'est d'utiliser, par exemple, le terme IA, ou autre chose, parce que c'est juste une analyse statistique. […] Ce que je veux dire par rapport à ça, je voulais donner un exemple de biais statistique facile juste de ce qu'on est en train de dire : si on se base sur le fait que c'est une application *deep learning*, ça veut dire que nécessairement il faut que ta machine ait, à un moment donné, une confirmation. C'est comme ça que ça fonctionne. […] Si tu te bases sur la confirmation du médecin, ça veut dire que nécessairement tu induis un biais systématique : le médecin va confirmer ce qu'il pensait au départ. Finalement, la machine n'est pas en train de dire de quoi, elle est en train d'apprendre ce que le médecin voulait que la machine dise.

3.4. Une transformation du rapport à la technologie

Comme mentionné précédemment, reconnaître une agentivité aux systèmes d’IA est assez inédit en ce qui a trait à la responsabilité face au développement technologique – celle-ci n’était jusqu’alors attribuée qu’aux humains (notamment, ceux qui les conçoivent). Ainsi, l’agentivité artificielle implique une certaine transformation du rapport à la technologie, notamment par le biais d’une anthropomorphisation, ce qui a été discuté lors de la coconstruction.

Une transformation du rapport à la technologie a en effet été soulignée lors des discussions. Les inquiétudes relatives à ce rapport portent sur le stress ajouté considérant la nature et la fréquence des échanges avec des systèmes d’IA (ex. notifications sur les téléphones intelligents, surveillance en continue), mais surtout sur l’idée que les systèmes d’IA viendraient remplacer les individus dans les échanges sociaux. Ce remplacement serait favorisé par une humanisation ou anthropologisation poussée des technologies en question, notamment celles capables d’affection et d’empathie qui pourraient être associées à des humains.

> P9 : La machine est en train de tellement bien converser avec elle que la famille vient moins la voir et en plus la dame se confie mieux à la machine qu'à ses proches. Et même qu'on est obligé de changer les vrais individus puisqu'elle rentre moins bien en relation.

P62 : Mais le robot est empathique. Il fait la *job* d'infirmière.

La transformation de la relation à la technologie est donc intrinsèquement liée à l'autonomie croissante des systèmes d'IA, qui risqueraient, dans certaines conditions, d'être considérés comme des humains à part entière. Selon les citoyens, l'impact d'une défaillance du système d'IA pourrait alors être d'autant plus important s'il implique également des aspects relationnels et émotifs. Le développement de liens affectifs envers les robots de soin pourrait également encourager l'isolement, en particulier considérant que les robots pourraient être perçus comme plus agréables que certains humains ou que les humains s'habitueraient à des interactions qui seraient spécifiques aux machines (qui seraient notamment plus simples). Est ressorti des discussions qu'un détachement du reste du monde serait ainsi à craindre avec l'apparition « d'amitiés artificielles » qui, en limitant les interactions sociales, pourrait conduire au « rejet de son environnement » et au « détachement » des autres.

> P1 : Ce qui m'a paru très grave, c'est l'amitié artificielle qu'elle créé avec le robot. Ça va la détacher encore plus, étant donné qu'elle a des problèmes cognitifs, ça va encore plus la détacher du monde.
> P47 : La première chose qui m'a tiquée, c'est le fait que ce soit… qu'il soit qualifié d'« empathique », alors je me suis demandée : sur quels critères on basait l'empathie du robot pis qu'à cause de ça, ça venait à une confusion des rôles pour Soline puis une fausse impression d'autonomie, ou une fausse impression de sécurité la présence de ce robot-là.

> P1 : C'est parce qu'à ce moment-là, j'avais marqué aussi qu'il y avait un détachement de la personnalité de l'individu, par rapport à ... qui devient contrôlé par le robot. C'est-à-dire, moi je m'habitue à ce robot que j'aime bien et tout, je me lie d'amitié avec lui. Ce qui va me permettre de … enfin, lier une amitié artificielle. Ça va me détacher de la voisine et de toi...

Certains s'inquiètent qu'une relation trop intime ou trop proche avec les technologies puisse conduire à une aliénation :

> P59 : Le deuxième enjeu que je vois qui est majeur, c'est l'aliénation. Dans le sens de : oui, la majorité de son temps, elle est avec un robot et elle finit par ne plus se confier à personne, à s'éloigner de tout le monde. Déjà, elle est géographiquement parlant éloignée de sa famille, mais maintenant elle est émotionnellement éloignée aussi. Ça c'est un enjeu important pour elle. Dans le cas de son Alzheimer, c'est … en tout cas, dans ma compréhension profane de l'Alzheimer, si on n'a pas la famille, au moins quelqu'un de la famille qui est là, on va finir par juste comparer son robot à son mari.

Les discussions relatives à la création de liens affectifs avec les machines ont parfois dérivé sur les séries de science-fiction qui mettent en scène des attachements affectifs très forts aux robots – voire

de l'amour (e.g. *Real Human*, *Her*, *Black Mirror*). Il a été souligné que ce genre de situations pourrait probablement se produire avec l'avènement de robots de soin « empathiques », ce qui serait « triste » et « dangereux » - sans pour autant que les citoyens précisent pourquoi.

Les citoyens ont souligné que ce nouveau rapport à la technologie – et par là-même l'agentivité qu'il serait possible d'accorder aux systèmes d'IA - est sous-tendu par la ***confiance*** que les individus pourront y porter : cette confiance est basée sur la fiabilité des systèmes, la qualité de leurs analyses ou leur transparence. Un minimum de confiance doit être garanti pour assurer l'acceptabilité des systèmes d'IA en santé mais ne doit pas être absolue, notamment car les analyses des algorithmes ne seront jamais fiables à 100%, qu'ils ne considèrent que des paramètres objectifs et qu'il existe un risque de surconsidérer leur précision ou de « surestimer le pouvoir des données ». Cette confiance est à actualiser en continue tant parce que les avancées technologiques sont en constante évolution que parce que les données sur lesquelles les algorithmes apprennent pourraient devenir obsolètes.

> P24 : Est-ce que la madame qui a 84 ans … Est-ce qu'elle est prête à accepter que ce soit l'intelligence artificielle qui lui pose son diagnostic ? […] Ça devient plus compliqué, c'est plus rapide…

L'ensemble des considérations relatives à la confiance qu'il est possible d'attribuer aux systèmes d'IA est intrinsèquement lié au niveau de confiance que l'on peut accorder au diagnostic, au contrôle et à l'efficacité des modèles à la base de leur fonctionnement. Par exemple, un certain scepticisme a été observé face aux jumeaux numériques, les participants dénonçant la « généralisation excessive » qui amènerait aux recommandations de santé. Ces préoccupations sont autant de raisons de limiter le poids que l'on accorde aux décisions algorithmiques et donc demandent de limiter l'agentivité artificielle. L'ensemble des éléments mentionnés précédemment (portrait incomplet et perte de l'individualisation) invitent alors à tempérer la confiance en l'algorithme et ses analyses, qui ne doit pas devenir « absolue » et risque de glisser vers un certain « déterminisme ». Certains mentionnent qu'il est alors nécessaire de garder un doute raisonnable face à la portée des recommandations des algorithmes :

> P61 : C'est ça que je comprends, c'est qu'on essaie [à tendre vers] une collecte de données vraiment exhaustive de la personne. Ça, c'est correct, dans le sens que ça serait le but pour essayer de réussir à prédire tout de la santé de la personne, mais le postulat qui devrait être mis sous-jacent, c'est qu'on ne va jamais réussir à avoir assez de données pour prédire tout sur l'être humain. Le bonhomme, il compile ses données depuis deux ans. Est-ce que

vraiment, en deux ans, il ne faudrait pas faire office d'une petite humilité pour dire que la précision est suffisante pour dire une prédiction sur toute sa vie.

3.5. Attentes citoyennes : limiter l'agentivité artificielle et favoriser l'agentivité humaine

Face au défi de l'agentivité artificielle, les citoyens participants ont formulé plusieurs attentes : 1) que les systèmes d'IA soient des outils et non des agents (décisionnels); 2) que les humains gardent la main sur les développements et les usages des systèmes d'IA en santé et 3) que les systèmes d'IA, si une certaine agentivité venait à leur être attribuée, soient transparents.

3.5.1. Les systèmes d'intelligence artificielle sont des outils

Afin de limiter les conséquences associées à l'attribution d'une certaine agentivité aux systèmes d'IA, les citoyens s'accordent pour défendre que la responsabilité doit toujours être attribuée aux humains – notamment car ce sont eux qui ont créé les systèmes. Ceci revient à limiter l'agentivité attribuée aux SIA et à considérer les systèmes comme des outils stricto sensu.

> P43 : Ce n'est pas un substitut d'humain, c'est une machine, c'est un grille-pain, un peu plus intelligent qu'un grille-pain.

> P43 : Moi, il y a quelque chose que je déteste dans cette phrase, c'est une IA. Ça n'existe pas, une IA. […] Le robot, il est connecté avec un système probablement hébergé sur un serveur à distance qui est... qui fonctionne grâce à des algorithmes d'apprentissage qui sont simplement des modèles mathématiques à qui on a donné des quantités de données [possibles], et voilà !

Pour les citoyens, ce ne sont pas les systèmes qui « posent l'acte » mais bien l'humain « en arrière ». Selon cette vision, ce « qui compte » est alors la façon dont l'algorithme a été développé et programmé, peu importe le comportement du système. Les citoyens ont à plusieurs reprises pris position sur ce point, en mentionnant que les systèmes d'IA ou les robots doivent rester « des outils », ne doivent pas « être décisionnels » mais seulement émettre des recommandations, ou ne peuvent porter seuls des décisions afin que les humains gardent toujours le contrôle.

> P43 : Mais le robot, c'est une machine. Il n'y a pas moyen que le robot soit imputable de quoi que ce soit.

> P4 : La machine propose, l'humain dispose.

P4 : C’est toujours consultatif, jamais décisionnel.

Considérer les systèmes d’IA comme des outils devrait permettre, selon les citoyens, de garantir que le médecin « conserve son pouvoir » en s’appuyant sur les algorithmes et prémunir contre « l’effet pervers » qui pourrait le conduire à se censurer si sa décision va à l’encontre de la recommandation algorithmique. Certains ont cependant souligné que l’utilisation de systèmes d’IA n’impacterait pas plus la responsabilité du médecin que les technologies traditionnelles d’aide à la décision :

> P23 : Ok c’est une supposition que je fais mais si l'algorithme dit " il y a 97% de chance que ce soit ça" et que lui il a 25 ans d'expérience, moi je me dis que lui il se dit que c'est minimum 75% de chance que ce soit ça. Fait que c'est son *extra feeling* qui fait qu'il se dit que c'est autre chose. Sa responsabilité reste la même qu'avant finalement. C'est juste les chiffres...

Si l’IA doit être considérée comme un « outil » et demeurer un support seulement, c’est également car elle ne sera jamais fiable à 100%. Ainsi, les citoyens ont reconnu que l’IA est un outil « fort intéressant et utile » mais un « outil d’aide à la décision point à la ligne », avec ses performances et ses limites.

3.5.2. L’humain garde la main

Face à l’autonomie croissante des systèmes d’IA, les citoyens ont recommandé qu’un professionnel de santé (ou un ensemble de professionnels de santé) soit toujours responsable du suivi des patients et que « ce ne soit pas l’algorithme qui prenne tout » ou que l’arbitrage des choix ne soit pas laissé à un système d’IA. Ils ont proposé d’encadrer le développement de l’IA dans l’optique de préserver la supervision humaine et de garantir que les patients aient toujours accès à un professionnel de santé lorsqu’ils font face à une décision algorithmique :

> L’utilisation de robots ne doit pas se faire sans supervision d’une autorité humaine institutionnelle soumis à une déontologie (Recommandation lue par l’animatrice, rédigée collectivement par la Table 2).

En particulier en ce qui a trait à la collecte des données :

> P20 : Où il [l’algorithme] va chercher ses données, ça devrait être encadré aussi, surtout dans le domaine de la santé. Il ne devrait pas pouvoir aller chercher des données n'importe où.

Pour la grande majorité des citoyens participants, la responsabilité de la décision médicale reste « sur les épaules du médecin » qui doit faire son travail et donner « son opinion » sur la décision algorithmique :

> P23 : La responsabilité du médecin doit toujours prévaloir sur l'IA qui n'est qu'un outil d'aide à la décision.

> P25 : C'est sa responsabilité qui doit être engagée, pas celle du robot. Je l'avais écrit... vu qu'il y a un risque d'erreur pour l'intelligence [artificielle] et pour le médecin...

L'utilisation des systèmes d'IA ne peut se faire sans la supervision d'une autorité médicale humaine ou une institution.

> Les données synthétisées par les algorithmes doivent être traitées et interprétées par un professionnel de la santé et le partage de cette analyse doit être fait par un professionnel de santé (Recommandation lue par l'animatrice, rédigée collectivement par la Table 3)

Les citoyens ont mentionné que l'absence de modération humaine est particulièrement problématique dans la situation ou des biais seraient perpétués. Pour certains participants, ce sont bien les humains qui doivent assurer de limiter les biais par la sélection des critères initiaux, des objectifs de l'analyse ou de la problématique.

Si la responsabilité humaine a fait consensus, il semble nécessaire de clarifier qui est responsable du système d'IA, qui contrôle le code, qui est imputable des décisions algorithmiques, des défauts de fabrication, des piratages et des erreurs qui en découlent (ce qui renvoient aux considérations relatives aux *mains multiples*). Les citoyens ont questionné de quelles manières étaient déterminées les « règles » qui permettent aux algorithmes de prendre des décisions ou « comment contrôler l'algorithme ».

> P64 : Moi, je me demandais, à qui il répond le robot ? Il est redevable à qui ? Il est embauché... ben il est embauché... il est acheté... Je suis en train de l'humaniser, mon dieu !

3.5.3. Des systèmes d'intelligence artificielle transparents

Les citoyens ont soulevé que l'autonomie croissante des systèmes d'IA ne peut se faire sans assurer leur transparence. Ils ont défendu que les systèmes ne devaient pas être « des boites noires » mais être « transparents quant à leur processus de décision » afin qu'il soit possible de « connaitre les facteurs décisionnels ». Même si l'ouverture des codes n'est pas « suffisante » pour permettre de comprendre le cheminement de l'algorithme, assurer l'explicabilité des systèmes permettrait de

responsabiliser les humains qui les utilisent. En effet, pour que la responsabilité puisse rester celle du médecin il est nécessaire de garantir la transparence des algorithmes, afin qu'ils puissent remettre en question leurs décisions :

> P21 : Mais pour pouvoir s'assurer que la responsabilité soit celle du médecin, il doit avoir accès et il doit y avoir une certaine transparence des algorithmes quand ils font des recommandations pour qu'il soit capable de challenger un peu ce qui lui est recommandé.

Pour assurer cette responsabilité, il est nécessaire que les algorithmes soient « capables d'expliquer la décision » ou la recommandation, de « montrer comment l'algorithme bâtit son raisonnement » et ce au médecin tout comme au patient. Les citoyens ont souligné l'importance que le médecin soit toujours mis au courant lorsqu'une décision de santé est prise sur la base d'une recommandation algorithmique et qu'il comprenne le « cheminement » qui a conduit l'algorithme à sa recommandation.

Les citoyens ont souligné que ce manque de transparence ou d'explicabilité pourrait influencer le partage des responsabilités entre médecins et algorithmes, induire des doutes quant à la décision de santé, et impacter la décision du médecin.

> P55 : C'est que tu ramasses une boite noire. Il rentre des choses là-dedans et on ne sait pas comment ça se passe, pour que cette décision-là soit prise. Ça va être important que l'IA puisse aussi justifier la rationnelle qui l'a amené à cette décision-là, parce que là il n'y a pas … aucune rationnelle. C'est juste une boite noire avec toutes les données. On ne sait pas le poids accordé à chacune des données.

Certains soulèvent cependant que ce devoir de transparence devrait « fonctionner dans les deux sens », soit qu'il est important de justifier la décision qu'il s'agisse de celle de l'algorithme ou de celle du médecin :

> P28 : J'ai réalisé à la fin que le questionnement pourrait fonctionner des deux sens. C'est-à-dire qu'on prend pour acquis qu'on voudrait que l'humain puisse questionner l'IA, mais peut-être que ce serait intéressant et instructif que l'IA puisse questionner pourquoi le médecin rend un avis différent. Mais ultimement, ça revient quand même à la même chose.

La nécessité de mettre en place des solutions (techniques) a été soulignée, afin de pallier ce manque de transparence et de renforcer le pouvoir de l'humain face aux décisions algorithmiques. Les citoyens ont par exemple proposé d'imposer que la décision médicale se base sur plusieurs types d'algorithmes conduisant à plusieurs recommandations « en attendant » que les systèmes

deviennent transparents. Ceci permettrait de limiter les impacts sur la décision du médecin et lui laisser le choix de quelles recommandations suivre.

> P28 : Avant qu'un algorithme puisse être utilisé en santé il faut qu'il soit muni d'un mécanisme par lequel on est capable de l'interroger pour justifier sa décision ou sa recommandation. Je pense que ça devrait être une obligation qu'il y ait un mécanisme.

4. Entre incapacitation humaine et agentivité artificielle : crainte du remplacement et attente de coopération humain-machine

À mi-chemin entre les préoccupations relatives au développement d'une agentivité artificielle et à l'incapacitation des humains par les systèmes d'IA, la crainte d'un remplacement des humains par les machines a été soulevée à plusieurs reprises. Découle de cette crainte de remplacement celle d'une déshumanisation des soins. En réponse à cette crainte, l'attente du développement d'une coopération humain-machine a été formulée.

4.1. Crainte du remplacement des humains par les machines

Les citoyens ayant participé à la consultation ont souligné une crainte relative au ***remplacement*** des humains par les machines et à la perte d'emploi associée. Il s'agit, dans le contexte de la santé, du remplacement du personnel médical - en particulier des médecins spécialistes ou des infirmières. L'utilisation de systèmes d'IA dans les zones où l'accès aux soins est déjà limité pourrait favoriser le remplacement (venant ici complémenter les professionnels de santé déjà en sous-effectifs). Certains craignent que l'avènement de l'IA en santé modifie le type de travail accompli par des professionnels de santé, certaines tâches étant vouées à disparaitre, tandis que d'autres craignent que l'automatisation amène à une réallocation des ressources entre main-d'œuvre et technologies qui ne serait pas nécessairement dans l'intérêt des humains :

> P23 : Si on a ça à offrir aux patients, on enlève(-tu) des infirmières, on enlève(-tu) des préposés ? Parce que là on en a moins besoin, ou on a moins d'argent pour ça. Je trouve ça crée tout un dilemme autour de, comment tu distribues ton argent en santé...

Cette crainte du remplacement est liée aux capacités des systèmes d'IA – se basant notamment sur la supposition que les algorithmes vont être meilleurs que les humains pour certaines tâches qui leur seront alors déléguées.

> P65 : Est-ce qu'on va assister à un espèce de grand remplacement ? Est-ce que les machines seront suffisamment intelligentes pour communiquer entre elles ?

Les critères qui permettent, en santé, de définir une tâche comme automatisable (soit, remplaçable) sont cependant questionnés.

> P31 : Si la machine est capable de lire le scan ou le *RMI* […] est-ce que ça va re-questionner certaines professions à l'intérieur pour la vitesse de lecture d'une radiographie en radiologie ?

Certains défendent que cette transformation prendrait du temps, en particulier pour les emplois qui nécessitent une interaction humaine (comme en santé) :

> P38 : D'ailleurs, quand on parlait de disruption dans le marché du travail, les *jobs* qui vont rester le plus c'est les interfaces humaines. Les *jobs*, on aime les humains. Ça va prendre du temps avant que ça, ça soit un ordi.

> P56 : Même si on remplace des radiologistes par l'IA, ça va prendre quand même des radiologistes pour suivre les résultats de l'IA.

La pertinence du remplacement est alors questionnée. D'un côté, certains soulignent les problèmes relatifs au remplacement des humains par les machines : notamment, dans les cas qui appellent à des réponses nuancées ou qui nécessitent un jugement subtil. Le risque de ne pas détecter des problèmes qui n'auraient pas été préprogrammés est également soulevé, créant des situations où la présence d'une personne sur place s'avèrerait de toute façon nécessaire - comme dans le cas des robots de soin à domicile. La nécessité d'une présence humaine est d'autant plus pertinente qu'il y aura toujours une marge d'erreur associée aux décisions et recommandations algorithmiques, qui pourraient ne pas être sensibles aux différents éléments contextuels :

> P52 : L'IA est à l'heure actuelle incapable de contextualiser ses données alors que le médecin est (capable) de garder le *big picture* et ... de prendre la décision basée sur le *big picture*, la contextualisation ce n'est pas le fort de l'IA pour l'instant.

D'un autre côté, le remplacement des humains par les machines peut s'avérer bénéfique dans certaines conditions : l'IA serait plus sécuritaire, plus fiable, plus efficace et plus rentable.

> P33 : Les algorithmes comme ils ont accès à toutes les données eux ils sont beaucoup plus experts.

> P4 : Mais aussi il peut y avoir des gains de l'autre côté. Peut-être que les médecins sur 100 examens qu'ils demandent, y'en a peut-être 80 qui sont inutiles. Peut-être qu'avec l'IA on va pouvoir demander moins d'examens et trouver le problème plus rapidement et donc il y aurait une économie de ressources ...

> P43 : Ça peut être des gens qui peuvent détester travailler pour des aînés et on n'a pas nécessairement les ressources. Là, l'avantage, c'est qu'on a une machine qui est précise, c'est-à-dire, qui va pouvoir dire : « Ah! Il y a un petit accroissement de tel problème, il faudrait augmenter un tout petit peu la dose de ça, etc. » C'est quand même plein de côtés positifs.

Déléguer certaines tâches aux machines permettrait alors aux professionnels de santé d'allouer plus de temps à d'autres tâches qui, elles, ne sont pas automatisables. Certains citoyens ont également souligné que ces enjeux ne sont pas nouveaux : une délégation des tâches aux machines s'observait déjà avant l'avènement de l'IA.

4.2. La déshumanisation des soins

Des inquiétudes relatives à la diminution du contact humain dans la relation de soin sont également apparues face à l'autonomie croissante des systèmes d'IA, notamment car celle-ci peut conduire au remplacement des humains par les machines. Les participants ont souligné que la présence « d'un lien » est nécessaire à une bonne relation de soin; car le transfert d'informations se fait mieux d'humain à humain (versus un algorithme en ligne qui ne pourrait peut-être pas répondre à toutes les questions des patients). Il est ainsi essentiel de ne pas limiter les échanges entre humains pour assurer une bonne réception et une bonne compréhension des informations de santé (ex. ne pas seulement les transmettre aux patients *via* une notification) qu'il s'agisse de l'annonce d'un diagnostic ou de l'obtention d'un consentement. Également, si l'algorithme peut être potentiellement plus efficace dans la précision du diagnostic, il reste qu'un professionnel de santé est nécessaire, considérant « la façon de l'annoncer ». Il ne s'agit pas en effet de seulement transmettre l'information pour qu'elle soit comprise mais également de prendre en compte tout le côté émotionnel et rassurant de parler à un professionnel de santé.

> P40 : Moi j'ai besoin d'avoir un contact avec un médecin, une infirmière, quelqu'un, qui comprenne comment je me sens. Peut-être que j'ai peur, peut-être que je n'ai personne à la maison pour s'occuper de moi. Je trouve qu'on parle beaucoup de communication, mais on est majoritairement émotionnel.

Accorder une trop grande agentivité aux systèmes d'IA pourrait, en effet, conduire à leur laisser une part non négligeable des tâches relatives au soin. Or, pour les citoyens, le maintien du lien et de la présence humaine est essentiel pour ne pas limiter les échanges entre humains et les conséquences négatives associées comme, par exemple : le sentiment de solitude, l'isolement, la

négligence, la « perte de relationnel » ou un relationnel qui se ferait avec les systèmes d'IA au détriment des humains.

> P42 : Pour moi, l'outil diagnostic devrait aider le médecin, mais tu ne dois pas recevoir le diagnostic sur ton téléphone. Tu dois toujours avoir une interface humaine parce que les humains ne sont pas nécessairement aptes à interpréter les données diagnostiques.
>
> P31 : Il ne faut pas que le médecin ne travaille qu'avec l'algorithme et oublie la personne.

Cette perte de contact humain est également problématique car elle revêt un caractère anxiogène. Certains ont souligné que la nécessité d'une humanité réconfortante est suffisante pour justifier de prévenir la perte de lien associée à l'usage croissant des systèmes d'IA : « Au lieu d'être juste alimenté par le risque, ça peut aussi être alimenté par le réconfort » (P36). Elle permettrait de transmettre les informations de manière « agréable » et « humanisante ». La nécessité de préserver le dialogue médecin-patient et la « place de l'humain » a été soulevée, notamment car il est possible de contredire ou négocier avec un humain (ce qui semble être plus difficile si l'on est obligé de s'adresser à un algorithme).

> P41 : Avant tout, je pense que ça prend une recommandation qui dit : « Peu importe ce que vous allez sortir... » Je dirais : « Ne donnez pas un diagnostic par texto à tous les lundis matin. » Je ne sais pas, ça prend un minimum d'humanité et de relationnel.

Si ces conséquences sont considérées comme négatives, c'est aussi qu'elles risquent de nuire à la qualité des soins : « L'isolement social va pas aider à ralentir la maladie » (P13). Par exemple, le maintien au domicile des personnes âgées grâce à des robots de soin est-il vraiment plus bénéfique que la mise en habitation de groupe adaptée – qui pourrait éventuellement bénéficier du soutien de systèmes d'IA ? L'autonomie des robots de soins à domicile pourraient conduire à un isolement social des patients, sous une nouvelle forme : elle serait issue du fait que l'on s'appuie de plus en plus sur « la machine » et pourrait avoir des conséquences sur la perception des autres humains et amener à moins communiquer avec eux ou mener à un délaissement.

> P11 : Moi j'ai identifié un double isolement social. Donc au niveau de sa famille qui n'a pas besoin de passer beaucoup de temps avec elle et elle, finalement, qui aussi préfère s'isoler au bout d'un moment et juste communiquer avec le robot.
>
> P46 : Je relisais le scénario et je remplaçais le robot par une personne humaine, et je me disais : la seule chose spécifique, peut-être que je vois au robot, c'est l'accroissement de l'isolement de la personne, le délaissement qu'on va attribuer à cette personne. Parce que la famille ne va pas venir la voir, l'équipe soignante n'a plus besoin de venir la voir...

Mais pour certains, l'IA pourrait aider à contrer l'isolement qui s'observe déjà dans plusieurs situations chez les patients (notamment les personnes âgées, en particulier celles atteintes de la maladie d'Alzheimer mises en scène dans le scénario). Si l'isolement n'est peut-être pas attribuable à la présence du robot, les systèmes d'IA pourraient au contraire aider à renforcer les liens entre humains :

> P43 : Imaginez que le robot soit connecté avec tous les robots de ce CLSC, dans toute la [région] environnante […] L'écosystème pourrait recommander que ces deux personnes se rencontrent parce qu'ils auraient des affinités.

De plus, des inquiétudes relatives à une vision appauvrie du patient ont été soulevées, notamment si la perception des individus venait à être réduite à leurs seules données. Cette préoccupation est relative au profilage opéré par les algorithmes qui risquent de dépeindre un portrait incomplet du patient, et le risque prend de l'ampleur si les systèmes d'IA sont des agents – ce qui mènerait à donner de l'importance dans la décision de santé à cette vision appauvrie.

Relativement à la création de profils de santé, plusieurs dénoncent le ***portrait incomplet*** des individus, qui ne tient pas assez compte du « contexte », une partie de la « vie » des patients ne pouvant apparaître dans les ensembles de données. L'évaluation issue du profilage est ainsi considérée comme présentant de « grosses lacunes ». Si pour certains citoyens participants, ce portrait incomplet est trop « subjectif », pour d'autres c'est justement la perte de l'évaluation subjective du sujet, en tant qu'humain à part entière qui pose problème. Certains dénoncent une perte de la « complexité de l'humain » issue des analyses « trop rationnelles » des systèmes d'IA, qui se font toujours selon des paramètres objectifs évinçant par la même la « nature de l'individu ». Ainsi, si les algorithmes peuvent « prendre en compte l'aspect technique », il manque « tout l'aspect émotif » qui permet de mieux comprendre qui « est » le patient, qui ne se retrouve pas dans des données objectivables comme l'activité physique ou les données biologiques. Certains mentionnent les recherches qui permettraient de coder des interfaces plus « émotionnelles » afin de palier la déshumanisation des soins :

> P38 : Il y a des études, des recherches qui se font aussi pour faire de l'interface émotionnelle, et des trucs comme ça. En rajoutant de l'irrationalité, tu vas être un peu plus humain, un peu plus connectée.

Ainsi, la rationalité constitutive des systèmes d'IA demanderait selon les citoyens à limiter l'agentivité qu'on pourrait leur accorder.

Ce portrait incomplet des patients pourrait amener à un glissement vers une perte « d'identité », de « l'individualisation », ou une « dépersonnalisation » des soins; notamment parce que la « psychologie du sujet » est « évincée ».

> P6 : L'individu perd son identité, il est comme noyé. […] C'est plus un individu. Une perte de l'individualisation.

> P52 : Ça part d'une approche personnalisée mais c'est une communication de l'information qui est hautement dépersonnalisée.

Certains voient cette ***perte d'individualisation*** comme une « dissociation du corps et de l'esprit » car il faut « traiter le patient dans sa globalité » :

> P7 : Donc il y a une dissociation selon moi du corps et de l'esprit. C'est sûr que ça rentre dans plein d'autres discussions mais chaque maladie est liée à des notions de psychologie. Il faut traiter le patient dans sa globalité. Puis faire un copier-coller d'une maladie d'un patient à l'autre ça ne fonctionne pas. Moi je ne crois pas à ça. Peut-être que c'est une bonne idée de se servir de la machine pour observer des tendances, mais après ça prend un être humain. Donc je trouve qu'on évince la psychologie et la relation du patient.

Un glissement vers une déshumanisation des soins est à craindre, dans le sens qu'on ne considère plus les patients comme des « humains », ou qu'il manque l'aspect « humain » quand les décisions de santé sont laissées aux algorithmes. Un détachement des médecins par rapport au diagnostic qu'ils posent risque aussi de nuire au bien-être des patients :

> P20 : Parce que si les médecins ça suffit qu'il y ait un algorithme, il y a un risque de détachement aussi par rapport au diagnostic qu'ils posent pis qui peut avoir un impact sur la santé mentale des patients.

Certains s'inquiètent que le médecin finisse par « oublier la personne » en ne travaillant qu'avec des algorithmes, voire s'inquiètent que les humains finissent par être considérés « comme des machines » ou des agrégats de données :

> P52 : Oui mais le patient et la machine ... Il ne peut pas y avoir de… Remettre en question le rôle de l'humain… dans le sens où on devient un peu un robot, un rassemblement d'informations et de données simplement.

> P34 : On touche presque sur le fondement de l'humain. Des fois, on prend des décisions non-rationnelles alors que l'IA veut que toutes les décisions soient rationnelles et rationnelles sur des points qui sont des fois obscures. Fondamentalement, ça va loin comme question philosophique.

Cette déshumanisation-dépersonnalisation risque de causer du tort au patient, de ne pas le respecter ou de l'insécuriser. Les interfaces « déshumanisantes » risqueraient de créer du stress et de l'anxiété « dans la relation à soi-même », ne prenant pas assez en compte le choix des utilisateurs. Les citoyens ont défendu que l'annonce d'un diagnostic par notification est extrêmement minimale comme protection du patient en termes de risques (comment va-t-elle être reçue par la personne ?).

> P41 : Je voulais amener un point que finalement, l'intelligence numérique en santé, et je vais mettre ça simple, c'est qu'on prend les dossiers de patients, qui ont été écrits par des choses, on les a numérisés et on les met dans le *blender* et on sort avec une donnée numérique qu'on va dire : « Voici le pronostic ». Et l'humanisation de ce pronostic-là va … Donc, pour moi, ce n'est même plus de l'autonomie, c'est de la déshumanisation du soin de santé.

Une numérisation trop poussée pose également problème au niveau de la déshumanisation des soins, car elle « dépersonnalise » les soins et pourrait « culpabiliser » le patient sur ses habitudes de vie. Le patient pourrait alors se sentir « laisser à lui-même », en particulier si à terme, on se fie trop au « numérique » :

> P65 : Est-ce qu'à terme, on va se fier à 100% au diagnostic numérique sans avoir à toucher le patient, à l'approcher, à aller chez lui ? Est-ce qu'on va pouvoir faire ça d'un autre pays ? Par Skype, je ne sais pas...

4.3. Attente d'une coopération humain-machine

Selon les citoyens, l'avènement de l'IA en santé se ferait idéalement selon le développement d'une collaboration entre humains et systèmes d'IA (ex. entre médecins et algorithmes), parfois nommée « coopération » ou « partenariat », qui permettrait de valoriser les compétences humaines nécessaires à l'exercice de la médecine tout en potentialisant les compétences de l'IA. Celle-ci pourrait prendre la forme d'un « feedback IA-médecin-patient », ou chacun pourrait justifier sa décision. La complémentarité entre capacités humaines et capacités algorithmiques pourrait ainsi être valorisée et permettrait au personnel de santé de se concentrer sur d'autres tâches non-automatisables, sur des cas exceptionnels, ou sur la dispensation de soins et de traitements.

> P15 : L'idée ce n'est pas d'automatiser tout le système de A à Z. C'est comment faciliter la tâche du personnel de santé.
>
> P42 : La machine, elle est entraînée et elle est très bonne souvent dans une boîte qui est définie et le médecin, lui il y a eu 20 ans de choses qu'il a vu et beaucoup plus courantes. C'est ça qui est important que ce partenariat machine-humain se fasse, c'est que l'humain sait très bien que la machine est très bonne sur un truc particulier, mais il ne doit pas remettre en question son jugement. S'il se dit : « Mon intuition me dit que peut-être pour ce patient-là, il faudrait investiguer plus. » Qu'il passe le temps pour le faire.

Selon cette vision, l'IA n'exercerait plus une pression concurrentielle mais viendrait complémenter le professionnel de santé. Ces derniers possèdent des connaissances théoriques, pratiques et expérientielles qu'il n'est pas possible de remplacer. Ces connaissances relèvent du « jugement clinique » mais surtout de « l'intuition ». Les systèmes d'IA seraient quant à eux de très bons appuis en ce qui a trait à l'analyse statistique (ex. des risques) pouvant traiter des millions de données. Les systèmes d'IA pourraient également servir de « deuxième lecture », permettant de valider l'intuition du médecin et la justesse de l'interprétation ou de réduire l'incertitude, en particulier pour l'analyse de situations standardisables.

Il a cependant été souligné que l'avènement de cette collaboration impliquerait une adaptation des professionnels de santé à ces nouvelles technologies; qu'il est essentiel de reconnaitre les capacités des systèmes d'IA mais également leurs limites, et que les systèmes d'IA à eux seuls ne pourront jamais suffire à remplacer les professionnels de santé car il manquera toujours « quelque chose ». Certains voient alors cette collaboration comme un « mariage arrangé » qui devra permettre que les professionnels de santé puissent choisir d'utiliser les systèmes d'IA ou non.

5. Conclusion

Les citoyens participants à la coconstruction de la Déclaration de Montréal ont ainsi discuté de trois grands défis de l'exercice de la responsabilité face au développement des systèmes d'IA en santé. L'ensemble de ces défis soulève des craintes et des attentes, résumées dans le Tableau 7. Le premier défi relève de la préservation des capabilités humaines, soit ce que les humains sont réellement capables de faire et d'être, qui sont essentielles à l'exercice de la responsabilité morale. Les citoyens ont ici manifesté des craintes relatives à l'incapacitation des professionnels de santé

et des patients, et des attentes relatives à la préservation de l'autonomie décisionnelle, notamment par l'entremise de formations et d'éducation.

Le second défi relève du problème des *mains multiples*, soit le fait que la multiplication des acteurs impliqués dans le parcours de soin complique l'attribution de la responsabilité morale et expose à un risque de déresponsabilisation en renvoyant la responsabilité à d'autres acteurs de la chaine causale. Les citoyens ont exprimé des craintes relatives aux conséquences de la multiplication des acteurs sur le soin et sur la santé, notamment concernant le partage de la propriété et de la gestion des données, une perte de lien naturel et de potentiels conflits d'intérêts entre les différentes parties prenantes. Ils ont formulé des attentes qui prennent la forme d'un contrat social définissant le partage des responsabilités des différents acteurs en fonction de leur rôle, et ont manifesté des attentes normatives relatives à ce partage : les institutions devraient être transparentes, les données devraient demeurer la propriété des usagers et accessibles aux professionnels de santé, et l'exercice de la responsabilité devrait se faire selon un certain principe de précaution. Ils ont également identifié les mécanismes existants qui pourraient répondre aux enjeux du développement responsable des systèmes d'IA en santé.

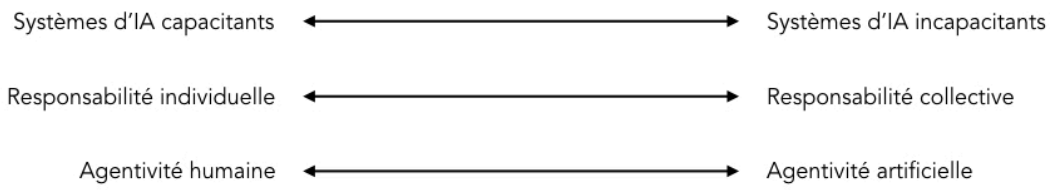
Le dernier défi est associé à l'autonomie croissante (ou la perception d'autonomie croissante) des systèmes d'IA. L'agentivité artificielle viendrait défier l'agentivité humaine en responsabilisant la technologie à la place des humains qui l'utilisent. Si les citoyens craignent l'apparition d'une agentivité artificielle, c'est surtout relativement aux biais que les algorithmes pourraient perpétuer dans leur analyse et à une certaine transformation du rapport à la technologie qui pourrait nuire à la confiance et aux rapports entre humains. Ils ont manifesté des attentes claires : les systèmes d'IA sont des outils, l'humain doit garder la main et s'il advenait que les systèmes d'IA deviennent des agents (décisionnels), ils devraient être transparents.

À mi-chemin entre agentivité artificielle et incapacitation humaine, une crainte du remplacement des humains par les machines a été soulevée, notamment car elle pourrait participer à la déshumanisation des soins, et une attente de coopération entre systèmes d'IA et professionnels

de santé a été formulée. Cette coopération devrait assurer que les systèmes d'IA soutiennent les professionnels de santé sans nuire à l'exercice de leur agentivité.

Ainsi, face au développement des systèmes d'IA en santé, il existerait, selon l'analyse faite des propos des citoyens consultés, trois principales manières de se déresponsabiliser, qui correspondent à trois principales tensions relatives à l'exercice de la responsabilité : une tension entre systèmes d'IA capacitants et incapacitants, entre la responsabilité individuelle et collective et entre l'agentivité humaine et artificielle (voir Schéma 4).

Schéma 4 - Les trois principales tensions relatives aux défis de l'exercice de la responsabilité dans le cadre de l'utilisation des systèmes d'IA en santé.



Afin d'assurer un développement éthique et responsable des systèmes d'IA en santé, il est essentiel de prendre en considération ces trois tensions afin de mettre en place un encadrement adapté et de répondre aux risques et enjeux éthiques de leurs utilisations. Cet encadrement doit favoriser le développement de technologies capacitantes, définir au mieux le partage des responsabilités et favoriser l'agentivité humaine en limitant l'agentivité artificielle.

Références bibliographiques

Alexiou, Athanasios, Maria Psixa, et Panagiotis Vlamos. 2011. « Ethical Issues of Artificial Biomedical Applications ». Dans *Artificial Intelligence Applications and Innovations*, 297-302. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-23960-1_36.

Allen, C., W. Wallach, et I. Smit. 2006. « Why Machine Ethics? » *IEEE Intelligent Systems* 21 (4): 12-17. https://doi.org/10.1109/MIS.2006.83.

Bostrom, Nick, et Eliezer Yudkowsky. 2011. « The Ethics of Artifical Intelligence ». Dans *The Cambridge Handbook of Artificial Intelligence*, 316-35. Cambridge University Press.

Bucher, Taina. 2016. « Neither Black Nor Box: Ways of Knowing Algorithms ». Dans *Innovative Methods in Media and Communication Research*, édité par Sebastian Kubitschko et Anne Kaun, 81-98. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-40700-5_5.

CCNE. 2019. « Données massives (big data) et santé : une nouvelle approche des enjeux éthiques ». Avis 130. Comité Consultatif National d'Éthique français. https://www.ccne-ethique.fr/sites/default/files/avis_130.pdf.

Coeckelbergh, Mark. 2010. « Health Care, Capabilities, and AI Assistive Technologies ». *Ethical Theory and Moral Practice* 13 (2): 181-90. https://doi.org/10.1007/s10677-009-9186-2.

———. 2015. « Artificial Agents, Good Care, and Modernity ». *Theoretical Medicine and Bioethics* 36 (4): 265-77. https://doi.org/10.1007/s11017-015-9331-y.

ÉPTC2 : Conseil de recherches en sciences humaines du Canada, Conseil de recherches en sciences naturelles et en génie du Canada, et Instituts de recherche en santé du Canada. 2018. « Énoncé de politique des trois Conseils : Éthique de la recherche avec des être humains ». http://www.ger.ethique.gc.ca/fra/policy-politique_tcps2-eptc2_2018.html.

Devillers, Laurence. 2017. *Des robots et des hommes: Mythes, fantasmes et réalité*. Plon.

Dixon-Woods, Mary, et Peter J. Pronovost. 2016. « Patient Safety and the Problem of Many Hands ». *BMJ Qual Saf*, février, bmjqs-2016-005232. https://doi.org/10.1136/bmjqs-2016-005232.

Floridi, Luciano, et Mariarosaria Taddeo. 2016. « What Is Data Ethics? » *Philosophical Transactions of the Royal Society A-Mathematical Physical and Engineering Sciences* 374 (2083): 20160360. https://doi.org/10.1098/rsta.2016.0360.

Gordon, John-Stewart. 2019. « Building Moral Robots: Ethical Pitfalls and Challenges ». *Science and Engineering Ethics*, janvier. https://doi.org/10.1007/s11948-019-00084-5.

Hoven, Jeroen van den. 2012. « Human Capabilities and Technology ». Dans *The Capability Approach, Technology and Design*, édité par Ilse Oosterlaken et Jeroen van den Hoven, 27-36. Philosophy of Engineering and Technology. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-3879-9_2.

Kononenko, Igor. 2001. « Machine learning for medical diagnosis: history, state of the art and perspective ». *Artificial Intelligence in Medicine* 23 (1): 89-109. https://doi.org/10.1016/S0933-3657(01)00077-X.

Lahlou, Saadi. 2008. « Identity, Social Status, Privacy and Face-Keeping in Digital Society ». *Social Science Information* 47 (3): 299-330. https://doi.org/10.1177/0539018408092575.

Moor, J. H. 2006. « The Nature, Importance, and Difficulty of Machine Ethics ». *IEEE Intelligent Systems* 21 (4): 18-21. https://doi.org/10.1109/MIS.2006.80.

Noorman, Merel. 2008. « Limits to the Autonomy of Agents ». Dans *Current Issues in Computing and Philosophy*, édité par P. Brey, A. Briggle, et K. Waelbers, 65–75. Ios Press.

———. 2016. « Computing and Moral Responsibility ». Dans *The Stanford Encyclopedia of Philosophy*, édité par Edward N. Zalta, Winter 2016. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2016/entries/computing-responsibility/.

Nussbaum, M.C. 2011. *Creating capabilities: The human development approach*. Harvard: The Belknap Press of Harvard University Press.

Oosterlaken, Ilse. 2015. *Technology and Human Development*. Routledge.

Proudfoot, Diane. 2011. « Anthropomorphism and AI: Turing's much misunderstood imitation game ». *Artificial Intelligence*, Special Review Issue, 175 (5): 950-57. https://doi.org/10.1016/j.artint.2011.01.006.

Scheutz, Matthias. 2016. « The need for moral competency in autonomous agent architectures ». Dans *Fundamental issues of artificial intelligence*, Springer International Publishing Switzerland, 517-27. V.C. Müller.

Schlosser, Markus. 2015. « Agency ». Dans *The Stanford Encyclopedia of Philosophy*, édité par Edward N. Zalta, Fall 2015. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2015/entries/agency/.

Selgelid, Michael J. 2013. « Dual-Use Research ». *The International Encyclopedia of Ethics*. http://onlinelibrary.wiley.com/doi/10.1002/9781444367072.wbiee607/full.

Sen, Amartya. 1992. *Inequality reexamined*. Cambridge: Harvard University Press.

———. 2005. « Human Rights and Capabilities ». *Journal of Human Development* 6 (2): 151-66. https://doi.org/10.1080/14649880500120491.

Sharkey, Noel. 2008. « The Ethical Frontiers of Robotics ». *Science* 322 (5909): 1800-1801. https://doi.org/10.1126/science.1164582.

Shulman, Carl, Henrik Jonsson, et Nick Tarleton. 2009. « Machine ethics and superintelligence ». Dans , 95–97. Tokyo, Japan: Carson Reynolds and Alvaro Cassinelli.

Thompson, Dennis F. 2004. « The problem of many hands ». Dans *Restoring Responsibility : Ethics in Govern1m1ent, Business, and Healthcare*, Cambridge University Press.

Zwitter, Andrej. 2014. « Big Data Ethics ». *Big Data & Society* 1 (2): 2053951714559253. https://doi.org/10.1177/2053951714559253

Chapitre 6 – Vers une responsabilité pragmatique pour l'innovation numérique en santé responsable

Les pages suivantes visent à mettre en lien les différents constats qui se dégagent des chapitres précédents afin de répondre à la question de recherche qui guide la présente thèse, soit de déterminer quels sont les défis du partage et de l'exercice de la responsabilité face aux risques et enjeux éthiques soulevés par l'utilisation de systèmes d'intelligence artificielle en santé en vue d'informer une innovation responsable. Dans la perspective de favoriser une innovation responsable, il semble nécessaire d'adapter les mécanismes existants de l'encadrement éthique des systèmes de santé. Afin d'adapter ces mécanismes, il est essentiel de tenir compte, sans pour autant s'y limiter, des trois tensions qui émergent des craintes et attentes citoyennes exprimées lors de la coconstruction de la Déclaration de Montréal, afin de répondre aux différents risques et enjeux éthiques de l'avènement des systèmes d'IA en santé. Si ces tensions renvoient à un risque réel de déresponsabilisation, une vision pragmatique de la responsabilité, parce qu'elle renvoie à différentes interpellations responsabilisantes, offre ici différents éléments pertinents afin de réfléchir aux défis de l'exercice de la responsabilité en vue de la promotion d'un encadrement éthique de l'utilisation des systèmes d'IA en santé.

1. L'innovation numérique en santé responsable selon une vision pragmatique de la responsabilité

Tel que présenté dans le Chapitre 1, le concept d'innovation et de recherche responsable (IRR) permet de mobiliser la responsabilité de différents acteurs de l'innovation et de la recherche (notamment, celle des chercheurs) en vue du développement d'une recherche et d'une innovation qui intègrent les attentes et les valeurs de la société. Cependant, selon Pellé et Reber (2016), aucune des approches de l'IRR ne semble s'être penchée en détail sur le concept de responsabilité en lui-même ni sur ses aspects polysémiques. Si les auteurs soulignent que la responsabilité morale semble être celle susceptible de guider au mieux les réflexions sur l'IRR[124], ils ont également mis en évidence que les discussions sur la définition de la responsabilité morale sont pratiquement inexistantes dans le discours qui l'entoure. Ayant passé en revue 10 conceptions de la responsabilité

[124] Comme précédemment mentionné, la responsabilité peut également renvoyer à une conception juridique ou sociale.

morale dans le cadre de l'IRR, Pellé et Reber (2016) concluent qu'il est possible de combiner ces différentes conceptions en vue d'interprétations riches de la responsabilité, qu'il devient alors possible d'adapter au contexte spécifique dans lequel elles s'appliquent. C'est également une approche sensible au contexte qui est sous-tendue par la conception pragmatique de la responsabilité ; car les conceptions de la responsabilité formelle ou de la responsabilité sollicitude (les deux pôles de la responsabilité morale présentés par Métayer) nous en disent « à la fois trop et pas assez sur la réalité de notre système » (Métayer 2001).

Si l'objectif n'est pas ici de proposer une nouvelle définition de la responsabilité en vue d'un apport théorique, il est cependant essentiel de se pencher brièvement sur certaines de ces conceptions afin de décrire et de comprendre ce que les craintes et les attentes citoyennes impliquent en termes de responsabilité. Les notions mobilisées dans le Chapitre 4 font en effet référence à des concepts associés à différentes conceptions de la responsabilité en philosophie morale, et plus particulièrement le concept d'agentivité qui est au cœur des trois tensions qui émergent des discussions citoyennes, soient celles entre : 1) technologies capacitantes et incapacitantes, 2) responsabilité individuelle et collective ; 3) agentivité artificielle et agentivité humaine. Si l'agentivité peut être définie très simplement comme la capacité d'agir (de manière intentionnelle) (Schlosser 2015), les craintes et attentes citoyennes renvoient à différentes de ces conceptions selon les tensions soulevées.

En effet, préserver l'agentivité humaine est un des éléments essentiels du mouvement des technologies appropriées qui se base sur l'approche des capabilités (Oosterlaken 2015) et qui a permis de mettre en évidence la première tension entre technologies capacitantes et incapacitantes. Le problème des *mains multiples* complique quant à lui l'attribution de la responsabilité face au nombre croissant d'acteurs impliqués dans le parcours de soins, notamment en dispersant le lieu de la capacité d'agir (Dixon-Woods et Pronovost 2016). L'apparition d'une agentivité artificielle (comprise selon une conception plus restreinte que celle de l'agentivité humaine) est crainte car elle pourrait nuire au contrôle des humains sur les machines – une des dimensions potentielles de l'exercice de leur agentivité – et conduire à une déresponsabilisation. Ces différentes façons d'entraver l'agentivité sont intrinsèquement liées : s'il est possible de considérer les systèmes d'IA

comme des « agents » ils deviennent par le fait même un des acteurs potentiels à qui renvoyer la responsabilité dans le cadre du problème des *mains multiples*. Les différentes situations décrites participent à l'incapacitation potentielle des humains (ex. en les remplaçant par des machines). Les tensions entre agentivité humaine et artificielle ou responsabilité individuelle et collective sont également liées au maintien des capabilités.

Les trois tensions mises en évidence demandent ainsi de préserver (voire de favoriser) l'agentivité humaine et, par extension, la responsabilisation. Le lien entre l'agentivité et l'exercice de la responsabilité se retrouve en partie dans les critères des conceptions formelles de la responsabilité morale (soit, le pôle négatif et rétrospectif), ce qui demande de s'attarder brièvement sur certains d'entre eux. Comme mentionné précédemment, la responsabilité formelle (ou « négative »), est associée à une vision antagonisante de la responsabilité et tente de définir les critères qui pourront déterminer qu'un « agent » est responsable, notamment dans la perspective de déterminer qui blâmer pour les conséquences négatives d'une action (Chartrand 2017; Métayer 2001; Pellé et Reber 2016). Elle peut encourager la suresponsabilisation selon la restriction de ces critères ou la déresponsabilisation des individus cherchant à éviter le blâme (Métayer 2001; Gotterbarn 2001). C'est dans la détermination desdits critères qu'il est possible de saisir l'importance de l'agentivité (Chartrand 2017) mais également de comprendre comment l'avènement des systèmes d'IA peut défier l'exercice de la responsabilité. Parmi ces critères, se trouvent : la **conscience** ou la connaissance des conséquences (un agent est responsable s'il a conscience de ses actes) ; la **causalité** ou le lien causal (un agent est responsable d'une action s'il l'a causée) ; le **contrôle** (un agent ne peut être considéré comme responsable d'un évènement que s'il a du contrôle sur celui-ci) ; la **raison** (un agent est responsable s'il peut évoquer les raisons qui ont motivé son action ou son jugement) (Chartrand 2017; Noorman 2016). À ces critères peut s'ajouter celui du **libre arbitre** comme condition essentielle à l'attribution de la responsabilité (un individu est responsable d'une action s'il l'a librement réalisée) (Noorman 2016).

Si ces critères sont définis dans le cadre de modèles théoriques relativement fictifs, différents auteurs ont mis en évidence que l'avènement des systèmes d'IA va avoir un impact sur nombre d'entre eux. Les avancées en IA compliquent effectivement la gestion du risque technologique en

venant modifier les différentes conditions nécessaires à l'attribution de la responsabilité selon les critères de la responsabilité formelle (Noorman 2016; Chartrand 2017; Bucher 2016). Si le **lien causal** est compliqué par les avancées en IA, c'est notamment parce que l'utilisation de systèmes d'IA implique de multiples acteurs et introduit une plus grande distance temporelle et physique entre l'action d'une personne et les conséquences de cette action – ce qui renvoie aux considérations associées au problème des *mains multiples* (Noorman 2016; Chartrand 2017). Les infrastructures complexes mobilisées en vue d'une seule et même application compliquent l'identification des différents éléments de la chaîne causale (Chartrand 2017) d'autant plus que les algorithmes évoluent dans le temps, diluant ainsi l'identification de leur potentielle agentivité (Bucher 2016). Également, il devient difficile d'envisager préalablement les conséquences des actions réalisées par l'entremise des systèmes d'IA, ce qui peut compliquer la connaissance des conséquences (ou **conscience**). En effet, les technologies numériques permettent aux humains des actions qu'ils ne pouvaient pas faire avant, introduisant des conséquences qui ne connaissent pas de précédents (Noorman 2016). Sans précédent, il est difficile de blâmer un individu qui n'aurait pas imaginé les conséquences négatives d'une de ses actions[125].

Le **contrôle** est défié par la division des pouvoirs relativement au nombre croissant d'acteurs impliqués (Chartrand 2017), mais également au manque de compréhension ou de moyens pour garantir une gestion des données massives et des systèmes d'IA. Les technologies numériques viennent également affecter la prise de décision humaine et influence la manière de faire des choix[126], pouvant alors potentiellement influencer l'exercice effectif du **libre arbitre** (Brundage 2016; Noorman 2016). Enfin, répondre au critère de **raison** est particulièrement difficile car les systèmes d'IA perturbent l'attribution de la responsabilité par leur rôle de médiation, en s'interposant entre le sujet et l'objet ou entre les humains entre eux, notamment relativement à leur fonction décisionnelle (Chartrand 2017). C'est-à-dire que lorsqu'un système d'IA pose « un acte épistémique » c'est selon Chartrand « souvent sur la base de raisons qui lui appartiennent en

[125] La conscience des conséquences demeure un critère difficile à expliquer et à reconnaître objectivement chez un agent – humain ou non (Chartrand 2017). Un exemple de conséquence négative non envisagée par les concepteurs de systèmes d'IA est celui du *chatbot* Tay de Microsoft. Lancée en mars 2016 sur Twitter, l'agent conversationnel initialement programmé pour modéliser le discours d'une adolescente est rapidement retiré du réseau social car il tient, de manière inattendue, de nombreux propos référés à des propos Nazis (Beran 2018).
[126] Voir à ce sujet les effets des *nudges* ou de bulles filtrantes, Section 1. du Chapitre 4.

propre »[127]. Ainsi, l'avènement des systèmes d'IA en santé risquerait bien de défier l'exercice de la responsabilité des différents acteurs qui prennent part au parcours de soins, et par là-même risquerait de défier la mise en place de mesures de gouvernance éthique effectives, en particulier celles qui visent à définir les balises de la responsabilité individuelle de certains d'entre eux (ex. un code de déontologie à l'attention des professionnels de santé).

S'il permette de mieux comprendre le rôle de l'agentivité dans l'exercice de la responsabilité, la mise en perspective des craintes et attentes citoyenne en vue d'une innovation responsable ne saurait cependant se limiter aux critères de la responsabilité formelle. Comme précédemment mentionné, la responsabilité morale peut également être conçue de manière positive (ce que Métayer dénomme la responsabilité sollicitude) (Pellé et Reber 2016; Métayer 2001) - bien qu'il existe de nombreuses nuances à apporter à cette division. La responsabilité positive renvoie à une responsabilité généralement prospective, comme un souci pour l'avenir et les conséquences de ce qui pourrait advenir sur « l'Autre » ou « un autre vulnérable » (Métayer 2001; Pellé et Reber 2016). Une conception positive de la responsabilité reconnaît un « lien indéfectible » entre les actions et la responsabilité des individus (Pellé et Reber 2016). Elle ne cherche pas à déterminer un individu à blâmer mais plutôt à identifier qui est en mesure de participer à la gestion des conséquences d'une action, ou qu'une personne jugée responsable s'assure qu'une action est effectuée ou évitée selon des degrés d'engagement variés (Gotterbarn 2001; Pellé et Reber 2016). Dans le contexte de la présente thèse, il s'agit d'identifier qui sont les acteurs qui peuvent et doivent participer à la gestion des risques et des bénéfices de l'utilisation des systèmes d'IA en vue d'une innovation responsable, selon un principe de précaution. La dimension positive de la responsabilité est ainsi également importante à considérer. Cependant, si cette conception permet d'inspirer les conduites actuelles en vue « d'un certain horizon normatif » (Pellé et Reber 2016), il est difficile de préciser les modalités de l'intervention responsable (notamment face à l'incertitude inexorablement associée à la prospective) (Métayer 2001). Elle peut également favoriser une certaine déresponsabilisation en renforçant les tendances au désengagement quand un individu choisit d'étendre sa responsabilité (Métayer 2001). Ces deux courants (responsabilité positive et négative), s'ils proposent des outils

[127] Cette considération renvoie en partie à celles de l'opacité des réseaux de neurones présentées dans le Chapitre 3.

théoriques fort utiles, entretiennent ainsi selon Métayer des « tendances concomitantes » à la déresponsabilisation et à la suresponsabilisation.

Selon Métayer (2001), si les codes sociaux permettaient de fixer et de délimiter clairement les responsabilités, ils

> ont été remplacés par un système de distribution des responsabilité ouvert et fluctuant, en constante redéfinition, tissant des réseaux d'interdépendance et des hiérarchies de niveaux de responsabilité d'une grande complexité. Les deux pôles de la responsabilité formelle rétrospective et de la responsabilité prospective font que le régime moral moderne déploie un espace flottant de responsabilité, livré aux initiatives des acteurs sociaux, qui devient le théâtre de multiples luttes et débats dont l'objet est l'appropriation, l'attribution et le partage des responsabilités (p. 46).

C'est bien le contexte et les circonstances qui détermineront alors si une action fera l'objet d'un jugement d'ordre moral; ceci fonctionnant avec les deux visions dichotomiques de la responsabilité car c'est là qu'elles pourront « y préciser leurs exigences » (Métayer 2001). Ainsi, plutôt que d'écarter les deux visions, la mise en perspective des craintes et attentes citoyennes selon une conception pragmatique de la responsabilité telle qu'entendue dans la présente thèse permet de considérer des exigences qui relèvent à la fois des deux conceptions. Comme le soutiennent Pellé et Reber (2016), il n'est pas nécessairement pertinent de choisir une conception de la responsabilité sur une autre dans le cadre de l'IRR, mais bien de toutes les considérer :

> En ayant à l'esprit la pluralité des conceptions de la responsabilité, les acteurs de l'innovation et de la recherche peuvent éviter les choix arbitraires, et mieux même, faire des choix appropriés (p. 135).

Selon les auteurs, chacune des conceptions de la responsabilité peuvent être considérées comme des parties de celle-ci en fonction des conditions de l'attribution (soit, le contexte), à la fois individuelle et collective, à la fois positive ou négative.

Or, comme souligné dans le Chapitre 4, la résolution de problèmes éthiques relatifs au développement responsable de l'IA en appelle également à des approches de l'éthique appliquée inductives et sensibles au contexte. Cleret de Langavant (2001) souligne la nécessité pour la bioéthique d'aborder les problèmes éthiques selon une approche intégrative et complexe, afin de tenir compte de l'ensemble des éléments contextuels liés à l'émergence de valeurs dans les

problèmes éthiques, de se prémunir d'un certain réductionnisme en considérant la multi dimensionnalité des phénomènes; incluant par exemple l'écologie de l'action[128] et la diversité de ses acteurs. L'écologie de l'action lorsqu'il est question de la responsabilité face à l'utilisation des systèmes d'IA en santé est particulièrement importante, considérant les défis soulignés par les citoyens : le nombre croissant d'acteurs impliqués (incluant possiblement comme acteurs les systèmes d'IA eux-mêmes) et les difficultés associées à l'exercice de la responsabilité qui en découlent. Également, il est possible de dégager la multi dimensionnalité des problèmes qui émaneraient du développement des systèmes d'IA en santé qui, comme présenté au fil des chapitres, demande de prêter une attention particulière aux considérations techniques, éthiques et sociales des nombreux contextes d'application de technologies et méthodes variées qui relèvent de l'IA.

Puisque nous nous intéressons ici à la dimension éthique de la responsabilité qu'invoque l'innovation responsable, on retiendra la nécessité, pour se prémunir d'une certaine « tyrannie des principes » (Toulmin 1981), de porter une attention particulière à la singularité des situations et aux contextes d'application des systèmes d'IA. Les considérations relatives à l'opérationnalisation des principes de l'éthique appliquée (notamment, celle relative à l'universalité potentielle de la portée des principes) sont d'ailleurs essentielles à prendre en compte en vue d'une innovation responsable, comme le discutent Pellé et Reber (2016) qui défendent un « pluralisme éthique » appliqué à la responsabilité. Le pragmatisme philosophique (reconnaissant la primauté de l'expérience pratique) suppose également un certain pluralisme (plutôt qu'un relativisme éthique) (Ralph 2018) notamment lorsqu'il est utilisé pour réfléchir à la responsabilité morale (Smiley 1992). Il est ainsi possible, avec une approche pragmatique de la responsabilité morale, de considérer à la fois des principes directeurs généraux et une application contextuelle, comme le souligne Smiley (1992) :

> While we may not be able to locate a universally correct understanding of moral responsibility, we can locate general rules according to which our practices of causal

[128] L'écologie de l'action correspond à une certaine déviation de l'action humaine, qui « échappe alors à la volonté et à l'entendement de son initiateur pour entrer dans un jeu d'interactions multiples qui la détourne de son but et lui donne parfois une destination contraire à celle qui était visée. » (Cleret de Langavant 2001 p. 142 citant Morin 1990)

> responsibility and blaming are now governed – rules which we inevitably apply according to our own particular interests and purposes (Smiley 1992 p. 22).

Une approche pragmatique est d'autant plus pertinente qu'elle autorise le doute et la remise en question (notamment, des normes en vigueur) sans d'ailleurs nécessairement restreindre la signification d'un concept à une fin particulière ou un objectif personnel (ce qui semble sous-entendu dans la citation de Smiley) (Ralph 2018). Ainsi, les défis de l'exercice de la responsabilité mis en évidence invitent ici à questionner les pratiques actuelles (et les valeurs et les normes qui les sous-tendent) dans le cadre de l'utilisation des systèmes d'IA en santé. De plus, comme le souligne Ralph (2018) l'engagement du pragmatisme envers l'expérimentalisme peut se traduire par un engagement démocratique visant l'amélioration du débat public (Ralph 2018). C'est ainsi en partie par l'entremise de l'intelligence collective et de processus délibératifs que le pragmatisme permet de tester les savoirs des experts et faire face au doute qui pourrait être créé par le pluralisme (Ralph 2018). Les tensions qui émergent des discussions citoyennes constituent alors un point de départ particulièrement pertinent si l'on envisage d'adapter les mécanismes existants de leur gouvernance éthique en santé d'un point de vue pragmatique.

En conséquence, il semble qu'en tout point l'éthique des systèmes d'IA en appelle à un certain pragmatisme sensible aux contextes d'application et à la singularité des situations où prennent naissance les enjeux éthiques. L'essence de l'innovation responsable se retrouve donc en partie dans les considérations contextuelles, tant en ce qui a trait à la détermination de la responsabilité qu'à la résolution des problèmes éthiques. Plus spécifiquement, selon la responsabilité pragmatique de Métayer, l'innovation responsable pourrait se définir sur la base d'une pratique d'interpellation (des « interpellations responsabilisantes ») qui met en relation un demandeur et un répondeur. Le demandeur est normativement actif dans la structure de la pratique d'interpellation par le fait qu'il « demande à l'autre des comptes, de répondre de ses actes ou de répondre de quelqu'un » (Métayer 2001). Face à l'interpellation, le répondeur est l'« acteur individuel ou collectif qui est le destinataire de l'interpellation de départ » et se doit de répondre à l'interpellation responsabilisante, quelle que soit la conception de la responsabilité mobilisée (Métayer 2001). Parce-que cette structure d'interpellation est comprise comme une pratique de responsabilisation, et que les éléments de cette pratique participent, de fait, à la construction du système moral, elle permet de

dessiner différentes pistes de réflexion en vue d'assurer une innovation numérique en santé responsable.

Les interpellations responsabilisantes sont celles qui émanent des différents risques et enjeux éthiques mentionnés dans le chapitre 3, que l'on retrouve dans les craintes et les attentes citoyennes présentées dans le Chapitre 5[129]. En effet, selon la responsabilité pragmatique de Métayer :

> Nous ne sommes pas seulement responsables en principe de toutes nos actions, mais aussi et toujours de quelque chose en particulier que nos actions affectent, que nous jugeons important et qui peut justifier l'expression d'un blâme ou d'un éloge (p. 20).

Ainsi, le respect du consentement des patients et des participants à la recherche interpelle, et renvoie à la nécessité de préserver les capabilités humaines. La déshumanisation pourrait être favorisée par le remplacement des humains par les machines ou la perte de lien naturel entre patients et professionnels de santé. De la même manière, les préoccupations relatives au respect de la vie privée et de la confidentialité, de la justice sociale, de la déshumanisation des soins ou du patient représentent toutes des interpellations responsabilisantes, encouragées ou concernées par certains des défis de l'exercice de la responsabilité.

Il est possible de dégager des éléments présentés dans les chapitres précédents quelques exemples pratiques d'usage des systèmes d'IA où l'exercice de la responsabilité pourrait être défié ou son attribution difficile à définir. Notamment, le respect du consentement libre et éclairé dans le cadre de l'utilisation secondaire de données (massives) représente une interpellation responsabilisante face à la préservation des *capabilités* des patients; soit de respecter leur décision en fonction de ce qu'ils valorisent et sur la base de ce qu'ils ont effectivement compris de la situation. Le consentement pourrait être mis à mal dans une situation ou l'individu consent dans le cadre de la collecte initiale (par exemple, par le biais d'une application mobile de bien-être) mais n'aurait pas consenti aux utilisations secondaires (par exemple, si les données été réutilisées par

[129] Sans entrer dans le détail, différents éléments soulevés dans le Chapitre 3 relativement aux cinq catégories de risques et enjeux éthiques ressortent des discussions citoyennes relativement aux trois défis identifiés. On note par exemple que le respect du consentement libre et éclairé se retrouve dans les préoccupations relatives à la préservation des capabilités humaines ; que la protection de la vie privée et de la confidentialité est une des conséquences du problème des *mains multiples*, ou que les préoccupations relatives à la justice sociale se retrouve en partie dans les craintes associées à la reconnaissance d'une agentivité artificielle (relativement aux biais algorithmiques). Retrouvant ces préoccupations dans les discussions de la coconstruction de la Déclaration de Montréal, l'interpellation responsabilisante n'en est que renforcée.

une compagnie d'assurance santé – comme mentionné dans les discussions citoyennes). S'il a été présenté qu'il est de plus en plus difficile pour les utilisateurs de garder un contrôle sur ce qu'il advient de leurs données et de comprendre l'étendue de leurs utilisations potentielles, qui pourrait alors être tenu pour responsable dans cette situation de l'atteinte à l'autonomie du patient – voire à sa vie privée si le couplage des données collectées à des données auxiliaires a permis de l'identifier à nouveau ? Il est difficile de déterminer qui pourrait être le *répondeur* d'une telle interpellation, comme cela a été souligné par les citoyens dans le contexte de la propriété des données p. 235 (*cf.* le défi des *mains multiples*) :

> Qui est imputable ? Les personnes qui ont entré les données ? Celle qui créé les algorithmes ? Le personnel médical ou le médecin prescripteur du robot ? Ceux qui développent les algorithmes ou ceux qui les exploitent ? Qui serait responsable si le système venait à être hacké ?

Considérant, comme présentés précédemment, les différentes manières dont l'informatisation de la société vient défier l'attribution de la responsabilité, les seuls critères de la responsabilité formelle ne saurait suffire ici pour identifier un responsable en vue de prévenir et gérer les risques de la situation. De même, les exigences de la responsabilité positive (déterminer qui devrait participer à la gestion des risques) ne permettent pas d'identifier l'étendue de la responsabilité de chacune des parties prenantes. Si l'utilisation secondaire était celle d'un chercheur, aurait-il le devoir d'informer le patient d'une découverte fortuite concernant sa santé ? Comment serait-ce possible sans réidentifier le patient ?

Nombreux des éléments contextuels seraient ici à prendre en compte afin de définir la responsabilité des parties prenantes engagées, des conditions de la collecte (ex. le niveau d'encadrement de cette dernière, l'accessibilité des clauses de confidentialité, l'expertise éthique ou relative à la santé des personnes qui la réalisent) à celle de l'analyse (ex. réalisée par un acteur qui opère dans le système publique ou privé, lesquels ne demandent pas de répondre aux mêmes normes déontologiques) en passant par les conditions qui permettent de préserver les capabilités des demandeurs et répondeurs (ex. informations claires pour les patients relativement à la réutilisation, connaissance pour les réutilisateurs des limites relatives au respect du consentement, préférences individuelles du patient ou préférences partagées des patients d'une même société).

Autre exemple d'interpellation responsabilisante, l'opacité des réseaux de neurones pourrait défier les *capabilités* des professionnels de santé comme des patients (comme mentionné dans le chapitre 3), en rendant inaccessibles les raisons qui sont à la base d'une recommandation algorithme – limitant par la même la compréhension de la recommandation, ou simplement ne rendant pas possible d'identifier la raison qui permet de préférer une solution sur une autre. Ceci pourrait, dans le contexte du diagnostic par un système expert, exposer à de sérieux dilemmes puisqu'il s'agit de décisions qui pourraient mettre en jeu la vie des patients (Castelvecchi 2016). D'un autre côté, la nécessité de transparence pourrait bien limiter l'utilisation de systèmes opaques potentiellement efficaces (Lipton 2016) qui pourraient détecter plus précisément l'apparition de maladie. S'il semble ici difficile de blâmer le professionnel de santé ou le patient selon les critères de responsabilité formelle (par exemple, le critère de **conscience**), la question se pose de savoir qui blâmer en cas d'erreur : le professionnel de santé ? le développeur ? ou l'algorithme lui-même (considérant le défi de *l'agentivité artificielle*) ? Écarter simplement ces critères ne permet cependant pas de saisir les difficultés de la situation relativement à l'attribution de la responsabilité.

En effet, relativement au défi de l'agentivité artificielle, une approche pragmatique de la responsabilité telle que défendue par Crnkovic et Persson (2008) permet de reconnaître une certaine forme de responsabilité aux robots et autres technologies qui relèvent de l'IA car celles-ci possèdent une forme d'agentivité (ceci est discuté plus en détail dans la section 2.1. du présent chapitre). Les systèmes d'IA selon cette perspective seraient responsables de « bien réaliser leurs tâches », ce qui renvoie à la conception de la responsabilité morale comme « rôle » (le rôle circonscrit l'étendue de la responsabilité aux tâches qui sont assignées à l'agent - Pellé et Reber 2016) ou plus simplement, renvoie à la notion de fiabilité des systèmes (Crnkovic et Persson 2008). Cette conception fonctionnelle ne saurait cependant suffire à répondre au problème posé par la situation. D'autres éléments contextuels (ex. l'importance pour le patient de comprendre les raisons qui motivent la décision médicale, le niveau de compréhension du système expert par le médecin, les autres alternatives possibles au diagnostic) seraient à considérer.

Dans ces exemples, l'identification des défis et celle des pistes de solution se dessinent non pas grâce à une conception particulière de la responsabilité morale mais bien une combinaison

pragmatique de leurs différentes exigences (cet apport est présenté dans les sections suivantes du présent chapitre). Défendre qu'il est nécessaire de tenir compte du contexte de chaque situation signifie cependant qu'il n'existe pas une mais plusieurs réponses (ou solutions) à ces différentes interpellations responsabilisantes. Déterminer les spécificités contextuelles de chacune des situations où, dans le contexte de la santé, les systèmes d'IA soulèveraient des difficultés éthiques sort néanmoins du cadre de l'analyse de la présente thèse[130]. Cependant, il est possible de dégager certains fondamentaux relativement aux défis de l'attribution de la responsabilité dans le contexte de l'innovation responsable en santé, lesquels relèvent du niveau « macro » de l'IRR tel que présenté par Barré (2011) (cf. p. 69 du Chapitre Méthodologie).

Partant de la structure d'interpellation responsabilisante présentée (un demandeur et un répondeur), les différents éléments qui risquent de défier l'exercice de la responsabilité dans le contexte des trois tensions qui émergent des craintes et attentes citoyennes sont ainsi explorés ci-après, confondant les notions qui relèvent de la responsabilité positive ou de la responsabilité négative[131], en vue de dessiner des pistes de réflexion pour une innovation numérique en santé responsable.

2. Les trois tensions émergentes selon une innovation responsable pragmatique

2.1. La tension entre agentivité humaine et artificielle

Les citoyens ont exprimé des craintes relatives au fait que les systèmes d'IA deviennent des agents et leur principale attente relativement à ce point est que ces systèmes demeurent des outils. Un des éléments qui ressort de ces craintes et de ces attentes est que l'agentivité, telle qu'entendue dans le discours citoyen, renvoie à une conception différente selon qu'il s'agisse de celle des systèmes d'IA ou de celle des humains[132]. En effet, les craintes citoyennes de la reconnaisse d'une agentivité artificielle renvoient principalement à la crainte que les systèmes d'IA deviennent des

[130] Voir la section *Limites relatives au projet en lui-même* du Chapitre 1 : Méthodologie.
[131] Cette exploration ne saurait cependant être exhaustive et ne tient pas compte de toutes les conceptions possibles de la responsabilité morale.
[132] Cette section s'attarde sur l'agentivité artificielle, l'agentivité humaine est discutée dans la dernière section de ce chapitre.

agents (moraux) décisionnels, en dehors d'autre considérations (ou critères) à la base de la capacité d'agir. De ce fait, l'agentivité crainte des systèmes d'IA renvoie à un type d'action bien précis : la décision. Par exemple, la dimension intentionnelle ressort peu des craintes et des attentes citoyennes (les citoyens, dans le contexte de la coconstruction de la Déclaration de Montréal, ne semblent pas s'être souciés de la véritable intention d'agir des systèmes d'IA).

La conception des systèmes d'IA comme des agents tel que discuté ne renvoie ainsi pas nécessairement à la conception relativement « classique » de l'agent intentionnel libre et autonome à qui l'on attribue traditionnellement une responsabilité formelle (Métayer 2001). C'est par exemple ce que soutient Johnson, qui défend que les algorithmes ont une pertinence morale mais pas une responsabilité en tant que telle car ils n'ont pas d'intentionnalité ou, s'ils en ont une, elle reste reliée à leur fonctionnalité (soit une absence de lien causal) (Johnson 2006). Comme présenté dans le Chapitre 5, les craintes relatives à l'agentivité artificielle demandent de limiter leur potentielle capacité « décisionnelle », qui doit rester celle des humains. Si l'on suit la logique associée à ces craintes et ces attentes, une première recommandation qui émerge d'un point de vue de l'innovation responsable pourrait ainsi être celle de limiter les capacités décisionnelles des systèmes d'IA ; mais pas forcément toutes formes d'agentivité.

En effet, il se pourrait qu'il y ait un intérêt normatif à reconnaître une potentielle agentivité aux systèmes d'IA. Comprendre ce point demande de se pencher sur les théories descriptives qui reconnaissent effectivement une certaine capacité d'agir aux agents artificiels. Dans une perspective Latourienne, l'agentivité se retrouve distribuée dans les réseaux d'acteurs, ces derniers rassemblant à la fois des humains et des technologies[133] (Latour 1996; Ihde 2006). Selon cette perspective, les systèmes d'IA peuvent être considérés comme des agents. Ils ont en effet une certaine capacité d'agir en recommandant des prescriptions, en posant des diagnostics ou en identifiants des facteurs de risques en santé, toutes pouvant être considérées comme des formes d'action. Selon Verbeek, les décisions et les actions morales se retrouvent bien dans un mélange

[133] Ihde (2006) décrit par exemple une « designer fallacy », idée fausse selon laquelle les concepteurs de technologies peuvent introduire dans leur conception des intentions d'usages reproduites par l'artéfact selon une certaine neutralité, ce qui ne tiendrait pas compte de la complexité de la conception technologique.

d'agentivité entre humains et technologies. Les technologies façonnent en partie les actions humaines et les actions humaines façonnent elles-mêmes l'agentivité des technologies (Verbeek 2006).

Dans une perspective contextuelle de l'éthique et de la responsabilité, ce point est primordial. Il serait problématique d'être aveugle à cet état de fait dans une perspective normative en vue d'une innovation responsable : en accordant une certaine capacité décisionnelle aux systèmes d'IA, les humains pourraient effectivement leur conférer une certaine « responsabilité » qui risque de détourner l'attention de celle des humains qui les conçoivent ou qui les utilisent. Reconnaitre une agentivité aux systèmes d'IA permet de reconnaître le risque de déresponsabilisation qui y est associé en se penchant sur cette spécificité qui émerge du contexte de l'utilisation des systèmes d'IA en santé. Cette idée est soutenue par Chartrand (2017), qui discute des moyens de répondre aux perturbations qui accompagnent l'utilisation des systèmes d'IA (dans un contexte de la responsabilité épistémique) :

> Une façon de faire consiste à donner aux agents artificiels un statut d'agents qui leur permette une certaine forme de responsabilité. Ce faisant, on peut s'attaquer en partie au problème de l'internalisme : si on admet une forme d'agencéité épistémique aux algorithmes qui font une partie de notre travail cognitif, on peut rendre compte de ce qui dilue notre responsabilité en imputant la source de cette dilution à une entité précise (p. 11).

Il existe ainsi un risque de ne pas reconnaître une agentivité aux systèmes d'IA : celui de passer à côté des perturbations qui accompagnent cette reconnaissance d'agentivité. En d'autres mots et pour reprendre l'attente normative formulée par les citoyens, limiter l'agentivité des systèmes d'IA suppose de reconnaître qu'ils détiennent une certaine capacité d'agir de manière autonome pouvant conduire à les rendre imputables dans certaines circonstances.

Reconnaissant une forme d'agentivité aux systèmes d'IA, il est nécessaire de distinguer cette capacité d'agir d'autre technologies ou dispositifs médicaux. On parle bien en effet du « mode d'action » d'un médicament, et la production d'images diagnostiques n'est-elle pas une « action » qu'opère un scanner ? Relativement à cet aspect, les craintes et attentes citoyennes ont mis en évidence qu'un des éléments qui distinguent les technologies médicales traditionnelles des systèmes d'IA est l'autonomie (ou la perception d'autonomie) de ces derniers – soit, notamment,

le fait que le système « apprend tout seul » (cf. Section 3.2. du Chapitre 5). Cependant, considérant les éléments présentés dans le Chapitre 2, il est essentiel de considérer ici les différentes méthodes, techniques et outils qui relèvent de l'IA car tous les systèmes d'IA n'ont pas le même niveau d'autonomie. Les différents types d'apprentissage automatique ne nécessitent pas le même degré de supervision ou d'intervention humaine. Par exemple, l'apprentissage supervisé nécessite un étiquetage humain des données alors que l'apprentissage profond apprend à partir des données brutes sans intervention humaine une fois que le modèle est créé. Les systèmes experts d'aide à la décision viennent appuyer la décision des professionnels de santé alors que les robots de soin peuvent réaliser certaines tâches, jusqu'ici réservées aux professionnels de santé, de manière relativement autonome.

En fonction de la conception que l'on a de l'autonomie, les limites de l'autonomie des agents artificiels sont nombreuses et flexibles (Noorman 2008). Noorman distingue deux conceptions de l'autonomie potentiellement utiles lorsqu'il est question des limites de l'agentivité artificielle : l'autonomie comprise comme autorégulation (*self-regulation*) et l'autonomie comprise comme délégation du contrôle.

La première, l'autonomie comprise comme autorégulation, peut être perçue comme l'objectif ultime de l'automatisation et renvoie au niveau de contrôle de la machine sur l'exécution d'un processus – soit, sa capacité à remplir les fonctions pour lesquelles elle a été créée sans qu'il y ait besoin de faire intervenir un humain (Noorman 2008). Définie ainsi, l'autonomie décrit une propriété observable et mesurable sans connotation morale ou normative (Noorman 2008). Cette conception ne réfère pas aux craintes qu'ont manifesté les citoyens, et on peut raisonnablement attendre d'un dispositif médical qu'il soit fiable et fonctionne de la manière attendue, soit de manière « autonome ». C'est ainsi plutôt l'autonomie dans sa deuxième conception, telle que présentée par Noorman, qui demande une attention particulière, soit celle comprise comme délégation du contrôle. Celle-ci revêt une dimension normative : elle place la technologie dans une relation de dépendance avec les humains au regard des décisions qui sont prises, et amène des questionnements relatifs à l'imputabilité et la responsabilité (soit, qui de l'humain ou de la machine est responsable, notamment en cas d'erreur) (Noorman 2008). C'est bien selon cette conception

que les craintes citoyennes se sont manifestées, et c'est en ce sens que l'agentivité artificielle demande à être limitée : elle doit inclure cette idée qu'il ne s'agit pas de limiter la réalisation, par les systèmes d'IA, de leur fonction de manière autonome (ce qui est d'ailleurs valable pour d'autres types de technologies en santé) mais bien de limiter la délégation du contrôle. La tension normative se situe plutôt dans le fait qu'ils puissent être tenus pour responsables des conséquences de leur décision, à la place des humains.

De plus, certains questionnent si les systèmes d'IA ont réellement la capacité de devenir des agents moraux et soulèvent de nombreuses limites à cette entreprise. Cette option se heurte à différents problèmes, le premier étant celui de faire en sorte que les algorithmes reconnaissent une situation chargée moralement, le second de déterminer sur quelle base définir ces critères moraux – ce qui revient aux difficultés de l'éthique appliquée soulevées dans le Chapitre 4 (Moor 2006; Scheutz 2016; Shulman, Jonsson, et Tarleton 2009). La tâche se complique également, pour certains, du fait du manque d'accès introspectif aux causes de nos intuitions morales : nous aurions, pour certains auteurs, une compréhension limitée de ce qu'est agir éthiquement (Moor 2006; Shulman, Jonsson, et Tarleton 2009). Certains soutiennent cependant que ces limitations pourraient rapidement être dépassées (ex. Davis 2015). Si les systèmes d'IA deviennent capables d'agir moralement, ils feraient cependant face à d'autres limites. En reconnaissant qu'ils sont basés sur des spéculations, Shulman *et al.* présentent quelques-uns des défis à venir dans la conception des agents artificiels si l'on suit cette idée : garantir la consistance de leurs objectifs (afin que les algorithmes ne se réécrivent pas entre eux); garantir qu'ils préfèrent toujours les mêmes fins que celles des humains (celles-ci pourraient référer aux valeurs universellement partagées, cf. Chapitre 4); qu'au fil des avancées les systèmes produisant des comportements bénins ne soient pas la cause de comportements catastrophiques; qu'une « explosion intelligente » ne permette pas aux agents artificiels d'agir librement peu importe les objectifs du système en place (Shulman, Jonsson, et Tarleton 2009).

Que l'on considère comme réelle ou non la capacité d'agir (moralement) des systèmes d'IA, le problème de la déresponsabilisation persiste car il réside dans le fait que certains leur accordent une forme d'autonomie qui s'approche de celle des humains. Cette distinction est importante

relativement à la réponse à apporter à cet éventuel problème de déresponsabilisation. Car, comme le mentionne Noorman (2008), que les agents artificiels atteignent un niveau d'autonomie similaire à celui des humains est en réalité un choix normatif :

> A levelling of humans and technologies in terms of their autonomy is therefore not an inevitable consequence of the development of increasingly intelligent autonomous technologies, but a result of normative choices (Noorman 2008, p. 65)

Ce choix normatif est donc bien celui des humains qui conçoivent, utilisent ou encadrent les systèmes d'IA – dans la mesure où ce choix est éclairé notamment relativement à la délégation du contrôle.

Il existe d'autres problèmes associés à l'extension des compétences algorithmiques à des compétences morales. L'opacité des réseaux de neurones en est un – comme l'ont d'ailleurs souligné les citoyens participants à la coconstruction. L'absence de transparence n'est pas un problème en soi, notamment lorsqu'il s'agit de décrire une propriété des dispositifs (comme souligné dans le Chapitre 3), mais devient problématique lorsqu'il est question de justifier la décision (au même titre que l'absence de transparence relative aux décisions humaines). Limiter la capacité décisionnelle permet d'éviter ici une déresponsabilisation mais également de renvoyer à l'absence de transparence des systèmes d'IA comme justification. La présence de biais potentiels dans les algorithmes demande également de conserver un esprit critique relativement aux décisions ou recommandations qui en émanent, et justifie encore une fois de limiter la reconnaissance de compétences morales aux algorithmes. Pour revenir aux aspects pragmatiques de la situation, il est alors important de limiter la portée des décisions algorithmiques, notamment car leurs règles de calculs sont procédurales et non substantielles (Cardon 2018). Cela signifie que les algorithmes n'ont pas d'accès sémantique aux informations qu'ils manipulent (c'est-à-dire qu'ils ne les « comprennent pas » d'un point de vue symbolique) (Cardon 2018) :

> Aussi, pour produire leurs résultats, doivent-ils [les algorithmes] trouver des *procédures* permettant de faire la meilleure approximation d'un principe que les utilisateurs vont interpréter de façon *substantielle* (p. 67).

C'est donc bien les humains qui utilisent les algorithmes qui prennent une décision.

Enfin, un autre problème réside dans le fait de placer les agents artificiels et les agents humains au même niveau ce qui, par extension, questionne les limites entre humains et machines et renvoient à des considérations plus profondes que la simple déresponsabilisation : l'agentivité artificielle questionne ce qui nous définit en tant qu'êtres humains et ce qui nous différencie des machines (Noorman 2008). C'est bien dans cette perspective que se sont manifestées les craintes citoyennes (cf. les discussions entourant la transformation du rapport à la technologie). En d'autres mots, les citoyens craignent que l'on considère les machines « comme des personnes », au risque de se déresponsabiliser et de déshumaniser le soin. Les projections d'intentions et la relative anthropomorphisation qui s'observe relativement aux systèmes d'IA est un élément essentiel de la perception de risque et du fait d'autoriser à reconnaître les systèmes d'IA comme responsables ; en particulier lorsqu'il est question du remplacement des humains par les machines. En effet, comme le reconnait Coeckelbergh (2015) dans le cadre de l'avènement de robots de soins qui remplacent les professionnels de santé : « this is not about the machine 'taking over'; it is about humans becoming machines » (p. 275). L'auteur dénonce ici la potentielle aliénation et la perception qu'auront les professionnels d'eux même et de leur expertise s'ils se voyaient massivement remplacés par des robots de soins (la pertinence de cette crainte est discutée dans la section 2.3 du présent chapitre).

Ainsi, si on en revient à l'innovation responsable selon une responsabilité pragmatique, il semble que la reconnaissance d'une agentivité artificielle constitue une interpellation responsabilisante (ce qui demande inévitablement de reconnaître l'existence d'une telle agentivité) qui demande de limiter cette agentivité relativement à la reconnaissance d'une capacité décisionnelle en ne déléguant pas le contrôle des décisions aux machines. Considérant qu'il s'agit d'une conception limitante de l'agentivité, reconnaitre une agentivité aux systèmes d'IA ne signifie pas forcément que l'humain perd la main, et ne revient pas à dire que les algorithmes ne jouent pas un rôle dans les actions morales. Si la reconnaissance d'une certaine agentivité aux systèmes d'IA peut amener à questionner certaines conceptions de la responsabilité en philosophie morale, il n'en est pas moins que, d'un point de vue pragmatique, les systèmes d'IA ne devraient jamais être considérés comme de potentiels *répondeurs* d'une interpellation responsabilisante, ces derniers

devant toujours être des humains. Cependant lesquels et de quelle manière ils peuvent y répondre renvoie aux deux autres tensions identifiées.

2.2. La tension entre responsabilité individuelle et collective

Les citoyens ont également manifesté des craintes relatives au grand nombre d'acteurs impliqués dans le parcours de soin, qui renvoie au problème des *mains multiples* : considérant ce grand nombre d'acteurs, il devient difficile de déterminer qui est responsable des conséquences d'une action et dans quelle mesure. Une tension sous-jacente au problème des *mains multiples* oppose la responsabilité individuelle de ceux qui participent (directement ou indirectement) au développement des systèmes d'IA et la responsabilité collective de certains groupes ou communautés (ex. scientifique, médicale) auxquels appartiennent ces individus. Pellé et Reber (2016) soulignent en effet en ce qui a trait à l'IRR que le problème des *mains multiples* est une des conséquences dommageables de la responsabilité négative, qui s'accompagne d'une dilution de la responsabilité. Trop d'emphase sur la responsabilité individuelle risquerait de conduire à la suresponsabilisation de certains individus plutôt que d'autres (Métayer 2001) ou de favoriser la déresponsabilisation des individus qui cherchent à éviter le blâme (Gotterbarn 2001).

Face au risque de déresponsabilisation en vue d'éviter le blâme, plusieurs défendent la nécessité de reconnaître une responsabilité collective (Dixon-Woods et Pronovost 2016) et/ou positive (Gotterbarn 2001). Cependant, trop d'emphase sur la responsabilité collective, au détriment de la responsabilité individuelle, expose au risque de dispersion de la responsabilité (Thompson 2004), et rejoint l'idée d'une irresponsabilité généralisée présentée par Beck (2001). En d'autres mots, trop d'emphase sur la responsabilité individuelle expose au risque que les acteurs, de plus en plus nombreux dans le parcours de soin, se déresponsabilisent en responsabilisant les autres acteurs de la chaine causale. Trop d'emphase sur la responsabilité collective expose au risque que plus personne ne se retrouve *in fine* responsable des conséquences (négatives ou positives) de l'utilisation des systèmes d'IA en santé.

Relativement à ce risque de déresponsabilisation, les craintes citoyennes font écho aux préoccupations de différents auteurs, lesquels ont souligné le problème des *mains multiples* dans le contexte de l'informatisation croissante de la société (Chartrand 2017; Noorman 2016; Nissenbaum 1994). Plusieurs auteurs ont souligné l'existence des phénomènes de suresponsabilisation ou de déresponsabilisation relativement à l'utilisation des systèmes d'IA en

santé. En effet, si certains mettent la responsabilité de l'innovation numérique responsable sur les épaules des développeurs – notamment par le biais d'une éthique *by design*, d'autres dénoncent le risque d'une surcharge morale de cette même communauté et demandent à reconnaître la responsabilité d'autres acteurs dans ces développements (IEEE 2017; Awad et al. 2018). Cette pression sur la responsabilité individuelle des développeurs est par exemple soulevée par le Center for Democracy and Technology :

> Principles established by academics, advocates, and policymakers are meant to demonstrate a philosophy that should be embedded throughout automated systems. This puts the burden on designers to understand how to integrate the goals of the principles into the technology itself (CDT 2017).

Plusieurs dénoncent également un phénomène de responsabilisation accrue des patients[134] (CCNE 2019; Coutellec et Weil-Dubuc 2017; Sharon 2017) qui peut s'observer, comme mentionné précédemment, par le biais des clauses de confidentialité ou conditions générales d'utilisation qui constitueraient des contrats de déresponsabilisation des entreprises qui les créent. Ces contrats tendent à responsabiliser les utilisateurs, à la place des entreprises, en tant que producteurs de données (Devillier 2017; Sharon 2016). Devillier (2017) dénonce dans cette situation les « pratiques abusives » liées aux objets connectés et aux applications mobiles de santé ; tandis que Sharon (2016) dénonce une pratique du partage des données « moralement douteuse », invalidant le consentement par là-même obtenu (qui n'est plus libre ni éclairé).

Alors que certains soulignent cette responsabilisation accrue des patients, d'autres comme Woolley dénoncent un « transfert du fardeau » des participants au comité d'éthique de la recherche dans la gestion des risques et enjeux éthiques associés aux données massives relatives à la santé. S'il existe des protections légales pour les utilisateurs concernant l'utilisation secondaire des données (le consentement étant le mécanisme habituel), la plupart des lois et agences de protection

[134] Ce phénomène dépasse l'avènement des systèmes d'IA en santé, soulevé par exemple par Rose (2003) relativement à la biomédicalisation de la société, qui tend à définir de plus en plus les problèmes en termes de santé et de maladie et à « pathologiser » les individus qui ne réussiraient pas à répondre aux exigences de vie saine et d'accomplissement personnel. Ce phénomène s'accompagne d'une certaine déresponsabilisation plus générale du système de santé en remettant les décisions et leurs conséquences sur les épaules des patients, comme dans le cas de la vaccination (Dubé et al. 2016).

des données qui s'appliquent aux *Big data* laissent la responsabilité de leur gestion éthique aux comités d'éthique, qui sont principalement centrés sur la protection individuelle des participants (boyd et Crawford 2012; Rumbold et Pierscionek 2017) et ne sauraient donc être tenus pour responsables de l'ensemble des conséquences positives ou négatives de l'utilisation des systèmes d'IA en santé.

Il est, enfin, possible de soulever des discours similaires relativement à la responsabilité scientifique. Il est parfois suggéré que les chercheurs ont une responsabilité plus grande par le fait qu'ils ont plus de pouvoir que d'autres (Johnson et Jameson 2008). Ce pouvoir vient, entre autres, de la connaissance hautement spécifique qu'ils détiennent, notamment car ils produisent les connaissances à la source de la menace (Beck 2001). Ce phénomène s'accompagne de l'avènement d'une certaine dépendance relative aux connaissances de ceux qui détiennent le savoir, conférant aux scientifiques un certain pouvoir face aux risques, soit une certaine responsabilité (Beck 2001). D'autres comme Murdock et Koepsell (2014) soulignent une asymétrie informationnelle qui divise le pouvoir entre chercheurs et décideurs : les premiers détiennent « la connaissance » et les seconds « la capacité décisionnelle ». On retrouve également dans la littérature sur le double-usage la reconnaissance d'une responsabilité scientifique relativement aux utilisations problématiques de la recherche, et ce malgré l'absence d'intention des chercheurs. Cette responsabilité correspond à une obligation morale de ne pas nuire, et donc de prévenir un mésusage raisonnablement prévisible dans la mesure des capacités et habilités du chercheur (Kuhlau et al. 2008). La responsabilité serait alors attribuée au chercheur par l'entremise d'une ignorance coupable, la faute étant de ne pas avoir cherché à réduire l'incertitude liée aux utilisations potentielles et les risques éventuels qui pourraient en découler (Kuhlau et al. 2011).

Cependant, la responsabilité qui incombe aux scientifiques mériterait, pour d'autres, d'être tempérée. En effet, il s'agirait d'une responsabilité de la communauté scientifique au sens large plutôt qu'une responsabilité individuelle du chercheur (Ehni 2008), et celle-ci suppose de participer à la prévention du dommage potentiel plutôt que d'y trouver une solution (Kuhlau et al. 2008). Dans le cadre de l'innovation responsable, Pellé et Reber défendent effectivement la pertinence de définir une responsabilité collective de la communauté scientifique, tant qu'il ne s'agirait pas de

considérer celle-ci comme étant « coupable » des conséquences néfastes de la recherche et de l'innovation ; mais plutôt comprise comme un engagement des professionnels de la recherche et de l'innovation relativement à leur communauté d'appartenance.

En réponse à la possible déresponsabilisation, Gotterbarn (2001) défend la nécessité d'envisager la responsabilité de manière positive, ce qui permettrait de reconnaître la responsabilité de plusieurs individus à des degrés différents, sans impliquer de lien causal proche ou direct entre un agent et les conséquences de ses actions :

> No matter which ethical theory is used to justify positive responsibility, the focus of positive responsibility is on what ought to be done rather than on blaming or punishing others for irresponsible behavior (p. 227).

Ceci ouvre alors la voie à la reconnaissance d'une responsabilité partagée, faisant écho aux attentes citoyennes, en vue de faire en sorte que tous soient parties prenantes de la gestion des risques et de l'innovation responsable (et non seulement concernés par un blâme potentiel). Les citoyens ont en effet soulevé, en réponse au problème des *mains multiples*, une certaine responsabilité partagée qui prend forme dans un contrat social. Considérant que l'innovation numérique semble être l'« affaire de tous » (comme souligné dans le Chapitre 4) il semble effectivement que le développement responsable des systèmes d'IA en santé se fait sur la base d'une responsabilité partagée. Cependant, la répartition des responsabilités selon Métayer demeure un idéal qui se heurte à l'équité du partage entre les différents acteurs. Cette responsabilité partagée demande ainsi d'identifier différents *répondeurs* qui, selon une responsabilité positive, participeraient à la gestion des risques et leur contrôle, selon un « horizon de causalité adéquat qui se retrouve entre responsabilité individuelle et collective » (Pellé et Reber 2016). S'il n'est pas possible ici de définir plus en détail le partage des responsabilités entre les différentes parties prenantes de l'utilisation des systèmes d'IA en santé, le problème des *mains multiples* éclaire sur deux éléments pertinents relativement à l'innovation responsable.

Premièrement, il semble ici qu'en responsabilisant les autres acteurs de la chaine causale, chacun devient à la fois *demandeur* et *répondeur* des interpellations responsabilisantes qui ont trait au développement des systèmes d'IA (en santé). L'interpellation responsabilisante demande de prêter attention à la fois au demandeur et au répondeur – car elle ne permet pas seulement de questionner la conduite de son destinataire mais aussi sert de véhicule aux convictions de son auteur

(Métayer 2001). Cet aspect demande de faire particulièrement attention aux asymétries de pouvoir qui pourraient exister entre chacun des acteurs (tous ne peuvent être reconnus responsables dans la même mesure) en particulier considérant le pouvoir que détiennent ceux qui deviennent propriétaires des données. Notamment, l'implication des secteurs à la fois public et privé dans ces développements, aux intérêts et aux normes différents complique la mise en place d'un encadrement éthique effectif en ce qui a trait à l'attribution des responsabilités – ce qui renvoie aux préoccupations relatives aux conflits d'intérêts potentiels soulevés par les citoyens. Le déséquilibre entre les pouvoirs demande d'adapter les mécanismes de gouvernance. Si les conflits d'intérêts financiers ne sont pas l'apanage des parties prenantes du développement des systèmes d'IA en ce qui a trait à la santé, Sharon (2016) qui dénonce une « Googlisation de la santé », soulève que les intérêts mercantiles des compagnies en jeu pourraient entrer en conflit avec les intérêts des patients :

> Entrance into the domain of health research secures companies access to a lucrative health market, with the promise of spin-offs down the line that will capitalize on newfound needs for health data generation and analysis. […] As DeepMind co-founder, Mustafa Suleyman has stated, 'Right now it is about building the tools and systems that are useful and once users are engaged then we can figure out how to monetize them' (Sharon 2016 p. 571).

En dehors des conflits d'intérêts financiers relatifs à l'implication d'acteurs privés (qui ne sont d'ailleurs pas nouveaux en santé), le manque d'intérêts communs entre les différentes parties prenantes de l'innovation numérique en santé (ex. scientifiques des données, médecins, patients) pourrait bien compliquer la mise en place de mesures de gouvernance effectives, notamment car il serait compliqué d'harmoniser ces intérêts.

Deuxièmement, il est à noter que les parties prenantes du développement des systèmes d'IA en santé dépassent les acteurs conventionnels du système de santé. On retrouve alors impliquées dans le parcours de soins des personnes extérieures au système de santé qui ne répondent pas des mêmes normes ni des mêmes codes de conduites. Par exemple, développeurs et scientifiques des données n'ont pas pour habitude de suivre les lignes directrices propres à la recherche avec des participants humains ou les codes de déontologie relatifs à la pratique médicale. La participation accrue de personnes extérieures demande ainsi, selon une vision pragmatique de la responsabilité, de ne pas se limiter aux acteurs de la recherche et de l'innovation ou du système de santé comme

seuls répondeurs des interpellations responsabilisantes qui émanent de l'utilisation des systèmes d'IA.

Ainsi, devant le problème des *mains multiples* se dessine la nécessité de reconnaître une responsabilité à la fois individuelle et collective, afin de tempérer les risques de déresponsabilisation et de suresponsabilisation qui émanent de l'utilisation des systèmes d'IA en santé. Ces deux conceptions ne sont pas antinomiques et semblent appropriées à la fois du point de vue de l'IRR (comme le soutiennent Pellé et Reber) mais également du point de vue des systèmes de santé, où chaque partie prenante peut être tenue responsable dans le cadre spécifique de ses fonctions, toujours en lien cependant avec les fonctions des autres acteurs (CSBE 2005).

2.3. La tension entre technologies capacitantes et incapacitantes

Les citoyens ont enfin manifesté des craintes relatives à une certaine incapacitation des patients et des professionnels de santé relativement à leur utilisation des systèmes d'IA. Ces craintes se sont manifestées en lien avec une atteinte à la liberté de choix ou à l'exercice de l'esprit critique des patients et une incapacitation des professionnels de santé par l'entremise d'une perte de compétences et d'une confiance en eux-mêmes ou d'une dépendance à la technologie. Les citoyens ont soulevé des craintes relatives au remplacement des humains par les machines, liées à la fois à la reconnaissance d'une agentivité artificielle (déjà discutée précédemment) et à la préservation des capabilités humaines.

Relativement au remplacement des professionnels de santé par les machines, certains auteurs mentionnent que si cette crainte s'est avérée dans certains secteurs, elle l'est moins en ce qui concerne les professions de la santé, qui requiert un niveau d'expertise élevé et dont la pratique nécessite de nombreuses tâches non-automatisables (Schwab 2016). Le risque de remplacement des professionnels de santé par les machines demande également de se pencher sur le type de systèmes d'IA en jeu. Par exemple, le remplacement du personnel médical semble plus à même d'arriver avec les robots de soin ou les applications mobiles que via les systèmes experts, soit plus à même d'arriver avec les dispositifs permettant un contact direct avec les patients. Également, cette crainte est sous-tendue par l'idée que les systèmes d'IA deviendraient plus performants que

les professionnels de santé pour réaliser certaines tâches. Or, comme le présente Oakden-Rayner, cette supposition se base généralement sur les résultats de deux études (Gulshan et al. 2016; Esteva et al. 2017) où les machines ont démontré des performances supérieures à celles des professionnels de santé relativement à la reconnaissance visuelle, mais seulement selon certains métriques d'évaluation (Oakden-Rayner 2017). Également, il est possible d'observer un certain paradoxe dans l'automatisation des actions humaines, connu sous le nom de paradoxe de Moravec (Rotenberg 2013). Le paradoxe de Moravec souligne qu'il est bien plus simple d'informatiser des tâches qui relèvent du raisonnement de haut niveau (comme jouer aux échecs) que celles qui relèvent de fonctions cognitives plus basiques comme la motricité ou la perception (Rotenberg 2013). Suivant ce paradoxe, on peut raisonnablement questionner dans quelle mesure les professionnels de santé pourront être intégralement remplacés dans un avenir proche – considérant que leur pratique requiert à la fois un raisonnement de haut niveau et des fonctions cognitives plus basiques. Également, ce qui relève des fonctions de « haut niveau » dans le paradoxe de Moravec ne semble pas englober toutes les fonctions que l'on peut considérer comme automatisables (ex. relationnelles et émotionnelles), nécessaires afin de dispenser un soin de qualité (Coeckelbergh 2015) mais également essentielles à la pratique d'une éthique inductive sensible au contexte.

Ces considérations sont d'autant plus importantes qu'il ne semble pas que l'objectif du développement des systèmes d'IA en santé relève explicitement du remplacement des humains par les machines. Par exemple, comme le mentionne Devillers (2017) :

> En robotique sociale, il ne s'agit pas de remplacer des humains par des machines, ni de ne s'entourer que de machines, mais de profiter de la grande différence entre les organismes biologiques et les machines électroniques pour leur faire faire ce en quoi elles excellent : des additions ou encore des recherches dans des millions de données (p. 117).

Ainsi, l'attente citoyenne relative à une coopération humain-machine fait écho à ce que les experts du domaine qualifient de « cobotique », qui se penche sur la manière dont les humains peuvent tirer profit des compétences des agents artificiels (Claverie, Blanc, et Fouillat 2013; Kleinpeter 2015), ces derniers demeurant ainsi des outils d'aide, entre autres à la décision. Il n'en est pas moins qu'en automatisant certains tâches, basiques ou non, et en ayant recours à des applications mobiles ou des robots de soin, le risque réduction du contact humain dans la relation de soins qui en découlerait est soulignée par plusieurs auteurs (Coeckelbergh 2012; Devillers 2017). Ainsi, bien

qu'il soit nécessaire de tempérer cette crainte, le remplacement des humains par les machines demeure une interpellation responsabilisante qui doit être prise en considération relativement à ses conséquences sur la relation médecin-patient et à l'impact sur la perception que les professionnels de santé pourraient avoir d'eux-mêmes (Coeckelbergh 2015; Sharon 2017).

Que les systèmes d'IA remplacent ou non les humains, le développement d'une « coopération humains-machines » demande de se pencher sur la possibilité que les systèmes d'IA incapacitent les patients ou les professionnels de santé, comme cela a été soulevé par les citoyens. Concernant l'incapacitation des patients, les craintes (atteinte au consentement éclairé et à la liberté de choix) et les attentes (préserver l'autonomie décisionnelle) se retrouvent toutes dans les considérations relatives au respect du consentement libre et éclairé largement discutées dans le Chapitre 3. Préserver les capabilités des patients s'avère ici essentiel considérant que la conception de l'agentivité qui sous-tend cette approche ne demande pas simplement de préserver la capacité d'agir mais également de le faire selon les objectifs que les individus valorisent (Oosterlaken 2015), donnant alors à l'agentivité une dimension normative importante qui renvoie à la conception que chacun pourrait avoir de la vie bonne.

Concernant les professionnels de santé, ces craintes font également écho à celles soulevées par différents auteurs. Par exemple, Jameson et Longo (2015) soulignent qu'un des grands défis de l'avènement de la médecine de précision est celui de gérer la complexité des algorithmes en jeu, alors même que les professionnels de santé se sentent déjà inadéquatement outillés face aux développements en génétique (Jameson et Longo 2015). Ils demandent ainsi de prêter une attention particulière au fait que la technologie dépasse actuellement le « *common clinician understanding* » (Jameson et Longo 2015). Dans une perspective plus large, cette incapacitation peut également provenir du fait de l'existence des nouvelles relations de pouvoir qui créent un déséquilibre au niveau de l'expertise des chercheurs et des professionnels de santé avec les acteurs des firmes internationales, amenant Sharon (2016) à questionner qui seront les experts de demain dans un monde où les GAFAM détiennent un monopole relativement à la propriété des données mais également relativement aux moyens de les traiter.

Les citoyens ont alors manifesté des attentes relatives à la capacitation des patients et des professionnels de santé, lesquelles renvoient notamment à la nécessité de préserver leur autonomie décisionnelle et de mettre en place différentes formations et mécanismes d'éducation sur l'IA et sur l'éthique de l'IA, ce qui est en adéquation avec plusieurs des recommandations de politiques publiques présentées dans le Chapitre 4 (ex. CERNA 2018; Déclaration de Montréal IA Responsable 2018; OSTP 2016). Ils n'ont cependant pas abordé, dans le cadre des discussions analysées pour les fins de la présente thèse, la nécessité de capaciter également des concepteurs de systèmes d'IA relativement à l'éthique. Cette nécessité a cependant été soulevée lors d'autres discussions de la coconstruction de la Déclaration de Montréal (Déclaration de Montréal IA Responsable 2018) et s'avère pertinente si l'on considère le lien intrinsèque entre les limites techniques des systèmes d'IA et leurs conséquences éthiques et sociales (mis en évidence dans le Chapitre 3). Considérant que la pratique scientifique et technique ont été longtemps considérées comme moralement neutres[135], cette capacitation et sa mise en pratique ne sont cependant pas sans défi. Elle s'avère toutefois essentielle si l'on considère le développement d'une éthique en pratique, comme soutenu par exemple par Borrett, Sampson, et Cavoukian (2017) relativement au développement d'une éthique de la recherche *by design* qui devrait relever d'une culture de l'éthique plutôt que de mécanismes de surveillance éthique. Concernant le secteur de la santé, l'approche des capabilités est pertinente pour adresser les enjeux de l'utilisation de technologies, qu'il s'agisse ou non de systèmes d'IA (Coeckelbergh 2012; 2010; Oosterlaken et van den Hoven 2012). Cette approche permet par exemple d'évaluer l'impact des technologies sur le soin, sur les patients ou la pertinence de la conception des systèmes d'IA (Oosterlaken et van den Hoven 2012).

Lorsqu'il est question de l'agentivité humaine, on notera que les craintes et attentes renvoient à une conception plus large de l'agentivité que lorsqu'il est question des agents artificiels, celle-ci incluant notamment l'agir libre et intentionnel. Cette conception de l'agentivité demande de tenir compte de différents aspects éthiques inhérents à la capacité d'agir des individus. Plus

[135] Cette considération est en partie issue des impératifs d'objectivité scientifique qui demandent que l'objet de la connaissance soit universellement accessible, indépendamment de la position du sujet connaissant (Keating 2015) mais renvoie également à la reconnaissance que la construction de systèmes socio-techniques va plus loin que la conception de dispositifs neutres que la société peut choisir ou non d'utiliser (Johnson et Jameson 2008).

précisément, selon l'approche des capabilités, il est possible de valoriser l'agentivité de trois manières (Crocker et Robeyns 2010 dans Oosterlaken 2015) :

> 1) selon sa valeur instrumentale, prendre en compte la perspective des individus augmentant les chances de succès de l'implémentation technologique ;
>
> 2) selon sa valeur constructive, soit la valorisation du processus qui amène à l'exercice de l'agentivité selon lequel une personne décide et façonne ses valeurs; et
>
> 3) selon sa valeur intrinsèque, l'agentivité ayant une valeur en elle-même et n'étant pas seulement un moyen d'atteindre une fin.

Ainsi, assurer que le développement des systèmes d'IA garantisse de préserver les capabilités humaines va dans le sens d'une innovation responsable relativement à plusieurs aspects essentiels. D'abord, comme présenté dans le Chapitre 1, l'IRR veut assurer l'acceptabilité sociale des avancées de la science et du développement technologique (selon une certaine « désirabilité sociale » (Barré 2011) en favorisant le développement d'une science *avec* la société (Owen, Macnaghten, et Stilgoe 2012). Garantir cette désirabilité renvoie à la première manière de valoriser l'agentivité : préserver l'agentivité humaine dans le développement technologique semble en accord avec l'idée d'un développement technologique « réussi ».

Deuxièmement, en valorisant le processus qui amène à l'exercice de l'agentivité (soit, la valeur constructive de cette dernière), l'approche des capabilités demande de prêter une attention particulière au contexte d'implémentation des technologies en question qui dépasse l'évaluation de caractéristiques intrinsèques à la technologie en elle-même :

> Throughout history, technology has been a powerful tool for development. The wheel allowed us – for example – to transport heavy loads and, more recently, mobile phones enabled us to communicate from any place in the world. Technology is also used for poverty reduction in many different ways – from water supply or electrification to developing long-distance education or telemedicine. However, over the past decades numerous technology-oriented development projects have failed. Little attention was given to processes of technological change, thus leaving out important issues such as participation or empowerment of people. These examples show that technology, despite being important, is not the only factor that ensures the success of a technological intervention. (Fernández-Baldor, Hueso, et Boni 2012 p. 135)

Ainsi, le développement technologique « approprié » en appelle à différents mécanismes participatifs et délibératifs afin de favoriser les capabilités humaines et trouver des consensus relativement à l'écart entre les intérêts individuels et collectifs (Oosterlaken et van den Hoven 2012), ce qui semble pertinent dans le contexte du développement des systèmes d'IA qui soulève de nombreux enjeux éthiques pour lesquels les réponses ne sauraient être généralisées. Ceci fait écho aux impératifs de l'IRR qui demandent, de plus en plus, de favoriser l'implication de la société civile dans les différents processus de l'innovation et de la recherche (Lehoux et al. 2018; Barré 2011; Owen, Macnaghten, et Stilgoe 2012) mais également à certaines des attentes citoyennes relativement à la gouvernance participative (*cf*. Section « Utilisateurs » dans 2.3.3. du Chapitre 5). Cet aspect est également essentiel en vue de préserver la composante « normative » des capabilités – celle qui renvoie à la vie bonne selon la conception que s'en font les individus et de ce qu'ils valorisent.

De plus, préserver l'agentivité humaine est important considérant sa valeur intrinsèque. Ceci est soutenu par les initiatives qui érigent des dimensions associées au concept de capabilité en principes directeurs. C'est par exemple le cas du IEEE qui promeut un « principe de compétence », soit de demander aux concepteurs de systèmes d'IA d'adhérer aux connaissances et aux expertises requises pour assurer que le fonctionnement de ces systèmes soit sûr et efficace (IEEE 2019). Les principes d'autonomie, de liberté, de réflexivité ou de participation démocratique (CNIL 2017; Déclaration de Montréal IA Responsable 2018; AI HLEG 2019; JSAI 2017) renvoient tous à des dimensions associées à la préservation des capabilités.

Enfin, du point de vue de l'innovation responsable, préserver les capabilités est également important parce-que la responsabilité peut elle-même être conçue comme une capacité[136] : une incapacitation des individus selon l'impact potentiel des systèmes d'IA sur les différentes

[136] Il est à noter que la notion de capabilité inclut, mais ne se limite pas, à la notion de capacité. Affirmer que la responsabilité constitue une capabilité demanderait une analyse plus approfondie. Par exemple, Ballet et Mahieu (2009) défendent que la notion d'agentivité dans les travaux de Sen renvoie à une conception de la responsabilité sociale liée à la notion de liberté individuelle (comprise comme liberté de choix de vie) mais qui écarte, selon les auteurs, l'interdépendance entre bien-être et responsabilité. Pour les fins de cette thèse, le lien inhérent entre capabilité et agentivité, et agentivité et responsabilité, suffit à défendre l'intérêt de considérer l'approche des capabilités en vue d'une innovation responsable.

dimensions qui relèvent des capabilités pourrait alors directement nuire à la capacité d'être responsable. En effet, comme présenté par Pellé et Reber (2016), la responsabilité peut, selon certaine des conceptions qui relèvent plutôt du courant « positif », renvoyer à la capacité d'agir de manière moralement appropriée (plutôt qu'une obligation d'agir d'une certaine manière) :

> L'individu responsable possède des compétences cognitives pour anticiper, interroger, évaluer les conséquences possibles de ses actions. Il peut reconnaître, discerner et ajuster de qu'il a à faire, dans un contexte particulier (p. 122).

Cette responsabilité est à la fois prospective et rétrospective, puisqu'elle réfère au savoir des individus et à leurs compétences normatives (Pellé et Reber 2016) ce qui renvoie à la vision pragmatique de la responsabilité de Métayer qui se base sur une de ses conceptions les plus ancienne, à savoir la capacité de réponse. La capacité ne représente plus alors une condition de la responsabilité mais une forme de responsabilité qui « désigne une aptitude particulière (individuelle, collective, institutionnelle ou systémique) à répondre de façon adéquate aux problèmes que soulève une situation » (Pellé et Reber 2016 p. 125).

Or, si l'on considère que le développement des systèmes d'IA fait face au problème des *mains multiples*, il est possible de souligner que ces acteurs ont des expertises variées et hautement spécifiques. C'est le cas des professionnels de santé, dont l'expertise est tellement spécialisée et multidisciplinaire que peu d'organisations ont les ressources ou les compétences pour concevoir des interventions et des évaluations globales afin de mitiger les risques (Dixon-Woods et Pronovost 2016). Il en est de même pour les scientifiques des données dont les spécialisations, expertises et compétences sont tellement variées qu'il est difficile de définir le rôle de chacun (Granville 2017). De s'assurer que chacun ait la capacité de répondre de sa responsabilité demande de prêter une attention particulière à cet aspect. En effet, les théories de Jonas (1979) soulignent que le prix interne du progrès scientifique est la spécialisation : la multiplication des matériaux du savoir, de ses subdivisions et de ses méthodes amène à une fragmentation de la connaissance disponible (Jonas 1979). Ainsi, le capital total du savoir augmente mais celui de l'individu (chercheur) devient fragmentaire. Ce savoir devient alors toujours plus ésotérique et toujours moins communicable, se voyant ainsi réservé à une élite (Jonas 1979), ce qui pourrait conduire à défier la possibilité d'une éthique en pratique effective, puisqu'il est difficile pour un seul et même acteur de tenir compte de la multi dimensionnalité des situations.

Dans ce contexte, l'interpellation responsabilisante issue de la préservation des capabilités humaines demande à minima de ne pas incapaciter les différents acteurs du développement des systèmes d'IA en santé, voire de les capaciter si l'on suit le mouvement des technologies appropriées. Il ne s'agit pas seulement de préserver l'agentivité (individuelle ou collective) des utilisateurs ou groupes d'utilisateurs dans les contextes concrets d'application (Oosterlaken 2015) mais bien de considérer l'ensemble des potentiels *demandeurs* et *répondeurs* du développement des systèmes d'IA en santé, qui ont des expertises variées ce qui implique de s'assurer que soient menées des interventions différentes en vue de préserver leurs capabilités. Préserver les capabilités est ainsi autant une interpellation responsabilisante qu'une nécessité en vue d'assurer que *demandeurs* et *répondeurs* remplissent adéquatement leur fonction normative. L'approche des capabilités demande alors de tenir compte à la fois du contexte et des préférences individuelles ou collectivement partagées, ce qui encourage la mise en place de processus délibératifs et inclusifs en vue d'une innovation numérique en santé responsable.

3. Quelques pistes relativement aux mécanismes à adapter

Avant de conclure, cette thèse finira sur quelques pistes qui démontrent que l'idée d'adapter les mécanismes existants, dont la pertinence a été mise en évidence dans le chapitre 4 et soutenue par les discussions citoyennes, est potentiellement plus prometteuse que la création de nouveaux mécanismes, notamment s'ils tiennent compte une fois adaptés des trois tensions précédemment présentées. Il est en effet soutenu dans le Chapitre 4 qu'il est nécessaire d'adapter la gouvernance éthique en santé aux aspects disruptifs de l'avènement des systèmes d'IA. C'est également ce qu'ont soutenu, voire proposé, les citoyens participants à la coconstruction de la Déclaration de Montréal. Les citoyens ont en effet formulé des recommandations concrètes relativement aux dispositifs à mettre en place, identifiés lors de l'analyse initiale des données de la Déclaration de Montréal mais également au fil de l'analyse des discussions présentées dans le Chapitre 5. Ces recommandations sont présentées en détail dans le rapport qui émane de l'analyse, p. 146 (Déclaration de Montréal IA Responsable 2018). Le Tableau 8 reprend les principales catégories de recommandations formulées dans le rapport de la Déclaration de Montréal et précise pour chacune d'entre elles les mécanismes (principalement canadiens et québécois) déjà en place, tel qu'identifiés par les citoyens. D'autres exemples ont été ajoutés en vue d'enrichir cette perspective. Cette liste ne saurait cependant être exhaustive et offre simplement des pistes de réflexion

relativement aux mécanismes à adapter. Elle vise avant tout à mettre en valeur la parole des citoyens ayant participé à la coconstruction plutôt que d'identifier l'ensemble des dispositifs de gouvernance existants.

Tableau 8. – Les mécanismes existants identifiés en vue de répondre au développement responsable de l'utilisation des systèmes d'IA en santé.

Propositions citoyennes pour le secteur de la santé (catégories issues des travaux de l'équipe de la Déclaration de Montréal)	**Mécanismes identifiés par les citoyens lors des discussions analysées dans la présente thèse**	**Exemples d'autres mécanismes existants**
Dispositions légales et juridiques	• Charte des droits et liberté • « Loi d'accès universel » • RGPD	• LPRP (Canada, gouvernement) • LPRPDE (Canada, entreprises)
Acteurs institutionnels et autres acteurs	• Ministère de la santé • Santé Canada • Ombudsman • Commissariat à la protection de la vie privée • Commission d'accès à l'information • INESSS	• Comités d'éthique de la recherche • Comités d'éthique clinique
Dispositifs d'évaluation de l'IA	• Approbation de la FDA (États-Unis) • Approbation de Santé Canada (Canada)	• Commission électrotechnique internationale • Normes ISO
Formations	• Formation des professionnels de santé • Formation initiale des patients	• Plan d'action numérique (Québec)
Codes de déontologie, d'éthique, ou de conduite	• ÉPTC • Serment d'Hippocrate[137] • Code de déontologie des médecins	
Mécanismes participatifs	• INESSS	• OBVIA

Les citoyens ont ainsi recommandé de mettre en place des « dispositions légales et juridiques » afin d'encadrer le développement des systèmes d'IA. On notera qu'ils ont relevé l'existence du RGPD européen en ce qui a trait à la gestion des données personnelles. Au Canada, ces aspects sont gérés sous l'égide de la Loi sur la protection des renseignements personnels (LPRP), qui régit le traitement des données par le gouvernement et la Loi sur la protection des renseignements personnels et les documents électroniques (LPRPDE), qui régit pour sa part le traitement qui en est fait par les entreprises[138]. Ces lois concernent principalement la gestion des données et ne répondent donc pas à tous les enjeux soulevés par l'utilisation des systèmes d'IA en

[137] Issu du rapport de la Déclaration de Montréal : certains citoyens ont proposé la création d'une « serment d'Hippocrate 2.0 ».
[138] Voir : www.priv.gc.ca/fr

santé. Les citoyens ont également souligné la potentielle nécessité d'adapter la Charte des droits et libertés ou la « Loi d'accès universel »[139].

Les citoyens ont également recommandé la mise en place d'acteurs institutionnels (ou autres acteurs) qui seraient les « garde-fous » du développement responsable de l'IA (Déclaration de Montréal IA Responsable 2018). Lors des discussions analysées dans le cadre de la présente thèse, de nombreux acteurs institutionnels existants ont été identifiés, comme l'Institut national d'excellence en santé et services sociaux (INESSS); le Ministère de la Santé ou Santé Canada. Les comités d'éthiques de la recherche et les comités d'éthiques cliniques déjà en place auraient également un rôle à jouer relativement à l'innovation numérique en santé.

Les citoyens ont aussi proposé de mettre en place des dispositifs d'évaluation de l'IA (principalement des certifications ou normes) (Déclaration de Montréal IA Responsable 2018). Lors des discussions analysées dans cette thèse, les citoyens ont identifié le rôle potentiel de Santé Canada dans l'approbation des systèmes d'IA s'il advenait qu'ils soient reconnus comme des dispositifs médicaux. Il est intéressant de souligner que des organismes comme ISO ou la Commission électrotechnique internationale font déjà référence en matière de normalisation.

Ils ont aussi soutenu la nécessité de mettre en place ou d'adapter les codes de déontologie et lignes directrices telles que l'Énoncé de politique des Trois Conseils (ÉPTC), qui a été récemment mis à jour, ou le code de déontologie des médecins. Ces recommandations vont dans le sens du travail réalisé par le Comité de travail sur les aspects d'éthique de la recherche dans les domaines du numérique, de l'intelligence artificielle et des données massives des Fonds de recherche du Québec (ci-après le Comité), qui souligne certaines des notions fondamentales de l'ÉPTC qui mérite une attention particulière en vue de leur application à l'heure de l'innovation numérique en santé. Pour ne citer qu'elles : la notion de **participant à la recherche** considérant notamment la dématérialisation de la participation et une nouvelle relation au temps avec les possibilités croissantes en termes de réutilisation des données (*cf.* la notion de « portabilité » dans les Chapitre 2 et 3); **la protection de la vie privée et de la confidentialité** considérant que l'anonymisation ne semble plus suffisante (*cf.* Section 2.1. du Chapitre 3), ou encore la gestion du **droit de se retirer en tout temps** à l'heure où les données circulent massivement et la nécessité

[139] Il est probable ici que les citoyens aient fait référence au principe d'accessibilité et d'universalité de la Loi canadienne sur la santé (voir : https://laws-lois.justice.gc.ca/fra/lois/C-6/)

de partage (*cf.* Section 1.2. du Chapitre 3)[140]. Également, les défis mis en évidence dans le Chapitre 5 demandent de reconsidérer les conditions de **l'utilisation secondaire des données** (celle-ci étant facilitée par les technologies qui relèvent de l'IA et la disponibilité de données massives, lesquelles défient l'application du **consentement libre, éclairé et continu** considérant le défi de *préserver les capabilités* mais également le **respect de la confidentialité et de la vie privée** dans sa conception traditionnelle comme mis en évidence dans la section relative au *problème des mains multiples*).

Enfin, les citoyens ont soutenu la nécessité de mettre en place des formations pour tous; en incluant notamment des dimensions pertinentes relatives à l'IA et à l'éthique dans le cursus scolaire. Concernant la formation, le gouvernement du Québec a mis en place un Plan d'action numérique en éducation et en enseignement, qui regroupe 33 mesures afin de soutenir les compétences numériques des québécois[141]. Les citoyens ont soutenu la nécessité de mettre en place des mécanismes délibératifs afin d'assurer un développement responsable des systèmes d'IA. Ils ont souligné le rôle de l'INESSS, dont c'est en partie la mission. L'Observatoire international sur les impacts sociétaux de l'IA et du numérique (OBVIA) récemment lancé au Québec a également pour objectif d'assurer une fonction délibérative[142].

Il est à noter que les citoyens ont également recommandé de mettre en place des outils numériques comme un consentement digital (cette dernière proposition est actuellement largement discutée (Jones, Kaufman, et Edenberg 2018; Schmietow 2016; Villani 2018; Woolley 2016; Christen et al. 2016). Ils ont également soutenu la nécessité de soutenir la recherche interdisciplinaire et intersectorielle, ce qui va dans le sens des initiatives canadiennes comme celle de l'Institut canadien de recherches avancées (CIFAR)[143].

Ainsi, en vue d'assurer une innovation numérique en santé responsable, il semble essentiel de soutenir et d'encourager les structures existantes face aux différents enjeux mis en évidence

---

[140] Voir la présentation de Deschênes, Mylène lors de la 9ème Journée d'étude des comités d'éthique de la recherche, le 24 octobre 2019 à Québec : « Le 'participant numérique' à l'heure de la recherche en intelligence artificielle et du numérique: revisitons nos bases en éthique de la recherche » disponible ici : https://www.msss.gouv.qc.ca/professionnels/documents/comites-d-ethique-de-la-recherche/journees-detude/2019-9e-journees/13h30-24oct_B1_MDeschenesMHirlte_JECER2019.pdf

[141] Voir : http://www.education.gouv.qc.ca/dossiers-thematiques/plan-daction-numerique/plan-daction-numerique/

[142] Voir : https://observatoire-ia.ulaval.ca/bienvenue/

[143] Voir : https://www.cifar.ca/fr/ia/strategie-pancanadienne-en-matiere-dintelligence-artificielle

dans la présente thèse. Dans cette perspective, il est également essentiel d'adapter les mécanismes en place, notamment en élargissant leur portée à des acteurs inusités du système de santé (ex. scientifiques des données), d'encourager les développements allant dans le sens de la prise en considération de ce que les citoyens sont effectivement capables de faire et d'être selon leur perspective, et de limiter la portée décisionnelle des systèmes d'IA.

Références bibliographiques

AI HLEG, (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.

Awad, Edmond, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, et Iyad Rahwan. 2018. « The Moral Machine Experiment ». *Nature* 563 (7729): 59. https://doi.org/10.1038/s41586-018-0637-6.

Ballet, Jérôme, et François-Régis Mahieu. 2009. « Capabilité et capacité dans le développement : repenser la question du sujet dans l'œuvre d'amartya sen ». *Revue Tiers Monde* n° 198 (2): 303-16.

Barré, Rémi. 2011. « Des concepts à la pratique de l'innovation responsable : à propos d'un séminaire franco-britannique ». *Natures Sciences Societes* Vol. 19 (4): 405-9.

Beck, Ulrich. 2001. *La société du risque. Sur la voie d'une autre modernité. trad. de l'allemand par L. Bernardi*. Paris, Aubier.

Beran, Ondřej. 2018. « An Attitude Towards an Artificial Soul? Responses to the "Nazi Chatbot" ». *Philosophical Investigations* 41 (1): 42-69. https://doi.org/10.1111/phin.12173.

Borrett, Donald S, Heather Sampson, et Ann Cavoukian. 2017. « Research Ethics by Design: A Collaborative Research Design Proposal ». *Research Ethics* 13 (2): 84-91. https://doi.org/10.1177/1747016116673135.

boyd, Danah, et Kate Crawford. 2012. « Critical Questions For Big Data Provocations for a Cultural, Technological, and Scholarly Phenomenon ». *Information Communication & Society* 15 (5): 662-79. https://doi.org/10.1080/1369118X.2012.678878.

Brundage, Miles. 2016. « Artifical intelligence and responsible innovation ». Dans *Fundamental issues of artificial intelligence*, Springer International Publishing Switzerland, 543-53. V.C. Müller.

Bucher, Taina. 2016. « Neither Black Nor Box: Ways of Knowing Algorithms ». Dans *Innovative Methods in Media and Communication Research*, édité par Sebastian Kubitschko et Anne Kaun, 81-98. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-40700-5_5.

Cardon, Dominique. 2018. « Le pouvoir des algorithmes ». *Pouvoirs* N° 164 (1): 63-73.

Castelvecchi, Davide. 2016. « Can We Open the Black Box of AI? » Nature News 538 (7623): 20. https://doi.org/10.1038/538020a.

CCNE. 2019. « Données massives (big data) et santé : une nouvelle approche des enjeux éthiques ». Avis 130. Comité Consultatif National d'Éthique français. https://www.ccne-ethique.fr/sites/default/files/avis_130.pdf.

CDT, (Center for Democracy and Technology). 2017. « Digital Decisions ». *Center for Democracy & Technology* (blog). 2017. https://cdt.org/issue/privacy-data/digital-decisions/.

CERNA. 2018. « La souveraineté à l'ère du numérique. Rester maîtres de nos choix et de nos valeurs ». http://cerna-ethics-allistene.org/digitalAssets/55/55708_AvisSouverainete-CERNA-2018.pdf.

Chartrand, Louis. 2017. « Agencéité et responsabilité des agents artificiels ». *Éthique publique. Revue internationale d'éthique sociétale et gouvernementale*, nº vol. 19, n° 2 (novembre). https://doi.org/10.4000/ethiquepublique.3068.

Christen, Markus, Josep Domingo-Ferrer, Bogdan Draganski, Tade Spranger, et Henrik Walter. 2016. « On the Compatibility of Big Data Driven Research and Informed Consent: The Example of the Human Brain Project ». Dans *The Ethics of Biomedical Big Data*, édité par Brent Daniel Mittelstadt et Luciano Floridi, 199-218. Law, Governance and Technology Series. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-33525-4_9.

Claverie, Bernard, Benoît Le Blanc, et Pascal Fouillat. 2013. « La cobotique. La robotique soumise ». *Communication et organisation*, nº 44 (décembre): 203-14. https://doi.org/10.4000/communicationorganisation.4425.

Cleret de Langavant, Ghislaine. 2001. *Bioéthique : Méthode et complexité*. Presses de l'Université du Québec. Québec.

CNIL (Commission nationale informatique et libertés). 2017. « Comment permettre à l'homme de garder la main ? Les enjeux éthiques des algorithmes et de l'intelligence artificielle ».

Coeckelbergh, Mark. 2010. « Health Care, Capabilities, and AI Assistive Technologies ». *Ethical Theory and Moral Practice* 13 (2): 181-90. https://doi.org/10.1007/s10677-009-9186-2.

———. 2012. « "How I Learned to Love the Robot": Capabilities, Information Technologies, and Elderly Care ». Dans *The Capability Approach, Technology and Design*, édité par Ilse

Oosterlaken et Jeroen van den Hoven, 77-86. Philosophy of Engineering and Technology. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-3879-9_5.

———. 2015. « Artificial Agents, Good Care, and Modernity ». *Theoretical Medicine and Bioethics* 36 (4): 265-77. https://doi.org/10.1007/s11017-015-9331-y.

Coutellec, Léo, et Paul-Loup Weil-Dubuc. 2017. « Chapitre 7. Big data ou l'illusion d'une synthèse par agrégation. Une critique épistémologique, éthique et politique ». *Journal international de bioethique et d'ethique des sciences* Vol. 28 (3): 63-79.

Crnkovic, Gordana Dodig, et Daniel Persson. 2008. « Sharing Moral Responsibility with Robots: A Pragmatic Approach. » Dans Frontiers in Artificial Intelligence and Applications Volume 173, édité par Anders Holst, Per Kreuger, et Peter Funk. IOS Press Books.

CSBE, Conseil de la santé et du bien-être. 2005. « Fiches de références pour l'élaboration d'un avant-projet de Déclaration des droits et des responsabilités en matière de santé et de bien-être ». Québec. https://www.csbe.gouv.qc.ca/fileadmin/www/Archives/ConseilSanteBienEtre/Rapports/200509_fiches_reference_declaration.pdf.

Davis, Ernest. 2015. « Ethical Guidelines for a Superintelligence ». *Artificial Intelligence* 220: 121-24. https://doi.org/10.1016/j.artint.2014.12.003.

Déclaration de Montréal IA Responsable. 2018. « Rapport de la Déclaration de Montréal pour un Développement Responsable de l'Intelligence Artificielle ». https://docs.wixstatic.com/ugd/ebc3a3_d806f109c4104c91a2e719a7bef77ce6.pdf.

Devillers, Laurence. 2017. *Des robots et des hommes: Mythes, fantasmes et réalité*. Plon.

Devillier, Nathalie. 2017. « Chapitre 5. Santé et Big data : l'émergence d'un droit d'infrastructure dans l'espace numérique ». *Journal international de bioethique et d'ethique des sciences* Vol. 28 (3): 51-56.

Dixon-Woods, Mary, et Peter J. Pronovost. 2016. « Patient Safety and the Problem of Many Hands ». *BMJ Qual Saf*, février, bmjqs-2016-005232. https://doi.org/10.1136/bmjqs-2016-005232.

Dubé, Eve, Dominique Gagnon, Manale Ouakki, Julie A. Bettinger, Maryse Guay, Scott Halperin, Kumanan Wilson, et al. 2016. « Understanding Vaccine Hesitancy in Canada: Results of a Consultation Study by the Canadian Immunization Research Network ». *PLOS ONE* 11 (6): e0156118. https://doi.org/10.1371/journal.pone.0156118.

Ehni, Hans-Jörg. 2008. « Dual Use and the Ethical Responsibility of Scientists ». *Archivum Immunologiae et Therapiae Experimentalis* 56 (3): 147-52. https://doi.org/10.1007/s00005-008-0020-7.

Esteva, Andre, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, et Sebastian Thrun. 2017. « Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks ». *Nature* 542 (7639): 115-18. https://doi.org/10.1038/nature21056.

Fernández-Baldor, Álvaro, Andrés Hueso, et Alejandra Boni. 2012. « From Individuality to Collectivity: The Challenges for Technology-Oriented Development Projects ». Dans *The Capability Approach, Technology and Design*, édité par Ilse Oosterlaken et Jeroen van den Hoven, 135-52. Philosophy of Engineering and Technology. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-3879-9_8.

Gotterbarn, Donald. 2001. « Informatics and Professional Responsibility ». *Science and Engineering Ethics* 7 (2): 221-30. https://doi.org/10.1007/s11948-001-0043-5.

Granville, Vincent. 2017. « Difference between Machine Learning, Data Science, AI, Deep Learning, and Statistics ». *Data science central* (blog). 2017. https://www.datasciencecentral.com/profiles/blogs/difference-between-machine-learning-data-science-ai-deep-learning.

Gulshan, Varun, Lily Peng, Marc Coram, Martin C. Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, et al. 2016. « Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs ». *JAMA* 316 (22): 2402-10. https://doi.org/10.1001/jama.2016.17216.

IEEE, Institute of Electrical and Electronics Engineers. 2017. « Ethically aligned design - Version 2 - For Public Discussion ». https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf.

———. 2019. « Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems ». First Edition. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf.

Ihde, Don. 2006. « The Designer Fallacy and Technological Imagination ». Dans *Defining Technological Literacy: Towards an Epistemological Framework*, édité par John R.

Dakers, 121-31. New York: Palgrave Macmillan US. https://doi.org/10.1057/9781403983053_9.

Jameson, J. Larry, et Dan L. Longo. 2015. « Precision Medicine--Personalized, Problematic, and Promising ». *The New England Journal of Medicine* 372 (23): 2229-34. https://doi.org/10.1056/NEJMsb1503104.

Johnson, Deborah G. 2006. « Computer Systems: Moral Entities but Not Moral Agents ». *Ethics and Information Technology* 8 (4): 195-204. https://doi.org/10.1007/s10676-006-9111-5.

Johnson, Deborah G., et Wetmore Jameson. 2008. « STS and Ethics: Implications for Engineering Ethics ». Dans *The Handbook of Science and Technology Studies, Third Edition*, MIT Press, 567-82. Edward J. Hackett Olga Amsterdamska Michael Lynch Judy Wajcman.

Jonas, Hans,. 1979. *Le principe responsabilité: une éthique pour la civilisation technologique*. Editions du Cerf.

Jones, M. L., E. Kaufman, et E. Edenberg. 2018. « AI and the Ethics of Automating Consent ». *IEEE Security Privacy* 16 (3): 64-72. https://doi.org/10.1109/MSP.2018.2701155.

JSAI. 2017. « The Japanese Society for Artificial Intelligence Ethical Guidelines ». Japon.

Keating, P. 2015. « Connaissance tacite. » Dans *Sciences, technologies et sociétés de A à Z*, Les presses de l'Université de Montréal, 53-56.

Kleinpeter, Édouard. 2015. « Le Cobot, la coopération entre l'utilisateur et la machine ». *Multitudes* n° 58 (1): 70-75.

Kuhlau, Frida, Stefan Eriksson, Kathinka Evers, et Anna T. Höglund. 2008. « Taking Due Care: Moral Obligations in Dual Use Research ». *Bioethics* 22 (9): 477-87. https://doi.org/10.1111/j.1467-8519.2008.00695.x.

Kuhlau, Frida, Anna T. Höglund, Kathinka Evers, et Stefan Eriksson. 2011. « A Precautionary Principle for Dual Use Research in the Life Sciences ». *Bioethics* 25 (1): 1-8. https://doi.org/10.1111/j.1467-8519.2009.01740.x.

Latour, Bruno. 1996. « On actor-network theory: A few clarifications ». *Soziale Welt* 47 (4): 369-81.

Lehoux, Pascale, Fiona A. Miller, Dominique Grimard, et Philippe Gauthier. 2018. « Anticipating Health Innovations in 2030–2040: Where Does Responsibility Lie for the Publics? » *Public Understanding of Science* 27 (3): 276-93. https://doi.org/10.1177/0963662517725715.

Lipton, Zachary C. 2016. « The Mythos of Model Interpretability ». arXiv:1606.03490 [cs, stat], juin. http://arxiv.org/abs/1606.03490.

Métayer, Michel. 2001. « Vers une pragmatique de la responsabilité morale ». *Lien social et Politiques*, nº 46: 19-30. https://doi.org/10.7202/000320ar.

Moor, J. H. 2006. « The Nature, Importance, and Difficulty of Machine Ethics ». *IEEE Intelligent Systems* 21 (4): 18-21. https://doi.org/10.1109/MIS.2006.80.

Murdock, Keelie Lyn Elektra, et David Koepsell. 2014. « Principals, agents, and the intersection between scientists and policy-makers: reflections on the H5N1 controversy ». *Infectious Diseases* 2: 109. https://doi.org/10.3389/fpubh.2014.00109.

Nissenbaum, Helen. 1994. « Computing and Accountability ». Communications of the ACM. 1 janvier 1994. https://link.galegroup.com/apps/doc/A15020194/AONE?sid=lms.

Noorman, Merel. 2008. « Limits to the Autonomy of Agents ». Dans *Proceedings of the 2008 Conference on Current Issues in Computing and Philosophy*, 65–75. Amsterdam, The Netherlands, The Netherlands: IOS Press. http://dl.acm.org/citation.cfm?id=1566234.1566244.

———. 2016. « Computing and Moral Responsibility ». Dans *The Stanford Encyclopedia of Philosophy*, édité par Edward N. Zalta, Winter 2016. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2016/entries/computing-responsibility/.

Oakden-Rayner, Luke. 2017. « Do machines actually beat doctors? ROC curves and performance metrics ». *Luke Oakden-Rayner* (blog). 2017. https://lukeoakdenrayner.wordpress.com/2017/12/06/do-machines-actually-beat-doctors-roc-curves-and-performance-metrics/.

Oosterlaken, Ilse. 2015. *Technology and Human Development*. Routledge.

Oosterlaken, Ilse, et Jeroen van den Hoven, éd. 2012. *The capability approach, technology and design*. Springer. Vol. 5. Philosophy of Engineering and Technology.

OSTP, (White House Office of Science and Technology Policy). 2016. « Preparing for the Future of Artificial Intelligence ».

Owen, Richard, Phil Macnaghten, et Jack Stilgoe. 2012. « Responsible Research and Innovation: From Science in Society to Science for Society, with Society ». *Science and Public Policy* 39 (6): 751-60. https://doi.org/10.1093/scipol/scs093.

Pellé, Sophie, et Bernard Reber. 2016. *Ethique de la recherche et innovation responsable*. ISTE editions. Vol. 2. Innovation et recherche responsables.

Ralph, Jason. 2018. « What Should Be Done? Pragmatic Constructivist Ethics and the Responsibility to Protect ». International Organization 72 (1): 173-203. https://doi.org/10.1017/S0020818317000455.

Rose, Nikolas. 2003. « Neurochemical Selves ». *Society* 41 (1): 46-59. https://doi.org/10.1007/BF02688204.

Rotenberg, Vadim S. 2013. « Moravec's Paradox: Consideration in the Context of Two Brain Hemisphere Functions ». *Activitas Nervosa Superior* 55 (3): 108-11. https://doi.org/10.1007/BF03379600.

Rumbold, John M. M., et Barbara K. Pierscionek. 2017. « A Critique of the Regulation of Data Science in Healthcare Research in the European Union ». *Bmc Medical Ethics* 18 (avril): 27. https://doi.org/10.1186/s12910-017-0184-y.

Scheutz, Matthias. 2016. « The need for moral competency in autonomous agent architectures ». Dans *Fundamental issues of artificial intelligence*, Springer International Publishing Switzerland, 517-27. V.C. Müller.

Schlosser, Markus. 2015. « Agency ». Dans *The Stanford Encyclopedia of Philosophy*, édité par Edward N. Zalta, Fall 2015. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2015/entries/agency/.

Schmietow, Bettina. 2016. « Ethical Dimensions of Dynamic Consent in Data-Intense Biomedical Research-Paradigm Shift, or Red Herring? » Dans *Ethics and Governance of Biomedical Research: Theory and Practice*, édité par D. Strech et M. Mertz, 4:197-209. Dordrecht: Springer.

Schwab, Klaus. 2016. *La quatrième révolution industrielle*. Dunod. Suisse: World Economic Forum.

Sharon, Tamar. 2016. « The Googlization of Health Research: From Disruptive Innovation to Disruptive Ethics ». *Personalized Medicine* 13 (6): 563-74. https://doi.org/10.2217/pme-2016-0057.

———. 2017. « Self-Tracking for Health and the Quantified Self: Re-Articulating Autonomy, Solidarity, and Authenticity in an Age of Personalized Healthcare ». *Philosophy & Technology* 30 (1): 93-121. https://doi.org/10.1007/s13347-016-0215-5.

Shulman, Carl, Henrik Jonsson, et Nick Tarleton. 2009. « Machine ethics and superintelligence ». Dans , 95–97. Tokyo, Japan: Carson Reynolds and Alvaro Cassinelli.

SMILEY, Marion. 1992. Moral Responsibility and the Boundaries of Community. Power and Accountability from a Pragmatic Point of View. Chicago, The University of Chicago Press. 296

Thompson, Dennis F. 2004. « The problem of many hands ». Dans *Restoring Responsibility : Ethics in Govern1m1ent, Business, and Healthcare*, Cambridge University Press.

Toulmin, Stephen. 1981. « The Tyranny of Principles ». *The Hastings Center Report* 11 (6): 31-39. https://doi.org/10.2307/3560542.

Verbeek, Peter-Paul. 2006. « Materializing Morality: Design Ethics and Technological Mediation ». *Science, Technology, & Human Values* 31 (3): 361-80. https://doi.org/10.1177/0162243905285847.

Villani, Cédric. 2018. « Donner un sens à l'intelligence artificielle. Pour une stratégie nationale et européenne. » https://www.aiforhumanity.fr/pdfs/9782111457089_Rapport_Villani_accessible.pdf.

Woolley, J. Patrick. 2016. « How Data Are Transforming the Landscape of Biomedical Ethics: The Need for ELSI Metadata on Consent ». Dans *The Ethics of Biomedical Big Data*, édité par Brent Daniel Mittelstadt et Luciano Floridi, 171-97. Law, Governance and Technology Series. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-33525-4_8.

Conclusion

Quels sont alors les défis du partage et de l'exercice de la responsabilité face aux risques et enjeux éthiques soulevés par l'utilisation de systèmes d'intelligence artificielle (IA) en santé en vue d'informer une innovation responsable ? À la lumière des différents éléments présentés dans les précédents chapitres, il est possible d'identifier que l'innovation numérique en santé responsable – qui concerne principalement l'utilisation des systèmes d'IA – risque sans doute de se heurter à différents enjeux et difficultés, considérant la complexité de l'écosystème à encadrer.

Pour répondre à la question de recherche, il a d'abord été nécessaire d'identifier les bénéfices, les risques et les enjeux éthiques qui accompagnent l'utilisation de systèmes d'IA en santé (***cf.* Objectif 1**). Tel que présenté dans le Chapitre 2, l'utilisation des systèmes d'IA en santé peut en effet représenter des bénéfices considérables pour améliorer les soins, la qualité de vie des patients, la prévention en santé publique ou encore pour l'avancée des connaissances (Torkamani et al. 2017; Peek et al. 2015; Jiang et al. 2017; Brownstein, Freifeld, et Madoff 2009). Ces bénéfices découlent des différentes applications des systèmes d'IA en vue de développer une médecine de précision de plus en plus ciblée (Bayer et Galea 2015; Ding et al. 2018; Jameson et Longo 2015; Ashley 2015), mais également dans l'objectif d'extraire des connaissances significatives des ensembles de données produites chaque jour pour améliorer les traitements, prévenir l'émergence de maladies ou encore orienter la recherche. Cependant, l'obtention de ces bénéfices grâce à l'utilisation de robots de soin, de systèmes experts ou de différents algorithmes pour valoriser les données relatives à la santé pourrait bien conduire à une transition du système de santé selon différents éléments disruptifs. Le décalage temporel et sémantique entre collecte et analyse de données représente un potentiel point de rupture important (Coutellec et Weil-Dubuc 2017). Les nouveaux lieux de collectes de données relatives à la santé (ex. les téléphones intelligents) ainsi que la réutilisation quasi infinie permise par leur stockage, défient la conception traditionnelle du consentement (Christen et al. 2016; Woolley 2016; Rumbold et Pierscionek 2017; Mittelstadt et Floridi 2016), mais aussi la conception de ce que représente un patient ou un participant à la recherche[144]. Également, les modes de collecte et d'analyse impliquent l'entrée en jeu de nouveaux

[144] Cet aspect a été mis de l'avant par les travaux Comité de travail sur les aspects d'éthique de la recherche dans les domaines du numérique, de l'intelligence artificielle et des données massives des Fonds de recherche du Québec,

acteurs dans le parcours de soins du patient (CCNE 2019; Sharon 2016), tels que les scientifiques des données, les développeurs mais également les multinationales de l'IA qui ne répondent pas des mêmes normes déontologiques que les professionnels de santé ou les chercheurs en santé.

Les bénéfices des systèmes d'IA doivent alors être confrontés aux différents risques décrits en vue d'un équilibre raisonnable. Le Chapitre 3 a permis de mettre en évidence cinq principales catégories de risques et enjeux éthiques, dont la nature n'est pas nouvelle quand on considère l'innovation en santé mais dont tous revêtent des aspects particuliers inhérents à l'avènement de l'IA et des données massives, voire sont nettement exacerbés considérant l'échelle de grandeur et la portée des méthodes et des technologies en jeu. Il s'agit ainsi d'atteintes potentielles relatives : 1) à la vie privée qui tend à devenir une atteinte à l'intimité (Azencott C.-A. 2018; Haddadi et Brown, s. d.; Iyengar, Kundu, et Pallis 2018; IEEE 2017; Déclaration de Montréal IA Responsable 2018) ; 2) au consentement libre et éclairé considérant que les mécanismes traditionnels ne sauraient suffire à assurer son respect (Christen et al. 2016; Woolley 2016; Rumbold et Pierscionek 2017; Mittelstadt et Floridi 2016); 3) à la justice sociale par le biais de différentes discriminations (Altman, Wood, et Vayena 2018; Friedler, Scheidegger, et Venkatasubramanian 2016; Zarsky 2016; Ganascia 2018); 4) à la déshumanisation potentielle des soins et du patient par différents mécanismes d'isolement ou de désubjectivation (Coeckelbergh 2015; 2012; Rouvroy 2014); et 5) à la sécurité considérant une certaine faillibilité des systèmes (Brundage et al. 2018).

Ces enjeux sont liés à trois principales caractéristiques inhérentes aux technologies en jeu lorsqu'il est question de systèmes d'IA en santé : les limites interprétatives et informationnelles de l'analyse des données massives, la nécessité de partage tant relative aux « données ouvertes » qu'à l'ouverture du code des algorithmes et l'opacité des réseaux de neurones qui conduit à un manque d'explicabilité relativement aux raisons qui motivent les décisions algorithmiques.

Selon un principe de précaution (Boisson de Chazournes 2002; Godard 2000; Lecourt 2007) ou une approche proportionnelle (ÉPTC2 2018), il est essentiel de considérer à la fois les bénéfices et les risques qui accompagnent l'utilisation de systèmes d'IA en santé. En effet, une innovation

présentés lors d'une conférence donnée par Mylène Deschênes dans le cadre de la Semaine sur la conduite responsable en recherche à l'Université Laval, le 8 mai 2019.

responsable ne saurait ignorer les avancées majeures qu'amènent les différentes technologies présentées. Plus encore, la collecte massive de données s'imposant comme mode de fonctionnement pour les années à venir fait que nous avons besoin des systèmes d'IA pour traiter ces ensembles gigantesques de données relatives à la santé qu'il n'est plus possible de traiter manuellement. Une innovation responsable demande cependant de limiter les risques inhérents aux usages présentés. De cette recherche d'un équilibre entre les bénéfices et les risques propres à l'innovation responsable découle la nécessité que les choix normatifs soient guidés par la recherche d'une pondération entre intérêts individuels et collectifs (ce qui a été mis en évidence dans le Chapitre 3 mais également par les citoyens participant à la coconstruction de la Déclaration de Montréal).

Pour répondre à ces enjeux, il a alors été nécessaire d'identifier les défis de la mise en place d'une gouvernance éthique adaptée, notamment en ce qui a trait à l'exercice de la responsabilité des parties prenantes de l'innovation numérique en santé (***cf.*** **Objectif 2**). Le Chapitre 3 a permis de mettre en évidence un lien intrinsèque entre les caractéristiques techniques des technologies en jeu (dont notamment, leurs limites) et les conséquences éthiques et sociales de leur usage, qui invitent à considérer une éthique *by design* (soit dès la conception). Une *éthique by design* se heurte cependant à des défis de taille relativement à l'opérationnalisation des principes de l'éthique de l'IA, et remet la responsabilité de l'innovation numérique en santé sur les seules épaules des développeurs. Si de nombreuses initiatives, de portée nationale ou internationale, se sont penchées sur la définition des principes éthiques directeurs pour un développement responsable de l'IA (IEEE 2019; Déclaration de Montréal IA Responsable 2018; JSAI 2017; AI HLEG 2019; Dawson et al. 2019; Amnesty International 2018), leur opérationnalisation se heurte aux difficultés inhérentes à toute démarche d'éthique appliquée. Ces difficultés résident dans la portée desdits principes et dans la manière de les mettre en pratique. En effet, les caractéristiques de l'avènement des systèmes d'IA ne permettant plus de circonscrire les utilisations à une région du monde ou à une juridiction particulière (Scherer 2015; Danaher 2015), les modes de gouvernance traditionnels sont mis à défi. Considérant que la collecte de données ne connaît pas de frontière, la nécessité d'une coordination internationale relativement à l'éthique de l'IA semble alors nécessaire. L'observation d'une convergence des principes de l'éthique de l'IA avec ceux de la bioéthique sous-tend l'idée d'une certaine universalité des valeurs ou principes mobilisés (Mittelstadt 2019; Jobin, Ienca, et Vayena 2019; Floridi et al. 2018), sans pour autant prétendre à l'universalité de

leur application. Si les principes sont pour certains trop abstraits pour être traduits en normes (ou autres mécanismes d'opérationnalisation) (Morley et al. 2019; Mittelstadt 2019), ce niveau d'abstraction autorise cependant une mise en pratique plurielle, apte à reconnaître les divergences morales relatives à leur interprétation. L'éthique appliquée à l'IA nécessite alors une approche inductive sensible au contexte et aux spécificités de chacune des situations où émergent les enjeux éthiques (Jobin, Ienca, et Vayena 2019; Morley et al. 2019; Cleret de Langavant 2001; Massé 2003).

C'est ainsi dans la manière de faire de l'éthique qu'il est possible de répondre, partiellement, aux impératifs d'une innovation responsable, dans laquelle l'exercice de la responsabilité de chacun des acteurs du développement des systèmes d'IA en santé se retrouve. L'analyse des discussions de la coconstruction de la Déclaration de Montréal ont mis en évidence trois principaux défis de l'exercice de la responsabilité face à l'utilisation des systèmes d'IA en santé : 1) la préservation des capabilités humaines ; 2) le problème des mains multiples et 3) l'agentivité artificielle. Concernant premier défi, les citoyens craignent que l'avènement de l'utilisation des systèmes d'IA incapacite les patients (par une atteinte à la liberté de choix, au consentement éclairé et une perte de l'esprit critique) et les professionnels de santé (par une perte de compétences et de confiance en eux-mêmes ainsi que la potentielle apparition d'une dépendance à la technologie). Ils ont manifesté des attentes relativement à la capacitation de ces acteurs, qui demandent formation et éducation pour préserver, entre autres, leur autonomie décisionnelle.

Le deuxième défi renvoi au problème des *mains multiples*, à savoir qu'un grand nombre d'acteurs étant impliqué dans le parcours de soins, il devient difficile de déterminer qui est responsable et dans quelle mesure. Ce nombre croissant d'acteurs entraîne différentes conséquences sur les soins, sur la gestion des données et le partage des responsabilités. Les citoyens craignent notamment que le problème des *mains multiples* encourage l'apparition d'une perte de lien naturel entre professionnels de santé et patients, de conflits d'intérêts potentiels, ou de différents enjeux relatifs à la propriété et à la protection des données. Ils ont soulevé des attentes relatives à un contrat social qui se dessinerait selon une responsabilité partagée des différentes parties prenantes du développement et de l'utilisation des systèmes d'IA en santé.

Le troisième défi émane de la potentielle reconnaissance d'une agentivité artificielle. Face à ce défi, les citoyens craignent que les systèmes d'IA deviennent des agents (décisionnels), notamment considérant les biais que pourraient contenir les algorithmes, et estiment important que les systèmes d'IA soient des outils (transparents) afin de permettre à l'humain de garder la main – pour reprendre l'expression de la CNIL (2017). À mi-chemin entre la préservation des capabilités humaines et la reconnaissance d'une agentivité artificielle, apparaît la crainte du remplacement des humains par les machines, et l'attente d'une coopération humain-machine afin d'éviter, notamment, la déshumanisation du soin.

À la lumière de ces différents constats, il a été possible de dégager plusieurs pistes de réflexion en vue d'informer la mise en place d'un encadrement éthique pertinent et effectif pour une innovation responsable en ce qui a trait au développement des systèmes d'IA en santé (***cf.* Objectif 3**). La nécessité d'adapter les mécanismes existants à la complexité de l'écosystème de l'innovation responsable (en particulier en ce qui a trait à l'utilisation des systèmes d'IA en santé) a été défendu à la fois par les citoyens et dans le Chapitre 4. Cette adaptation doit tenir compte des différents éléments disruptifs présentés afin de répondre aux enjeux éthiques que soulèvent l'utilisation des systèmes d'IA en santé. Elle demande également de se pencher sur les trois défis de l'exercice de la responsabilité qui émergent des discussions citoyennes ; défis qui mèneraient tous trois à une certaine tendance à la déresponsabilisation et demanderaient de préserver l'agentivité humaine. De ces défis se dégagent trois principales tensions relatives à l'exercice de la responsabilité, soit : 1) une tension entre agentivité artificielle et agentivité humaine ; 2) une tension entre responsabilité individuelle et collective et 3) une tension entre technologie capacitante et incapacitante.

Afin d'assurer un encadrement éthique effectif et pertinent, il est soutenu qu'une conception pragmatique de la responsabilité serait particulièrement adaptée à l'identification des enjeux de l'innovation numérique en santé responsable. Cette conception défendue par Métayer (2001), permet de se pencher sur la responsabilité et son partage sur la base d'une structure d'interpellation responsabilisante qui implique un demandeur et un répondeur. Parce-que cette conception permet de jongler entre les obligations qui émanent des conceptions positives et négatives de la responsabilité et qu'elle demande une approche sensible au contexte, elle s'accorde aisément aux

impératifs d'une innovation responsable qui demande une approche mixte en vue d'interprétations riches de la responsabilité (Pellé et Reber 2016). La conception pragmatique de la responsabilité est également cohérente avec une compréhension de l'éthique appliquée qui prend son sens dans la prise en compte de la spécificité des situations et demande une sensibilité contextuelle en vue de résoudre les problèmes et dilemmes éthiques (Massé 2003; Toulmin 1981; Cleret de Langavant 2001).

Cette conception de la responsabilité invite ainsi à reconnaître l'existence d'une certaine agentivité artificielle (qui renvoie à la capacité de réaliser des actions de manière autonome) afin d'en limiter l'étendue (relativement à la capacité décisionnelle). Ceci répond en partie à l'ambiguïté de la gouvernance algorithmique (Introna 2016; Musiani 2013; Déclaration de Montréal IA Responsable 2018) : une responsabilité pragmatique (Métayer 2001) demande la mise en place d'une gouvernance des humains sur les algorithmes afin d'empêcher que « les algorithmes nous gouvernent » (Déclaration de Montréal IA Responsable 2018). Devant le nombre croissant d'acteurs impliqués, chacun se retrouve potentiellement *demandeur* et *répondeur* des différentes interpellations responsabilisantes qui prennent naissance dans l'apparition des enjeux éthiques présentés. Si la solution semble être dans la mise en place d'une responsabilité partagée, qui implique à la fois la responsabilité individuelle de chacun (pour éviter la déresponsabilisation) et une responsabilité collective (pour tempérer la surcharge morale), il est nécessaire de se pencher sur la capacitation des parties prenantes du développement des systèmes d'IA en santé afin de garantir que tous soient en mesure de remplir ces responsabilités. Cette capacitation demande de favoriser l'agentivité humaine, pour sa valeur instrumentale (permettant d'assurer un développement technologique « réussi ») mais aussi pour sa valeur intrinsèque (l'agentivité humaine étant moralement valorisable) (Crocker et Robeyns 2010 dans Oosterlaken 2015). Ainsi, il semble essentiel que les impératifs de l'innovation responsable s'orientent vers la préservation des capabilités des individus, soit de ce qu'ils sont réellement capables de faire et d'être selon les objectifs qu'ils valorisent (Oosterlaken 2015; Oosterlaken et van den Hoven 2012).

On notera finalement qu'en filigrane des craintes et des attentes citoyennes, des différents éléments disruptifs présentés, et de l'émergence des enjeux éthiques de l'avènement des systèmes d'IA en santé, se dessine une crainte générale de « déshumanisation ». Celle-ci réfère en premier

lieu à la déshumanisation des soins et du patient, comme présentée dans le Chapitre 3 mais également telle que soulevée par les citoyens ayants participé à la coconstruction de la Déclaration de Montréal. Cette déshumanisation prend naissance dans le potentiel remplacement exponentiel des humains par les machines (ici, les systèmes d'IA) mais également par la perte de lien naturel entre patients et professionnels de santé (sous-tendu par l'apparition du problème des *mains multiples*). Ces deux constats ont pour conséquence de limiter le contact humain et d'augmenter la distance entre professionnels de santé et patients. On note également une déshumanisation relative à la « désubjectivation » (Rouvroy 2014; Rouvroy et Berns 2013; Ibekwe-Sanjuan 2014) des patients qui risqueraient, selon les préoccupations qui accompagnent l'apparition du phénomène de quantification du soi (Ajana 2017; Haddadi et Brown, s. d.; Swan 2013; Brouard 2017), d'être considérés comme des agrégats de données plutôt que comme des personnes à part entière (Coutellec et Weil-Dubuc 2017). Toutefois, à l'heure de l'automatisation du travail de l'humain sur l'humain (Lahlou 2015), cette crainte de déshumanisation ne saurait se restreindre au seul secteur de la santé. L'avènement des systèmes d'IA s'accompagne de potentielles transformations relationnelles importantes (Schwab 2016; Devillers 2017). Les capacités croissantes des systèmes d'IA questionnent les fondements de ce que nous considérons être l'essence de ce qui nous différencie des machines, et par extension de ce qui nous définit en tant qu'êtres humains (Coeckelbergh 2015; Noorman 2008; Rouvroy 2014; Ibekwe-Sanjuan 2014). La numérisation de la société encourage une tendance à considérer l'image digitale des individus plutôt que les individus dans toute la complexité de leur réalité non-digitale. Ce phénomène tend pour certains à donner une vision appauvrie de l'humain et de ses interactions (Coutellec et Weil-Dubuc 2017; Rouvroy 2014). Cette numérisation croissante pourrait alors conduire à évincer des éléments essentiels qui constituent nos façons de vivre et nos façons d'être, qui ne sauraient être transformées en langage numérique :

> À moins de confondre singularité et particularité, sentiments et émotions, les éléments contextuels de l'existence ne peuvent être numérisés (Coutellec et Weil-Dubuc 2017 p. 77).

Face à l'émergence de potentielles nouvelles identités (humaines et artificielles), de nouvelles relations (entre les humains par le biais des machines et entre humains et machines) et de nouvelles agentivités (elles aussi, humaines et artificielles), différentes conséquences émotionnelles et relationnelles sont à présager sur les individus, au-delà des parties prenantes du systèmes de

santé[145]. Parce qu'elles revêtent une composante émotionnelle importante, il semble difficile de transformer ces préoccupations en mécanismes concrets de gouvernance. Il n'en est pas moins vrai qu'elles sont essentielles à considérer du point de vue de l'innovation responsable et que l'aspect émotionnel ne saurait justifier de les écarter. En effet, selon la responsabilité pragmatique, la dimension morale ne peut se restreindre à un « héritage rationaliste », mais demande d'inscrire la responsabilité dans les différentes composantes narratives et émotionnelles de l'histoire des sujets responsables[146] (Métayer 2001). Elle fait écho à la responsabilité sollicitude, comme « souci de », qui demande de dépasser la seule analyse rationnelle des risques (Pellé et Reber 2016). De plus, une éthique inductive ne saurait être pertinente sans « humainement » équilibrer les conflits entre les principes (Toulmin 1981) notamment afin de se prémunir d'une certaine rationalité instrumentale de l'éthique déductive (Cleret de Langavant 2001). Ces aspects sont essentiels à considérer d'un point de vue de la gouvernance éthique et de l'encadrement des systèmes d'IA car ils sont à la base de transformations potentielles majeures et pourraient affecter la société dans son ensemble, mais aussi car ils vont dans le sens de la préservation de la confiance du public en la science, élément essentiel du point de vue de l'innovation responsable (Resnik 2011; Koepsell 2017).

Puisqu'il n'existe pas de développement technologique sans adhésion sociale, l'idée du contrat social entre les acteurs de l'innovation responsable demande ainsi de prêter attention aux dimensions déshumanisantes – quelle que soit la rationalité qu'on leur accorde. Les acteurs de l'innovation responsable doivent donc rester attentifs à la perception citoyenne des nouvelles

[145] Ces trois dimensions (nouvelles identités, nouvelles relations et nouvelles agentivités) sont celles qui composent le concept d'AIship, que nous avons développé avec Bélisle-Pipon et Couture en 2019 dans le cadre d'un projet d'exposition artistico-scientifique du même nom (voir : https://aiship.org/en/home/). Le projet explore notamment les conséquences émotionnelles et relationnelles de l'utilisation des systèmes d'IA en santé.

[146] De plus, la distinction entre raison et émotion est actuellement relativement contestée, ce qui renforce l'argument de tenir compte des émotions dans des perspectives de gouvernance éthique. Par exemple, selon le neuroscientifique Antonio Damasio, la raison ne peut être dissociée des émotions (Damasio 2010). En présentant l'aspect neurologique des émotions et leur impact dans la prise de décision et le comportement social, l'auteur démontre non seulement que les émotions participent à la raison, selon des mécanismes qui échappent parfois à notre conscience, mais qu'en plus leur rôle dans le processus de raisonnement est parfois positif. Dans la même veine, les études en neuroéconomie (discipline aux frontières des neurosciences et de l'économie comportementale (Pelloux, Rullière, et Van Winden 2009) confirment l'importance des émotions dans la prise de décision. En incluant celles-ci (jusqu'alors évincées) dans le raisonnement économique, les études démontrent comment l'empathie, la confiance ou la générosité ont un rôle nécessaire dans la prise de décision rationnelle (Henrich et al. 2001; Kosfeld et al. 2005; Zak, Kurzban, et Matzner 2004; Zak, Stanton, et Ahmadi 2007).

méthodes, des techniques et des technologies qui relèvent de l'IA. À cette fin, il est essentiel de créer des espaces de délibération qui permettraient la mise en commun des perspectives et des attentes normatives; afin d'assurer que perdure, entre autres, la relation qui se veut mutuellement bénéfique entre science et société (Koepsell 2017) et de favoriser le développement d'une compréhension commune et partagée de ce que peut représenter le bien commun dans le contexte de l'avancée des connaissances en IA.

Références bibliographiques

AI HLEG, (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.

Ajana, Btihaj. 2017. « Digital Health and the Biopolitics of the Quantified Self ». *DIGITAL HEALTH* 3 (janvier): 2055207616689509. https://doi.org/10.1177/2055207616689509.

Altman, M., A. Wood, et E. Vayena. 2018. « A Harm-Reduction Framework for Algorithmic Fairness ». *IEEE Security Privacy* 16 (3): 34-45. https://doi.org/10.1109/MSP.2018.2701149.

Amnesty International, Access Now. 2018. « The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems ». RightsCon Toronto. https://www.accessnow.org/toronto-declaration.

Ashley, Euan A. 2015. « The Precision Medicine Initiative: A New National Effort ». *JAMA* 313 (21): 2119-20. https://doi.org/10.1001/jama.2015.3595.

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): 20170350. https://doi.org/10.1098/rsta.2017.0350.

Bayer, Ronald, et Sandro Galea. 2015. « Public Health in the Precision-Medicine Era ». *The New England Journal of Medicine* 373 (6): 499-501. https://doi.org/10.1056/NEJMp1506241.

Boisson de Chazournes, Laurence. 2002. « Le principe de précaution: nature, contenu et limites ». http://archive-ouverte.unige.ch/unige:15028/ATTACHMENT01.

Brouard, Benoît. 2017. « Chapitre 2. Utilisation des Big Data en santé : le cas des objets connectés ». *Journal international de bioethique et d'ethique des sciences* Vol. 28 (3): 27-30.

Brownstein, John S., Clark C. Freifeld, et Lawrence C. Madoff. 2009. « Digital Disease Detection — Harnessing the Web for Public Health Surveillance ». *New England Journal of Medicine* 360 (21): 2153-57. https://doi.org/10.1056/NEJMp0900702.

Brundage, Miles, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, et Bobby Filar. 2018. « The malicious use of

artificial intelligence: Forecasting, prevention, and mitigation ». *arXiv preprint arXiv:1802.07228*.

CCNE. 2019. « Données massives (big data) et santé : une nouvelle approche des enjeux éthiques ». Avis 130. Comité Consultatif National d'Éthique français. https://www.ccne-ethique.fr/sites/default/files/avis_130.pdf.

Christen, Markus, Josep Domingo-Ferrer, Bogdan Draganski, Tade Spranger, et Henrik Walter. 2016. « On the Compatibility of Big Data Driven Research and Informed Consent: The Example of the Human Brain Project ». Dans *The Ethics of Biomedical Big Data*, édité par Brent Daniel Mittelstadt et Luciano Floridi, 199-218. Law, Governance and Technology Series. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-33525-4_9.

Cleret de Langavant, Ghislaine. 2001. *Bioéthique : Méthode et complexité*. Presses de l'Université du Québec. Québec.

CNIL (Commission nationale informatique et libertés). 2017. « Comment permettre à l'homme de garder la main ? Les enjeux éthiques des algorithmes et de l'intelligence artificielle ».

Coeckelbergh, Mark. 2012. « "How I Learned to Love the Robot": Capabilities, Information Technologies, and Elderly Care ». Dans *The Capability Approach, Technology and Design*, édité par Ilse Oosterlaken et Jeroen van den Hoven, 77-86. Philosophy of Engineering and Technology. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-3879-9_5.

———. 2015. « Artificial Agents, Good Care, and Modernity ». *Theoretical Medicine and Bioethics* 36 (4): 265-77. https://doi.org/10.1007/s11017-015-9331-y.

Coutellec, Léo, et Paul-Loup Weil-Dubuc. 2017. « Chapitre 7. Big data ou l'illusion d'une synthèse par agrégation. Une critique épistémologique, éthique et politique ». *Journal international de bioethique et d'ethique des sciences* Vol. 28 (3): 63-79.

Damasio, Antonio. 2010. *L'erreur de Descartes : la raison des émotions*. Poches Odile Jacob.

Danaher, John. 2015. « Philosophical Disquisitions: Is effective regulation of AI possible? Eight potential regulatory problems ». *Philosophical Disquisitions* (blog). 7 juillet 2015. http://philosophicaldisquisitions.blogspot.com/2015/07/is-effective-regulation-of-ai-possible.html.

Dawson, D, E Schleiger, J Horton, J McLaughlin, C Robinson, G Quezada, J Scowcroft, et S Hajkowicz. 2019. « Artificial Intelligence : Australia's Ethics Framework. A Discussion Paper. » Australia: Data61 CSIRO. https://consult.industry.gov.au/strategic-policy/artificial-intelligence-ethics-framework/supporting_documents/ArtificialIntelligenceethicsframeworkdiscussionpaper.pdf.

Déclaration de Montréal IA Responsable. 2018. « Rapport de la Déclaration de Montréal pour un Développement Responsable de l'Intelligence Artificielle ». https://docs.wixstatic.com/ugd/ebc3a3_d806f109c4104c91a2e719a7bef77ce6.pdf.

Devillers, Laurence. 2017. *Des robots et des hommes: Mythes, fantasmes et réalité*. Plon.

Ding, Michael Q., Lujia Chen, Gregory F. Cooper, Jonathan D. Young, et Xinghua Lu. 2018. « Precision Oncology beyond Targeted Therapy: Combining Omics Data with Machine Learning Matches the Majority of Cancer Cells to Effective Therapeutics ». *Molecular Cancer Research* 16 (2): 269-78. https://doi.org/10.1158/1541-7786.MCR-17-0378.

Floridi, Luciano, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, et al. 2018. « AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations ». *Minds and Machines* 28 (4): 689-707. https://doi.org/10.1007/s11023-018-9482-5.

Friedler, Sorelle A., Carlos Scheidegger, et Suresh Venkatasubramanian. 2016. « On the (im)possibility of fairness ». *arXiv:1609.07236 [cs, stat]*, septembre. http://arxiv.org/abs/1609.07236.

Ganascia, Jean-Gabriel. 2018. *Éthique, intelligence artificielle et santé*. ERES. https://www.cairn.info/traite-de-bioethique-iv--9782749260839-page-527.htm.

Godard, Olivier. 2000. « Le principe de précaution, une nouvelle logique de l'action entre science et démocratie » 11: 17-56.

Haddadi, Hamed, et Ian Brown. s. d. « Quantified Self and the Privacy Challenge », 2.

Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, et Richard McElreath. 2001. « In search of homo economicus: behavioral experiments in 15 small-scale societies ». *American Economic Review* 91 (2): 73–78.

Ibekwe-Sanjuan, Fidelia. 2014. « Big Data, Big machines, Big Science : vers une société sans sujet et sans causalité ? » Dans *XIXème Congrès de la Sfsic. Penser les techniques et les*

technologies : Apports des Sciences de l'Information et de la Communication et perspectives de recherches., 1-10. Toulon, France. https://hal.archives-ouvertes.fr/hal-01066202.

IEEE, Institute of Electrical and Electronics Engineers. 2017. « Ethically aligned design - Version 2 - For Public Discussion ». https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf.

———. 2019. « Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems ». First Edition. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf.

Introna, Lucas D. 2016. « Algorithms, Governance, and Governmentality , Algorithms, Governance, and Governmentality: On Governing Academic Writing , On Governing Academic Writing ». *Science, Technology, & Human Values* 41 (1): 17-49. https://doi.org/10.1177/0162243915587360.

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

Jameson, J. Larry, et Dan L. Longo. 2015. « Precision Medicine--Personalized, Problematic, and Promising ». *The New England Journal of Medicine* 372 (23): 2229-34. https://doi.org/10.1056/NEJMsb1503104.

Jiang, Fei, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, et Yongjun Wang. 2017. « Artificial intelligence in healthcare: past, present and future ». *Stroke and Vascular Neurology*. https://doi.org/10.1136/svn-2017-000101.

Jobin, Anna, Marcello Ienca, et Effy Vayena. 2019. « Artificial Intelligence: the global landscape of ethics guidelines ». *arXiv:1906.11668 [cs]*, juin. http://arxiv.org/abs/1906.11668.

JSAI. 2017. « The Japanese Society for Artificial Intelligence Ethical Guidelines ». Japon.

Koepsell, David. 2017. « Duties of Science to Society (and Vice Versa) ». Dans *Scientific Integrity and Research Ethics*, 85-95. SpringerBriefs in Ethics. Springer, Cham. https://doi.org/10.1007/978-3-319-51277-8_8.

Kosfeld, Michael, Markus Heinrichs, Paul J. Zak, Urs Fischbacher, et Ernst Fehr. 2005. « Oxytocin Increases Trust in Humans ». *Nature* 435 (7042): 673-76. https://doi.org/10.1038/nature03701.

Lahlou, Saadi. 2015. « Un monde numérique : le renversement du miroir ». Dans . Vol. 53. Variances.

Lecourt, Dominique. 2007. « L'étrange fortune du principe de précaution ». *Études de l'Observatoire du principe de précaution*. http://www.estig.ipbeja.pt/~ac_direito/etude1.pdf.

Massé, Raymond. 2003. « Valeurs universelles et relativisme culturel en recherche internationale: les contributions d'un principisme sensible aux contextes socioculturels ». *Autrepart* n° 28 (4): 21-35.

Métayer, Michel. 2001. « Vers une pragmatique de la responsabilité morale ». *Lien social et Politiques*, nº 46: 19-30. https://doi.org/10.7202/000320ar.

Mittelstadt, Brent. 2019. « AI Ethics – Too Principled to Fail? » SSRN Scholarly Paper ID 3391293. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=3391293.

Mittelstadt, Brent Daniel, et Luciano Floridi. 2016. « The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts ». *Science and Engineering Ethics* 22 (2): 303-41. https://doi.org/10.1007/s11948-015-9652-2.

Morley, Jessica, Luciano Floridi, Libby Kinsey, et Anat Elhalal. 2019. « From What to How. An Overview of AI Ethics Tools, Methods and Research to Translate Principles into Practices ». *arXiv:1905.06876 [cs]*, mai. http://arxiv.org/abs/1905.06876.

Musiani, Francesca. 2013. « Governance by algorithms ». *Internet Policy Review*, août. https://policyreview.info/articles/analysis/governance-algorithms.

Noorman, Merel. 2008. « Limits to the Autonomy of Agents ». Dans *Current Issues in Computing and Philosophy*, édité par P. Brey, A. Briggle, et K. Waelbers, 65–75. Ios Press.

Oosterlaken, Ilse. 2015. *Technology and Human Development*. Routledge.

Oosterlaken, Ilse, et Jeroen van den Hoven, éd. 2012. *The capability approach, technology and design*. Springer. Vol. 5. Philosophy of Engineering and Technology.

Peek, Niels, Carlo Combi, Roque Marin, et Riccardo Bellazzi. 2015. « Thirty years of artificial intelligence in medicine (AIME) conferences: A review of research themes ». *Artificial Intelligence in Medicine*, Artificial Intelligence in Medicine AIME 2013, 65 (1): 61-73. https://doi.org/10.1016/j.artmed.2015.07.003.

Pellé, Sophie, et Bernard Reber. 2016. *Ethique de la recherche et innovation responsable*. ISTE editions. Vol. 2. Innovation et recherche responsables.

Pelloux, Benjamin, Jean-Louis Rullière, et Frans Van Winden. 2009. « La neuroéconomie dans l'agenda de l'économie comportementale ». *Revue française d'économie* 23 (4): 3-36. https://doi.org/10.3406/rfeco.2009.1705.

Resnik, David B. 2011. « Scientific Research and the Public Trust ». *Science and Engineering Ethics* 17 (3): 399-409. https://doi.org/10.1007/s11948-010-9210-x.

Rouvroy, Antoinette. 2014. « Des données sans personne: le fétichisme de la donnée à caractère personnel à l'épreuve de l'idéologie des Big Data ». *Contribution en marge de l'Étude annuelle du Conseil d'État. Le numérique et les droits et libertés fondamentaux*.

Rouvroy, Antoinette, et Thomas Berns. 2013. « Gouvernementalité algorithmique et perspectives d'émancipation, Faced with algorithmic governmentality ». *Réseaux*, nº 177 (mai): 163-96. https://doi.org/10.3917/res.177.0163.

Rumbold, John M. M., et Barbara K. Pierscionek. 2017. « A Critique of the Regulation of Data Science in Healthcare Research in the European Union ». *Bmc Medical Ethics* 18 (avril): 27. https://doi.org/10.1186/s12910-017-0184-y.

Scherer, Matthew U. 2015. « Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies ». SSRN Scholarly Paper ID 2609777. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=2609777.

Schwab, Klaus. 2016. *La quatrième révolution industrielle*. Dunod. Suisse: World Economic Forum.

Sharon, Tamar. 2016. « The Googlization of Health Research: From Disruptive Innovation to Disruptive Ethics ». *Personalized Medicine* 13 (6): 563-74. https://doi.org/10.2217/pme-2016-0057.

Swan, Melanie. 2013. « The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery ». *Big Data* 1 (2): 85-99. https://doi.org/10.1089/big.2012.0002.

Torkamani, Ali, Kristian G. Andersen, Steven R. Steinhubl, et Eric J. Topol. 2017. « High-Definition Medicine ». *Cell* 170 (5): 828-43. https://doi.org/10.1016/j.cell.2017.08.007.

Toulmin, Stephen. 1981. « The Tyranny of Principles ». *The Hastings Center Report* 11 (6): 31-39. https://doi.org/10.2307/3560542.

Woolley, J. Patrick. 2016. « How Data Are Transforming the Landscape of Biomedical Ethics: The Need for ELSI Metadata on Consent ». Dans *The Ethics of Biomedical Big Data*, édité par Brent Daniel Mittelstadt et Luciano Floridi, 171-97. Law, Governance and Technology Series. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-33525-4_8.

Zak, Paul J., Robert Kurzban, et William T. Matzner. 2004. « The Neurobiology of Trust ». *Annals of the New York Academy of Sciences* 1032 (1): 224-27. https://doi.org/10.1196/annals.1314.025.

Zak, Paul J., Angela A. Stanton, et Sheila Ahmadi. 2007. « Oxytocin Increases Generosity in Humans ». *PLOS ONE* 2 (11): e1128. https://doi.org/10.1371/journal.pone.0001128.

Zarsky, Tal. 2016. « The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making ». *Science, Technology, & Human Values* 41 (1): 118-32. https://doi.org/10.1177/0162243915605575.

Annexes

Annexe 1 : Approbation éthique

Université de Montréal

N^O de certificat
18-106-CERES-D(1)

Comité d'éthique de la recherche en sciences et en santé (CERSES)

CERTIFICAT D'APPROBATION ÉTHIQUE

- 1er renouvellement -

Le Comité d'éthique de la recherche en sciences et en santé (CERSES), selon les procédures en vigueur et en vertu des documents relatifs au suivi qui lui a été fournis conclu qu'il respecte les règles d'éthique énoncées dans la Politique sur la recherche avec des êtres humains *de l'Université de Montréal.*

Projet	
Titre du projet	**Systèmes d'intelligence artificielle et santé : les enjeux d'une innovation responsable. Une analyse des craintes et attentes citoyennes face aux défis de l'exercice de la responsabilité**
Étudiante requérante	**Nathalie Voarino** (ND), Candidate au Ph. D. en sciences biomédicales (option bioéthique), École de santé publique - Département de médecine sociale et préventive
Sous la direction de	Béatrice Godard, professeure titulaire, École de santé publique - Département de médecine sociale et préventive, Université de Montréal & Ghislaine Cleret de Langavant, professeure associée, École de santé publique - Département de gestion, d'évaluation et de politique de santé, Université de Montréal.
Note :	Changement du titre du projet; Changement du directeur de recherche de Bryns Williams-Jones à Béatrice Godard; Octroi d'une bourse de fin d'études doctorales par la FESP (16 septembre 2019)
Financement	
Organisme	Faculté des études supérieures et post doctorales de l'Université de Montréal
Programme	Bourse de fin d'études doctorales
Titre de l'octroi si différent	
Numéro d'octroi	
Chercheur principal	
No de compte	

MODALITÉS D'APPLICATION

Tout changement anticipé au protocole de recherche doit être communiqué au Comité qui en évaluera l'impact au chapitre de l'éthique. Toute interruption prématurée du projet ou tout incident grave doit être immédiatement signalé au Comité.

Selon les règles universitaires en vigueur, un suivi annuel est minimalement exigé pour maintenir la validité de la présente approbation éthique, et ce, jusqu'à la fin du projet. Le questionnaire de suivi est disponible sur la page web du Comité.

16 septembre 2019
Date de délivrance du renouvellement ou de la réémission*

1er octobre 2020
Date du prochain suivi

23 juillet 2018
Date du certificat initial

1er octobre 2020
Date de fin de validité

3
A2

Annexe 2 : Informations sur les tables de coconstruction analysées

	Type d'activité et date	Durée	Nombre de participants	Scénarios utilisés
Table 1	Café citoyen 03-03-2018	1h30	9	Les jumeaux numériques L'assurance santé discriminante
Table 2	Café citoyen 10-03-2018	1h30	6 (+1 participant de la table 1)	Un robot pour maintenir une personne âgée à domicile
Table 3	Café citoyen 17-03-2018	1h30	5 (+ 1 participant de la table 1)	L'assurance santé discriminante
Table 4	Café citoyen 24-03-2018	1h30	6	Décision thérapeutique à l'hôpital
Table 5	Café citoyen 25-03-2018	1h30	8	Décision thérapeutique à l'hôpital
Table 6	Journée de co-construction 13-03-2018	3h	9	Les jumeaux numériques
Table 7*	Journée de co-construction 13-03-2018	1h30	7 (+ 1 participante de la table 1)	Un robot pour maintenir une personne âgée à domicile
Table 8	Journée de co-construction 06-04-2018	3h	9	Les jumeaux numériques
Table 9	Groupe de discussion 29-03-2018	1h30	9	Les jumeaux numériques L'assurance santé discriminante Décision thérapeutique à l'hôpital Un robot pour maintenir une personne âgée à domicile
Total		16h30	68	

*L'enregistrement de la moitié de l'activité seulement été disponible.

Annexe 3 : Scénarios

(Issus des travaux réalisés dans le cadre de la coconstruction de la Déclaration de Montréal pour un développement responsable de l'IA)

Thème 1 : Santé prédictive

Scénario de départ : Les jumeaux numériques

10 mars 2025. Olivier reçoit une notification sur son téléphone lui indiquant qu'un de ses jumeaux numériques vient de recevoir un diagnostic de dépression.

Des jumeaux numériques sont des personnes qui partagent les mêmes caractéristiques biologiques et qui ont des profils de santé similaires. Toutes les données relatives à la santé d'Olivier sont collectées par Santé Canada depuis décembre 2023. Certaines proviennent de l'application santé de son téléphone - comme le nombre de pas qu'il effectue chaque jour ou ses heures de sommeil - et de ce qu'il partage publiquement sur les réseaux sociaux - données rachetées aux compagnies Alphabet et Baidu. Elles sont croisées avec les données qui proviennent directement du système de santé concernant son historique de maladies et ses prédispositions génétiques. Ces données sont mises en relation avec celles de l'ensemble de la population dans le « nuage de santé mondial », piloté depuis 2023 par l'Organisation mondiale de la Santé, qui permet de définir les profils de santé des individus, afin d'offrir à chacun une médecine de précision et une prévention ciblée et hautement personnalisée.

Olivier découvre donc ce matin-là qu'il est susceptible de développer la même pathologie qu'un de ses 126 jumeaux de santé numériques. Face à l'annonce de ce pronostic, l'algorithme de Santé Canada fait les recommandations suivantes à Olivier :

- Se rendre dans un centre spécialisé en santé mentale afin de recevoir un traitement préventif adapté;
- Diminuer sa charge de travail à moins de 40 heures par semaine;
- Augmenter son activité physique, en concordance avec les études sur les effets bénéfiques du sport sur la prévention de la dépression.

Olivier décide d'ignorer ces recommandations, car il travaille présentement à un contrat particulièrement déterminant pour sa carrière. Cependant, au cours des mois suivants, il apprend que 25 de ses jumeaux numériques ont reçu un diagnostic similaire.

Thème 1 : Santé prédictive

Scénario de départ : Un robot pour maintenir une personne âgée à domicile

12 janvier 2025. La famille Couture vient d'acquérir Vigilo, un robot empathique et sympathique d'assistance aux personnes âgées qui souffrent de troubles cognitifs, pour leur grand-mère Soline, 83 ans, atteinte de la maladie d'Alzheimer dans un stade précoce.

Ils sont tous très heureux de faire entrer Vigilo dans la famille. Grâce à lui, Soline va pouvoir demeurer à son domicile le plus longtemps possible et ne pas emménager immédiatement dans une maison de retraite, ce qui semblait inévitable au vu de sa perte croissante d'autonomie. De plus, la famille pourra maintenant espacer ses visites à Soline, qui vit dans un petit village de campagne, car Vigilo leur transmet quotidiennement un rapport sur l'état de santé de leur grand-mère.

Vigilo évalue l'évolution de la démence de Soline en lui administrant régulièrement différents tests neuropsychologiques. Il s'occupe également de préparer les piluliers de Soline et s'assure qu'elle prend bien ses médicaments en temps voulu, grâce à sa caméra intégrée. Il transmet ensuite l'information au personnel de soins en charge de son dossier.

Grâce à ses capteurs, Vigilo peut percevoir les émotions de Soline sur la base des expressions de son visage et réagir de manière appropriée. Il est également capable d'entretenir des conversations suivies avec elle, lui donner des conseils et lui rappeler des choses oubliées. L'aide familiale de Soline ne se rend à son domicile que deux fois par semaine pour faire sa toilette et lui préparer des repas.

Au fil des mois, Soline se sent de plus en plus proche de Vigilo, et elle se confie à lui. Soline n'échange désormais avec sa famille que par le biais de Vigilo. De plus, il semble qu'elle soit de moins en moins encline à échanger avec d'autres personnes, comme ses voisins et son aide familiale avec qui elle a de plus en plus tendance à se disputer. La discorde augmente à un point tel que la famille de Soline doit rechercher une nouvelle aide familiale après seulement deux mois.

Vigilo possède une IA entraînée sur de nombreux patients pour apprendre à détecter et à prédire des pertes de capacité cognitives ou des situations de dépression. En cas d'alerte, Vigilo peut même envoyer de courts extraits de ses conversations avec Soline, avec le risque de révéler des facettes de son intimité ou des secrets qu'elle souhaiterait peut-être ne pas dévoiler. Caroline, la fille de Soline, a accès à ces rapports et extraits de conversation et s'inquiète de ce qu'elle découvre.

Thème 1 : Santé prédictive

Scénario variante : Décision thérapeutique à l'hôpital

12 janvier 2026. Antoine, le petit-fils de Soline, amène sa grand-mère de 84 ans à l'hôpital pour réaliser une série d'examens médicaux. Le personnel de l'hôpital a invité Soline à venir en consultation car Vigilo, son robot de soins, a signalé, parmi différentes anomalies dans ses paramètres de santé depuis quelques semaines, que la patiente avait été victime d'un malaise vagal.

Soline et Antoine arrivent donc au Centre hospitalier. Après que l'algorithme de triage ait récolté l'ensemble des informations de Soline en quelques secondes en lisant son *insigne santé*, Soline et Antoine sont dirigés par le robot d'assistance vers la salle d'attente du Dr. Khan, en cardiologie. Antoine est impatient de rencontrer le médecin, car il a de nombreuses questions à lui poser sur l'état de santé de sa grand-mère. Questions auxquelles l'algorithme de conversation du site internet de Santé Canada n'a pu répondre.

Les examens de Soline sont entre les mains du Dr. Khan, expert cardiologue de renommée internationale. Le Dr. Khan exerce depuis plus de 25 ans; il a réussi à traiter de nombreux cas particulièrement difficiles qui semblaient insolubles au premier abord, et a suivi plus de 10 000 patients dans sa carrière.

Depuis 2024, l'ensemble de la profession médicale de l'hôpital passe par un algorithme d'aide au diagnostic. L'algorithme n'est là que pour soutenir le médecin quant au diagnostic et au traitement du patient. Le Dr. Khan a cependant du mal à se prononcer sur le cas de Soline : selon son analyse, il est clair qu'elle souffre d'un trouble cardiaque qui s'accompagne d'un risque de syncope élevé et qui nécessite la pose d'un *pacemaker*. Cependant, l'algorithme indique les résultats suivants :

97,23 % Bradycardie sinusale[147] légère

1,72 % Bloc auriculo-ventriculaire supra-hissien du premier degré[148]

1,05 % Trouble cardiaque avec risque de syncope élevé

Selon les résultats émis par l'algorithme, la pose d'un *pacemaker* semble inutile.

L'algorithme en question, considéré comme extrêmement fiable lors de la dernière conférence des cardiologues du Canada, a posé plus de 25 millions de diagnostics à ce jour et ses prédictions ne se sont pratiquement jamais avérées fausses. Mais le Dr. Khan est plus que persuadé de son analyse personnelle. Cependant, il ne connait pas le chemin qu'a emprunté l'algorithme pour donner ses prévisions et il sait qu'il n'est pas à l'abri de commettre lui-même une erreur. Il se demande s'il doit en informer sa patiente, au risque de l'inquiéter et d'initier inutilement une thérapie, ou ne pas suivre les recommandations de l'algorithme et la renvoyer chez elle.

Thème 1 : Santé prédictive

Scénario variante : L'assurance santé discriminante[149]

15 novembre 2025. Olivier reçoit une notification sur son téléphone intelligent lui indiquant qu'il lui manque trois heures d'activité physique cette semaine pour que les conditions de son contrat d'assurance santé privée complémentaire soient respectées. Si Olivier ne respecte pas les conditions, sa franchise sera majorée de 10 %.

La semaine dernière, les examens d'Olivier ont révélé la présence d'une tumeur, bénigne mais à surveiller, sur son poumon droit. Il sait que les frais associés à un éventuel traitement, s'il développait un cancer, ne seront pas complètement remboursés par son assurance complémentaire car il cumule deux facteurs de risque qui engagent sa responsabilité : il fume un paquet de cigarettes par jour depuis plus de 10 ans et il ne pratique pas de sport régulièrement. En maintenant ces comportements, Olivier va à l'encontre des recommandations de Santé Canada émises depuis son application santé.

[147] Trouble relativement bénin assez fréquent chez les sportifs et les personnes âgées, caractérisé par un rythme cardiaque anormalement lent

[148] Défaut de la conduction cardiaque aux conséquences moins importantes que la pathologie décelée par le Dr. Khan et qui ne nécessite pas d'intervention lourde.

[149] Ce scénario, s'il est incarné par un Québécois ou un Canadien, part d'une hypothèse que le système de santé a subi une réforme de la couverture universelle des services médicalement requis (conformément aux lois en vigueur en 2018 au Canada et au Québec) vers une couverture à la fois du secteur public et du secteur privé, par le biais d'assurances contractées par les individus pour des services couverts par le secteur public. Ceci étant précisé, les questions éthiques demeurent les mêmes.

En effet, afin d'éviter les discriminations liées au genre, à l'origine ethnique ou au statut social, et pour faire face au problème croissant d'allocation des ressources en santé, la priorité d'accès aux soins est déterminée par un algorithme. Cet algorithme est programmé pour respecter différents paramètres objectifs, comme la gravité de la maladie, la probabilité que le traitement soit efficace, les comportements de santé des patients et les conduites à risque ayant des conséquences sur l'efficacité du traitement.

Les derniers mois ont été éprouvants pour Olivier qui n'a pas trouvé le courage de s'adonner à une activité physique et qui ne parvient toujours pas à arrêter de fumer. Même si sa situation financière lui permet d'assumer la majoration de 10 %, il trouve cette situation particulièrement injuste. Il décide donc de préserver la confidentialité de ses données d'activité physique et de ne plus les partager avec le *Nuage de santé mondial*, afin que son assurance privée ne majore pas sa franchise. Il sait toutefois que cette décision l'expose au risque d'être éventuellement défavorisé au moment de s'inscrire à une liste d'attente pour accéder à des soins de santé.

Annexe 4 : Informations sur le recrutement

(Issus des travaux réalisés dans le cadre de la coconstruction de la Déclaration de Montréal pour un développement responsable de l'IA)

Recrutement

1- Parties prenantes (professionnels des secteurs)

Les professionnels des secteurs ont été joints par courriel, de façon ciblée et personnalisée, via nos réseaux de contacts que sont nos centres de recherche, ou chercheurs, qui travaillent avec des professionnels de secteurs, ou des organismes comme la Ville de Montréal, par exemple. Parmi ces centres :

- CRDP – juristes
- ESPUM – santé publique
- CRIMT – travail et économie
- CIRANO – travail et économie
- CRDM – techno et santé
- CEPPP - santé
- IVADO et MILA

Des parties prenantes ont également été recrutées à partir des recherches réalisées par secteur, par les auxiliaires de recherche.

2- Experts académiques

Les professeurs et chercheurs ont également été joints par courriel, via le réseau UdeM – Facultés de md, infirmière, arts et sciences, sciences des données, droit - mais aussi via les différents Centres de recherche interuniversitaires, notamment :

- CRÉ
- CIRANO
- CRIMT
- IVADO et MILA

Des experts académiques ont également été recrutés à partir des recherches réalisées par domaine, par les auxiliaires de recherche.

3- Citoyens

Les citoyens, quant à eux, ont été recrutés plus largement en utilisant les moyens suivants :

- Site web de la Déclaration
- Twitter de la Déclaration
- Publicité Facebook (compte de l'UdeM)
- Infolettre de Sainte-Julie
- Facebook des différentes bibliothèques participantes
- Affiches dans les bibliothèques participantes
- Site web de la Ville de Montréal
- Site web de la Ville de Québec
- Site web du Musée de la civilisation à QC

Le concours de Marie Martel, professeure à l'École de bibliothéconomie et des sciences de l'information, à l'Université de Montréal et chercheure au Centre de recherche Design Société de l'UdeM, fut crucial pour le recrutement initial des bibliothèques.

Annexe 5 : Questionnaire sociodémographique

(Issu des travaux réalisés dans le cadre de la coconstruction de la Déclaration de Montréal pour un développement responsable de l'IA)

Questions sociodémographiques

Vous êtes :

☐ Un homme (H) ☐ Une femme (F) ☐ Genre neutre (X)

Votre âge :

☐ 12 ans ou moins ☐ 35-44 ans ☐ 65-74 ans
☐ 13-18 ans ☐ 45-54 ans ☐ 75+ ans
☐ 19-34 ans ☐ 55-64 ans

Votre plus haut niveau de scolarité atteint :

☐ Aucun certificat, diplôme ou grade
☐ Diplôme d'études secondaires ou l'équivalent
☐ Titre d'études postsecondaire (certificat ou diplôme d'une école de métiers, certificat d'apprenti inscrit)
☐ Diplôme d'études collégiales
☐ Certificat universitaire inférieur au baccalauréat
☐ Baccalauréat
☐ Certificat universitaire supérieur au baccalauréat
☐ Diplôme en médecine
☐ Doctorat acquis

Dans quel secteur d'activité œuvrez-vous? (Plusieurs réponses possibles)

☐ Administration publique ☐ Arts, spectacles et loisirs

- ☐ Commerce de détail
- ☐ Construction
- ☐ Énergie et ressources
- ☐ Enseignement
- ☐ Finance et assurances
- ☐ Gestion de sociétés et d'entreprises
- ☐ Hébergement et services de restauration
- ☐ Information et culture
- ☐ Recherche (industrielle ou universitaire)
- ☐ Services administratifs, services de soutien
- ☐ Services professionnels, scientifiques et techniques
- ☐ Soins de santé, biotechnologies et assistance sociale
- ☐ Technologies de l'information
- ☐ Transport et entreposage
- ☐ Autre : _____