L'IMPACT DE L'INTELLIGENCE ARTIFICIELLE SUR LES RELATIONS MEDECIN-PATIENT

Rapport commandé par le Comité directeur pour les droits de l'Homme dans les domaines de la biomédecine et de la santé (CDBIO)

Auteur: Brent Mittelstadt

COUNCIL OF EUROPE
CONSEIL DE L'EUROPE

L'IMPACT DE L'INTELLIGENCE ARTIFICIELLE SUR LES RELATIONS MEDECIN-PATIENT *

Par Brent Mittelstadt, Chercheur principal et directeur de recherche à l'Oxford Internet Institute, Université d'Oxford, Royaume-Uni.

* traduction non vérifiée

TABLE DES MATIERES

1. ELEMENTS ESSENTIELS

1. En réponse à un appel du Comité de bioéthique (DH-BIO)[1] à travailler sur la confiance, la sécurité et la transparence, ce rapport étudie les impacts connus et potentiels des systèmes d'IA sur la relation médecin-patient. Cet impact est encadré par les principes des droits de l'homme mentionnés dans la Convention européenne sur les droits de l'homme et la biomédecine de 1997, autrement connue sous le nom de « Convention d'Oviedo », et ses amendements ultérieurs.

2. La mise en œuvre de l'IA dans le domaine des soins cliniques n'en est encore qu'à ses balbutiements. Une efficacité clinique n'a été établie que pour relativement peu de systèmes par rapport aux nombreuses activités de recherche menées sur les applications de l'IA en médecine. Bien souvent, les travaux de recherche, le développement et les essais pilotes ne se traduisent pas par une efficacité clinique avérée, une mise sur le marché ou un déploiement massif. Globalement, la généralisation de la performance des essais cliniques à la pratique reste à démontrer

3. L'une des caractéristiques de la médecine est la "relation thérapeutique" entre les cliniciens et les patients. Cette relation est renforcée par l'introduction de l'IA. Cependant, le rôle du patient, les facteurs qui poussent les gens à consulter un médecin et la vulnérabilité du patient ne sont pas modifiés par l'introduction de l'IA en tant que médiateur ou fournisseur de soins médicaux. Ce qui change plutôt, ce sont les moyens de fournir des soins, la manière dont ils peuvent être fournis et par qui. Le transfert de l'expertise et des responsabilités en matière de soins vers des systèmes d'IA peut être perturbant à bien des égards.

4. L'impact potentiel de l'IA sur les droits de l'homme dans la relation médecin-patient peut être classé selon six thèmes : (1) Inégalité dans l'accès à des soins de santé de qualité ; (2) Transparence vis-à-vis des professionnels de la santé et des patients ; (3) Risque de biais social dans les systèmes d'IA ; (4) Dilution de la prise en compte du bien-être du patient, (5) Risques de biais d'automatisation, de perte de compétences et de déplacement de la responsabilité ; (6) Conséquences sur le droit à la vie privée

5. Concernant le point (1), le déploiement des systèmes d'IA – en tant que technologie émergente – ne sera ni immédiat ni universel dans tous les états membres ou systèmes de santé. L'ampleur du déploiement, sa vitesse, mais aussi la hiérarchisation des priorités sera inévitablement hétérogène dans les établissements et les régions.

6. Les conséquences de l'IA sur les soins cliniques et la relation médecin-patient demeurent incertaines et varieront certainement selon les technologies et les situations. Les systèmes d'IA peuvent s'avérer plus efficaces que les soins prodigués par des êtres humains, mais aussi fournir des soins de moindre qualité avec moins d'interactions personnelles.

[1] Remplacé depuis par le Comité directeur pour les droits de l'Homme dans les domaines de la biomédecine et de la santé (CDBIO).

7. Le déploiement hétérogène des systèmes d'IA, dont les conséquences sur l'accès et la qualité des soins sont incertaines, risque de créer de nouvelles inégalités en matière de santé dans les états membres.

8. L'article 4 de la Convention d'Oviedo renvoie à l'obligation qu'ont les professionnels de la santé de prodiguer des soins en respectant des normes professionnelles. Or, il n'est pas clair si les développeurs, les fabricants et les fournisseurs de services pour les systèmes d'IA seront tenus de respecter les mêmes normes professionnelles.

9. Au moment d'intégrer des systèmes d'IA qui interagissent directement avec les patients, il convient d'examiner attentivement le rôle joué par les professionnels de la santé, lesquels sont tenus de respecter des normes professionnelles.

10. Concernant le point (2), la transparence et le consentement éclairé sont des valeurs clés dans la relation médecin-patient médiée par l'IA. La complexité de l'IA soulève une question : comment les systèmes d'IA doivent-ils s'expliquer, ou être expliqués, aux médecins et aux patients ? Cette question a de nombreuses significations possibles : (i) Comment fonctionne un système ou un modèle d'IA ? Comment un système d'IA produit-il un résultat spécifique ? (ii) Comment un système d'IA a-t-il été conçu et testé ? Comment est-il régi ? (iii)Quelles informations sont nécessaires pour enquêter sur le comportement des systèmes d'IA ? Les réponses à chacune de ces questions peuvent être nécessaires pour obtenir un consentement éclairé dans le cadre de soins facilités par l'IA.

11. Dans les cas où les systèmes d'IA fournissent une certaine forme d'expertise clinique, par exemple en recommandant un diagnostic particulier ou en interprétant des images médicales, l'obligation d'expliquer la décision serait apparemment transférée du médecin au système d'IA, ou du moins au fabricant du système d'IA. Les difficultés à expliquer comment les systèmes d'IA transforment les données d'entrée en données de sortie posent un défi épistémologique fondamental pour le consentement éclairé. Si l'on met de côté la capacité du patient à comprendre les fonctionnalités des systèmes d'IA, dans de nombreux cas, les patients n'ont tout simplement pas un niveau de connaissances suffisant pour donner un consentement libre et éclairé. Les systèmes d'IA utilisent des volumes de données sans précédent pour prendre leurs décisions et interprètent ces données à l'aide de techniques statistiques complexes. Dès lors, il est de plus en plus difficile de mesurer l'ampleur du traitement des données utilisées pour les diagnostics et les traitements.

12. Les systèmes d'IA qui interagissent directement avec les patients devraient leur indiquer qu'ils sont des systèmes artificiels. La question de savoir si l'utilisation de systèmes d'IA dans le secteur de la santé doit être toujours communiquée aux patients par les cliniciens et les établissements de santé est plus difficile.

13. Concernant (3), les systèmes d'IA sont largement reconnus comme souffrant de biais dans leurs entrées, leur traitement et leurs sorties. Dans les systèmes d'IA, les décisions biaisées et injustes ne sont souvent pas le produit de raisons techniques ou réglementaires, mais traduisent plutôt des biais sociaux et les inégalités sociales sous-jacentes. Par exemple, les échantillons des essais cliniques et des études sur la santé ont toujours été biaisés en faveur des

hommes blancs, ce qui signifie que les résultats sont moins susceptibles de s'appliquer aux femmes et aux personnes de couleur.

14. Les biais sociaux peuvent conduire à une distribution inégale des résultats entre les populations ou les groupes démographiques protégés. Les sociétés occidentales ont longtemps été marquées par d'importantes inégalités sociales. Ces tendances historiques et contemporaines influencent la formation des futurs systèmes. Sans intervention, les systèmes d'IA apprendront et renforceront ces modèles préexistants qui favorisent l'inégalité des chances et l'inégal accès aux ressources dans la société.

15. La détection des biais dans les systèmes d'IA n'est pas simple. Des règles biaisées en matière de prise de décision peuvent être cachées dans des modèles de type « boîte noire ». La simple anonymisation des données de santé peut ne pas être une solution adéquate pour atténuer les biais en raison de l'influence de l'inégalité historique et de l'existence de substituts puissants pour les attributs protégés (par exemple, le code postal comme substitut de l'ethnicité). À en juger par ces divers problèmes de biais social, de discrimination et d'inégalité, les professionnels et les établissements de santé font face à une tâche difficile, à savoir veiller à ce que leur utilisation des systèmes d'IA n'aggrave pas les inégalités existantes et ne crée pas de nouvelles formes de discrimination.

16. Concernant (4), le développement de la confiance dans une relation médecin-patient peut être inhibé par la médiation technologique. En tant que médiateur placé entre le médecin et le patient, les systèmes d'IA peuvent empêcher la compréhension tacite de la santé et du bien-être du patient, et encourager le clinicien et le patient à discuter de la santé uniquement en termes de quantités mesurables ou interprétables par une machine.

17. Concernant (5), pour assurer la sécurité des patients et remplacer la protection offerte par l'expertise clinique humaine, des normes de test et de validation robustes sont nécessaires dans le pré-déploiement des systèmes d'IA dans le contexte de soins cliniques. Il n'existe pas encore de preuves de l'efficacité clinique de nombreuses technologies d'IA dans le domaine des soins de santé, ce qui a constitué, à juste titre, un obstacle à leur déploiement à grande échelle.

18. Concernant (6), l'IA pose plusieurs problèmes spécifiques liés au droit à la vie privée et aux réglementations complémentaires en matière de protection des données.). Ces droits visent à faire bénéficier les individus d'une plus grande transparence et d'un meilleur contrôle sur les formes automatisées de traitement des données. Ils apporteront sans aucun doute une protection précieuse aux patients dans toute une série d'utilisation de l'IA médicale.

19. La Convention d'Oviedo prévoit une application spécifique du droit à la vie privée (article 8 de la CEDH) qui reconnaît la nature particulièrement sensible des informations personnelles relatives à la santé et établit un devoir de confidentialité pour les professionnels de la santé.

20. Des normes éthiques doivent s'appuyer sur la transparence, les biais, la confidentialité et l'efficacité clinique afin de protéger les intérêts des patients en matière de consentement éclairé, d'égalité, de vie privée et de sécurité. De telles normes pourraient servir de base à un déploiement de l'IA dans le secteur

de la santé qui favoriserait la relation de confiance entre les médecins et les patients plutôt que de l'entraver.

21. Dans les cas où on observe un impact évident de l'IA sur les droits et les protections exposés dans la convention d'Oviedo, il est opportun que le Conseil de l'Europe présente des recommandations et des exigences contraignantes pour les signataires concernant la façon d'utiliser et de gérer l'IA. Les recommandations devraient se concentrer sur un niveau plus élevé de soins positifs en ce qui concerne la relation médecin-patient, afin de s'assurer qu'elle ne soit pas exagérément perturbée par l'introduction de l'IA dans les environnements de soins.

22. Le Conseil de l'Europe pourrait établir des normes sur le contenu et la manière dont les informations sur la recommandation d'un système d'IA concernant le diagnostic et le traitement d'un patient devraient être communiquées au patient. Ces normes devraient elles aussi aborder le rôle que joue le médecin pour expliquer aux patients les recommandations en matière d'IA et la façon dont les systèmes d'IA peuvent être conçus pour l'accompagner dans ce rôle.

23. La capacité de l'IA à remplacer ou à augmenter l'expertise clinique humaine par des analyses très complexes sur des données d'une ampleur et d'une diversité sans précédent, pourrait bien modifier la relation médecin-patient dans des proportions jamais atteintes auparavant.

24. La mesure dans laquelle un système d'IA fait obstacle à la « bonne » pratique de la médecine dépend du modèle de service. Si l'IA vient seulement compléter l'expertise des professionnels de la santé, lesquels sont liés par le devoir de loyauté vis-à-vis à du patient, ses effets sur la fiabilité et la qualité humaine des entretiens cliniques peuvent se révéler minimes. D'un autre côté, dans le cas où l'IA est utilisée pour étoffer largement l'expertise clinique humaine ou pour la remplacer, son impact sur la relation de soins est plus difficile à prévoir. Avec le recours croissant aux systèmes d'IA, de nouvelles normes, largement admises, en matière de « bons » soins verront sans doute le jour, les cliniciens passant plus de temps en face à face avec leurs patients tout en s'appuyant largement sur des recommandations issues de systèmes automatiques. L'impact de l'IA sur la relation médecin-patient reste très incertain. Il est peu probable que nous assistions dans les cinq prochaines années à une reconfiguration radicale des soins, c'est-à-dire à un remplacement de l'expertise humaine par l'intelligence artificielle.

25. Une reconfiguration radicale de la relation médecin-patient telle que certains l'imaginent, où des systèmes artificiels diagnostiqueraient et traiteraient les patients directement, les cliniciens humains intervenant a minima, reste, semble-t-il, une perspective lointaine.

26. À l'avenir, le modèle idéal de soins cliniques et de déploiement de l'IA dans les soins de santé est celui qui utilise les meilleurs aspects de l'expertise clinique humaine et des diagnostics de l'IA.

27. La relation médecin-patient est une pierre angulaire de la « bonne » pratique médicale, et pourtant, elle semble évoluer vers une relation médecin-patient-IA. Le défi auquel sont confrontés les fournisseurs d'intelligence artificielle, les autorités de réglementation et les décideurs est de définir des normes et des

exigences solides pour ce nouveau type de relation thérapeutique, afin que les intérêts des patients et l’intégrité morale de la médecine en tant que profession ne soient pas fondamentalement endommagés par l'introduction de l'IA.

2. INTRODUCTION

Les solutions technologiques telles que l'intelligence artificielle (IA) sont de plus en plus souvent considérées comme susceptibles d'apporter une réponse aux pressions croissantes qui pèsent sur les ressources dans les domaines de la médecine, des soins de santé et de la recherche biomédicale. Les systèmes d'IA promettent d'apporter des moyens innovants en matière d'évaluation et d'amélioration de la qualité des soins cliniques, de mise en œuvre de la recherche biomédicale et d'étude de thérapies et produits pharmaceutiques inédits, et d'extension de l'offre de soins à des populations auparavant mal desservies[2]. La conviction que l'IA pourrait soulager les professionnels de santé « de certaines tâches administratives fastidieuses et libérer du temps pour les soins de santé[3] » constitue l'un des principaux moteurs de l'innovation et de son adoption. Des systèmes experts et robotiques viennent à l'appui de la prise de décision et des soins médicaux, leur apportant une aide pour la gestion des dossiers et le catalogage des documents, le diagnostic, la planification des traitements et la réalisation des interventions. Des transformations similaires touchent les soins à domicile et les services de la protection sociale, avec la mise en place de systèmes de suivi et de gestion à distance. Complétant ou remplaçant de plus en plus souvent les comptes rendus verbaux et les soins physiques en face à face, des représentations des patients reposant sur des données numériques permettent le suivi, la modélisation et la gestion de la santé[4].

La portée très spécifique de l'IA et d'autres technologies émergentes, algorithmiques et à grand volume de données, consiste en leur capacité à compléter, améliorer et faciliter la prise de décision par un être humain, en recommandant la meilleure mesure à prendre dans une situation donnée, la meilleure interprétation des données, etc.[5] Mais ces systèmes peuvent aussi être utilisés pour remplacer purement et simplement la prise de décision humaine, l'expertise et les soins cliniques en face à face. Les applications de traitement du langage naturel telles que GPT-3 d'OpenAI, par exemple, laissent entrevoir un avenir dans lequel le contact initial avec le patient et même le triage pourront être partiellement ou entièrement traités par des agents conversationnels artificiels. Des systèmes d'IA sont déjà utilisés par des cliniciens et des hôpitaux dans la prise de décisions cliniques et opérationnelles, par exemple dans la prédiction des risques, la planification des sorties, les diagnostics et les systèmes

[2] Organisation mondiale de la santé, *Éthique et gouvernance de l'intelligence artificielle pour la santé : orientations de l'OMS* (2021) ; Comité italien de bioéthique, *Artificial Intelligence and Medicine: Some Ethical Aspects* (2020), http://bioetica.governo.it/en/opinions/joint-opinions-icbicbbsl/artificial-intelligence-and-medicine-some-ethical-aspects/ (consulté le 30 nov. 2021).

[3] Conseil de l'Europe, *Intelligence artificielle et santé : défis médicaux, juridiques et éthiques à venir* (2020).

[4] Brent Mittelstadt *et al.*, « The Ethical Implications of Personal Health Monitoring », *International Journal of Technoethics* nº 5, p. 37-60 (2014).

[5] George A. Diamond, Brad H. Pollock et Jeffrey W. Work, « Clinician Decisions and Computers », *Journal of The American College of Cardiology* nº 9, p. 1385-1396 (1987) ; James G. Mazoué, « Diagnosis Without Doctors », *J MED PHILOS* nº 15, p. 559-579 (1990).

d'aide à la décision[6]. De même, les progrès de l'apprentissage profond laissent envisager un avenir dans lequel des systèmes informatiques capables d'adopter un comportement intelligent piloteront la mise au point de nouveaux médicaments et la recherche biomédicale[7]. Les récentes avancées réalisées dans le traitement pharmaceutique d'une forme rare de cancer du cerveau ou le progrès décisif enregistré par Deepmind dans le domaine du repliement des protéines avec son programme AlphaFold illustrent déjà le potentiel des techniques les plus récentes en matière d'IA médicale[8].

Si les promesses de l'IA sont évidentes, une importante zone d'incertitude n'en subsiste pas moins concernant ses répercussions sur la pratique des soins de santé, et en particulier sur la relation médecin-patient. L'expertise médicale n'est plus le domaine réservé de professionnels de santé diplômés et de chercheurs qualifiés ; au contraire, les technologies de l'IA offrent à tout un ensemble de parties prenantes – publiques et privées, professionnelles et non professionnelles, humaines et technologiques –, la possibilité de fournir des soins de santé.

Pour répondre, d'une part, à la prise de conscience croissante par le Conseil de l'Europe des perspectives offertes par l'IA sur la pratique de la médecine et des soins cliniques, mais aussi des risques qu'elle fait peser sur eux, et, d'autre part, à l'appel du Comité de bioéthique (DH-BIO) de travailler sur les questions de confiance, de sécurité et de transparence dans ce contexte[9], le présent rapport étudie les répercussions connues et potentielles des systèmes d'IA sur la relation médecin-patient. Ces répercussions sont encadrées par les principes des droits de l'homme mentionnés par la Convention européenne sur les droits de l'homme et la biomédecine (1997), également connu sous le nom de « Convention d'Oviedo », et ses amendements ultérieurs. Les principes des droits de l'homme relatifs à la santé peuvent exiger le respect de certaines normes dans les relations médecin-patient, qui peuvent être déstabilisées, supplantées ou du moins améliorées par le recours à l'IA en matière de prise en charge clinique.

Le rapport est structuré comme suit.

- **Le chapitre 2** rappelle l'historique et évoque le contexte des définitions de l'IA et des technologies connexes, des défis éthiques communs que représentent les systèmes d'IA, et donne un bref aperçu historique des principes des droits de l'homme en matière de santé dans le cadre de la Convention d'Oviedo.

[6] Rebecca Robbins et Erin Brodwin, « Patients Aren't Being Told About the AI Systems Advising Their Care », *Stat* (2020), https://www.statnews.com/2020/07/15/artificial-intelligence-patient-consent-hospitals/ (consulté le 9 nov. 2021).
[7] Organisation mondiale de la santé, cf. *supra* note 1.
[8] Diana M. Carvalho *et al.*, « Repurposing Vandetanib Plus Everolimus for the Treatment of ACVR1-mutant Diffuse Intrinsic Pontine Glioma », *Cancer Discov* (2021), https://cancerdiscovery.aacrjournals.org/content/early/2021/09/20/2159-8290.CD-20-1201 (consulté le 30 nov. 2021) ; John Jumper *et al.*, « Highly Accurate Protein Structure Prediction with AlphaFold », *Nature* nº 596, p. 583-589 (2021).
[9] Conseil de l'Europe, cf. *supra* note 2.

- **Le chapitre 3** passe en revue les types de technologies d'IA utilisées en médecine, en traitant plus particulièrement des systèmes d'IA qui visent à améliorer les soins cliniques et le vécu du patient.
- **Le chapitre 4** propose un cadre théorique applicable à la relation médecin-patient, fondé sur les droits de l'homme et rattachant les objectifs de la médecine aux normes de bonne pratique médicale élaborées par la médecine en tant que profession réglementée.
- **Le chapitre 5** distingue ensuite plusieurs catégories de conséquences actuelles et potentielles imputables aux systèmes d'IA et touchant la relation médecin-patient, en mettant en lumière les questions de biais, d'inégalité dans l'accès aux soins, d'opacité et de transparence, d'autonomie et de sécurité du patient, de responsabilité du clinicien et de biais d'automatisation, ainsi que le droit fondamental à la protection de la vie privée.
- **Le chapitre 6** vient conclure le présent rapport par des recommandations visant à renforcer la protection des droits de l'homme dans le contexte de l'IA et de la relation médecin-patient.

3. HISTORIQUE ET CONTEXTE

Des concepts comme l'intelligence artificielle (IA), l'apprentissage automatique, les algorithmes et les systèmes d'IA recèlent une grande diversité de significations selon qu'ils sont employés dans un contexte universitaire, politique ou public. On regrettera qu'ils soient souvent utilisés de façon interchangeable[10]. Dans un souci de clarté, quelques définitions et distinctions sont proposées ci-après.

Le concept d'intelligence artificielle fait référence à une intelligence telle qu'elle est manifestée par une machine, l'intelligence étant comprise dans l'optique de son expression chez les humains et les animaux. En tant que discipline universitaire, l'intelligence artificielle étudie les « agents intelligents » ou l'« intelligence informatique », entendus comme des systèmes qui perçoivent leur environnement et entreprennent des actions qui maximisent leurs chances d'atteindre leurs objectifs[11]. L'apprentissage automatique peut être envisagé comme un type spécialisé d'IA dans lequel l'agent, ou le programme informatique, améliore grâce à l'expérience ses performances dans la réalisation d'une tâche donnée. Les systèmes d'apprentissage automatique font appel à « des connaissances préalables ainsi que des données d'entraînement pour guider l'apprentissage[12] ».

Pour le dire simplement, on peut concevoir l'apprentissage automatique comme un type de logiciel qui apprend à partir d'un ensemble de données d'entraînement, dans lequel des étiquettes sont créées et appliquées par des étiqueteurs humains conformément à des connaissances préalables. Un exemple classique est fourni par les programmes de reconnaissance d'images à qui l'on apprend à distinguer des classes d'objets. Dans ce cas, l'ensemble de données d'entraînement consiste en une série d'images préétiquetées à partir de laquelle le système peut déduire des règles de classification à appliquer à de nouvelles images ou à de nouveaux ensembles de données.

Les algorithmes peuvent être considérés comme les composants essentiels des systèmes d'apprentissage automatique et d'intelligence artificielle qui guident les processus d'apprentissage et de transformation des données d'entrée en données de sortie. En termes mathématiques, on peut concevoir un algorithme comme une construction mathématique dotée d'« une structure de contrôle finie, abstraite, efficace, composée, qui est impérativement donnée et réalise un objectif donné en vertu de règles données[13] ». Dans un souci de clarté, nous proposerons une définition plus simple : un algorithme est une succession d'opérations bien définies qui produit en sortie un résultat à partir d'un ensemble d'éléments fournis en entrée.

Un algorithme d'apprentissage automatique peut être entendu comme un type d'algorithme dans lequel une partie de la succession d'opérations a été apprise plutôt que prédéfinie. Par exemple, un algorithme d'apprentissage automatique utilisé pour réaliser des tâches de classification élabore des classes qui peuvent être généralisées

[10] Robin K. Hill, « What an Algorithm Is », *Philos. Technol.* nº 29, p. 35-59, 36 (2015).
[11] David Poole, Alan Mackworth et Randy Goebel, *Computational Intelligence* (1998).
[12] Tom Mitchell, *Machine learning* (1997).
[13] Hill, cf. *supra* note 9, p. 47.

au-delà des données d'entraînement[14]. L'algorithme crée un modèle en vue de classer de nouvelles données d'entrée. Un modèle d'apprentissage automatique est formé des données internes de l'algorithme qui sont adaptées aux données d'entrée afin d'améliorer les performances.

Les technologies de reconnaissance d'images, par exemple, peuvent décider quels types d'objets apparaissent dans une image. L'algorithme « apprend » en définissant des règles permettant de déterminer comment seront classées les nouveaux éléments fournis en entrée. Le modèle peut être enseigné à l'algorithme par des entrées étiquetées manuellement (apprentissage supervisé) ; dans d'autres cas, l'algorithme lui-même définit les modèles les mieux adaptés pour interpréter un ensemble de données d'entrées (apprentissage non supervisé)[15]. Dans les deux cas, l'algorithme définit des règles de prise de décision pour traiter les nouveaux éléments fournis en entrée. Un utilisateur humain n'est ordinairement pas en mesure de comprendre de manière critique la logique qui préside aux règles de prise de décision produites par l'algorithme[16].

Les définitions grand public et politique de ces termes ne se conforme pas à ces définitions formelles, ce qui peut être source de confusion. L'Organisation mondiale de la Santé (OMS), par exemple, définit l'intelligence artificielle comme « l'exécution par des programmes informatiques de tâches généralement associées à des êtres intelligents ». Les définitions de ce type se révèlent problématiques car trop générales, dans la mesure où elles reposent sur la définition de l'« intelligence » et de l'étendue des comportements des « êtres intelligents », et ne peuvent donc pas être utilisées en tant que telles pour classer un système particulier soit comme IA soit comme non-IA uniquement. Cela dit, le caractère ouvert de la définition peut également se révéler utile sur le plan politique, en permettant la prise en compte de systèmes supplémentaires allant au-delà des plus récentes réalisations lors de la rédaction de la première version.

Quelles que soient leurs limites, les définitions politiques de l'IA sont sans doute plus importantes que les définitions techniques si l'on se soucie de l'harmonisation des cadres réglementaires et politiques. La législation sur l'intelligence artificielle (« Artificial Intelligence Act »), un cadre réglementaire horizontal fondé sur les risques proposés par la Commission européenne, donne une définition particulièrement large de l'IA qui promet d'infléchir la politique internationale à l'avenir[17] :

> « On entend par "système d'intelligence artificielle" (système d'IA) un logiciel qui est développé au moyen d'une ou plusieurs des techniques et approches

[14] Pedro Domingos, « A Few Useful Things to Know About Machine Learning », *Communications of the ACM*, nº 55, p. 78-87 (2012).
[15] Bart W. Schermer, « The Limits of Privacy in Automated Profiling and Data Mining », *Computer Law & Security Review* nº 27, p. 45-52 (2011) ; Martijn Van Otterlo, « A Machine Learning View on Profiling », *Privacy, Due Process and The Computational Turn – Philosophers of Law Meet Philosophers of Technology,* p. 41-64 (2013).
[16] Andreas Matthias, « The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata », *Ethics Inf Technol* nº 6, p. 175-183, 179 (2004).
[17] Commission européenne, *Proposition de Règlement du parlement européen et du conseil établissant des règles harmonisées concernant l'intelligence artificielle (législation sur l'intelligence artificielle) et modifiant certains actes législatifs de l'union*, COM/2021/0106 final (2021), https://eur-lex.europa.eu/legal-content/en/txt/?uri=celex%3a52021pc0206 (consulté le 27 oct. 2021).

> énumérées à l'annexe I et qui peut, pour un ensemble donné d'objectifs définis par l'homme, générer des résultats tels que des contenus, des prédictions, des recommandations ou des décisions influençant les environnements avec lesquels il interagit. »

L'annexe à la proposition de législation sur l'intelligence artificielle comporte une liste non exhaustive de techniques et de moyens pouvant être considérés comme relevant de l'IA, qui englobe l'apprentissage automatique, les voies d'approches fondées sur la logique et les connaissances, ainsi que diverses méthodes statistiques :

> « (a) Approches d'apprentissage automatique, y compris d'apprentissage supervisé, non supervisé et par renforcement, utilisant une grande variété de méthodes, y compris l'apprentissage profond ;
>
> (b) Approches fondées sur la logique et les connaissances, y compris la représentation des connaissances, la programmation inductive (logique), les bases de connaissances, les moteurs d'inférence et de déduction, le raisonnement (symbolique) et les systèmes experts ;
>
> (c) Approches statistiques, estimation bayésienne, méthodes de recherche et d'optimisation. »

Comme le montre cette définition, un « système d'IA » tel que l'entend la proposition de législation sur l'intelligence artificielle ne s'aligne pas strictement sur les définitions techniques proposées ci-dessus. Par exemple, dans cette définition, l'apprentissage automatique est traité moins comme une composante de l'IA que comme un type spécialisé d'IA. Pour éviter toute ambiguïté et aux fins du présent rapport, nous proposons de définir spécifiquement un « système d'intelligence artificielle » de la manière suivante :

> Les « systèmes d'intelligence artificielle » désignent des logiciels autonomes ou intégrés au matériel de traitement de l'information, qui agissent comme un agent intelligent ou font preuve d'intelligence informatique. Un « système d'IA » peut consister en un ou plusieurs algorithmes ou modèles, mais désigne habituellement des systèmes complexes dans lesquels plusieurs algorithmes ou modèles fonctionnent ensemble pour effectuer une tâche complexe.

Le discours public est à l'heure actuelle dominé par des préoccupations relatives à une classe particulière de systèmes d'IA, les systèmes qui prennent des décisions et formulent des recommandations sur des sujets importants ayant trait à la vie quotidienne. Ces systèmes complètent, améliorent ou remplacent l'analyse et la prise de décision humaines et sont fréquemment utilisés en raison de la portée ou de l'ampleur des données et des règles concernées. Le nombre de fonctionnalités prises en compte dans les tâches de classification peut atteindre plusieurs millions. Ces tâches reproduisent des traitements de données précédemment effectués par des

travailleurs humains, mais à une échelle bien plus considérable et en faisant appel à une logique décisionnelle qualitativement distincte. Ces systèmes prennent des décisions généralement fiables (mais pas nécessairement correctes) sur la base de règles complexes qui mettent au défi ou en échec les capacités humaines d'action et de compréhension[18]. Autrement dit, le présent rapport traite des systèmes d'IA dont les actions sont difficiles à prévoir pour l'homme ou dont la logique décisionnelle est difficile à expliquer après coup.

Les défis éthiques communément rencontrés en matière d'IA

Des études préalables portant sur les problèmes éthiques auxquels l'IA est confrontée ont permis d'identifier six types de préoccupations qui trouvent leur source dans les paramètres opérationnels des algorithmes décisionnels et des systèmes d'IA. Le schéma reproduit et adapté dans la figure 1 tient compte du fait que :

> « les algorithmes décisionnels (1) transforment les données en preuves pour un *résultat donné* (ci-après appelé conclusion), et que ce résultat est ensuite utilisé pour (2) déclencher *et* motiver une action qui (en tant que telle, ou lorsqu'elle est associée à d'autres actions) peut ne pas être neutre d'un point de vue éthique. Ce travail est effectué de manière complexe et (semi-) autonome, ce qui (3) complique l'attribution de la responsabilité pour les effets des actions dictées par des algorithmes[19]. »

À partir de ces caractéristiques opérationnelles, il est possible d'identifier trois types épistémologiques et deux types normatifs de préoccupations éthiques, selon la manière dont les algorithmes traitent les données pour produire des preuves et motiver des actions. Les cinq types de préoccupations proposés sont susceptibles d'être à l'origine de défaillances impliquant de multiples agents humains, organisationnels et technologiques. Ce panachage d'acteurs humains et technologiques soulève de difficiles questions à propos de l'attribution des responsabilités concernant les conséquences des comportements de l'IA. Ces difficultés sont cernées dans le concept de traçabilité en tant que type principal et définitif de préoccupation.

[18] Brent Mittelstadt *et al.*, « The Ethics of Algorithms: Mapping the Debate », *Big Data & Society* nº 3, (2016), http://bds.sagepub.com/lookup/doi/10.1177/2053951716679679 (consulté le 15 déc. 2016). La suite du chapitre 2.1 emprunte largement aux conclusions et au cadre éthique proposé par cette étude cartographique.
[19] *Id.*

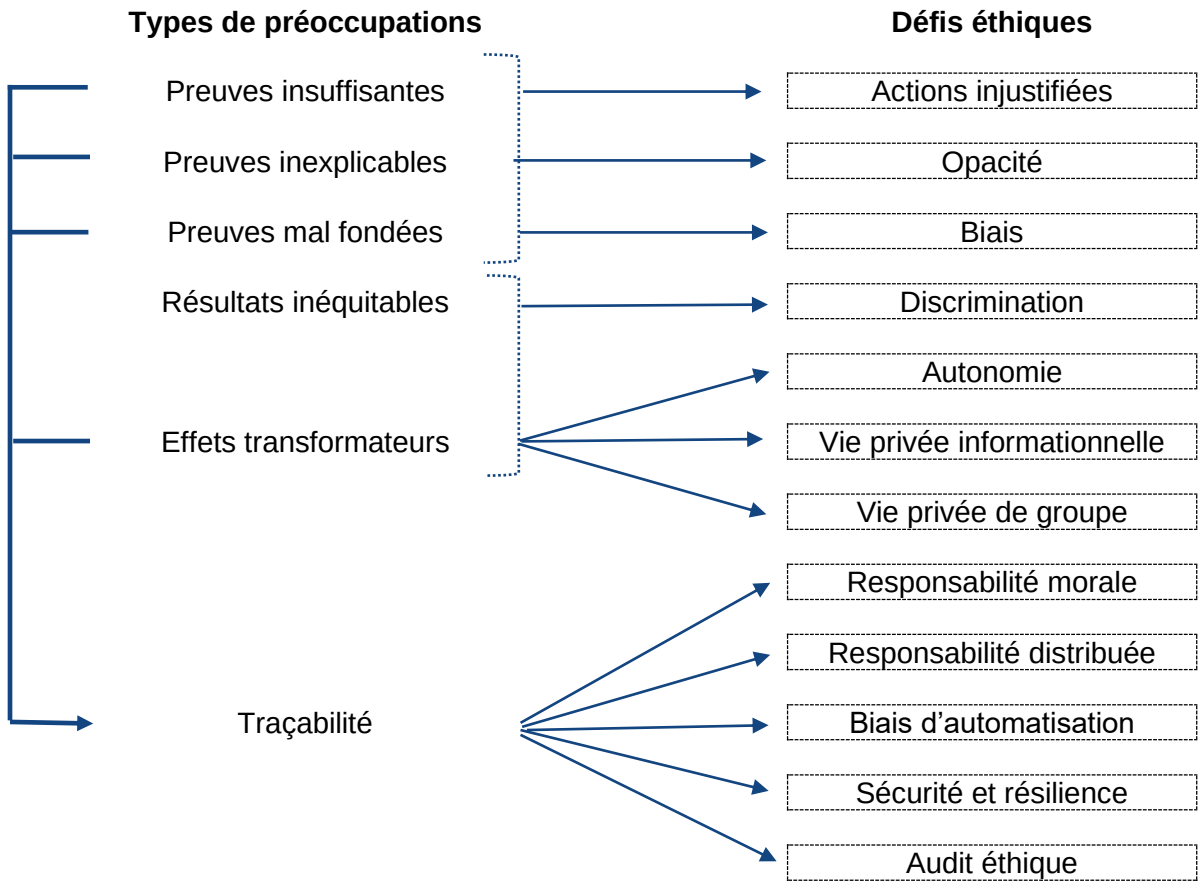

Figure 1 – Types de préoccupations et de défis éthiques soulevés par les algorithmes (adapté de Mittelstadt *et al.*, 2016)

Les trois préoccupations épistémologiques mentionnées ci-dessus concernant les algorithmes décisionnels et les systèmes d'IA peuvent être définies comme suit :

- **Preuves insuffisantes** – Lorsque les algorithmes tirent des conclusions des données qu'ils traitent à l'aide de statistiques inférentielles et/ou de techniques d'apprentissage automatique, ils produisent des connaissances probables[20] mais inévitablement incertaines. La théorie de l'apprentissage statistique[21] et la théorie de l'apprentissage informatique[22] s'intéressent toutes deux à la caractérisation et à la quantification de cette incertitude. Les méthodes statistiques peuvent identifier des corrélations significatives, mais les corrélations ne sont généralement pas suffisantes pour apporter la démonstration d'une causalité[23], et sont donc susceptibles de ne pas constituer une raison suffisante pour motiver une action sur la base de la connaissance d'une telle relation. Le concept de « données exploitables »

[20] La notion de « connaissances probables » est utilisée ici au sens que lui donne Ian Hacking dans son ouvrage *The Emergence of Probability: A Philosophical Study of Early Ideas About Probability, Induction and Statistical Inference* (2006), qui l'associe à l'émergence des sciences de la probabilité et au développement de la pensée statistique (par exemple, dans le contexte des assurances) à partir du XVII^{e} siècle.
[21] Gareth James *et al.*, *An Introduction to Statistical Learning* (2013).
[22] Leslie G. Valiant, « A theory of the Learnable », *Communications of the ACM* n^{o} 27, p. 1134-1142 (1984).
[23] Peter Grindrod, *Mathematical Underpinnings of Analytics: Theory and Applications* (2014).

prend en compte l'incertitude inhérente aux corrélations statistiques ainsi que le caractère normatif du choix d'agir sur celles-ci[24].

- **Preuves inexplicables** – Lorsque des données sont utilisées en tant qu'informations probantes (ou traitées pour produire de telles informations) destinées à appuyer une conclusion, il est raisonnable de s'attendre à ce que le lien entre les données et la conclusion soit intelligible et se prête à un examen rigoureux[25]. Étant donné la complexité et l'ampleur de nombreux systèmes d'IA, l'intelligibilité et le contrôle ne vont pas de soi. Des limitations pratiques et de principe résultent du défaut d'accès aux ensembles de données et des difficultés inhérentes à la détermination de la façon dont la multitude de données et de fonctionnalités prises en compte par un système d'IA aboutit à des conclusions et des résultats spécifiques[26].

- **Preuves mal fondées** – Les algorithmes traitent des données et sont par conséquent soumis à une limitation commune à tous les types de traitement de données, à savoir que les résultats ne peuvent jamais excéder les données d'entrée. Le principe informatique informel *garbage in, garbage out* (principe GIGO qui peut être rendu par « à données inexactes, résultats erronés ») illustre ce phénomène et sa signification : la fiabilité des conclusions (mais également leur neutralité) ne peut être supérieure à celle des données sur lesquelles elles sont fondées[27].

Les trois préoccupations épistémiques exposées jusqu'ici intéressent la qualité des *preuves* produites par un algorithme motivant une action particulière. De telles actions peuvent également susciter des préoccupations normatives. Ces préoccupations normatives potentielles sont au nombre de deux :

- **Résultats inéquitables** – Toute action guidée par des algorithmes peut être examinée sous différents angles, critères et principes éthiques. L'acceptabilité normative de l'action et de ses effets est tributaire de l'observateur et peut être évaluée indépendamment de sa qualité épistémologique. Une action peut être considérée comme discriminatoire, par exemple, du seul fait de ses conséquences sur une catégorie de personnes protégées, même si elle est effectuée sur la base d'éléments d'appréciation probants, vérifiables et bien fondés.

- **Effets transformateurs** – L'impact des systèmes d'IA ne peut pas toujours être imputable à des défaillances épistémiques ou éthiques. En l'absence d'effets dommageables manifestes, leurs répercussions sur le plan éthique peuvent paraître neutres de prime abord. Un ensemble distinct de

[24] Boaz Miller et Isaac Record, « Justified Belief in a Digital Age: on the Epistemic Implications of Secret Internet Technologies », *Episteme* nº 10, p. 117-134 (2013).
[25] Hilary Kornblith, *Epistemology: Internalism and Externalism* (2001).
[26] Miller et Record, cf. *supra* note 23.
[27] Pour une approche formelle du principe *« garbage in, garbage out »*, voir : Claude E. Shannon et Warren Weaver, *The Mathematical Theory of Communication* (1998).

conséquences, qui peuvent être qualifiées d'effets transformateurs, porte sur des changements subtils dans la conceptualisation et l'organisation du monde[28].

Une dernière préoccupation majeure intéresse la nécessité de spécifier des caractéristiques communes aux systèmes d'IA et aux conditions environnementales afin d'assurer une répartition équitable de l'obligation de rendre des comptes et de la responsabilité entre tous les acteurs et parties prenantes impliqués dans le développement, la mise en place et l'utilisation des systèmes d'IA :

- **Traçabilité** – Les systèmes d'IA impliquent souvent de multiples agents qui peuvent être des concepteurs et des utilisateurs humains, des fabricants et des organismes de déploiement, ainsi que les systèmes et les modèles eux-mêmes. Les systèmes d'IA sont également susceptibles d'interagir directement en formant des réseaux multi-agents caractérisés par une rapidité comportementale qui échappent à la surveillance et à la compréhension de leurs homologues humains en raison de leur vitesse, de leur ampleur et de leur complexité. Comme le suggèrent Mittelstadt *et al.* dans leur étude cartographique de l'IA, « les algorithmes sont des artefacts logiciels utilisés dans le traitement des données, qui héritent en tant que tels des défis éthiques associés à la conception et à la disponibilité des nouvelles technologies et de ceux associés à la manipulation de grands volumes de données à caractère personnel et autre[29]. » L'ensemble de ces facteurs a pour conséquence qu'il est difficile de détecter les effets dommageables, d'en remonter à la cause et d'en imputer la responsabilité lorsque des systèmes d'IA se comportent de manière inattendue. Les problèmes suscités par l'un des cinq types de préoccupations cités ci-dessus peuvent donc engendrer une difficulté connexe touchant à la traçabilité et à la nécessité d'établir à la fois la cause et la responsabilité de comportements aux effets dommageables[30].

Comme le montre la figure 1, les préoccupations de ce type que soulèvent les algorithmes décisionnels et les systèmes d'IA peuvent être imputées à des défis et des concepts éthiques très largement débattus. Selon cette approche, on énumérera brièvement quelques-uns des principaux défis éthiques posés par les caractéristiques opérationnelles des algorithmes décisionnels et les six types de préoccupations décrites ci-dessus[31] .

- **Actions injustifiées** – La prise de décision algorithmique et l'exploration de données reposent pour l'essentiel sur des connaissances inductives et des corrélations identifiées au sein d'un ensemble de données. Des corrélations

[28] Luciano Floridi, *The Fourth Revolution: How the Infosphere is Reshaping Human Reality* (2014).
[29] Mittelstadt *et al.*, cf. *supra* note 17.
[30] G. O. Mohler *et al.*, « Self-Exciting Point Process Modeling of Crime », *Journal of the American Statistical Association* nº 106, p. 100-108 (2011) ; Luciano Floridi, « Faultless Responsibility: On the Nature and Allocation of Moral Responsibility for Distributed Moral Actions », *Philosophical Transactions of the Royal Society: Mathematical, Physical and Engineering Sciences A* nº 374 : 20160112 (2016).
[31] Cette liste est adaptée d'une analyse documentaire effectuée par l'auteur, dont le compte rendu est publié dans Mittelstadt *et al.*, cf. *supra* note 17.

fondées sur un volume « suffisant » de données sont souvent considérées comme suffisamment crédibles pour diriger une action sans établir au préalable une causalité[32]. Or des actions inspirées de corrélations peuvent se révéler doublement problématiques. Des corrélations fallacieuses sont susceptibles d'être découvertes à la place de véritables connaissances causales. Même si des corrélations fortes ou des connaissances causales sont repérées, ces connaissances peuvent ne concerner que des populations alors que des actions ayant un impact personnel significatif visent des individus[33].

- **Opacité** – Elle désigne le problème de la « boîte noire » de l'IA : la logique qui sous-tend la transformation des données d'entrée en résultats peut être inconnue des observateurs ou des parties concernées, ou fondamentalement impénétrable ou inintelligible. Dans le domaine des algorithmes d'apprentissage automatique, l'opacité est le produit de la dimensionnalité élevée des données, de la complexité du code et du caractère changeant de la logique décisionnelle[34]. La transparence et l'intelligibilité sont généralement souhaitées car les algorithmes peu prévisibles ou peu interprétables sont difficiles à contrôler, surveiller et corriger[35]. La transparence est souvent naïvement considérée comme une panacée pour les questions éthiques que soulèvent les nouvelles technologies[36].

Il est fréquemment et à dessein difficile d'accéder à une information relative à la fonctionnalité des algorithmes[37]. Outre son accessibilité, l'information doit être compréhensible pour être considérée comme transparente[38]. Les efforts visant à rendre les algorithmes transparents sont confrontés à un défi de taille, qui est de rendre

---

[32] Mireille Hildebrandt, « Who Needs Stories if You Can Get the Data? ISPs in the Era of Big Number Crunching », *Philos. Technol.* nº 24, p. 371-390 (2011) ; Mireille Hildebrandt et Bert-Jaap Koops, « The Challenges of Ambient Law and Legal Protection in the Profiling Era », *The Modern Law Review* nº 73, p. 428-460 (2010) ; Viktor Mayer-Schönberger et Kenneth Cukier, *Big Data : A Revolution That Will Transform How We Live, Work and Think* (2013) ; Tal Zarsky, « The Trouble with Algorithmic Decisions An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making », *Science Technology Human Values* nº 41, p. 118-132 (2016).

[33] Phyllis Mckay Illari et Federica Russo, *Causality: Philosophical Theory Meets Scientific Practice* (2014).

[34] Jenna Burrell, « How the Machine "Thinks:" Understanding Opacity in Machine Learning Algorithms », *Big Data & Society* (2016).

[35] Andrew Tutt, « An FDA for Algorithms » (2016), http://papers.ssrn.com/abstract=2747994 (consulté le 13 avril 2016).

[36] Anjanette Raymond, « The Dilemma of Private Justice Systems: Big Data Sources, the Cloud and Predictive Analytics », *Northwestern Journal of International Law & Business, Forthcoming* (2014), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2469291 (consulté le 22 juil. 2015) ; Kate Crawford, « Can an Algorithm be Agonistic? Ten Scenes from Life in Calculated Publics », *Science Technology Human Values* nº 41, p. 77-92 (2016) ; Daniel Neyland, « Bearing Account-able Witness to the Ethical Algorithmic System », *Science Technology Human Values* nº 41, p. 50-76 (2016).

[37] Tasha Glenn et Scott Monteith, « New Measures of Mental State and Behavior Based on Data Collected From Sensors, Smartphones, and the Internet », *Curr Psychiatry Rep* nº 16, p. 1-10 (2014) ; Meredith Stark et Joseph J. Fins, « Engineering Medical Decisions », *Cambridge Quarterly of Healthcare Ethics* nº 22, p. 373-381 (2013) ; Rob Kitchin, « Thinking critically about and researching algorithms », *Information, Communication & Society* p. 1-16 (2016) ; Matthias Leese, « The new profiling: Algorithms, black boxes, and the failure of anti-discriminatory safeguards in the European Union », *Security Dialogue* nº 45, p. 494-511 (2014).

[38] Matteo Turilli et Luciano Floridi, « The Ethics of Information Transparency », *Ethics Inf Technol* nº 11, p. 105-112 (2009).

des processus décisionnels complexes à la fois accessibles et compréhensibles. Le problème ancien de l'interprétabilité des algorithmes d'apprentissage automatique met en évidence le défi de l'opacité dans le domaine des algorithmes[39]. Dans le contexte de la médecine, l'Organisation mondiale de la santé (OMS) a reconnu l'importance cruciale de la lutte contre l'opacité, en adoptant des dispositions visant à garantir la transparence, 'l'explicabilité' et l'intelligibilité dans la conception et l'utilisation de l'IA en matière de soins de santé[40].

- **Biais** – L'automatisation de la prise de décision humaine est souvent justifiée par une prétendue absence de partialité dans l'IA et les algorithmes[41]. Cette croyance n'est pas défendable, car les systèmes d'IA prennent inévitablement des décisions biaisées[42]. La conception et les fonctionnalités d'un système reflètent les valeurs de son concepteur et les utilisations projetées, ne serait-ce que dans la mesure où une conception particulière serait préférée car considérée comme étant l'option la meilleure ou la plus efficace. L'élaboration ne s'inscrit pas dans une trajectoire neutre et linéaire[43]. En conséquence, « les valeurs de l'auteur, consciemment ou non, sont figées dans le code, ce qui a pour effet de les institutionnaliser[44] ». L'inclusion et l'équité tant dans la conception que dans l'utilisation de l'IA sont donc essentielles pour combattre les biais implicites[45]. Friedman et Nissenbaum précisent que les biais résultent (1) des valeurs sociales préexistantes incluses dans les « institutions, pratiques et mentalités sociales » sur lesquelles se constitue la technologie, (2) des contraintes techniques et (3) des aspects émergents d'un contexte d'utilisation[46].

- **Discrimination –** Les biais présents dans les systèmes d'IA peuvent avoir pour conséquence une discrimination exercée à l'encontre d'individus et de groupes. Une analyse discriminatoire peut favoriser l'existence de prédictions se réalisant d'elles-mêmes et la stigmatisation de groupes ciblés, portant ainsi

---

[39] Hildebrandt, cf. *supra* note 31 ; Leese, cf. *supra* note 36 ; Burrell, cf. *supra* note 33 ; Tutt, cf. *supra* note 34.
[40] Organisation mondiale de la santé, cf. *supra* note 1, p. XIII.
[41] Engin Bozdag, « Bias in Algorithmic Filtering and Personalization », *Ethics Inf Technol* nº 15, p. 209-227 (2013) ; Gauri Naik et Sanika S. Bhide, « Will the Future of Knowledge Work Automation Transform Personalized Medicine? », *Applied & Translational Genomics* nº 3, p. 50-53 (2014).
[42] Kevin Macnish, « Unblinking Eyes: The Ethics of Automating Surveillance », *Ethics Inf Technol* nº 14, p. 151-167 (2012) ; Sue Newell et Marco Marabelli, « Strategic Opportunities (and Challenges) of Algorithmic Decision-Making: A Call for Action on the Long-term Societal Effects of 'Datification' », *The Journal of Strategic Information Systems* nº 24, p. 3-14, 6 (2015) ; Bozdag, cf. *supra* note 40 ; Batya Friedman et Helen Nissenbaum, « Bias in computer systems », *ACM Transactions on Information Systems (TOIS)* nº 14, p. 330-347 (1996) ; Omer Tene et Jules Polonetsky, *Big Data for All: Privacy and User Control in the Age of Analytics* (2013), http://heinonlinebackup.com/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/nwteintp11&section=20 (consulté le 2 oct. 2014) ; Felicitas Kraemer, Kees van Overveld et Martin Peterson, « Is There an Ethics of Algorithms? », *Ethics and Information Technology* nº 13, p. 251-260 (2011).
[43] Jeffrey Alan Johnson, « Technology and Pragmatism: From Value Neutrality to Value Criticality » (2006), http://papers.ssrn.com/abstract=2154654 (consulté le 24 août 2015).
[44] Macnish, cf. *supra* note 41, p. 158.
[45] Organisation mondiale de la santé, cf. *supra* note 1, p. XIII.
[46] Friedman et Nissenbaum, cf. *supra* note 41.

atteinte à leur autonomie et leur participation à la société[47]. Bien qu'il n'existe pas de définition unique de la discrimination, les cadres juridiques internationaux ont depuis longtemps engendré une jurisprudence abondante examinant les formes de discrimination (directe ou indirecte, par exemple), les objectifs des législations relatives à l'égalité (égalité formelle et substantielle, par exemple), et les seuils appropriés en matière de répartition des résultats entre groupes. Dans ce contexte, il est particulièrement difficile d'intégrer dans les systèmes d'IA des considérations de non-discrimination et d'équité[48]. Il peut être concevable d'orienter les algorithmes de sorte qu'ils ne tiennent pas compte d'attributs sensibles favorisant la discrimination[49], tels que le sexe ou l'origine ethnique[50], en fonction de l'apparition de discrimination dans un contexte donné. Toutefois, il n'est pas aisé de prévoir ou détecter des indicateurs indirects d'attributs protégés[51], plus particulièrement quand les algorithmes accèdent à des ensembles couplés de données[52].

- **Autonomie** – Les décisions porteuses de valeur prises par les algorithmes peuvent également faire peser une menace sur l'autonomie. La personnalisation du contenu par des systèmes d'IA, comme les systèmes de recommandation, se révèle à cet égard particulièrement problématique. La personnalisation peut être entendue comme l'élaboration d'architectures du choix qui ne sont pas identiques entre elles dans un même échantillon[53]. En filtrant l'information, l'IA peut influencer le comportement des personnes concernées et la prise de décision par des humains[54]. Différentes informations, des prix et autres contenus peuvent être proposés pour cataloguer des groupes ou des publics au sein d'une population définie par un ou plusieurs attributs, par

[47] Macnish, cf. *supra* note 41 ; Leese, cf. *supra* note 36 ; Solon Barocas, « Data Mining and the Discourse on Discrimination » (2014), https://dataethics.github.io/proceedings/DataMiningandtheDiscourseOnDiscrimination.pdf (consulté le 20 déc. 2015).

[48] Sandra Wachter, Brent Mittelstadt et Chris Russell, « Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI », *Computer Law & Security Review* nº 41, 105567 (2021) ; Sandra Wachter, Brent Mittelstadt et Chris Russell, « Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law », *W. Va. L. Rev.* nº 123, p. 735 (2020).

[49] Solon Barocas et Andrew D. Selbst, « Big Data's Disparate Impact » (2015), http://papers. ssrn. com/abstract=2477899 (consulté le 16 oct. 2015).

[50] Toon Calders, Faisal Kamiran et Mykola Pechenizkiy, « Building Classifiers with Independency Constraints », ICDMW'09. *IEEE International Conference on Data Mining Workshops*, 2009, p. 13-18 (2009) ; Faisal Kamiran et Toon Calders, « Classification with no Discrimination by Preferential Sampling », *Proc. 19th Machine Learning Conf. Belgium and the Netherlands* (2010), http://wwwis.win.tue.nl/~tcalders/pubs/benelearn2010 (consulté le 24 août 2015) ; Schermer, cf. *supra* note 14.

[51] Zarsky, cf. *supra* note 31 ; Andrea Romei et Salvatore Ruggieri, « A Multidisciplinary Survey on Discrimination Analysis », *The Knowledge Engineering Review* nº 29, p. 582-638 (2014).

[52] Barocas et Selbst, cf. *supra* note 48.

[53] Omer Tene et Jules Polonetsky, « Big Data for All: Privacy and User Control in the Age of Analytics », *NW. J. Tech. & Intell. Prop*. (2013), http://heinonlinebackup.com/hol-cgi-bin/get_pdf. cgi?handle=hein.journals/nwteintp11&section=20 (consulté le 2 oct. 2014).

[54] Mike Ananny, « Toward an Ethics of Algorithms Convening, Observation, Probability, and Timeliness », *Science Technology Human Values* nº 41, p. 93-117 (2016).

exemple la solvabilité, ce qui peut aboutir à une discrimination. La personnalisation réduit la diversité des informations rencontrées par l'utilisateur en excluant des contenus jugés non pertinents ou contradictoires avec les croyances ou les désirs de l'utilisateur[55]. Ce point se révèle problématique dans la mesure où la diversité de l'information peut être considérée comme une condition favorisant l'autonomie[56]. Il est contrevenu à l'autonomie décisionnelle du sujet lorsque le choix désiré reflète les intérêts d'un tiers avant ceux de la personne concernée[57].

En matière d'autonomie, un défi connexe concerne l'intelligibilité, ou compréhensibilité, des systèmes algorithmiques et de leurs résultats. Les professionnels de santé qui intègrent des recommandations fondées sur l'IA dans l'ordinaire de leur prise en charge clinique de patients, par exemple, peuvent se heurter à une perte d'autonomie si l'assise de ces recommandations est mal comprise. De même, les patients sont confrontés à un défi analogue lorsqu'ils prennent sur la base des recommandations de l'IA des décisions éclairées portant sur leurs soins. Consciente de ces risques, l'OMS reconnaît la nécessité de « protéger l'autonomie de l'être humain » en tant que principe éthique fondamental régissant la conception, l'utilisation et la gouvernance de l'IA dans les soins de santé, en raison du risque de transfert aux machines du pouvoir décisionnel des êtres humains[58].

- **Vie privée informationnelle et vie privée de groupe** – Les algorithmes transforment également la notion de protection de la vie privée. Il est souvent fait appel au concept de vie privée informationnelle[59], ou au droit des personnes concernées de « protéger les données à caractère personnel contre tout accès illicite par des tiers » pour faire face à la discrimination, à la personnalisation et à la diminution de l'autonomie dues à l'opacité. La vie privée informationnelle se rapporte à la capacité d'un individu à contrôler les informations le concernant[60] et aux efforts que devraient déployer des tiers pour obtenir ces informations. Un droit à l'identité dérivé de la vie privée informationnelle sous-entend le caractère problématique d'un profilage opaque ou secret lorsqu'il est effectué par un tiers. Dans un environnement de soins de santé, cela pourrait inclure les assureurs, les prestataires de soins à distance (par exemple, les chatbots et les prestataires de services de triage), les entreprises de technologie grand public et autres. L'opacité décisionnelle empêche la surveillance et une prise de décision éclairée concernant le partage des

[55] Eli Pariser, *The Filter Bubble : What The Internet is Hiding From You* (2011) ; Belinda A. Barnet, « Idiomedia: The Rise of Personalized, Aggregated Content », *Continuum* nº 23, p. 93-99 (2009).
[56] Jeroen van den Hoven et Emma Rooksby, « Distributive Justice and the Value of Information: A (Broadly) Rawlsian Approach », *Information Technology And Moral Philosophy* nº 376, (2008).
[57] Stark et Fins, cf. *supra* note 36 ; Sally A. Applin et Michael D. Fischer, « New Technologies and Mixed-Use Convergence: How Humans and Algorithms Are Adapting to Each Other », *in* 2015 *IEEE International Symposium On Technology And Society* (ISTAS), p. 1-6 (2015).
[58] Organisation mondiale de la santé, cf. *supra* note 1, p. XII.
[59] Schermer, cf. *supra* note 14.
[60] L. Van Wel et L. Royakkers, « Ethical Issues in Web Data Mining », *Ethics and Information Technology* nº 6, p. 129-140 (2004).

données[61]. Les personnes concernées se trouvent dans l'impossibilité de définir des normes en matière de protection de la vie privée pour régir de manière générique tous types de données, car la valeur ou la perspicacité des données ne sont établies qu'au moyen de leur traitement[62].

Les protections de la vie privée fondées sur l'identifiabilité sont mal adaptées à la limitation de la gestion externe des identités au moyen de l'analytique. Les protections réglementaires actuelles peinent à faire face aux risques que fait peser l'analytique sur la vie privée informationnelle en raison du lien établi entre la définition des « données à caractère personnel » et une personne identifiée ou identifiable ; l'identification d'un utilisateur n'est en effet souvent pas nécessaire aux fins de profilage et de prise de décision algorithmiques. Les connaissances produites concernent moins des individus précisément identifiables que des groupes organisés par des algorithmes. Les cadres réglementaires existants en matière de protection de la vie privée et des données ne reflètent pas l'importance du profilage et des groupes au regard des procédés modernes d'analyse des données et d'automatisation décisionnelle[63].

- **Responsabilité morale et responsabilité distribuée** – Quand une technologie est défaillante, il convient d'établir les responsabilités et de les assortir de sanctions[64]. L'imputabilité de la responsabilité ne peut être justifiée que si l'acteur dispose d'une certaine marge de contrôle et d'intentionnalité dans l'exécution de l'action[65]. Classiquement, les développeurs et les ingénieurs informaticiens exercent « dans ses moindres détails le contrôle du comportement de la machine », pour autant qu'ils sont capables d'expliquer à un tiers sa conception et sa fonction globales[66]. Cette notion traditionnelle de la responsabilité en matière de création de logiciels suppose que le développeur est en mesure d'envisager les effets probables de la technologie et ses défauts de fonctionnement potentiels[67], et par conséquent de faire des choix de conception orientés vers les résultats les plus souhaitables conformément aux spécifications fonctionnelles[68].

L'imputation justifiée de la responsabilité morale n'est pas aisée dans le cas des algorithmes et systèmes d'IA dotés de capacités d'apprentissage. Le modèle traditionnel d'attribution de la responsabilité en informatique exige que le système soit bien défini, compréhensible et prévisible ; les systèmes complexes et évolutifs (c'est-

---

[61] Hojung Kim, Joseph Giacomin et Robert Macredie, « A Qualitative Study of Stakeholders' Perspectives on the Social Network Service Environment », *International Journal Of Human-Computer Interaction* nº 30, p. 965-976 (2014).
[62] Van Wel et Royakkers, cf. *supra* note 59 ; Hildebrandt, cf. *supra* note 31.
[63] Brent Mittelstadt, « From Individual to Group Privacy in Big Data Analytics », *Philosophy & Technology* nº 30, p. 475-494 (2017) ; 126 Linnet Taylor, Luciano Floridi et Bart Van Der Sloot, *Group Privacy: New Challenges of Data Technologies* (2017), http://link.springer.com/book/10.1007/978-3-319-46608-8 (consulté le 18 janv. 2017).
[64] Kraemer, van Overveld et Peterson, cf. *supra* note 41, p. 251.
[65] Matthias, cf. *supra* note 15.
[66] *Id.*
[67] Luciano Floridi, Nir Fresco et Giuseppe Primiero, « On Malfunctioning Software », *Synthese* nº 192, p. 1199-1220 (2014).
[68] Matthias, cf. *supra* note 15.

à-dire comportant d'innombrables règles de décision et lignes de code) font obstacle à une surveillance holistique des voies et dépendances décisionnelles. À cet égard, les algorithmes d'apprentissage automatique suscitent des difficultés considérables, comme le montrent par exemple les algorithmes génétiques qui se programment eux-mêmes[69]. Le modèle traditionnel de responsabilité échoue car « personne n'exerce un contrôle suffisant sur les actions de la machine pour être en mesure d'assumer la responsabilité desdites actions[70] ». La responsabilité distribuée soulève donc une difficulté particulière pour les systèmes d'IA, mais elle pourrait être éludée par l'application de la responsabilité objective ou de systèmes similaires de responsabilité sans faute.

- **Biais d'automatisation –** Un problème connexe concerne la diffusion du sentiment de responsabilité et de l'obligation de rendre des comptes chez les utilisateurs de systèmes d'IA, et la tendance liée à faire confiance aux résultats des systèmes sur la base de leur objectivité, précision ou complexité perçues[71]. Les décideurs humains peuvent se détourner de leur responsabilité lorsqu'ils délèguent à l'IA la prise de décision. Des effets similaires peuvent être observés dans les réseaux mixtes composés de systèmes humains et de systèmes informatiques : ils sont caractérisés par une diminution du sentiment de responsabilité personnelle et l'exécution d'actions par ailleurs autrement injustifiables, comme cela a déjà été étudié dans les milieux bureaucratiques[72]. Par exemple, des algorithmes impliquant des parties prenantes représentant plusieurs disciplines différentes peuvent inciter chaque partie à supposer que les autres assumeront la responsabilité éthique des actions de l'algorithme[73]. L'apprentissage automatique ajoute une strate de complexité supplémentaire entre les concepteurs et les actions dirigées par l'algorithme, ce qui peut affaiblir à juste titre la responsabilité imputée aux premiers.

- **Sécurité et résilience –** La nécessité d'établir les responsabilités se ressent avec acuité en cas de défaut de fonctionnement des algorithmes. Les algorithmes contraires à l'éthique peuvent être envisagés comme des artefacts logiciels défaillants qui ne fonctionnent pas comme prévu. Des distinctions ont été utilement établies entre les erreurs de conception (types) et les erreurs opérationnelles (instances), ainsi qu'entre l'incapacité de fonctionner comme prévu (fonctionnement non conforme) et la présence d'effets secondaires non voulus (fonctionnement non voulu)[74]. Le fonctionnement non voulu se distingue des effets secondaires purement négatifs par son « évitabilité », c'est-à-dire la mesure avec laquelle des types comparables de systèmes ou d'artefacts accomplissent la fonction prévue sans qu'apparaissent les effets en question.

[69] Burrell, cf. *supra* note 33 ; Matthias, cf. *supra* note 15 ; Zarsky, cf. *supra* note 31.
[70] Matthias, cf. *supra* note 15, p. 177.
[71] Zarsky, cf. *supra* note 31, p. 121.
[72] Hannah Arendt, *Eichmann in Jerusalem: A Report on the Banality of Evil* (1971).
[73] Michael Davis, Andrew Kumiega et Ben Van Vliet, « Ethics, Finance, and Automation: A Preliminary Survey of Problems in High Frequency Trading », *Science and Engineering Ethics* nº 19, p. 851-874 (2013).
[74] Floridi, Fresco et Primiero, cf. *supra* note 66.

Ces distinctions font la lumière sur les aspects éthiques des systèmes d'IA strictement liés à leur fonctionnement, soit dans l'abstrait (par exemple lorsque sont examinées des performances brutes), soit dans le cadre d'un système décisionnel plus large, et révèlent l'interaction multidimensionnelle qui existe entre le comportement voulu et le comportement réel. L'apprentissage automatique, notamment, pose des problèmes particuliers, car l'obtention du comportement voulu ou « correct » n'implique pas l'absence d'erreurs ni d'actions préjudiciables et de boucles de rétroaction[75].

Les deux types de défaut de fonctionnement impliquent une imputabilité distincte des responsabilités, selon qu'il s'agit de concepteurs d'algorithmes et de logiciels, d'utilisateurs et d'artefacts informatiques. En cas de fonctionnement non conforme et de fonctionnement non voulu, l'imputation équitable des responsabilités entre de grandes équipes de conception et dans des contextes d'utilisation complexes représente un défi redoutable. Il est nécessaire de spécifier les impératifs à respecter en matière de résilience considérée comme idéal éthique dans la conception d'algorithmes pour parer aux défauts de fonctionnements, afin de garantir que les systèmes d'IA sont à la fois sûrs et résilients face aux fonctionnements non conformes et aux fonctionnements non voulus. Ces conditions reflètent l'importance éthique du bien-être humain et la façon dont il peut être affecté par l'IA. Compte tenu de ce qui précède, l'OMS a explicitement reconnu le primat de la protection du bien-être et de la sécurité des personnes en le consacrant comme un principe éthique central en matière d'utilisation de l'IA dans les soins de santé[76].

- **Audit éthique –** La question reste ouverte de savoir comment, de manière optimale, rendre opérationnels ces défis éthiques et fixer des normes de contrôle, s'agissant en particulier de l'apprentissage automatique. Se contenter de rendre transparent le code d'un algorithme ne suffit pas pour garantir un comportement éthique. Une voie envisageable pour réaliser l'interprétabilité, l'équité et atteindre d'autres objectifs éthiques dans les systèmes d'IA pourrait passer par un contrôle effectué par des processeurs de données[77], des régulateurs externes[78] ou des chercheurs empiriques[79], en utilisant des études d'audit ex-post[80], des études ethnographiques réflexives portant sur la conception et les tests[81], ou des mécanismes de rapport conçus dans

[75] À l'exception des cas triviaux, il est inévitable que soient présents des faux positifs et des faux négatifs dans le travail des algorithmes, plus particulièrement dans le cadre de l'apprentissage automatique.
[76] Organisation mondiale de la santé, cf. supra note 1, p. XIII.
[77] Zarsky, cf. *supra* note 31.
[78] Tutt, cf. *supra* note 34 ; Zarsky, cf. *supra* note 31 ; Frank Pasquale, *The Black Box Society: The Secret Algorithms that Control Money and Information* (2015).
[79] Neyland, cf. *supra* note 35 ; Kitchin, cf. *supra* note 36.
[80] Christian Sandvig *et al.*, « Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms », *Data and Discrimination: Converting Critical Concerns Into Productive Inquiry* (2014), http://social.cs.uiuc.edu/papers/pdfs/ICA2014-Sand vig.pdf (consulté le 13 fév. 2016); Philip Adler *et al.*, « Auditing Black-box Models by Obscuring Features », arXiv: 1602.07043 [CS, stat] (2016), http://arxiv.org/abs/1602.07043 (consulté le 5 mars 2016) ; Romei et Ruggieri, cf. *supra* note 50 ; Kitchin, cf. *supra* note 36 ; Nicholas Diakopoulos, « Algorithmic Accountability: Journalistic Investigation of Computational Power Structures », *Digital Journalism* nº 3, p. 398-415 (2015).
[81] Neyland, cf. *supra* note 35.

l'algorithme lui-même[82]. Quel que soit le type d'IA, l'audit est une condition préalable nécessaire pour *vérifier* le bon fonctionnement. S'agissant des systèmes ayant un impact prévisible sur l'homme, l'audit peut créer un enregistrement procédural *ex-post* d'une prise de décision automatisée complexe, afin de rendre intelligibles les décisions problématiques ou inexactes, ou détecter des discriminations ou des préjudices analogues.

La Convention d'Oviedo et les principes des droits de l'homme en matière de santé

La Convention européenne pour la protection des Droits de l'Homme et de la dignité de l'être humain à l'égard des applications de la biologie et de la médecine (1997) ou « Convention d'Oviedo » assure la protection des droits de l'homme en matière de biomédecine au niveau transnational. La Convention d'Oviedo est un instrument-cadre, c'est-à-dire qu'elle contient des principes généraux destinés à être appliqués en droit interne par ses signataires. Elle comporte de nombreux principes et exigences novateurs, fondés sur les principes et les droits contenus dans « les précédents traités internationaux relatifs aux droits de l'homme, tels que le Pacte international relatif aux droits civils et politiques de 1966 et la Convention européenne des droits de l'homme (CEDH) de 1950 (par exemple, les droits à la vie, à l'intégrité physique et au respect de la vie privée, l'interdiction des traitements inhumains ou dégradants et de toute forme de discrimination)[83] ». La Convention d'Oviedo s'inspire des droits à la vie, à l'intégrité physique et au respect de la vie privée, et de l'interdiction de la discrimination édictés par la CEDH, et se fonde sur ces droits. La Cour européenne des droits de l'homme a eu recours à la Convention d'Oviedo en tant que cadre interprétatif pour élucider et mieux comprendre la portée et l'importance de ces droits dans le contexte de la biomédecine[84].

On ne saurait surestimer l'importance de ces droits de l'homme constitutifs de la Convention d'Oviedo. Dans son ensemble, la Convention a pour finalité de « protéger l'être humain dans sa dignité et son identité et de garantir à toute personne, sans discrimination, le respect de son intégrité et de ses autres droits et libertés fondamentales à l'égard des applications de la biologie et de la médecine » (article 1). Certaines valeurs et finalités sont explicitement défendues et protégées par la Convention, tandis que d'autres peuvent être déduites d'exigences particulières. Avant tout, la dignité humaine et la primauté du patient sont des éléments centraux de la Convention :

> « À l'évidence, la notion de dignité de l'être humain est le fondement de la Convention d'Oviedo. Selon le rapport explicatif, "la notion de dignité de l'être humain […] désigne la valeur essentielle à maintenir. Elle constitue le

[82] Alfredo Vellido, José David Martín-Guerrero et Paulo JG Lisboa, « Making Machine Learning Models Interpretable », *in* ESANN nº 12, p. 163-172 (2012).
[83] Roberto Andorno, *The Oviedo Convention: A European Legal Framework at the Intersection of Human Rights and Health Law*, 2 p. 133-143, 133 (2005).
[84] Francesco Seatzu et Simona Fanni, « The Experience of the European Court of Human Rights with the European Convention on Human Rights and Biomedicine », *Utrecht J. Int'l & Eur. L.* nº 31, p. 5-16 (2015).

> fondement de la plupart des valeurs défendues par cette Convention". Rappelant l'historique de l'instrument, l'un des membres du groupe de rédaction reconnaît qu'"il a été rapidement décidé que les notions de dignité, d'identité et d'intégrité de l'être humain/individu devaient être à la fois la base et le cadre de l'ensemble des autres principes et notions à inclure dans la convention[85]" ».

Dans toute la Convention d'Oviedo, il est fait référence à d'autres valeurs et droits, tels que les droits à la vie, à l'intégrité physique et au respect de la vie privée, et l'interdiction de la discrimination. Par exemple, l'article 10 réaffirme le droit au respect de la vie privée énoncé par l'article 8 de la CEDH et repris dans la Convention pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel :

> 1. « Toute personne a droit au respect de sa vie privée s'agissant des informations relatives à sa santé.
>
> 2. Toute personne a le droit de connaître toute information recueillie sur sa santé. Cependant, la volonté d'une personne de ne pas être informée doit être respectée. »

Faisant suite aux exigences de transparence qu'implique le droit au respect de la vie privée prévu à l'article 10, l'article 5 de la Convention d'Oviedo affirme l'impératif bien établi du consentement éclairé en médecine :

> « Une intervention dans le domaine de la santé ne peut être effectuée qu'après que la personne concernée y a donné son consentement libre et éclairé.
>
> Cette personne reçoit préalablement une information adéquate quant au but et à la nature de l'intervention ainsi que quant à ses conséquences et ses risques.
>
> La personne concernée peut, à tout moment, librement retirer son consentement. »

Selon le rapport explicatif, l'exigence du consentement « fait apparaître l'autonomie du patient dans sa relation avec les professionnels de santé et conduit à restreindre les approches paternalistes qui ignoreraient la volonté du patient. » Les paragraphes 35 et 36 du rapport apportent des détails supplémentaires sur les exigences particulières nécessaires afin que le consentement soit considéré comme libre et éclairé, y compris les contraintes pesant sur l'influence du médecin sur la décision du patient et les obligations concernant la qualité, l'étendue et la clarté des informations fournies :

[85] Andorno, cf. *supra* note 82.

> « 35. Le consentement du patient ne peut être libre et éclairé que s'il est donné à la suite d'une information objective du professionnel de la santé responsable, quant à la nature et aux conséquences possibles de l'intervention envisagée ou de ses alternatives et en l'absence de toute pression de la part d'autrui. L'article 5, alinéa 2, mentionne ainsi les éléments les plus importants relatifs à l'information qui doit précéder l'intervention, mais il ne s'agit pas d'une énumération exhaustive : le consentement éclairé peut impliquer, selon les circonstances, des éléments supplémentaires. Pour que son consentement soit valide, la personne doit avoir reçu des informations sur les faits pertinents concernant l'intervention envisagée. Elle doit être renseignée sur l'objectif, la nature et les conséquences de l'intervention et sur les risques qu'elle peut comporter. S'agissant des risques de l'intervention ou de ses alternatives, l'information devrait porter non seulement sur les risques inhérents au type d'intervention envisagé, mais également sur les risques qui sont propres aux caractéristiques individuelles de chaque personne, telles que l'âge ou l'existence d'autres pathologies. Il doit être répondu de manière adéquate aux demandes d'information complémentaire formulées par les patients.
>
> 36. En outre, cette information doit être formulée dans un langage compréhensible par la personne qui va subir l'intervention. Le patient doit être mis à même de mesurer, par un langage qui soit à sa portée, l'objectif et les modalités de l'intervention au regard de sa nécessité ou de sa simple utilité mises en parallèle avec les risques encourus et les incommodités ou souffrances provoquées. »

L'article 10 prévoit pour toute personne un « droit de connaître » son état de santé et toute information recueillie sur sa santé, mais aussi un « droit de ne pas en être informé ». Ces droits constituent des éléments essentiels des relations médecin-patient envisagée par la Convention d'Oviedo. Si les patients sont en droit de prendre une décision éclairée concernant leurs soins de santé, il s'ensuit qu'ils ont qualité à recevoir toute information adéquate pour être en mesure de prendre cette décision en toute connaissance de cause[86].

En ce qui concerne la discrimination, l'article 11 interdit explicitement toute forme de discrimination à l'encontre d'une personne en raison de son patrimoine génétique. De même, l'article 3 prévoit un accès équitable à des soins de santé de qualité appropriée :

> « Les Parties prennent, compte tenu des besoins de santé et des ressources disponibles, les mesures appropriées en vue d'assurer, dans leur sphère de juridiction, un accès équitable à des soins de santé de qualité appropriée. »

L'inégalité de l'accès aux soins ou de la qualité des soins pourrait être considérée comme une violation de l'interdiction de la discrimination énoncée par l'article 14 de la CEDH, notamment en ce qui concerne la discrimination dans « l'appartenance à une minorité nationale, la fortune, la naissance ou toute autre situation » (voir chapitre

[86] *Id.*

intitulé « Inégalité dans l'accès à des soins de santé de qualité). De même, l'article 4 traite de la qualité des soins et des normes professionnelles dans le domaine de la santé et de la recherche :

> « Toute intervention dans le domaine de la santé, y compris la recherche, doit être effectuée dans le respect des normes et obligations professionnelles, ainsi que des règles de conduite applicables en l'espèce. »

À juste titre, la Convention d'Oviedo ne précise pas les normes de qualité à respecter en matière de soins de santé et de recherche, mais confie aux organismes professionnels et au droit interne de ses signataires la détermination des normes en fonction des besoins locaux en matière de santé et des ressources disponibles. Cela dit, comme la Convention d'Oviedo prescrit des normes minimales en matière de protection des droits de l'homme, les états membres conservent, en incorporant la Convention dans leur droit interne, la possibilité d'adopter des normes plus élevées. En ce qui concerne les normes applicables à la qualité des soins, cette possibilité peut être réalisée dans le contexte des articles 3 et 4. Le paragraphe 30 du Rapport explicatif fournit des précisions sur les parties envisagées pour fixer ces obligations et normes professionnelles :

> « 30. Les interventions doivent être effectuées dans le respect tout d'abord de l'ordre juridique général, complété et développé par les règles professionnelles. Ces règles prennent, selon les pays, tantôt la forme de codes professionnels d'éthique, tantôt celle de la déontologie médicale (code de déontologie de source étatique ou professionnelle), du droit de la santé, de l'éthique médicale ou de tout autre moyen garantissant les droits et les intérêts des patients et pouvant prendre en compte tout droit à l'objection de conscience des professionnels de la santé. »

Les paragraphes 31 et 32 développent la nature de la médecine en tant que profession, les variations de normes d'un pays à l'autre, l'engagement des médecins à respecter les normes éthiques et juridiques, ainsi que le contenu et l'évolution des normes au fil du temps :

> « 31. Le contenu des normes, des obligations professionnelles et des règles de conduite n'est pas uniforme dans tous les pays. Les mêmes devoirs médicaux peuvent comporter des nuances selon la société concernée. En revanche, les principes fondamentaux de l'activité médicale s'appliquent dans tous les pays. Le médecin, et d'une manière générale tout professionnel qui concourt à la réalisation d'un acte de santé, est soumis à des impératifs juridiques et éthiques. Il doit agir avec soin et compétence et tenir compte attentivement des besoins de chacun des patients.
>
> 32. Le médecin a pour tâche essentielle non seulement de guérir les malades, mais aussi de prendre des mesures propres à maintenir et promouvoir la santé

> et de soulager les douleurs, en tenant compte du bien-être psychique du patient. La compétence du médecin doit d'abord être reliée aux connaissances scientifiques et à l'expérience clinique propres à la profession ou à la spécialité à un moment donné. L'état actuel de la science détermine le niveau professionnel et le perfectionnement que l'on peut attendre des professionnels de la santé dans l'exercice de leur activité. En suivant les progrès de la médecine, il évolue au gré des nouveaux développements et élimine les méthodes qui ne reflètent pas l'état actuel de la science. Il est cependant admis que les normes professionnelles n'imposent pas nécessairement une conduite comme étant la seule possible : les "règles de l'art" peuvent offrir plusieurs voies d'intervention possibles, ménageant ainsi une certaine liberté de méthode ou de technique. »

Le paragraphe 33 du rapport explicatif fournit ensuite une brève indication du modèle idéal de relation médecin-patient en ce qui concerne le choix des interventions :

> « 33. Une intervention doit ensuite être jugée au regard du problème de santé spécifique posé par un patient déterminé. En particulier, l'intervention doit répondre aux critères de pertinence et de proportionnalité entre le but poursuivi et les moyens mis en œuvre. En plus, un facteur important de la réussite d'un traitement médical est la confiance du patient en son médecin. Cette confiance détermine également les devoirs du médecin par rapport à son patient. Un élément important de ces devoirs est le respect des droits du patient, qui crée et accroît la confiance mutuelle. L'alliance thérapeutique sera renforcée si les droits des patients sont pleinement respectés. »

La Convention d'Oviedo précise par conséquent un certain nombre de droits et d'exigences relatifs aux droits de l'homme protégés dans d'autres contextes ou dérivés de ceux-ci. Des valeurs et intérêts clés peuvent avoir pour origine les sujets abordés dans la convention. Ces valeurs ancrées dans les principes des droits de l'homme en matière de santé sont susceptibles de guider l'élaboration d'un cadre théorique pour les relations médecin-patient. Plus particulièrement, la Convention d'Oviedo prévoit et fait valoir les valeurs suivantes :

- **La dignité humaine**
- **La primauté des intérêts du patient sur les intérêts sociétaux et scientifiques**
- **Le droit à la vie**
- **L'intégrité physique**
- **La protection de la vie privée et de l'identité**
- **Le consentement éclairé**
- **Le droit de connaître et le droit de ne pas être informé**

- **L'interdiction de la discrimination et de l'inégalité dans l'accès aux soins de santé**
- **Les normes de qualité des soins**

Ces valeurs seront examinées au chapitre intitulé « Cadre théorique de la relation médecin-patient », dans le contexte des objectifs de la médecine en tant que profession et bien sociétal, et utilisées comme base d'élaboration d'un cadre théorique pour la relation médecin-patient. Ce cadre, de même que les valeurs qui ont pour origine la Convention d'Oviedo et sur lesquelles il repose, suggère que certains biens doivent exister dans la relation médecin-patient. De même, différents modèles de rencontres cliniques et de relation médecin-patient seront plus ou moins mis en conformité avec ces valeurs. Toutes ces questions seront abordées dans le chapitre susmentionné, après un bref aperçu des systèmes d'IA dans le domaine médical.

Pour situer le présent rapport dans le cadre des travaux politiques en cours du Conseil de l'Europe, il convient de mentionner brièvement les rapports ayant récemment traité d'autres domaines de travail pertinents en matière d'impact de l'IA sur les soins de santé. Le « Protocole d'amendement à la Convention pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel (STCE n° 223) » a été ouvert en octobre 2018 et devrait être ratifié en octobre 2023. Le Protocole porte modification de la Convention STE n° 108. L'article 8 révisé de la convention (désormais article 9) est d'un intérêt tout particulier pour l'IA dans le domaine médical, car il accorde aux personnes un certain nombre de droits à la protection des données :

1. « Toute personne a le droit :

    a. de ne pas être soumise à une décision l'affectant de manière significative, qui serait prise uniquement sur le fondement d'un traitement automatisé de données, sans que son point de vue soit pris en compte ;
    b. d'obtenir, à sa demande, à intervalle raisonnable et sans délai ou frais excessifs, la confirmation d'un traitement de données la concernant, la communication sous une forme intelligible des données traitées, et toute information disponible sur leur origine, sur la durée de leur conservation ainsi que toute autre information que le responsable du traitement est tenu de fournir au titre de la transparence des traitements, conformément à l'article 8, paragraphe 1 ;
    c. d'obtenir, à sa demande, connaissance du raisonnement qui sous-tend le traitement de données, lorsque les résultats de ce traitement lui sont appliqués ;
    d. de s'opposer à tout moment, pour des raisons tenant à sa situation, à ce que des données à caractère personnel la concernant fassent l'objet d'un traitement, à moins que le responsable du traitement ne démontre des motifs légitimes justifiant le traitement, qui prévalent sur les intérêts ou les droits et libertés fondamentales de la personne concernée ;

e. d'obtenir, à sa demande, sans frais et sans délai excessif, la rectification de ces données ou, le cas échéant, leur effacement lorsqu'elles sont ou ont été traitées en violation des dispositions de la présente Convention ;
f. de disposer d'un recours, conformément à l'article 12, lorsque ses droits prévus par la présente Convention ont été violés ;
g. de bénéficier, quelle que soit sa nationalité ou sa résidence, de l'assistance d'une autorité de contrôle au sens de l'article 15 pour l'exercice de ses droits prévus par la présente Convention. »

Un grand nombre de ces droits fait écho aux protections énoncées par le règlement général sur la protection des données (RGPD), un cadre normatif de protection des données à caractère personnel mis en œuvre par la Commission européenne en 2018, notamment un droit assorti de limitations à ne pas faire l'objet d'une décision fondée exclusivement sur un traitement automatisé, un droit d'obtenir des informations sur le traitement des données, des droits d'obtenir la rectification et l'effacement des données à caractère personnel[87]. Ces droits sont susceptibles de fournir un soutien important à la défense de l'idéal du consentement éclairé dans les applications médicales de l'IA, en donnant accès à une information relative à la portée et la nature du traitement automatisé.

Le rapport de la commission des questions sociales, de la santé et du développement durable, « Intelligence artificielle et santé : défis médicaux, juridiques et éthiques à venir » publié en octobre 2020 par l'Assemblée parlementaire du Conseil de l'Europe comporte un projet de recommandation visant à répondre à l'impact grandissant de l'IA sur les soins de santé[88]. L'exposé des motifs du rapport se livre à un examen détaillé des diverses répercussions de l'IA envisagées sur les plans médicaux, juridiques et éthiques, notamment :

- **Nécessité d'un examen éthique de la recherche biomédicale et limitations des compétences et capacités des organes d'examen éthique en matière d'évaluation des risques et opportunités très spécifiques de l'IA**
- **Responsabilité des fournisseurs d'IA dans le domaine du médicament et des soins de santé**
- **Protection des données à caractère personnel dans le cadre de l'harmonisation des systèmes de collecte de données et du soutien à l'innovation et à la recherche en matière d'IA, plus particulièrement en Europe**

[87] Sandra Wachter, Brent Mittelstadt et Luciano Floridi, « Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation », *International Data Privacy Law* nº 7, p. 76-99 (2017) ; Sandra Wachter et B. D. Mittelstadt, « A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI », *Columbia Business Law Review* nº 2019 (2019).
[88] Conseil de l'Europe, cf. *supra* note 2.

- **Garantir la légalité, l'équité, la spécification des finalités, la proportionnalité, la prise en compte du respect de la vie privée dès la conception et par défaut, la responsabilité, la conformité, la transparence, la sécurité des données et la gestion des risques**
- **Difficultés à résoudre de manière à garantir un contrôle véritable et le consentement éclairé des patients et des autres personnes concernées**
- **Obligations positives des États à protéger la vie et la santé par l'intermédiaire de mécanismes nationaux de signalement**
- **Résolution des conflits entre la « liberté d'innover » et une véritable protection des droits de l'homme**

Ces questions et d'autres soulevées par des rapports antérieurs du Conseil de l'Europe ne seront pas examinées ici en détail, mais évoquées dans le chapitre intitule « Les conséquences potentielles de l'IA sur la relation médecin-patient » consacré à l'examen de l'impact potentiel de l'IA sur la relation médecin-patient.

4. VUE D'ENSEMBLE DES TECHNOLOGIES DE L'IA EN MEDECINE

Comme indiqué au chapitre intitulé « Historique et contexte », les technologies pouvant être décrites comme de l'IA sont très diverses. Les définitions de haut niveau des concepts pertinents, notamment de l'intelligence artificielle, des algorithmes et de l'apprentissage machine, étant établies, il convient d'examiner plus en détail les types d'applications possibles de l'IA dans le domaine médical. Dans la mesure où le présent rapport est consacré aux effets de l'IA sur la relation médecin-patient, ses applications médicales ne seront pas toutes abordées. Pour commencer, nous pouvons opérer une distinction entre trois types d'IA en fonction des utilisateurs visés :

- **l'IA pour les chercheurs en biomédecine**
- **l'IA pour les patients**
- **l'IA pour les professionnels de santé**

> Ces deux dernières catégories sont les plus pertinentes dans le cadre du présent rapport compte tenu de l'attention particulière accordée à la relation médecin-patient.

D'autres taxonomies sont évidemment possibles, comme en témoigne un rapport récent de l'OMS, qui distingue les applications de l'IA pouvant être utilisées dans :

- **Les soins de santé**
- **La recherche en santé et le développement de médicaments**
- **La gestion et la planification des systèmes de santé**
- **La santé publique et la surveillance de celle-ci**

La taxonomie adoptée ici s'appuie sur les utilisateurs ciblés par les systèmes d'IA, car les solutions adaptées aux défis éthiques que présentent ces systèmes varient généralement en fonction des intérêts, du niveau de compétence et des exigences des différents groupes de parties prenantes.

Bien que cela ne concerne pas directement la relation médecin-patient, il est intéressant de passer en revue quelques exemples d'utilisation de l'IA pour la recherche médicale. L'une de ses applications les plus courantes dans la recherche biomédicale est la découverte de médicaments. Par exemple, des informaticiens et des cancérologues de l'Institute of Cancer Research et de la fondation Royal Marsden NHS ont récemment découvert un nouveau traitement médical pour une forme grave de cancer du cerveau chez les enfants (gliome diffus pontique

intrinsèque)[89]. De même, les dernières avancées réalisées par Deepmind à travers le programme AlphaFold sur le repliement des protéines indiquent que l'IA est porteuse d'avenir pour la recherche fondamentale[90]. Elle peut également être utilisée pour structurer et étiqueter des ensembles de données médicales désorganisés ou hétérogènes ou pour y effectuer des recherches. Les classifieurs d'images peuvent notamment traiter des volumes de données d'imagerie médicale très importants bien plus rapidement que les étiqueteurs manuels. Ces systèmes peuvent également servir à des fins administratives ou opérationnelles, comme expliqué plus loin.

Il faut relever un domaine d'utilisation de l'IA qui rend les frontières entre la recherche et les soins cliniques floues, celui du diagnostic génétique préimplantatoire, dans lequel un algorithme résume « les effets estimés de centaines ou de milliers de variations génétiques associées au risque pour un individu d'avoir un problème de santé ou une caractéristique spécifique ». Cette pratique fait planer le spectre de l'eugénisme, car elle pourrait permettre aux parents de sélectionner des embryons qui présentent des avantages sur la santé, mais aussi des caractéristiques socialement souhaitables et non associées à des maladies[91].

De nombreuses applications d'IA sont en cours de développement pour être utilisables directement par les patients, souvent en collaboration avec un professionnel de santé ou un agent artificiel. Elles comprennent les applications de télémédecine permettant l'observation et les rencontres cliniques à distance ainsi que le traitement par observation vidéo ; les assistants virtuels ou les chat bots d'information ou de tri ; les applications pour la gestion des maladies chroniques, comme les maladies cardiovasculaires ou l'hypertension ; les applications de santé et de bien-être ; les systèmes de suivi personnel de la santé, notamment les dispositifs portables comprenant des fonctions analytiques intégrées et prodiguant des recommandations comportementales ; les systèmes de suivi à distance pour la reconnaissance faciale, la détection de l'allure, la biométrie et les comportements liés à la santé[92].

L'un des avantages présumés des systèmes d'IA destinés aux patients est de « [leur] permettre de mieux maîtriser les soins de santé dont ils bénéficient et de mieux comprendre l'évolution de leurs besoins »[93]. Les systèmes de suivi de la santé et de télémédecine pourraient par exemple aider les patients à gérer eux-mêmes leurs problèmes de santé chroniques, tels que le diabète, l'hypertension ou les maladies cardiovasculaires[94]. Les « chat bots » thérapeutiques peuvent également constituer

[89] Andrew Gregory, « Scientists use AI to create drug regime for rare form of brain cancer in children », *The Guardian* (2021), https://www.theguardian.com/science/2021/sep/23/scientists-use-ai-to-create-drug-regime-for-rare-form-of-brain-cancer-in-children (consulté le 26 sept. 2021) ; Carvalho *et al.*, *supra* note 7.
[90] Jumper *et al.*, cf. *supra* note 7.
[91] Sheetal Soni et Julian Savulescu, « Polygenic Embryo Screening: Ethical and Legal Considerations », *The Hastings Center* (2021), https://www.thehastingscenter.org/polygenic-embryo-screening-ethical-and-legal-considerations/ (consulté le 23 novembre 2021).
[92] Mittelstadt *et al.*, cf. *supra* note 3.
[93] Organisation mondiale de la santé, cf. *supra* note 1.
[94] Mittelstadt *et al.*, cf. *supra* note 3 ; Ministère britannique de la Santé et de la Protection sociale, *The Topol Review: Preparing the healthcare workforce to deliver the digital future* (2019), https://topol.hee.nhs.uk/.

une aide pour la gestion des problèmes de santé mentale[95]. D'aucuns affirment, par exemple, que l'application de traitement de langage naturel GPT-3 pourrait, à terme, servir de base aux agents conversationnels qui travaillent directement avec des patients, notamment en faisant office de point de contact initial ou en effectuant un tri entre les patients nécessitant une intervention d'urgence et les autres (ce qui suscite davantage de controverse) [96]. Ces utilisations semblent fort probables compte tenu du déploiement de « médecins généralistes virtuels » sous forme de chat bots, qui orientent les demandes d'intervention et fournissent des informations aux patients[97] ; il faut toutefois noter que ces applications de l'IA font l'objet d'importants débats, en ce qui concerne leur acceptabilité sur le plan éthique et leur réglementation[98]. Par ailleurs, elles pourraient entraîner une réduction de l'accès aux soins humains[99].

Enfin, un large éventail d'applications de l'IA sont destinées aux professionnels de santé. Trois grandes catégories peuvent être définies :

- **Les applications conçues à des fins de diagnostic, thérapeutiques et pour d'autres formes de soins cliniques**
- **Les applications conçues à des fins opérationnelles ou administratives**
- **Les applications conçues pour la surveillance de la santé publique**

La distinction entre chacune d'entre elles n'est pas toujours tranchée, comme nous le verrons plus loin. Afin de limiter l'objet du présent rapport aux effets potentiels de l'IA sur la relation médecin-patient, seules les deux premières seront étudiées. La surveillance de la santé publique pourrait également être considérée comme une extension de l'expérience clinique ou de la relation médecin-patient, dans la mesure où les patients peuvent être contactés de façon proactive par les responsables de la santé publique pour un suivi clinique. Néanmoins, le présent rapport porte principalement sur l'expérience clinique immédiate et sur la relation entre les professionnels de santé et leurs patients.

Les systèmes d'IA destinés aux soins cliniques sont conçus pour remplir des tâches très variées, qui comprennent la formulation de recommandations en matière de diagnostic, l'optimisation des protocoles de traitement, et plusieurs autres formes d'aide à la prise de décision.

D'après l'OMS :

> « l'IA fait actuellement l'objet d'évaluations pour être utilisée pour le diagnostic radiologique en oncologie (imagerie thoracique, imagerie abdominale et

[95] Ministère britannique de la Santé et de la Protection sociale, cf. *supra* note 93.
[96] Diane M. Korngiebel et Sean D. Mooney, « Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery », *NPJ Digital Medicine* n°4, pp. 1–3 (2021).
[97] Weiyu Wang et Keng Siau, « Trust in health chatbots » (2018) ; Claire Woodcock *et al.*, « The Impact of Explanations on Layperson Trust in Artificial Intelligence–Driven Symptom Checker Apps: Experimental Study », *Journal of Medical Internet Research* n ° 23, p. e29386 (2021).
[98] Gareth Iacobucci, « Row over Babylon's chatbot shows lack of regulation » (2020) ; Wang et Siau, cf. *supra* note 96.
[99] Organisation mondiale de la santé, cf. *supra* note 1.

pelvienne, coloscopie, mammographie, imagerie cérébrale et optimisation des doses du traitement radiologique), pour des applications non radiologiques (dermatologie, pathologie), pour le diagnostic de la rétinopathie diabétique, pour l'ophtalmologie et pour le séquençage de l'ARN et de l'ADN visant à guider l'immunothérapie »[100].

Parmi les futures applications de l'IA en cours de développement (mais pas encore déployées dans la pratique clinique) figurent des systèmes permettant de dépister « les accidents vasculaires cérébraux, les pneumonies, le cancer du sein par l'imagerie[101], les maladies coronariennes par échocardiographie[102] et le cancer du col de l'utérus[103] », notamment des systèmes conçus spécifiquement pour être utilisés dans les pays à revenu faible et intermédiaire[104]. Des systèmes sont actuellement mis au point pour prévoir le risque de maladies liées au mode de vie, comme les maladies cardiovasculaires[105] et le diabète[106].

Le développement de systèmes de classification d'images médicales s'est fortement accru ces dernières années. Des travaux antérieurs, par exemple, ont montré que les réseaux neuronaux peuvent atteindre une sensibilité systématiquement plus élevée dans la détection de signes pathologiques en radiologie[107]. Des systèmes de classification d'images peuvent également être utilisés pour permettre un meilleur dépistage de la tuberculose[108], de la Covid-19 et d'autres maladies grâce à

[100] Wenya Linda Bi *et al.*, « Artificial intelligence in cancer imaging: clinical challenges and applications », *CA: a cancer journal for clinicians* n ° 69, pp. 127–157 (2019) ; Organisation mondiale de la santé, cf. *supra* note 1.

[101] Pranav Rajpurkar *et al.*, « Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists », *PLOS Medicine* n° 15, p. e1002686 (2018) ; Babak Ehteshami Bejnordi *et al.*, « Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer », *Jama* n° 318, pp. 2199–2210 (2017).

[102] Maryam Alsharqi *et al*., « Artificial intelligence and echocardiography », *Echo research and practice* n° 5, R115–R125 (2018).

[103] « Using Artificial Intelligence to Detect Cervical Cancer », NIH Director's Blog (2019), https://directorsblog.nih.gov/2019/01/17/using-artificial-intelligence-to-detect-cervical-cancer/ (consulté le 1er décembre 2021).

[104] Organisation mondiale de la santé, cf. *supra* note 1 ; « Dépistage et traitement novateurs et abordables pour prévenir le cancer du col de l'utérus », Unitaid, https://unitaid.org/project/innovative-affordable-screening-and-treatment-to-prevent-cervical-cancer/#fr (consulté le 1er décembre 2021).

[105] Rui Fan *et al.*, « AI-based prediction for the risk of coronary heart disease among patients with type 2 diabetes mellitus », *Scientific reports* n°10, pp. 1–8 (2020) ; Yang Yan *et al.*, « The primary use of artificial intelligence in cardiovascular diseases: what kind of potential role does artificial intelligence play in future medicine? », *Journal of geriatric cardiology: JGC* n°16, p. 585 (2019).

[106] Jyotismita Chaki *et al.*, « Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review », *Journal of King Saud University-Computer and Information Sciences* (2020).

[107] Ohad Oren, Bernard J Gersh et Deepak L Bhatt, « Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints », *The Lancet Digital Health* n° 2, pp. e486–e488 (2020).

[108] Yan Xiong *et al.*, « Automatic detection of mycobacterium tuberculosis using artificial intelligence », *Journal of thoracic disease* n° 10, p. 1936 (2018).

l'interprétation d'images en couleurs[109] ou à rayons X[110]. Les systèmes de « jumeaux numériques », qui simulent des organes ou des systèmes multiorganes de patients individuels à des fins de modélisation et de prévention des maladies[111], constituent un autre phénomène nouveau.

De manière générale, la mise en œuvre de l'IA dans le domaine des soins cliniques n'en est encore qu'à ses balbutiements. Une efficacité clinique n'a été établie que pour relativement peu de systèmes par rapport aux nombreuses activités de recherche menées sur les applications de l'IA en médecine. Bien souvent, les travaux de recherche, le développement et les essais pilotes ne se traduisent pas par une efficacité clinique avérée, une mise sur le marché ou un déploiement massif. Globalement, la généralisation de la performance des essais cliniques à la pratique reste à démontrer[112].

En 2019, une méta-analyse portant sur les classifieurs d'images d'apprentissage profond dans les soins de santé a révélé que malgré les allégations selon lesquelles les systèmes d'IA seraient d'une efficacité équivalente à celle des professionnels de santé humains,

> « Peu d'études présentent des résultats ayant fait l'objet d'une validation externe ou comparent la performance des modèles d'apprentissage profond à celles des professionnels de santé en utilisant un même échantillon ». De même, « les études sur l'apprentissage profond font rarement l'objet de rapports, ce qui limite la fiabilité d'interprétation de la précision diagnostique indiquée »[113].

Les données probantes relatives à l'efficacité clinique des systèmes d'apprentissage profond ont pu être améliorées depuis ce constat, mais une adoption généralisée dépendra de toute évidence de rapports normalisés sur la précision permettant aux instances de réglementation de la médecine et aux organismes d'excellence en matière de soins cliniques d'évaluer leur efficacité clinique.

Sur le court terme, un défi à relever en ce qui concerne les classifieurs d'images est de mettre au point des systèmes capables de traiter en même temps plusieurs types d'images ou de scanners, tels que les rayons X ou les tomodensitomètres, qui sont souvent étudiés ensemble par les radiologues alors que les systèmes d'IA ne peuvent en général que les interpréter séparément. Il en va de même pour le dépistage

---

[109] *Id.*

[110] Apoorva Mandavilli, « These Algorithms Could Bring an End to the World's Deadliest Killer », *The New York Times* (2020), https://www.nytimes.com/2020/11/20/health/tuberculosis-ai-apps.html (consulté le 1er décembre 2021).

[111] Matthias Braun, « Represent me: please! Towards an ethics of digital twins in medicine », *J Med Ethics* (2021).

[112] Organisation mondiale de la santé, cf. *supra* note 1 p. 6.

[113] Xiaoxuan Liu *et al.*, « A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis », *The Lancet Digital Health* n° 1, pp. e271–e297 (2019).

simultané de problèmes de santé ou de pathologies, les classifieurs existants n'étant souvent entraînés que pour déceler un seul type d'anomalie[114].

Enfin, de nombreux systèmes d'IA sont aussi élaborés pour être utilisés à des fins administratives ou opérationnelles. Ils peuvent fournir une aide sur diverses tâches liées à l'administration des hôpitaux et sur l'évaluation des activités. Par exemple, les outils de planification de sortie peuvent estimer les dates de sortie des patients hospitalisés et les obstacles à celle-ci, et signaler aux professionnels les patients qui sont cliniquement (presque) prêts à quitter l'hôpital, tout en fournissant une liste des mesures à prendre avant leur sortie. Certains dispositifs peuvent même fixer des dates pour les rendez-vous et les soins nécessaires[115]. Des systèmes de traitement en langage naturel pourraient servir à automatiser les tâches courantes ou à forte intensité de main-d'œuvre, comme la recherche et la navigation dans les registres de dossiers médicaux électroniques ou la préparation automatique de documents médicaux et d'ordonnances[116]. Selon l'OMS, « les professionnels pourraient avoir recours à l'IA pour synthétiser les dossiers des patients pendant les consultations, identifier les patients à risque et les groupes vulnérables, l'utiliser comme aide pour prendre des décisions difficiles concernant le traitement, ou encore pour déceler les erreurs cliniques. Dans les pays à revenu faible et intermédiaire, par exemple, l'IA pourrait être utilisée dans la gestion des traitements antiviraux pour prévoir la résistance aux médicaments contre le VIH et l'évolution de la maladie, afin d'aider les médecins à optimiser le traitement »[117].

Il est souvent difficile de distinguer les applications de l'IA pour les soins cliniques et la recherche et ses utilisations à des fins opérationnelles et d'amélioration par les hôpitaux et les systèmes de santé. Nombre de ces systèmes ont été mis au point pour identifier les patients à risque. Par exemple, les structures appartenant au réseau de santé de l'UCLA ont recours à un outil qui a repéré, dans les établissements de soins primaires, les patients présentant un risque élevé d'être hospitalisés ou d'aller fréquemment aux urgences dans l'année à venir. De la même manière, l'Université de la santé et des sciences de l'Oregon utilise un algorithme de régression pour déceler les signes de septicémie chez les patients de l'hôpital[118]. Ces deux types d'IA sont considérés comme un type d'outil opérationnel permettant d'assurer un suivi de la prise en charge et d'accorder la priorité à la qualité des soins, et ne relèvent pas des soins cliniques ou de la recherche.

[114] Stephanie Price, « Technological innovations of AI in medical diagnostics », *Health Europa* (2020), https://www.healtheuropa.eu/technological-innovations-of-ai-in-medical-diagnostics/103457/ (consulté le 6 septembre 2021).
[115] Robbins et Brodwin, cf. *supra* note 5.
[116] Korngiebel et Mooney, cf. *supra* note 95.
[117] Organisation mondiale de la santé, cf. *supra* note 1 ; Jerome Amir Singh, « Artificial Intelligence and global health: opportunities and challenges », *Emerging Topics in Life Sciences* n° 3, pp. 741–746 (2019).
[118] Robbins et Brodwin, cf. *supra* note 5.

5. CADRE THEORIQUE DE LA RELATION MEDECIN-PATIENT

La santé est un bien fondamental valorisé dans de nombreux contextes, notamment dans la vie privée, sociale et économique, et lié à la subsistance et au bien-être de la personne dans son ensemble. Quand la santé fait défaut, il n'est pas possible de faire des projets personnels, de poursuivre des objectifs ou de se créer une identité qui ne soit pas restreinte par une limitation physique, mentale ou sociale[119]. Il s'agit donc d'une condition préalable à la jouissance d'autres biens humains.

De manière générale, la finalité de la médecine est de garantir la santé d'une société et des individus qui la composent[120]. Bien qu'il soit difficile de définir la santé et la maladie en tant que concepts, la médecine est largement reconnue comme une pratique visant à promouvoir la santé, œuvrant ainsi en faveur d'un bien fondamental[121]. L'absence de consensus sur une définition « correcte » de la santé, qui ressort des débats à ce sujet, ne remet pas en cause la valeur fondamentale de la santé pour la vie humaine[122]. La réalisation des objectifs de la médecine passe par de « bonnes » rencontres médicales avec les patients[123]. Dans la poursuite de ces objectifs dans le cadre de la relation médecin-patient, les capacités morales et techniques doivent être réunies dans l'intérêt du patient, car les activités médicales touchent des individus ayant des valeurs et des intérêts moraux.

Comme indiqué au chapitre intitulé « La Convention d'Oviedo et les principes des droits de l'homme en matière de santé », la Convention d'Oviedo définit les valeurs suivantes :

- **La dignité humaine**
- **La primauté de l'intérêt du patient sur l'intérêt de la société et de la science**
- **Le droit à la vie**

---

[119] Andrew Edgar, « The expert patient: Illness as practice », *Medicine, Health Care and Philosophy* n° 8, pp. 165–171 (2005).
[120] Organisation mondiale de la santé, *Préambule à la Constitution de l'Organisation mondiale de la Santé* (1948) ; Kenneth William Musgrave Fulford, *Moral theory and medical practice* (1989).
[121] Fulford, cf. *supra* note 119 ; Edmund D Pellegrino et David C Thomasma, *The virtues in medical practice* (1993) ; Paul Schotsmans, Bernadette Dierckx de Casterle et Chris Gastmans, « Nursing considered as moral practice: a philosophical-ethical interpretation of nursing », *Kennedy Institute of Ethics Journal* n° 8, pp. 43–69 (1998).
[122] Fulford, cf. *supra* note 119 ; Alan Petersen, « Risk, governance and the new public health », *in* Foucault: *Health and Medicine* pp. 189–206 (éd. Alan Petersen & Robin Bunton, 1997) ; Adele E. Clarke *et al.*, « Biomedicalization: Technoscientific transformations of health, illness, and U.S. biomedicine », *American Sociological Review* n° 68, pp. 161–194 (2003).
[123] Alasdair MacIntyre, *Après la vertu : étude de théorie morale* (3e édition révisée, 2007) ; Pellegrino et Thomasma, cf. *supra* note 120 ; General Medical Council, « Good Medical Practice » (2013), http://www.gmc-uk.org/static/documents/content/GMP_2013.pdf_51447599.pdf.

- **L'intégrité physique**
- **La vie privée et l'identité**
- **Le consentement éclairé**
- **Le droit de savoir et celui de ne pas savoir**
- **L'interdiction de la discrimination et des inégalités dans l'accès aux soins de santé**
- **Les normes relatives à la qualité des soins de santé**

Ces valeurs, ainsi que les divers objectifs de la médecine en tant que pratique, peuvent être mis en œuvre à travers différents types de relations médecin-patient. Les modèles de la relation médecin-patient (idéale) se sont adaptés au fil du temps pour tenir compte de l'importance croissante de l'autonomie du patient et de son juste équilibre avec d'autres obligations éthiques incombant au médecin en matière de bienfaisance, de non-malfaisance et de justice[124]. Dans un article fondateur, Emanuel et Emanuel (1992) ont proposé quatre modèles de relations médecin-patient.

- **Modèle paternaliste** – Ce modèle confère la plus grande partie du pouvoir de décision au médecin. Il suppose l'existence de valeurs ou de critères objectifs et partagés pour définir la meilleure stratégie favorisant la santé et le bien-être du patient. Le rôle du médecin est celui d'un expert, d'un praticien compétent chargé de « promouvoir le bien-être du patient indépendamment des préférences que celui-ci exprime à ce moment-là ». Le médecin agit comme « tuteur du patient, en exposant et en mettant en œuvre ce qu'il y a de mieux pour le patient ». L'autonomie du patient se traduit uniquement par son consentement à la mise en œuvre de ce que le médecin a défini comme la meilleure stratégie.
- **Modèle informatif** – Ce modèle confère en revanche la plus grande partie du pouvoir de décision au patient. L'objectif des interactions cliniques est « que le médecin fournisse au patient toutes les informations pertinentes, que le patient choisisse les interventions médicales et que le médecin mette en œuvre les interventions choisies ». Ce modèle ne suppose pas l'existence de valeurs objectives ; les valeurs et l'intérêt du patient sont considérés comme connus ou établis par le patient, et non par le médecin. Le rôle de ce dernier est de présenter des faits qui contribuent à ce que le patient prenne une décision correspondant le mieux à ses intérêts.
- **Modèle interprétatif** – Très proche du modèle informatif, ce modèle accorde toutefois un plus grand rôle au médecin consistant à aider le patient à comprendre ses valeurs et ses intérêts, ainsi que les conséquences possibles

[124] Tom L. Beauchamp & James F. Childress, *Les Principes de l'éthique biomédicale* (2009) ; E. J. Emanuel & L. L. Emanuel, « Four models of the physician-patient relationship », *JAMA: The Journal Of The American Medical Association* n° 267, pp. 2221–2226 (1992).

de différentes interventions sur ces éléments. Le médecin agit comme conseiller qui aide le patient à « expliciter et rendre cohérentes » ses valeurs, sans les évaluer, ni essayer de les hiérarchiser à la place du patient. Le choix final de l'intervention revient encore au patient selon ce modèle, mais le médecin oriente plus activement la prise de décision que dans le modèle informatif

- **Modèle délibératif** – Très proche du modèle informatif, ce modèle accorde toutefois un rôle plus important au médecin consistant à évaluer et hiérarchiser les valeurs du patient et à « mettre en évidence les types de valeurs que renferme chaque possibilité… en évoquant les raisons pour lesquelles certaines valeurs liées à la santé méritent davantage d'être recherchées ». La délibération entre médecin et patient se limite « aux valeurs liées à la santé, c'est-à-dire celles qui ont une influence sur la maladie et la prise en charge du patient ou qui sont influencées par celles-ci ; le médecin reconnaît que nombre d'éléments moraux ne sont pas liés à la maladie et à la prise en charge du patient et sortent du cadre de leur relation professionnelle ». La délibération vise la persuasion morale, et non la contrainte, le patient décidant en définitive de la validité et de la priorité appropriées de ces valeurs dans sa vie. Alors que le médecin est un conseiller selon le modèle interprétatif, il agit « en tant qu'enseignant ou ami, en impliquant le patient dans le dialogue sur la meilleure stratégie à adopter » dans le modèle délibératif. Il indique à la fois au patient ce qu'il pourrait faire et, dans le cadre de sa compréhension de la vie et des valeurs du patient, l'intervention qu'il devrait choisir selon lui. La décision finale revient toujours au patient mais est soumise à plus de persuasion et à l'argumentation normative du médecin. Dans ce modèle, l'autonomie du patient est conçue comme un outil de développement personnel moral ; « le patient est en mesure de ne pas simplement suivre des préférences non examinées ou des valeurs examinées, mais d'envisager, à travers le dialogue, d'autres valeurs non liées à la santé, leur honorabilité et leurs répercussions sur la prise en charge ».

Un cinquième modèle, le « modèle instrumental », est évoqué par Emanuel et Emanuel dans leur analyse des relations médecin-patient, mais a rapidement été écarté pour des raisons morales. Aucune importance n'y est accordée aux valeurs du patient ; au lieu de cela, le médecin prend une décision ou convainc le patient de choisir un protocole de traitement particulier en se fondant sur des valeurs externes, comme le bien de la société ou de la science. Bien que ce modèle ait été moralement condamné, à juste titre, il convient de noter qu'il reste potentiellement pertinent pour servir d'avertissement pour le déploiement de l'IA. Lorsque l'IA est recherchée non pas pour le bien du patient mais par souci d'efficacité ou de réduction des coûts, il serait possible de faire valoir que la relation médecin-patient est instrumentalisée. L'influence de ces valeurs externes sur les relations médecin-patient est détaillée plus loin.

Chacun de ces modèles de relations médecin-patient traduit différents degrés de respect accordés à l'autonomie et au développement personnel moral du patient. Les droits et les valeurs inscrits dans la Convention d'Oviedo donnent une certaine idée de l'acceptabilité générale de ces modèles. L'adoption d'un modèle paternaliste

semble susceptible d'entraîner une violation de l'exigence de consentement éclairé définie à l'article 5. De même, un modèle délibératif ne respecterait pas un aspect spécifique de cette exigence de consentement, qui est développé dans le rapport explicatif à la convention : le consentement du patient devrait se fonder sur « une information objective » donnée « en l'absence de toute pression de la part d'autrui ». La difficulté à fournir des informations objectives sera à nouveau abordée dans une partie du chapitre intitulé « Les conséquences potentielles de l'IA sur la relation médecin-patient » portant sur la transparence des soins cliniques facilités par l'IA.

L'éthique professionnelle dans le domaine de la médecine

La Convention d'Oviedo appelle expressément les états membres et les associations professionnelles à définir des normes et des obligations de qualité dans son article 4. Comment la médecine, en tant que profession, fixe-t-elle ses propres normes en matière de soins cliniques et de relations médecin-patient, et en fonction de quels objectifs ou de quelles valeurs ? Le présent chapitre propose à cet égard un cadre théorique pour comprendre la médecine en tant que profession autonome. Ce cadre est conforme à nombre des valeurs prescrites dans la Convention d'Oviedo ; cet aspect sera examiné plus en détail au chapitre intitulé « Les conséquences potentielles de l'IA sur les relations médecin-patient ».

Une approche influente qui définit des fins idéales (et par voie de conséquence des normes et des biens internes) pour la médecine, et fondée sur l'éthique des vertus d'Alasdair MacIntyre[125], a été proposée par Pellegrino et Thomasma[126]. En vertu de cette approche, la médecine peut être considérée comme une « pratique morale »[127], avec des vertus décrivant les qualités requises des médecins, en plus des « connaissances scientifiques et médicales, des aptitudes pratiques et de l'expérience qui garantit que le médecin fait ce qu'il faut en adoptant la bonne attitude pour atteindre les objectifs de la médecine »[128]. Selon la définition de MacIntyre, la médecine est une pratique morale, car en tant que profession autonome, elle définit et respecte des normes internes de qualité pour la prise en charge médicale ainsi que des processus de certification visant à maintenir ces normes[129].

La finalité d'une pratique peut être appréhendée à travers un examen critique de ses éléments internes ou de ses normes d'évaluation ; pour la médecine, il s'agit de celles qui régissent les relations médecin-patient[130]. Comme le montrent ces relations, « les finalités de la médecine sont… le rétablissement ou l'amélioration de la santé et, plus immédiatement, la guérison, c'est-à-dire l'élimination de la maladie ou, lorsque celle-

[125] MacIntyre, cf. *supra* note 122.
[126] Pellegrino et Thomasma, cf. *supra* note 120 p. 52.
[127] Pellegrino et Thomasma, cf. *supra* note 120.
[128] Petra Gelhaus, « The desired moral attitude of the physician (I) empathy », *Medicine, Health Care and Philosophy* n° 15, pp. 103–113, 104 (2012).
[129] Pellegrino et Thomasma, cf. *supra* note 120 ; Paul Starr, *La Transformation sociale de la médecine américraine (Édition révisée) : l'essor d'une profession souveraine et la naissance d'une vaste industrie* (2e édition révisée, 2017) ; General Medical Council, *Consent Guidance* (2008), http://www.gmc-uk.org/guidance/ethical_guidance/consent_guidance_index.asp ; General Medical Council, cf. *supra* note 122.
[130] Pellegrino et Thomasma, cf. *supra* note 120 p. 52.

ci n'est pas possible, la prise en charge et l'assistance au patient pour qu'il vive avec la douleur, l'inconfort ou l'invalidité chronique »[131]. La relation médecin-patient, entendue comme un type de « relation thérapeutique », est le principal mécanisme permettant d'atteindre ces fins.

Considérer la médecine comme une pratique morale régie par des normes de bonnes pratiques qui sont mises en œuvre dans le cadre d'une relation thérapeutique, ce n'est pas adopter une vision archaïque de la médecine comme d'une relation paternaliste patient-prestataire. Au contraire, la relation thérapeutique implique aussi bien des interventions cliniques que la fourniture d'informations ou de services aux patients visant à leur apporter des connaissances, à les autonomiser ou à les aider à prendre soin d'eux-mêmes. Même dans les rencontres cliniques modernes dans lesquelles les patients sont « autonomisés » grâce à un accès démocratisé aux données médicales, des valeurs personnelles et une expérience vécue de la maladie[132], le « rôle » idéal du médecin, qui requiert certaines compétences techniques et une formation professionnelle, est incontestable – la question est plutôt de savoir s'il faut se fier à ces compétences sans se poser de question.

Obligations déontologiques (*fudiciary duties*) et relation thérapeutique

Les principes de droits de l'homme en matière de santé et les droits complémentaires consacrés par des instruments de politique comme la Charte des droits fondamentaux de l'Union européenne reflètent les obligations morales et de confidentialité de la profession médicale. Comme évoqué précédemment, ces obligations peuvent être rattachées aux objectifs fondamentaux ou aux finalités de la médecine en tant que pratique et à de nombreux fondements théoriques possibles, notamment les droits humains, l'éthique des soins et l'éthique féministe, ainsi que dans l'éthique des vertus.

Le reste du présent chapitre rend compte des concepts de relation thérapeutique et d'obligations déontologiques (*fudiciary duties*) en médecine développés dans le cadre de l'éthique des vertus. Une approche fondée sur les vertus confère une grande importance à la prise en charge du patient dans son ensemble et à la promotion de son bien-être à travers la mise en place de bonnes pratiques. Des normes sont définies en fonction de critères comme la compassion, qui « garantit que le patient n'est pas considéré comme un simple numéro »[133], la compréhension contextuelle des valeurs, des antécédents et des préoccupations du patient, un « intérêt pour les processus internes du patient… une compétence appropriée pour répondre de manière non verbale et en établissant un dialogue habile et tenant compte de ses besoins » [134], en plus des compétences spécialisées pour « résoudre » le problème du patient ou gérer un problème de santé persistant. Cela dit, ces objectifs fondamentaux sont communs à de nombreuses autres approches qui ne relèvent pas de l'éthique des vertus. Par exemple, les approches de l'éthique des soins et de l'éthique féministe se concentrent sur des éléments connexes, comme le rôle d'aidant du professionnel de santé, les

[131] *Id.* pp. 52–3.
[132] Emanuel et Emanuel, cf. *supra* note 123 ; Edgar, cf. *supra* note 118.
[133] Petra Gelhaus, « The desired moral attitude of the physician: (II) compassion », *Medicine, Health Care and Philosophy* n° 15, pp. 397–410, 405 (2012).
[134] Gelhaus, cf. *supra* note 127 p. 108.

relations et les responsabilités en matière de soins (par opposition à la justice et aux droits)[135], la connaissance tacite et la prise en charge prenant en compte le contexte, l'intérêt et les besoins du patient en tant qu'individu unique intégré dans la société, ainsi que les déséquilibres de pouvoir et la contrainte découlant de la position vulnérable du patient.

Plusieurs caractéristiques de la relation thérapeutique créent pour les professionnels des obligations morales de protéger l'intérêt des patients[136]. Il peut s'agir plus précisément des éléments suivants :

- **Vulnérabilité et inégalité –** Les patients ont le sentiment de perdre le contrôle sur la définition et la poursuite de leurs objectifs personnels et peuvent ressentir un stress émotionnel, de la peur, de l'inquiétude et de l'angoisse[137]. Leur priorité immédiate est alors de retrouver la santé et un état de bien-être, ce qui passe par le soulagement ou l'élimination des symptômes. Une relation déséquilibrée est créée dans laquelle le patient, pour recouvrer la santé, est contraint de demander l'aide d'une personne ayant une expertise médicale privilégiée. Les médecins sont tenus de ne pas utiliser leur expertise ou leur position de pouvoir privilégiée pour exploiter le patient « vulnérable »[138].

- **Caractère confidentiel –** Le patient place expressément ou tacitement sa confiance en un médecin déterminé et révèle des aspects de lui-même ou de sa vie pour permettre le diagnostic et guérir, renonçant ainsi à une partie de sa vie privée pour que « d'autres personnes aient accès à des informations personnelles ou à [leur] corps »[139]. Les médecins ont l'obligation morale d'utiliser les informations et l'accès que le patient leur donne dans le cadre d'une relation de confiance, dans l'intérêt supérieur du patient et non dans leurs propres intérêts[140].

- **Nature des décisions médicales** – Les décisions médicales présentent un mélange de caractéristiques techniques et morales. Le diagnostic du médecin et la prise en charge du patient doivent être précis d'un point de vue technique pour contribuer à sa santé physique[141]. Toutefois, les décisions prises devraient également favoriser le bien-être moral ou l'autonomie du patient en tant qu'être

[135] Carol Gilligan, *Une voix différente. Pour une éthique du care* (1993).
[136] Pellegrino et Thomasma, cf. *supra* note 120 pp. 35–6, 42–4 ; Schotsmans, Dierckx de Casterle et Gastmans, cf. *supra* note 120.
[137] Pellegrino et Thomasma, cf. *supra* note 120 ; David B. Morris, « About suffering: Voice, genre, and moral community », *Daedalus* n° 125, pp. 25–45 (1996) ; Keith Bauer, « Cybermedicine and the moral integrity of the physician–patient relationship », *Ethics and information technology* n° 6, pp. 83–91 (2004) ; Deborah Lupton, « The digitally engaged patient: self-monitoring and self-care in the digital health era », *Social Theory & Health* n° 11, pp. 256–270, 263 (2013).
[138] Pellegrino et Thomasma, cf. *supra* note 120 pp. 35–6 ; Gilligan, cf. *supra* note 134.
[139] Beauchamp et Childress, cf. *supra* note 123 p. 298.
[140] Pellegrino et Thomasma, cf. *supra* note 120 pp. 35–6, 42–4 ; Bauer, cf. *supra* note 136 ; John Heritage *et al.*, « Problems and Prospects in the Study of Physician-Patient Interaction: 30 Years of Research », *Annual Review of Sociology* n° 32, pp. 351–374, 355 (2006) ; O. Karnieli-Miller & Z. Eisikovits, « Physician as partner or salesman? Shared decision-making in real-time encounters », *Social Science & Medicine* n° 69 pp. 1–8, 2 (2009).
[141] Pellegrino et Thomasma, cf. *supra* note 120 pp. 35–6, 42–4.

doué de valeur morale, dans le sens où elles devraient être alignées avec ses valeurs[142].

- **Caractéristiques des connaissances médicales** – Les connaissances médicales ne sont pas privées. En vue d'assurer la disponibilité d'un nombre suffisant de professionnels de santé, les sociétés fournissent aux médecins les connaissances privilégiées et l'accès aux corps humains qui leur permettent d'acquérir l'expertise médicale nécessaire, et peuvent limiter la reconnaissance du statut de praticien de la médecine aux seules personnes formées de cette manière. Les médecins ont l'obligation morale d'agir en tant que gardiens de ces connaissances, en veillant à ce qu'elles soient facilement accessibles pour les autres et utilisées de manière éthique pour la prise en charge des patients, et non dans un but purement personnel[143].

- **Complicité morale –** Le médecin est le canal par lequel les interventions médicales parviennent jusqu'au patient, dans la mesure où il doit donner son accord avant que toute intervention ne soit menée. Du fait de sa position, il a l'obligation morale d'agir comme « point de filtrage », en protégeant le bien-être du patient et en reconnaissant sa complicité dans toute intervention mise en œuvre[144].

Ces caractéristiques ne sont pas indéniables ; par exemple, on peut critiquer le fait de parler de vulnérabilité et d'inégalité pour qualifier l'expérience du malade, car cela ne semble s'appliquer qu'à de graves problèmes de santé pour lesquels des traitements sont possibles[145]. Bien que l'approche de la « relation thérapeutique » décrive un modèle idéal de relations médecin-patient (et, par voie de conséquence, de la médecine elle-même), la notion sous-jacente selon laquelle être médecin implique des obligations morales envers le patient est communément acceptée[146]. Le caractère fondamental de la relation médicale en tant que relation dans laquelle un patient dans le besoin recherche des connaissances, une expertise ou une prise en charge médicales est incontestable. Lorsqu'il sollicite l'aide d'un professionnel, le patient accepte de façon tacite de se révéler et de dévoiler des aspects privés de sa vie au médecin disposant des compétences en la matière, afin de recouvrer la santé. La relation constitue un échange de biens sensibles pour l'amélioration de la qualité de vie, échange dans lequel le patient est contraint de s'engager par la maladie, s'il souhaite guérir. Les médecins ne sont pas consultés en tant que simples « encyclopédies du savoir », mais plutôt à titre d'experts « de confiance » qui sont en

---

[142] Beauchamp et Childress, cf. *supra* note 123 ; Karnieli-Miller et Eisikovits, cf. *supra* note 139.
[143] Pellegrino et Thomasma, cf. *supra* note 120 pp. 35–6, 42–4.
[144] *Id.* pp. 35–6, 42–4.
[145] Martha C. Nussbaum, *Frontiers of Justice Disability, Nationality, Species Membership (OIP): Disability, Nationality, Species Membership* (Tanner Lectures on Human Values) (New Ed., éd. 2007) ; Barbara Page-Hanify, « Intellectual Handicap - Achievement of Potential », *Australian Occupational Therapy Journal* n° 27, pp. 53–60 (1980).
[146] Beauchamp et Childress, cf. supra note 123 ; Andrew Edgar & Stephen Pattison, « Integrity and the moral complexity of professional practice », *Nursing Philosophy* n° 12, pp. 94–106 (2011) ; Gelhaus, cf. *supra* note 127 ; Y. M. Barilan & M. Brusa, « Deliberation at the hub of medical education: beyond virtue ethics and codes of practice », *Medicine, Health Care and Philosophy* n° 16, pp. 3–12 (2013).

mesure de procéder à une évaluation subjective et de comprendre le patient en tant que personne intégrée dans la société ayant une histoire et des valeurs[147].

Être un professionnel de santé ou faire partie de la médecine entendue comme une profession réglementée exige de s'engager à satisfaire les obligations morales découlant de la relation thérapeutique[148]. A cet égard, la médecine peut être considérée comme une « pratique morale », car ses membres forment une communauté qui partage des objectifs communs et des obligations morales[149], ce qui signifie qu'ils sont « guidés par une même source de morale – des règles, principes ou traits de caractère fondamentaux qui définiront une vie morale conforme aux fins, aux objectifs et aux buts de la médecine »[150]. D'un point de vue critique, cette vision oppose les normes et obligations s'appliquant aux praticiens à celles des institutions dans lesquelles la prise en charge a lieu. Alors que la première obligation du professionnel de santé est envers le patient, les institutions ont d'autres intérêts (légitimes) qui ont trait à la dotation en ressources et à la qualité de la prise en charge dans l'ensemble de l'institution. Du point de vue de l'éthique des vertus, les vertus médicales et les normes internes de bonnes pratiques peuvent contribuer à garantir que les fins de la médecine, et en définitive les obligations envers les patients contractées dans le cadre de la relation thérapeutique, sont respectées au fil du temps et résistent à la dégradation due à l'influence délétère des institutions et d'éléments extérieurs[151]. Les vertus spécifiques des bonnes pratiques dans le domaine médical sont précisées dans l'Annexe.

Nouveaux défis des relations médecin-patient

D'aucuns pourraient soutenir que le modèle de la relation thérapeutique est dépassé, car « la notion de patients se plaçant entre les mains d'un médecin et sollicitant son avis d'expert est remplacée par le concept de patients produisant des connaissances de santé et acquérant des connaissances d'expert pour gérer leur maladie par eux-mêmes »[152]. Cette caractérisation de la médecine suggère que les relations médecin-patient ont évolué et peuvent faire sans difficulté une place à l'IA, sans que cela n'ait d'influence sur la nature de la prise en charge médicale.

Avec l'évolution de la pratique de la médecine face aux nouvelles technologies, « une partie du passé est inévitablement perdue, pas toujours pour le pire »[153]. La médecine connaît depuis longtemps des avancées technologiques qui bousculent le modèle traditionnel dans lequel les soins sont prodigués au patient de manière individuelle et en personne par le médecin. Internet a, par exemple, donné aux patients les moyens d'accéder à un plus grand volume de données médicales, mais a entraîné de

[147] Emanuel et Emanuel, cf. *supra* note 123 p. 2225 ; Gelhaus, cf. *supra* note 127 p. 110.
[148] Starr, cf. *supra* note 128.
[149] Pellegrino et Thomasma, cf. *supra* note 120 p. 3 ; Morris, cf. *supra* note 136 ; Schotsmans, Dierckx de Casterle, et Gastmans, cf. *supra* note 120.
[150] Pellegrino et Thomasma, cf. *supra* note 120 p. 3.
[151] *Id.* p. 32 ; MacIntyre, cf. *supra* note 122.
[152] Deborah Lupton, « M-health and health promotion: The digital cyborg and surveillance society », *Social Theory & Health* n° 10, pp. 229–244, 233 (2012).
[153] Pellegrino et Thomasma, cf. *supra* note 120 p. 32.

nouveaux risques découlant d'informations trompeuses ou inexactes. L'intégration de nouvelles parties prenantes dans les relations de soins n'est pas problématique en elle-même, mais doit être mesurée en termes de conséquences sur la relation thérapeutique et sur les fins de la médecine, autrement dit, en termes de conséquences sur la prise en charge du patient.

La relation thérapeutique doit être comprise comme un cadre idéaliste régissant la relation entre les médecins « experts » et les patients « vulnérables ». En tant qu'idéal, ce modèle ne reflète pas celui des soins au « patient autonomisé », qui est apparu en parallèle ces dernières décennies[154]. En partant du principe que la médecine moderne se caractérise par des patients « autonomisés » qui affaiblissent la position privilégiée des médecins comme « experts », on ne saurait présumer que confiance et guérison vont de pair.

Cependant, la relation thérapeutique décrit les motivations du patient à solliciter une prise en charge professionnelle ou à rechercher des connaissances et des technologies lui permettant de se soigner. Qu'elle soit abordée à travers des soins professionnels ou autonomes, la vulnérabilité du patient n'est pas éliminée. Les obligations en matière de confidentialité découlant de cette vulnérabilité ne changent pas non plus lorsqu'elles sont réparties entre plusieurs sources d'expertise, qu'il s'agisse de professionnels de santé, de bases de données regroupant des connaissances médicales et des conseils, ou d'autres technologies et systèmes d'appui aux soins personnels comme la télémédecine ou les informations médicales mises à disposition sur Internet.

Dans ce contexte et avec le déploiement à venir de l'IA en médecine, il est d'autant plus important de trouver de nouveaux moyens permettant de respecter la confidentialité dans la pratique. Des questions pertinentes ont été posées, par exemple en ce qui concerne la validité et l'efficacité des connaissances médicales disponibles sur les portails en ligne. Par ailleurs, bien que les informations médicales soient de plus en plus accessibles par d'autres moyens, le rôle de l'expert en tant qu'indicateur de la fidélité à la confiance ne change pas[155]. Les fournisseurs de conseils, d'informations ou de soins médicaux de mauvaise qualité, quel qu'en soit le format, peuvent être critiqués.

Compte tenu de ce qui précède, la relation thérapeutique peut être entendue comme descriptive du caractère moral et des obligations liées à la pratique médicale, traditionnellement incarnés par les professionnels de santé mais de plus en plus diffusés à travers différentes plateformes et personnes, notamment des portails en ligne, des fabricants de dispositifs pour les consommateurs, des fournisseurs de services de bien-être, et d'autres. Même si la médecine moderne a dépassé le modèle médecin-patient unique décrit dans la relation thérapeutique, les obligations qui y sont liées n'ont pas disparu. La diffusion et le déplacement de ces obligations par les nouveaux acteurs technologiques en médecine suscitent au contraire des préoccupations sur la manière de régir au mieux l'introduction de l'IA en médecine. Notre concept de la relation thérapeutique pourrait bien évidemment être révisé pour

[154] Emanuel et Emanuel, cf. *supra* note 123.
[155] *Id.*

faire primer l'autonomie du patient sur tout le reste. Cela risque cependant de réduire le médecin à un simple prestataire de services, incapable d'appliquer pleinement l'ensemble des vertus médicales et des normes internes à la pratique.

Le choix de l'unité de mesure revêt une importance cruciale dans l'évaluation des conséquences de l'IA et des technologies algorithmiques sur les relations médecin-patient. Si l'on considère uniquement les coûts et les avantages, ou l'utilité, la justification de la facilitation et de l'augmentation des soins par l'IA est simple. Toutefois, si les technologiques algorithmiques peuvent permettre de prendre en charge un plus grand nombre de patients de manière plus efficace ou à un plus faible coût, leur utilisation peut en même temps nuire aux aspects non mécaniques des soins. Une distinction peut être établie entre ces effets des systèmes algorithmiques (et les éléments constituant leur utilité) pour le bien du patient ou de la médecine en tant que pratique régie par des normes internes et une déontologie bien établies, et ceux qui profitent aux institutions médicales et aux services de santé.

La complicité morale qui caractérise les relations médecin-patient, dans lesquelles la prise en charge est idéalement guidée par l'évaluation de l'état du patient effectuée par le professionnel en tenant compte du contexte et des antécédents du patient, ne peut être aisément reproduite dans les interactions avec les systèmes d'IA. L'introduction de l'IA comme facilitatrice ou accélératrice des soins ne modifie pas le rôle du patient, les facteurs qui le motivent à consulter un médecin, ni sa vulnérabilité. Ce qui change, c'est le mode de prestation des soins, la manière dont ils peuvent être prodigués et les personnes qui peuvent le faire. Le déplacement de l'expertise et des responsabilités des soins vers les systèmes d'IA peut entraîner de nombreux effets perturbateurs, qui sont évoqués au chapitre intitulé « Les conséquences potentielles de l'IA sur la relation médecin-patient ».

6. LES CONSEQUENCES POTENTIELLES DE L'IA SUR LA RELATION MEDECIN-PATIENT

L'IA laisse présager une variété de possibilités et d'avantages pour l'exercice de la médecine, mais également des risques. En s'appuyant sur les défis éthiques de l'IA et du contexte politique qui ont été développés dans le chapitres intitulé « Historique et contexte », « Vue d'ensemble des technologies de l'IA en médecine », et « Cadre théorique de la relation médecin-patient », le présent chapitre identifie six conséquences potentielles de l'IA sur la relation médecin-patient.

Inégalité dans l'accès à des soins de santé de qualité

Le déploiement des systèmes d'IA – en tant que technologie émergente – ne sera ni immédiat ni universel dans tous les états membres ou systèmes de santé. L'ampleur du déploiement, sa vitesse, mais aussi la hiérarchisation des priorités sera inévitablement hétérogène dans les établissements et les régions. Les systèmes de télémédecine, par exemple, sont bien adaptés à la prestation de soins dans des endroits éloignés ou inaccessibles, ou en cas de pénurie de personnel de santé ou de spécialistes[156]. Ils permettent de combler les lacunes dans la couverture des soins de santé, mais pas nécessairement avec des soins de qualité équivalente aux soins traditionnels en face à face. À court terme, les conséquences sur la relation médecin-patient pourraient donc être beaucoup plus importantes dans les zones où les pénuries de personnel étaient déjà un problème ou bien dans celles touchées par de nouvelles pénuries dues à la pandémie de Covid-19. La nature et l'intensité de ces conséquences restent à envisager.

Le déploiement inévitablement hétérogène de l'IA risque de provoquer des déséquilibres géographiques dans la performance des systèmes de soins de santé et des inégalités dans l'accès à des soins de qualité. Cela va dans les deux sens : si les systèmes d'IA améliorent la qualité des soins, en fournissant un diagnostic plus précis ou efficace, en élargissant l'accès aux soins ou en développant de nouvelles interventions pharmaceutiques et thérapeutiques, les patients traités dans les régions ou les établissements de santé qui feront partie des pionniers de l'IA bénéficieront de ces systèmes avant les autres. Les systèmes d'IA peuvent également être utilisés pour libérer les cliniciens des tâches subalternes qui demandent beaucoup de travail, comme la saisie de données, et leur permettre ainsi de consacrer plus de temps aux patients qu'auparavant[157].

Toutefois, ces avantages ne vont pas de soi. Les conséquences de l'IA sur les soins cliniques et la relation médecin-patient demeurent incertaines et varieront certainement selon les technologies et les situations. Les systèmes d'IA peuvent s'avérer plus efficaces que les soins prodigués par des êtres humains, mais aussi

[156] Organisation mondiale de la santé, cf. *supra* note 1.
[157] *Id.* cf. section 8.

fournir des soins de moindre qualité avec moins d'interactions personnelles. Dans de nombreux domaines, l'IA est considérée comme un moyen prometteur pour réduire les coûts et les temps d'attente, ou pour combler les lacunes existantes dans la couverture des soins, lorsque l'accès aux professionnels et aux établissements de santé est limité[158]. Les patients traités dans les « zones pionnières » recevront à tout le moins un type de soins différent, qui ne sera peut-être pas de la même qualité que les soins traditionnels prodigués par les professionnels de la santé.

Le déploiement hétérogène des systèmes d'IA, dont les conséquences sur l'accès et la qualité des soins sont incertaines, risque de créer de nouvelles inégalités en matière de santé dans les états membres. Il se peut que les régions qui ont toujours connu des problèmes relatifs à la faible qualité des soins ou à l'inégal accès à ces derniers soient considérées comme des bancs d'essai clés pour la prestation de soins facilités par l'IA. Les patients de ces régions pourraient accéder plus aisément aux systèmes d'IA, tels que les robots conversationnels ou la télémédecine, mais continuer à avoir un accès limité aux soins prodigués par des êtres humains ou aux consultations cliniques en face à face. La probabilité que ce risque se matérialise dépend largement du rôle stratégique accordé aux systèmes d'IA. Si l'on considère que les systèmes d'IA peuvent être un moyen de remplacer les soins en face-à-face plutôt qu'un moyen de libérer du temps aux cliniciens, l'augmentation des inégalités dans l'accès aux soins fournis par des êtres humains semble inévitable.

L'article 4 de la Convention d'Oviedo renvoie à l'obligation qu'ont les professionnels de la santé de prodiguer des soins en respectant des normes professionnelles. Or, il n'est pas clair si les développeurs, les fabricants et les fournisseurs de services pour les systèmes d'IA seront tenus de respecter les mêmes normes professionnelles. Le rapport explicatif de la Convention soulève cette question indirectement, en notant que « de l'expression "normes et obligations professionnelles", [il ressort que l'article 4] ne concerne pas les personnes qui, sans être des professionnels de la santé, sont appelées, par exemple dans une situation d'urgence, à exécuter des actes de nature médicale. » Ainsi, un robot conversationnel conçu pour le triage initial des patients peut-il être considéré comme une « personne » exécutant un « acte médical »[159] ? Si ce n'est pas le cas, comment garantir l'intervention d'un professionnel de la santé qui respecte ces normes de façon appropriée ?

Toute réduction des actes de surveillance ou de soins cliniques prodigués par des professionnels de la santé en raison du déploiement de systèmes d'IA pourrait donc potentiellement être considérée comme une violation de l'article 4. En particulier, les modèles de soins qui intègrent des robots conversationnels ou d'autres agents artificiels conçus pour fournir des soins ou soutenir directement les patients sembleraient présenter ce risque. Au moment d'intégrer des systèmes d'IA qui interagissent directement avec les patients, il convient d'examiner attentivement le rôle joué par les professionnels de la santé, lesquels sont tenus de respecter des normes professionnelles.

---

[158] Department of Health, *Innovation Health and Wealth: Accelerating Adoption and Diffusion in the NHS* (2011) ; *DEPARTMENT OF HEALTH, EQUITY AND EXCELLENCE: LIBERATING THE NHS*. (2010).
[159] Korngiebel et Mooney, cf. *supra* note 95, 3.

Transparence vis-à-vis des professionnels de la santé et des patients

L'IA pose la question de notions de responsabilité en termes connus et nouveaux. Les systèmes auxquels nous confions de plus en plus le soin de prendre des décisions et de formuler des recommandations susceptibles de changer le cours de nos vies, reposent sur les avancées technologiques dans le temps, mais ils sont numériques, distribués et souvent imperceptibles. Lorsque sont prises des décisions importantes dont les effets se répercutent sur les moyens d'existence et le bien-être des personnes, on s'attend pour le moins à ce que le fondement rationnel ou les raisons de ces décisions soient compréhensibles.

Cette attente se reflète dans l'article 5 de la Convention d'Oviedo qui réaffirme le droit des patients au consentement éclairé avant toute intervention ou recherche médicale. Le Rapport explicatif de la Convention d'Oviedo dresse une liste non exhaustive des informations à fournir à la personne concernée. L'une des exigences primordiales est que les informations doivent être communiquées aux patients d'une manière aisément compréhensible afin de garantir qu'elles puissent éclairer utilement leurs décisions. Habituellement, cela imposerait que soient posées des exigences sur la manière dont les professionnels de la santé expliquent leurs décisions et leurs recommandations aux patients. Dans les cas où les systèmes d'IA fournissent une certaine forme d'expertise clinique, par exemple en recommandant un diagnostic particulier ou en interprétant des images médicales, l'obligation d'expliquer la décision serait apparemment transférée du médecin au système d'IA, ou du moins au fabricant du système d'IA.

Les difficultés à expliquer comment les systèmes d'IA transforment les données d'entrée en données de sortie posent un défi épistémologique fondamental pour le consentement éclairé. Si l'on met de côté la capacité du patient à comprendre les fonctionnalités des systèmes d'IA, dans de nombreux cas, les patients n'ont tout simplement pas un niveau de connaissances suffisant pour donner un consentement libre et éclairé. Les systèmes d'IA utilisent des volumes de données sans précédent pour prendre leurs décisions et interprètent ces données à l'aide de techniques statistiques complexes. Dès lors, il est de plus en plus difficile de mesurer l'ampleur du traitement des données utilisées pour les diagnostics et les traitements[160].

En pratique, les exigences de transparence permettant de s'assurer d'un consentement éclairé peuvent s'obtenir de plusieurs manières. En supposant que le médecin reste le principal point de contact du patient, il peut être considéré comme un médiateur entre ce dernier et le système d'IA. Dans ce modèle de médiation, le médecin peut recevoir une explication du système d'IA et agir ensuite comme un « traducteur » pour le patient, en « traduisant » l'explication du système dans un format adapté et facilement compréhensible. Lorsque les médecins n'agissent pas en tant que médiateurs, par exemple lorsque les robots conversationnels assurent le triage des patients ou leur fournissent un diagnostic directement, on pourrait alors

[160] CONSEIL DE L'EUROPE, cf. *supra* note 2.

exiger que les systèmes d'IA expliquent leur prise de décision directement aux patients.

Ces deux modèles posent des difficultés pour l'explication des comportements complexes de systèmes de type « boîte noire » à des utilisateurs experts ou non. Au minimum, les systèmes d'IA qui interagissent directement avec les patients devraient leur indiquer qu'ils sont des systèmes artificiels. La question de savoir si l'utilisation de systèmes d'IA dans le secteur de la santé doit être communiquée aux patients par les cliniciens et les établissements de santé est plus difficile[161].

Une préoccupation qui revient souvent concernant l'IA utilisée à des fins opérationnelles par les hôpitaux – qui recourent par exemple à des outils de stratification du risque et de planification des sorties –, est le fait que les patients ne sont pas informés de l'utilisation de l'IA dans le cadre de leur prise en charge[162].

D'une part, les professionnels de la santé consultent régulièrement de nombreuses sources d'information pour diagnostiquer la maladie et traiter les patients, comme des modèles, des graphiques, des radiographies, etc., qu'ils ne communiqueraient pas ou dont ils ne discuteraient pas de manière proactive dans le cadre du consentement éclairé. D'autre part, les systèmes d'IA qui fournissent une expertise clinique, par exemple en interprétant des images médicales et en recommandant une classification des anomalies, peuvent constituer un type d'information qualitativement différent des sources qui entrent traditionnellement en ligne de compte dans la prise de décision clinique.

Néanmoins, dans la pratique, les systèmes d'IA utilisés pour aider les soins cliniques et stratifier les risques parmi les patients sont souvent traités comme des technologies purement opérationnelles plutôt que cliniques. De nombreux établissements de santé affirment y avoir recours pour améliorer la qualité et l'efficacité des soins, et non pour assoir la prise de décision clinique. À cet égard, les systèmes d'IA peuvent être considérés comme équivalents aux systèmes administratifs utilisés dans les hôpitaux pour le traitement des données des patients, et non pour leurs soins immédiats[163]. Bien entendu, tous les établissements de santé ne considèrent pas les systèmes de prédiction du risque basés sur l'IA comme purement opérationnels ; dans certains cas, il est demandé aux patients de consentir explicitement à l'utilisation d'un système d'IA conçu pour identifier les patients qui risquent de mourir dans les 48 heures[164]. Nous reviendrons sur les recommandations concernant la communication d'informations sur l'utilisation des systèmes d'IA dans le chapitre intitulé « Registre public des systèmes d'IA médicale pour la transparence ».

Indépendamment de la question de savoir si certaines technologies d'IA doivent être classées comme cliniques ou opérationnelles/administratives, il existe des questions pertinentes concernant l'intelligibilité des systèmes de type « boîte noire » à un niveau plus fondamental. Par rapport à la prise de décision humaine et organisationnelle, l'IA

[161] I. Glenn Cohen, « Informed Consent and Medical Artificial Intelligence: What to Tell the Patient? » *Symposium: Law and the Nation's Health*, 108 GEO. L.J. 1425–1470 (2019) ; Robbins et Brodwin, cf. *supra* note 5.
[162] Cohen, cf. *supra* note 160 ; Robbins et Brodwin, cf. *supra* note 5.
[163] Robbins et Brodwin, cf. *supra* note 5.
[164] *Id.*

pose un défi unique. La structure interne d'un modèle d'apprentissage automatique peut être constituée de millions de caractéristiques reliées dans un réseau complexe de comportements dépendants les uns des autres. Faire comprendre cette structure interne et ces relations de dépendance d'une manière intelligible pour l'homme est extrêmement difficile[165]. La façon dont les systèmes d'IA prennent des décisions peut donc s'avérer trop complexe pour que des êtres humains puissent bien comprendre l'ensemble des critères de décision ou le raisonnement sur lesquels ils reposent.

En supposant que l'exigence de transparence qui sous-tend le consentement éclairé est une valeur essentielle dans la relation médecin-patient facilitée par l'IA, le défi que pose l'opacité soulève une question : comment les systèmes d'IA doivent-ils rendre des comptes aux médecins et aux patients ? Pour répondre à cette question, commençons par examiner différents types de questions sur les systèmes d'IA qui vont nous permettre de les rendre compréhensibles.

- **Comment fonctionne un système ou un modèle d'IA ? Comment un système d'IA produit-il un résultat spécifique ?** Ce sont des questions d'*intelligibilité*. Elles concernent les fonctionnalités internes et les comportements externes d'un système d'IA. Un modèle entièrement intelligible est un modèle compréhensible par l'homme, ce qui signifie qu'un être humain peut comprendre l'ensemble des causes d'un résultat donné[166]. Les modèles faiblement intelligibles « sont opaques en ce sens que si l'on reçoit le résultat de l'algorithme – par exemple, une décision de classification –, on a rarement une idée concrète de comment ou pourquoi cette classification particulière a été obtenue à partir des données d'entrée »[167]. L'intelligibilité peut également être définie en termes de prévisibilité du modèle : un modèle est intelligible si une personne bien informée peut prévoir de manière cohérente ses résultats et ses comportements[168]. Les questions relatives au comportement d'un modèle portent essentiellement sur la manière dont un résultat ou un comportement particulier du modèle s'est produit[169]. Cependant, le comportement d'un modèle peut également être interprété de manière large de façon à y inclure les effets sur les établissements et les utilisateurs dignes de confiance et leurs décisions influencées par l'IA ; par exemple, il est pertinent de se demander comment le

[165] Jenna Burrell, « How the Machine "Thinks:" Understanding Opacity in Machine Learning Algorithms », *BIG DATA & SOCIETY* (2016) ; Zachary C. Lipton, *The Mythos of Model Interpretability*, ARXIV:1606.03490 [CS, STAT] (2016), http://arxiv.org/abs/1606.03490 (consulté pour la dernière fois le 15 octobre 2016).

[166] Paulo JG Lisboa, « Interpretability in Machine Learning–Principles and Practice », *in FUZZY LOGIC AND APPLICATIONS* 15–21 (2013), http://link.springer.com/chapter/10.1007/978-3-319-03200-9_2 (consulté pour la dernière fois le 19 décembre 2015) ; Tim Miller, « Explanation in artificial intelligence: Insights from the social sciences », 267 *ARTIFICIAL INTELLIGENCE* 1–38 (2019).

[167] Burrell, cf. *supra* note 164, 1.

[168] Been Kim, Rajiv Khanna & Oluwasanmi O. Koyejo, « Examples are not enough, learn to criticize! criticism for interpretability », *in ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS* 2280–2288 (2016).

[169] On parle parfois d'« *explicabilité* » d'un modèle pour se référer au degré auquel les causes du comportement spécifique d'un modèle peuvent être expliquées. Elle est traitée ici comme une composante de l'*intelligibilité*, à côté de la compréhensibilité intrinsèque du modèle.

diagnostic d'un médecin a pu être influencé par la recommandation d'un système expert[170].

- **Comment un système d'IA a-t-il été conçu et testé ? Comment est-il régi ?** Ce sont des questions de *transparence*. Contrairement à l'intelligibilité, la transparence ne concerne pas les fonctionnalités ou le comportement du système d'IA lui-même, mais plutôt les processus impliqués dans la conception, le développement, les tests, le déploiement et la réglementation du système d'IA. La transparence exige principalement que soient communiquées des informations sur les institutions et les personnes qui créent et utilisent les systèmes d'IA, ainsi que sur les structures de réglementation et de gouvernance qui contrôlent à la fois les institutions et les systèmes. Ici, l'intelligibilité joue un rôle complémentaire de soutien. Des modèles ou des explications compréhensibles relatives aux décisions spécifiques prises par un système peuvent, par exemple, être nécessaires aux organes de contrôle pour auditer efficacement l'IA et s'assurer que les exigences réglementaires sont respectées dans chaque contexte d'utilisation.
- **Quelles informations sont nécessaires pour enquêter sur le comportement des systèmes d'IA ?** Il s'agit d'une question de *traçabilité*. Pour évaluer le comportement des systèmes d'IA, certains éléments sont nécessaires, notamment « les jeux de données et les processus permettant au système d'IA de rendre une décision, y compris les processus de collecte et d'étiquetage de données, ainsi que les algorithmes utilisés »[171]. Ces données doivent être systématiquement enregistrées lors du fonctionnement du système pour permettre une gouvernance efficace. La traçabilité est donc une exigence fondamentale pour l'audit a posteriori et les explications du comportement du modèle ; sans les bonnes données, les explications ne peuvent être dégagées après qu'un modèle a produit une décision ou un autre résultat[172].

Les réponses à chacune de ces questions peuvent être nécessaires pour obtenir un consentement éclairé dans le cadre de soins facilités par l'IA. Cela ne signifie pas que chacune des questions doit obtenir des réponses de la part du patient et du professionnel de la santé ; il se peut plutôt que certaines questions soient mieux orientées vers le patient ou vers le professionnel de la santé. Par exemple, les patients peuvent être plus directement concernés par des questions sur la démarche utilisée pour décider de leur cas ou sur la manière dont une recommandation ou un diagnostic a été établi[173]. Les questions concernant la façon dont les systèmes d'IA ont été conçus et testés, et comment ils sont sécurisés et validés au fil du temps, peuvent être plus pertinentes pour les professionnels de la santé et les administrateurs qui doivent

---

[170] GROUPE D'EXPERTS DE HAUT NIVEAU SUR L'INTELLIGENCE ARTIFICIELLE, *Lignes directrices en matière d'éthique pour une IA digne de confiance* (2019).
[171] *Id.*
[172] Mittelstadt *et al*., cf. *supra* note 17.
[173] Sandra Wachter, Brent Mittelstadt & Chris Russell, « Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR », 3 *HARVARD JOURNAL OF LAW & TECHNOLOGY* 841–887 (2018).

évaluer la fiabilité d'un système en termes d'intégration de celui-ci dans les processus de décisions cliniques et opérationnelles existants[174]. Comme nous l'avons indiqué dans le chapitre intitulé « Cadre théorique de la relation médecin-patient », l'idéal du consentement éclairé est une composante de la relation médecin-patient qui nécessite une discussion entre le patient et le professionnel de la santé sur les options de traitement possibles, les valeurs, etc. Le fait de diriger certaines explications vers les parties les mieux à même de les comprendre, ou les plus directement intéressées, ne sape pas obligatoirement les idéaux de transparence ou de consentement éclairé. Au contraire, cela peut favoriser un dialogue constructif entre le patient et le médecin sur les options de soins facilités par l'IA.

Risque de biais social dans les systèmes d'IA

Comme nous l'avons vu au chapitre intitulé « Les défis éthiques communément rencontrés en matière d'IA », en un sens, les systèmes d'IA sont inéluctablement biaisés. De nombreux biais sont dus à des raisons techniques. Un décalage peut ainsi exister entre l'étape d'entraînement et celle des tests[175]. Les développeurs et les fabricants de systèmes d'IA conçoivent inévitablement des systèmes qui reflètent leurs valeurs ou les exigences réglementaires; cela peut aussi être considéré comme un type de biais qui variera selon les fabricants et les états membres[176]. Toutefois, dans les systèmes d'IA, les décisions biaisées et injustes ne sont souvent pas le produit de raisons techniques ou réglementaires, mais traduisent plutôt des biais sociaux et les inégalités sociales sous-jacentes[177].

Ces types de biais sociaux sont préoccupants pour plusieurs raisons.

- Premièrement, ils peuvent nuire à la précision des modèles d'une population ou d'un groupe démographique à l'autre. De nombreux biais peuvent être attribués à des jeux de données qui ne sont pas représentatifs de la population visée par un système. En médecine, il existe des lacunes cruciales en matière de données qui pourraient être comblées si les ressources, l'accès ou la motivation n'étaient pas limités. Les essais cliniques et les études sur la santé sont principalement réalisés sur des hommes blancs, ce qui signifie que les résultats sont moins susceptibles de s'appliquer aux femmes et aux personnes de couleur[178]. Il existe un manque de données grave et dangereux car de nombreux modèles cliniques traitent les femmes comme des « petits hommes »[179] et ne tiennent donc pas compte des différences biologiques entre

[174] CONSEIL DE L'EUROPE, cf. *supra* note 2.
[175] Friedman et Nissenbaum, cf. *supra* note 41; Wachter, Mittelstadt, et Russell, cf. *supra* note 47.
[176] CONSEIL DE L'EUROPE, cf. *supra* note 2.
[177] *Id.* ; Wachter, Mittelstadt, et Russell, cf. *supra* note 47.
[178] CAROLINE CRIADO PEREZ, *INVISIBLE WOMEN: EXPOSING DATA BIAS IN A WORLD DESIGNED FOR MEN* 115–116 (2019) ; sur comment remédier aux biais dans le contexte médical, voir Timo Minssen *et al.*, « Regulatory responses to medical machine learning », *JOURNAL OF LAW AND THE BIOSCIENCES* (2020) ; et Mirjam Pot, Wanda Spahl & Barbara Prainsack, « The Gender of Biomedical Data: Challenges for Personalised and Precision Medicine », 9 *SOMATECHNICS* 170–187 (2019).
[179] ANGELA SAINI, *INFERIOR: HOW SCIENCE GOT WOMEN WRONG AND THE NEW RESEARCH THAT'S REWRITING THE STORY* 59 (2017).

les hommes et les femmes[180]. Par exemple, un pourcentage différent de graisse corporelle, une peau plus fine, un système hormonal différent, des niveaux d'hormones changeant tout au long du cycle menstruel, avant la puberté et après la ménopause, sont des facteurs qui déterminent l'efficacité des médicaments ou la mesure dans laquelle nous sommes touchés par les toxines ou les impacts environnementaux[181].

- Deuxièmement, les biais sociaux peuvent conduire à une distribution inégale des résultats entre les populations ou les groupes démographiques protégés. Ce type d'inégalité est particulièrement grave dans le contexte de la médecine, car il a une incidence sur les biens fondamentaux : « toute erreur dans le fonctionnement d'un algorithme peut amener à prescrire un traitement inadapté et mettre en danger non seulement la santé, mais aussi la vie de groupes entiers de la population »[182]. De vastes pans des sociétés occidentales font actuellement face à des préjugés et à des inégalités de taille que l'on retrouve dans des décisions historiques et qui peuvent influencer la formation de futurs systèmes. Les tendances observées en matière de prise de décision ont jusqu'à présent conduit à renforcer l'inégalité des chances entre certains groupes[183]. Sans intervention, les systèmes d'IA apprendront et renforceront ces modèles préexistants qui favorisent l'inégalité des chances et l'inégal accès aux ressources dans la société.

Comme indiqué, l'article 14 de la CEDH interdit la discrimination. L'égalité est une valeur essentielle qui sous-tend les droits de l'homme. Toutefois, il est extrêmement difficile de parvenir à une égalité réelle ou à des « règles du jeu équitables » dans la pratique. En ce qui concerne l'IA, la partialité des données et les boucles de rétroaction sont des défis majeurs auxquels il faut s'attaquer pour faire en sorte que les systèmes n'exacerbent pas les inégalités existantes et ne créent pas de nouvelles formes de discrimination qui iraient à l'encontre de l'article 14. L'Assemblée parlementaire du Conseil de l'Europe a reconnu le risque de biais à cet égard, notant que « les États membres du Conseil de l'Europe devraient participer plus activement au développement d'applications de l'IA pour les services de santé, ou du moins faire valoir leur souveraineté pour imposer un mécanisme de filtrage et d'autorisations préalable à leur déploiement. L'implication des états aiderait aussi à garantir que ces applications soient alimentées par un volume suffisant de données sans parti pris et dûment protégées »[184].

En ce qui concerne la partialité des jeux de données, le fait de concevoir le biais uniquement comme une propriété des jeux de données est insuffisant pour parvenir à

[180] PEREZ, cf. *supra* note 177, 116. Cette prise en compte des différences biologiques est notamment omise parce qu'elle est plus complexe (par exemple, à cause de la fluctuation des niveaux d'hormones pendant le cycle menstruel), plus risquée (les participantes peuvent être enceintes), et plus exigeante en temps et en ressources. SAINI, cf. *supra* note 178, 58.
[181] SAINI, cf. *supra* note 178, 62; PEREZ, cf. *supra* note 177, 116.
[182] CONSEIL DE L'EUROPE, cf. *supra* note 2.
[183] Voir par exemple ANGELA Y. DAVIS, *WOMEN, RACE, & CLASS* (2011).
[184] CONSEIL DE L'EUROPE, cf. *supra* note 2.

une égalité réelle dans la pratique[185]. En supposant qu'il soit possible de créer un jeu de données qui restitue parfaitement les biais et les inégalités existant dans la société, l'entraînement d'un modèle avec ce jeu de données ne ferait rien pour corriger les inégalités qu'il met en évidence. Au contraire, de telles garanties ne peuvent être fournies que si l'on procède également à l'examen, au test et, peut-être, à la correction des biais dans le système d'IA qui a été formé et ses résultats.

Concernant les boucles de rétroaction, un système d'IA peut aggraver considérablement la situation de groupes déjà défavorisés en renforçant les biais sociaux qu'il a appris. Cependant, éviter simplement le renforcement des biais et des inégalités, ou veiller à ce que les systèmes d'IA n'aggravent pas le statu quo, ne permet pas d'atteindre une égalité réelle dans la pratique[186]. Il faut au contraire examiner de manière critique l'acceptabilité des inégalités existantes et prendre des mesures pour améliorer positivement la situation des groupes défavorisés. De même, les systèmes d'IA peuvent créer de nouvelles formes de discrimination plutôt que de simplement renforcer les inégalités et les biais existants[187]. Le déploiement de l'IA en médecine doit tenir compte à la fois de la nécessité d'une action critique positive et de la possibilité de nouvelles formes de discrimination alimentées par l'IA.

La détection des biais dans les systèmes d'IA n'est pas simple. Des règles biaisées en matière de prise de décision peuvent être cachées dans des modèles de type « boîte noire ». D'autres biais peuvent être détectés en examinant les résultats des systèmes d'IA pour y déceler des distributions inégales entre les groupes démographiques ou les populations concernées. Cependant, accéder à l'ensemble des décisions ou des résultats d'un système n'est pas forcément simple, ne serait-ce qu'en raison des normes de protection des données ; « certaines restrictions dans l'utilisation des données de santé à caractère personnel peuvent empêcher d'établir des rapprochements essentiels entre les données et causer des distorsions, voire des erreurs, dans les analyses reposant sur l'IA »[188]. Cela nous permet au moins d'affirmer que la simple anonymisation des données de santé peut ne pas être une solution adéquate pour atténuer les biais ou corriger leurs effets en aval. Même lorsque des ensembles de décisions sont accessibles, il se peut aussi que l'on manque de données démographiques pour certaines populations, ce qui signifie que la recherche de biais ne peut pas mesurer la distribution des groupes protégés par la loi[189].

À en juger par ces divers problèmes de biais social, de discrimination et d'inégalité, les professionnels et les établissements de santé font face à une tâche difficile, à savoir veiller à ce que leur utilisation des systèmes d'IA n'aggrave pas les inégalités existantes et ne crée pas de nouvelles formes de discrimination. La lutte contre les biais sociaux est un défi à multiples facettes. Cette lutte doit reposer sur des normes robustes de détection et de test des biais, des normes de collecte et de conservation de haute qualité pour les jeux de données de formation et de test, et des tests au

[185] Wachter, Mittelstadt, et Russell, cf. *supra* note 47.
[186] *Id.*
[187] Wachter, Mittelstadt, et Russell, cf. *supra* note 47.
[188] CONSEIL DE L'EUROPE, cf. *supra* note 2.
[189] Wachter, Mittelstadt, et Russell, cf. *supra* note 47 ; Sandvig *et al.*, cf. *supra* note 79 ; Brent Mittelstadt, « Automation, Algorithms, and Politics| Auditing for Transparency in Content Personalization Systems », 10 *INTERNATIONAL JOURNAL OF COMMUNICATION* 12 (2016).

niveau individuel pour veiller à ce que les résultats et les recommandations fournis aux patients ne soient pas principalement déterminés par des caractéristiques protégées par la loi[190]. Si l'on ne parvient pas à mettre en œuvre des normes rigoureuses en matière de test des biais, on risque d'exacerber les inégalités en matière de soins assistés par l'IA et de compromettre leur fiabilité. Ces risques sont particulièrement graves dans le contexte actuel des inégalités en matière d'accès à des soins de qualité, où le déploiement de l'IA peut être accéléré pour des raisons d'efficacité et d'allocation des ressources plutôt que pour des considérations purement cliniques.

Dilution de la prise en compte du bien-être du patient

Traditionnellement, les soins cliniques et la relation médecin-patient reposent idéalement sur une évaluation de l'état du patient par le médecin, qui tient compte de son environnement et de ses antécédents médicaux. Ce type d'évaluation est difficilement reproduit dans les soins facilités par les technologies. Les représentations des données du patient le réduisent nécessairement à des caractéristiques chiffrées. Des problèmes peuvent surgir lorsque les évaluations cliniques s'appuient de plus en plus sur des représentations des données, construites par exemple par des technologies de télésurveillance, ou sur d'autres données qui n'ont pas été recueillies lors de rencontres en face à face. Les représentations des données du patient peuvent être considérées comme une mesure « objective » de la santé et du bien-être de ce dernier, mais elles réduisent l'importance des facteurs contextuels de la santé ou de la vision du patient en tant que personne socialement incarnée. Les représentations des données peuvent créer un semblant de certitude, dans lequel les données de suivi « objectives » sont considérées comme une représentation fidèle de la situation du patient, alors qu'elles perdent de vue le contexte interpersonnel du patient et d'autres connaissances tacites[191].

Les professionnels de la santé font face à cette difficulté lorsqu'ils tentent d'intégrer les systèmes d'IA dans les soins de santé courants. La quantité et la complexité des données et des recommandations issues des technologies en ce qui concerne l'état d'un patient font qu'il est difficile de détecter l'absence d'informations contextuelles importantes. S'en remettre aux données recueillies par les « applications de santé » ou les technologies de surveillance (par exemple, les montres intelligentes) et traiter ces données comme la principale source d'information sur la santé d'un patient, par exemple, peut faire passer sous silence des aspects de sa santé qui sont difficiles à surveiller. Ces aspects incluent des éléments essentiels de la santé mentale et du bien-être du patient, tels que son état social, mental et émotionnel. Il peut en résulter une « décontextualisation » de l'état du patient, qui perd un certain contrôle sur la manière dont son état est présenté et compris par les cliniciens et les soignants[192].

[190] Wachter, Mittelstadt, et Russell, cf. *supra* note 47 ; Wachter, Mittelstadt, et Russell, cf. *supra* note 47 ; Matt J. Kusner *et al*., *Counterfactual Fairness* (2017).
[191] Mark Coeckelbergh, « E-care as craftsmanship: virtuous work, skilled engagement, and information technology in health care », 16 *MEDICINE, HEALTH CARE AND PHILOSOPHY* 807–816 (2013).
[192] Mittelstadt *et al.*, cf. *supra* note 3.

Toutes ces possibilités laissent supposer que la prestation de soins au moyen des technologies risque de limiter les rencontres entre le médecin et le patient, grâce auxquelles se crée habituellement la confiance nécessaire à la relation médecin-patient. Les technologies qui nuisent à la transmission des « signaux psychologiques et des émotions » peuvent empêcher le médecin de connaître l'état du patient, ce qui compromet « l'établissement d'une relation médecin-patient fondée sur la confiance et la guérison »[193]. Les prestataires de soins, en plus d'appliquer leurs connaissances de la médecine au cas du patient, peuvent être moins à même de faire preuve de compréhension, de compassion et d'autres traits souhaitables que l'on retrouve dans les « bonnes » interactions médicales. De par leur rôle de médiateur entre le médecin et le patient, les systèmes d'IA modifient les liens entre les cliniciens et les patients en confiant à un système technologique une partie des soins continus du patient. L'augmentation de la distance entre les professionnels de la santé et les patients semble réduire les possibilités de développer une compréhension tacite de la santé et du bien-être du patient[194].

Risques de biais d'automatisation, de perte de compétences et de déplacement de la responsabilité

Comme nous l'avons vu au chapitre intitulé « Les défis éthiques communément rencontrés en matière d'IA », l'introduction de systèmes d'IA dans les soins cliniques présente un risque de biais d'automatisation, ce qui signifie que les cliniciens peuvent faire confiance aux résultats ou aux recommandations des systèmes d'IA non pas en raison de leur efficacité clinique prouvée, mais plutôt sur la base de l'objectivité, de la précision ou de la complexité perçues[195]. Tout déploiement de systèmes d'IA conçus pour améliorer la prise de décision humaine par des recommandations, des avertissements ou des interventions similaires risque d'introduire un biais d'automatisation. Les travaux empiriques sur le phénomène sont quelque peu naissants, mais une étude récente a montré comment même les décideurs experts peuvent être enclins à des biais d'automatisation au fil du temps pour des raisons problématiques (par exemple, le coût d'un système d'IA en tant qu'indicateur d'exactitude ou d'égalité)[196]. Le Conseil de l'Europe a clairement reconnu le risque de biais d'automatisation en appelant les états membres à garantir que « les applications de santé reposant sur l'IA ne remplacent pas complètement le jugement humain, et donc que les décisions prises avec l'IA dans le cadre des soins de santé professionnels sont toujours validées par des professionnels de santé dûment formés »[197].

Les systèmes d'IA utilisés comme des prestataires de soins cliniques ou des systèmes de diagnostic experts peuvent freiner le développement des compétences, des communautés professionnelles et des normes de « bonnes pratiques » en

[193] Bauer, cf. *supra* note 136, 84.
[194] Coeckelbergh, cf. *supra* note 190.
[195] Zarsky, cf. *supra* note 31, 121.
[196] Daniel N. Kluttz & Deirdre K. Mulligan, « Automated Decision Support Technologies and the Legal Profession », 34 *BERKELEY TECH. L.J.* 853 (2019).
[197] CONSEIL DE L'EUROPE, cf. *supra* note 2.

médecine[198]. Ce phénomène de « perte de compétences » va à l'encontre de ce que l'OMS appelle « l'IA centrée sur l'humain », qui appuie et accroît l'expertise humaine et le développement des compétences, plutôt que de les saper ou de les remplacer[199]. Les professionnels de la santé développent des normes éthiques et de bonne pratique à travers leurs expériences de l'exercice de la médecine. Pour définir les normes, les praticiens peuvent s'appuyer sur la sagesse pratique développée grâce à leur expérience. Les membres du corps médical forment une communauté qui partage des obligations morales et des objectifs communs[200]. Les normes éthiques ou internes d'une pratique contribuent à garantir la réalisation de ses objectifs dans le temps en luttant contre l'influence des institutions et de critères externes. L'élaboration, le maintien et l'application de ces normes peuvent être déplacés par la prestation de soins au moyen des technologies.

Il s'ensuit que le développement, le maintien et l'application des normes internes nécessaires pour respecter les obligations morales envers les patients peut être compromis lorsque les soins sont facilités par des technologies, et donc prodigués en partie par des institutions et des individus non professionnels. Il est possible que les systèmes algorithmiques déplacent les responsabilités traditionnellement assumées par les professionnels de la santé, tout en fournissant des soins plus efficaces ou « meilleurs » en termes de coût-bénéfice uniquement. Afin d'éviter la détérioration des soins médicaux qui sont globalement de qualité, et pas seulement techniquement « efficaces », ces obligations morales de servir et de respecter les patients en premier lieu doivent être prises au sérieux par les nouveaux prestataires de soins et de services qui ne font pas partie des communautés médicales traditionnelles. En d'autres termes, les soins reposant sur l'IA peuvent créer un fossé en matière de compétences professionnelles et de responsabilité.

La perte de compétences et le biais d'automatisation présentent également des risques directement pour les patients. L'un des rôles de l'expertise clinique humaine est de protéger les intérêts et la sécurité des patients. Les risques pour la sécurité proviennent de sources diverses, notamment les « attaques malveillantes sur les logiciels, [la] conception non éthique du système ou [sa] défaillance involontaire, la perte de contrôle humain et l' « exercice irresponsable du pouvoir conféré par le numérique » peuvent entraîner des « dommages matériels pour la santé humaine, les biens et l'environnement »[201].

Si cette expertise humaine est amoindrie par la perte de compétences ou déplacée par le biais d'automatisation, les tests et les preuves d'efficacité clinique doivent venir combler les lacunes pour garantir la sécurité des patients. Un compromis similaire existe en ce qui concerne l'opacité et la précision ; certains chercheurs ont fait valoir que les systèmes de santé reposant sur l'IA ne doivent pas nécessairement être explicables si leur précision et leur efficacité clinique peuvent être validées de manière

[198] *Id.* ; Coeckelbergh, cf. *supra* note 190.
[199] Organisation mondiale de la santé, cf. *supra* note 1.
[200] MACINTYRE, cf. *supra* note 122.
[201] CONSEIL DE L'EUROPE, cf. *supra* note 2 ; CONSEIL DE L'EUROPE, *Responsabilité et IA* (2019), https://rm.coe.int/responsability-and-ai-fr/168097d9c6.

fiable[202]. Dans les deux cas, la protection des intérêts vitaux des patients, ou les obligations déontologiques (*fudiciary duties*) généralement assumées par les professionnels de la santé, sont transférées aux fournisseurs de systèmes d'IA ou aux systèmes eux-mêmes.

Par conséquent, en vue de remplacer la protection offerte par l'expertise clinique humaine tout en continuant à garantir la sécurité des patients, des normes de test et de validation robustes sont nécessaires dans le pré-déploiement des systèmes d'IA dans le contexte de soins cliniques. Ces normes devraient également aborder les aspects non cliniques complémentaires de la sécurité, tels que la cybersécurité, les dysfonctionnements et la résilience[203]. Bien qu'il s'agisse d'une conclusion apparemment évidente, l'existence de telles normes et les preuves de leur respect ne peuvent être considérées comme allant de soi. Comme nous l'avons vu au chapitre intitulé « Vue d'ensemble des technologies de l'IA en médecine », il n'existe pas encore de preuves de l'efficacité clinique de nombreuses technologies d'IA dans le domaine des soins de santé, ce qui a constitué, à juste titre, un obstacle à leur déploiement à grande échelle.

Un sujet connexe mais tout aussi important concerne la responsabilité en cas de dysfonctionnement et d'effets néfastes de l'IA. Comme nous l'avons vu au chapitre intitulé « Vue d'ensemble des technologies de l'IA en médecine », la responsabilité distribuée est un défi à la fois moralement et juridiquement difficile à relever. L'Assemblée parlementaire du Conseil de l'Europe a reconnu la nécessité de clarifier la responsabilité des parties prenantes de l'IA, notamment « des développeurs aux autorités de réglementations, et des intermédiaires aux utilisateurs (notamment les pouvoirs publics, les professionnels de santé, les patients et le grand public). » Les états membres du Conseil de l'Europe sont appelés « à élaborer un cadre juridique pour clarifier la responsabilité des parties prenantes dans la conception, le déploiement, l'entretien et l'utilisation des applications de l'IA en rapport avec la santé (y compris pour les dispositifs médicaux implantables et portables) dans le contexte national et pan-européen, redéfinir la responsabilité des acteurs pour les risques et préjudices émanant et assurer que les structure de gouvernance et des forces de l'ordre soient en place pour garantir la mise en œuvre de ce cadre juridique »[204]. Le Comité d'experts du Conseil de l'Europe sur la dimension droits de l'homme des traitements automatisés de données et différentes formes d'intelligence artificielle (MSI-AUT), dans un rapport datant de 2019, a exploré les défis spécifiques liés à la responsabilité et aux lacunes en matière de responsabilité dans l'IA de manière beaucoup plus détaillée que ce qui est possible ici[205].

Conséquences sur le droit à la vie privée

L'IA pose plusieurs problèmes spécifiques liés au droit à la vie privée et aux réglementations complémentaires en matière de protection des données. Comme

[202] Boris Babic *et al.*, « Beware explanations from AI in health care », 373 *SCIENCE* 284–286 (2021).
[203] CONSEIL DE L'EUROPE, cf. *supra* note 2.
[204] *Id.*
[205] CONSEIL DE L'EUROPE, cf. *supra* note 200.

indiqué dans le chapitre intitulé « La Convention d'Oviedo et les principes des droits de l'homme en matière de santé », le Conseil de l'Europe est en train de ratifier les amendements à la Convention pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel (STE n° 108 et STCE n° 223). Ces droits supplémentaires visent à faire bénéficier les individus d'une plus grande transparence et d'un meilleur contrôle sur les formes automatisées de traitement des données. Ces droits apporteront sans aucun doute une protection précieuse aux patients dans toute une série d'utilisation de l'IA médicale.

L'utilisation des données des patients pour l'entraînement et les tests des systèmes d'IA constitue un défi particulier, propre à l'IA et qui mérite d'être examiné de plus près. La confidentialité dans la relation médecin-patient est une valeur essentielle pour protéger le droit à la vie privée. Dans le même temps, le développement, le déploiement et l'utilisation accrus des systèmes d'IA dans le domaine des soins de santé peut accroître la nécessité de créer ou de conserver des jeux de données de patients réels de haute qualité pour entraîner et tester les systèmes. L'innovation peut menacer la vie privée et la confidentialité de deux manières. Premièrement, il pourrait y avoir une plus grande pression pour réaffecter et accorder à des tiers l'accès à des données de patients (dépersonnalisées) et à des dossiers médicaux électroniques pour tester et développer des systèmes d'IA.

Deuxièmement, les cliniciens peuvent être encouragés à prescrire des examens et des analyses supplémentaires non pas pour leur valeur clinique, mais plutôt en raison de leur utilité pour entraîner ou tester des systèmes d'IA. Cela a des implications à la fois en termes d'augmentation des coûts des soins de santé, mais aussi d'exposition des patients à des risques inutiles de fuite de données ou d'autres atteintes à la vie privée. La Convention d'Oviedo prévoit une application spécifique du droit à la vie privée (article 8 de la CEDH) qui reconnaît la nature particulièrement sensible des informations personnelles relatives à la santé et établit un devoir de confidentialité pour les professionnels de la santé. Toute production de données dont la valeur clinique est discutable ou qui est clairement motivée par son utilité de test ou de développement de systèmes d'IA violerait apparemment la définition du droit à la vie privée de la Convention.

On peut en déduire que lorsqu'il existe un besoin légitime de données réelles pour tester et entraîner des systèmes d'IA, les intérêts en matière d'innovation et d'efficacité ou de qualité des soins doivent être mis en balance avec les intérêts individuels du patient en matière de vie privée et de confidentialité. Si cet équilibre n'est pas atteint, la confiance entre les patients et les prestataires de soins risque d'être ébranlée. Cette perte de confiance ne serait alors pas le fait d'une utilisation inadéquate de l'IA lors de rencontres cliniques individuelles, mais plutôt d'une incapacité institutionnelle à protéger les intérêts du patient en matière de vie privée et de confidentialité au niveau institutionnel. Au minimum, toute utilisation de dossiers médicaux de patients à des fins d'entraînement et de test des systèmes d'IA devrait être assortie de solutions techniques adaptées pour renforcer la confidentialité et l'anonymisation, telles que la

confidentialité différentielle (qui introduit un « bruit » statistique pour empêcher l'identification d'une personne particulière dans un jeu de données)[206].

[206] Cynthia Dwork, « Differential Privacy », *in* *AUTOMATA, LANGUAGES AND PROGRAMMING* 1–12 (Michele Bugliesi *et al*. éd., 2006), http://link.springer.com/chapter/10.1007/11787006_1 (consulté pour la dernière fois le 4 avril 2016) ; Paul Ohm, « Broken promises of privacy: Responding to the surprising failure of anonymization », 57 *UCLA LAW REVIEW* 1701 (2010).

7. RECOMMANDATIONS CONCERNANT DES NORMES ETHIQUES COMMUNES POUR UNE IA FIABLE

La discussion du chapitre intitulé « Les conséquences potentielles de l'IA sur la relation médecin-patient » a conclu que le développement des normes éthiques doit s'appuyer sur la transparence, les biais, la confidentialité et l'efficacité clinique afin de protéger les intérêts des patients en matière de consentement éclairé, d'égalité, de vie privée et de sécurité. Ensemble, de telles normes pourraient servir de base à un déploiement de l'IA dans le secteur de la santé qui favoriserait la relation de confiance entre les médecins et les patients plutôt que de l'entraver. Ces normes peuvent aborder aussi bien la manière dont les systèmes sont conçus et testés avant leur déploiement que la manière dont ils sont mis en œuvre dans les soins cliniques habituels et les processus institutionnels de prise de décision.

La Convention d'Oviedo agit comme une norme minimale de protection des droits de l'homme qui nécessite d'être transposée dans les législations nationales. Sur cette base, il est possible de formuler des recommandations spécifiques et positives concernant le niveau de soins qui doit être atteint dans les soins faisant appel à l'IA. Ces recommandations ne doivent pas entraver l'exercice de la souveraineté nationale dans l'élaboration des normes via la législation nationale et les organisations professionnelles, ainsi qu'il est expliqué en détail dans l'article 4 de la Convention d'Oviedo. Il est toutefois également possible d'établir des normes qui n'interfèrent pas avec l'article 4 et peuvent être considérées comme directement applicables. En particulier, comme l'indique Andorno :

> « Les normes communes établies par le Conseil de l'Europe fonctionneront principalement par l'intermédiaire des États. Cela n'exclut évidemment pas que certaines normes figurant dans la Convention soient directement applicables dans le droit national des États qui l'ont ratifiée. C'est par exemple le cas de certaines normes relatives à des droits individuels tels que le droit à l'information, l'exigence du consentement éclairé et le droit de ne pas être discriminé en raison de caractéristiques génétiques. Les normes d'interdiction peuvent également être considérées comme ayant une efficacité immédiate, mais en l'absence de sanctions légales, qu'il revient à chaque État de définir (article 25), leur efficacité est limitée aux recours civils et administratifs. »

Dans les cas où on observe un impact évident de l'IA sur les droits et les protections exposés dans la Convention d'Oviedo, il est opportun que le Conseil de l'Europe présente des recommandations et des exigences contraignantes pour les signataires concernant la façon d'utiliser et de gérer l'IA. Les recommandations devraient se concentrer sur un niveau plus élevé de soins positifs en ce qui concerne la relation médecin-patient, afin de s'assurer qu'elle ne soit pas exagérément perturbée par l'introduction de l'IA dans les environnements de soins. De telles normes devraient

évidemment permettre une certaine interprétation locale sur des questions normatives essentielles telles que les degrés acceptables de biais d'automatisation, les compromis acceptables entre les résultats d'un groupe de patients à l'autre et les domaines similaires influencés par les normes locales.

Les exemples suivants de recommandations détaillent les éventuelles exigences et recommandations essentielles pour une norme d'intelligibilité qui vise à protéger le consentement éclairé dans les soins faisant appel à l'IA, une norme de transparence pour l'intelligibilité publique et une norme pour la collecte de données sensibles à des fins de vérification de l'existence de biais. Chacune de ces normes devrait être considérée comme un exemple du type de recommandations qui peuvent être élaborées à partir de la discussion précédente sur les impacts éthiques potentiels de l'IA sur la relation médecin-patient.

Exigences d'intelligibilité du consentement éclairé

D'après le Rapport explicatif, l'article 5 de la Convention d'Oviedo comporte une liste non exhaustive d'informations qui devraient être partagées dans le cadre d'un processus de consentement éclairé. Cette liste étant non exhaustive, le Conseil de l'Europe pourrait établir des normes sur le contenu et la manière dont les informations sur la recommandation d'un système d'IA concernant le diagnostic et le traitement d'un patient devraient être communiquées au patient. Étant donné le rôle traditionnel du médecin dans la diffusion de ce type d'informations lors des consultations cliniques et dans les discussions qui s'ensuivent, ces normes devraient elles aussi aborder le rôle que joue le médecin pour expliquer aux patients les recommandations en matière d'IA et la façon dont les systèmes d'IA peuvent être conçus pour l'accompagner dans ce rôle.

De nombreux concepts sont communs à l'ensemble des questions et critères qui motivent l'interprétabilité dans le domaine de l'IA. Les méthodes d'interprétation visent à expliquer la fonctionnalité ou le comportement des modèles d'apprentissage automatique dits à effet « boîte noire », qui sont des composants essentiels des systèmes de prise de décision des IA. Les modèles d'apprentissage automatique entraînés sont qualifiés de « boîtes noires » lorsqu'ils ne sont pas compréhensibles pour les observateurs humains parce que leur logique interne et leur fondement sont inconnus ou inaccessibles à l'observateur, ou connus mais impossibles à interpréter du fait de leur complexité.[207] L'interprétabilité dans le sens étroit du terme utilisé ici désigne la capacité à comprendre la fonctionnalité et la signification d'un phénomène donné, dans ce cas d'un modèle d'apprentissage automatique entraîné et de ses résultats, et à l'expliquer en termes humains compréhensibles.[208]

[207] Riccardo Guidotti *et al.*, « A Survey of Methods for Explaining Black Box Models », n° 51 *ACM Comput. Surv.* 93:1-93:42 (2018) ; Bureau britannique du Commissaire à l'information et The Alan Turing Institute, « Explaining decisions made with AI » (2020), https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-ai/.

[208] Finale Doshi-Velez et Been Kim, « Towards A Rigorous Science of Interpretable Machine Learning », arXiv:1702.08608 [cs, stat] (2017), http://arxiv.org/abs/1702.08608 (consulté le 22 sep. 2017).

« L'explication » est également un concept clé de l'interprétabilité. De façon générique, les explications dans le domaine de l'IA relient « les valeurs tangibles d'une occurrence à la prédiction du modèle d'une manière qui est humainement compréhensible ».[209] Cette définition approximative cache des nuances importantes. Le terme englobe de multiples façons d'échanger des informations avec différentes parties prenantes à propos d'un phénomène, ici la fonctionnalité d'un modèle ou la justification et les critères permettant une prise de décision.[210]

Deux distinctions essentielles sont pertinentes afin de comprendre comment « l'explication » peut être concrétisée en médecine.

- Premièrement, les méthodes peuvent se distinguer en fonction de ce qu'elles cherchent à expliquer. Les explications ayant trait à la fonctionnalité du modèle traitent de la logique globale suivie par le modèle lorsqu'il produit des résultats à partir de données d'entrée. Les explications sur le comportement du modèle, en revanche, cherchent à expliquer comment ou pourquoi un comportement donné du modèle est survenu, par exemple comment ou pourquoi un résultat spécifique a été produit à partir d'une entrée spécifique. Les explications sur la fonctionnalité du modèle visent à expliquer ce qui se passe à l'intérieur du modèle, tandis que les explications sur le comportement du modèle visent à expliquer ce qui a mené à un comportement ou à un résultat spécifique, en référençant les caractéristiques essentielles ou les éléments qui influencent ce comportement. Il n'est pas strictement nécessaire de comprendre l'ensemble des relations, des dépendances et du poids des critères au sein du modèle pour expliquer son comportement.

- Deuxièmement, les méthodes d'interprétation peuvent se différencier par leur façon de concevoir « l'explication ». De nombreuses méthodes conçoivent les explications comme des modèles d'approximation, qui sont un type de modèle plus simple, qui peut être interprété par l'homme, et qui est créé pour estimer de manière fiable la fonctionnalité d'un modèle « boîte noire » plus complexe. Souvent, et cela peut porter à confusion, on fait référence au modèle d'approximation lui-même en tant qu'explication du modèle boîte noire. Cette approche contraste avec le traitement de « l'explication » dans la philosophie des sciences et l'épistémologie, dans lesquelles le terme désigne habituellement des déclarations qui expliquent les causes d'un phénomène donné.[211]

Cette utilisation du terme « explication » peut porter à confusion. Il est préférable d'envisager les modèles d'approximation comme des outils à partir desquels on peut établir des déclarations explicatives à propos du modèle original.[212] Les déclarations

---

[209] Christoph Molnar, *Interpretable Machine Learning* p. 32 (2020), https://christophm.github.io/interpretable-ml-book/ (consulté le 31 janv. 2019).
[210] Lipton, cf. *supra* note 164 ; Miller, cf. *supra* note 165.
[211] Brent Mittelstadt, Chris Russell et Sandra Wachter, « Explaining Explanations in AI », *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19* p. 279-288 (2019).
[212] *Id.*

explicatives elles-mêmes peuvent être textuelles, quantitatives ou visuelles et rendre compte de divers aspects du modèle et de ses comportements.

D'autres distinctions permettent de classifier les différents types d'explications et de méthodes d'interprétation. Une distinction de base peut être établie en interprétabilité entre les interprétabilités globale et locale. La distinction concerne la portée du modèle ou des résultats qu'une méthode donnée d'interprétation ou d'explication tente de rendre compréhensible par les humains. Les méthodes globales visent à expliquer la fonctionnalité d'un modèle dans son ensemble ou au sein d'une série spécifique de résultats en fonction de l'importance des critères, de leurs dépendances ou interactions, et de leur effet sur les résultats. Au contraire, les méthodes locales peuvent par exemple évaluer l'influence de zones spécifiques de l'espace d'entrée ou de variables données sur un ou plusieurs résultats spécifiques du modèle.

Les modèles peuvent être interprétés de manière globale à un niveau holistique ou modulaire.[213] L'interprétation holistique globale désigne des modèles qui sont compréhensibles par un observateur humain en ce sens que l'observateur peut suivre l'ensemble de la logique ou des étapes opérationnelles prises par le modèle et qui mènent à tous les résultats possibles du modèle.[214] Il devrait être possible pour une seule personne de comprendre dans leur intégralité les modèles interprétables de manière holistique.[215] Un observateur aurait « une vision holistique de ses critères et de chacun de ses composants appris tels que les poids, les autres paramètres et les structures ».[216]

Étant donné les limites de la compréhension humaine et de la mémoire à court terme, l'interprétation holistique globale n'est actuellement réalisable en pratique que sur des modèles relativement simples avec peu de critères, d'interactions ou de règles, ou avec une forte linéarité ou monotonicité.[217] Pour des modèles plus complexes, l'interprétabilité globale au niveau modulaire peut être faisable. Ce type d'interprétabilité implique de comprendre une caractéristique ou un segment particulier du modèle, par exemple les poids dans un modèle linéaire, ou les prédictions des divisions et des nœuds terminaux dans un arbre de décision.[218]

En ce qui concerne l'interprétabilité locale, un résultat unique peut être considéré comme interprétable si les étapes qui y ont mené peuvent être expliquées. L'interprétabilité locale ne nécessite pas impérativement d'expliquer toute la série des étapes ; au contraire, il peut suffire d'expliquer un ou plusieurs aspects du modèle qui ont mené au résultat, comme une valeur de critère ayant une influence critique.[219] Un groupe de résultats est considéré comme interprétable au niveau local si les méthodes qui permettent d'expliquer les résultats individuels peuvent être appliquées au groupe.

---

[213] Molnar, cf. *supra* note 208.
[214] Guidotti *et al.*, cf. *supra* note 206.
[215] Lipton, cf. *supra* note 164.
[216] Molnar, cf. *supra* note 208, p. 27.
[217] Guidotti *et al.*, cf. *supra* note 206.
[218] Molnar, cf. *supra* note 208.
[219] *Id.* ; Wachter, Mittelstadt et Russell, cf. *supra* note 172.

Les groupes peuvent également être expliqués grâce à des méthodes qui produisent une interprétabilité globale au niveau modulaire.[220]

Ces distinctions permettent de tirer quelques conclusions initiales sur la façon dont l'IA peut le mieux être expliqué aux médecins et aux patients. Au moment de l'adopter, il semble approprié de fournir des explications globales sur la fonctionnalité du modèle afin de garantir une adéquation fiable entre l'utilisation prévue du système d'IA dans un contexte de soins donné et la performance réelle du système. Lorsqu'il s'agit d'expliquer des résultats ou des recommandations spécifiques aux patients, les explications sur le comportement du modèle formulées sous forme de notes explicatives semblent constituer la meilleure option pour expliquer la logique décisionnelle du système tout en restant compréhensible pour les utilisateurs experts comme non experts. Dans ce contexte, des méthodes telles que les « explications contrefactuelles » peuvent être préférables puisqu'elles facilitent le débogage et l'évaluation de la performance du système par des utilisateurs experts tout en restant compréhensibles par les patients non experts au niveau d'une explication individuelle.[221] Pour résumer, afin de rendre les systèmes d'IA intelligibles pour les patients, les explications simples, locales et contrastives sont préférables aux explications d'approximation globales, qui peuvent être difficiles à comprendre et à interpréter.

L'utilisation exclusive dans les soins cliniques de modèles qui peuvent être interprétés intrinsèquement constitue une approche alternative mais complémentaire qui permet aux professionnels de santé de comprendre les systèmes de manière holistique et de mieux les expliquer à leurs patients.[222] La mise en œuvre de cette approche créerait toutefois des exigences supplémentaires en matière d'expertise technique des professionnels de santé dans les domaines de l'informatique, des statistiques et de l'apprentissage automatique, auxquelles il pourrait s'avérer très difficile voire excessif de répondre en pratique.

Registre public des systèmes d'IA médicale pour la transparence

Pour revenir à la question de la divulgation aux patients des informations concernant l'utilisation des systèmes d'IA à des fins opérationnelles et cliniques discutée au chapitre intitulé « Transparence vis-à-vis des professionnels de la santé et des patients », l'Assemblée parlementaire du Conseil de l'Europe a reconnu l'importance de sensibiliser la population aux utilisations de l'IA dans le secteur de la santé afin de construire une relation de confiance avec les patients et de s'assurer de la possibilité d'obtenir un consentement éclairé dans les soins faisant appel à l'IA. Son rapport d'octobre 2020 suggère notamment que la transparence des systèmes d'IA dans le secteur de la santé « peut impliquer la création d'un dispositif national de gouvernance de données relatives à la santé qui pourrait s'appuyer sur les propositions des institutions internationales. Parmi celles-ci, notons la Recommandation "Décoder

[220] Molnar, cf. *supra* note 208.
[221] Wachter, Mittelstadt et Russell, cf. *supra* note 172.
[222] Cynthia Rudin, « Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead », *Nat Mach Intell* n° 1, p. 206-215 (2019).

l'intelligence artificielle : 10 mesures pour protéger les droits de l'homme", publiée par la Commissaire aux droits de l'homme du Conseil de l'Europe (mai 2019), les Lignes directrices en matière d'éthique pour une IA digne de confiance énoncées par l'UE (avril 2019), la Recommandation et les principes de l'OCDE sur l'IA (mai 2019) et les Principes du G20 sur une Intelligence artificielle centrée sur l'être humain (juin 2019). »[223]

> Suivant ces propositions et recommandations, une base de données publique est vue comme un élément essentiel pour améliorer « l'alphabétisation algorithmique » parmi le grand public, ce qui constitue un précurseur fondamental à l'exercice de nombreux droits humains et juridiques.[224]

Dans la mesure où le dispositif proposé est conçu pour accroître la sensibilisation de la population aux systèmes d'IA dans le secteur de la santé, il peut être considéré comme un type de registre public pour les systèmes d'IA dans ce secteur. Les registres sont des listes publiques des systèmes actuellement utilisés qui contiennent une description normalisée de chaque système. Les informations qui figurent dans les registres varient mais peuvent inclure des éléments tels que l'utilisation ou le but prévu du système ; son fabricant ou distributeur ; la/les méthode(s) sous-jacente(s) (par ex. l'apprentissage profond, la régression) ; toute évaluation effectuée aussi bien en termes de précision que de biais et d'autres dimensions éthiques et juridiques ; une description d'ensembles de données d'entraînement et d'essai ; et une explication de la façon dont les prédictions ou les résultats du système sont utilisés par les dirigeants humains ou autrement intégrés aux services et processus décisionnels existants.[225] En outre, les registres comportent souvent une fonction de rétroaction qui permet aux citoyens d'apporter leur contribution sur les utilisations actuelles et proposées de l'IA par les organismes et services publics.[226]

Il existe plusieurs exemples de registres existants dans des organismes publics municipaux, nationaux et internationaux. En 2020, les villes d'Amsterdam et d'Helsinki ont lancé des registres publics pour les systèmes d'IA et algorithmiques utilisés dans les services municipaux.[227] En novembre 2021, le « Bureau central du numérique et des données » du Bureau du Cabinet britannique a lancé une norme nationale de

[223] Conseil de l'Europe, cf. *supra* note 2.

[224] *Id.*

[225] Corinne Cath et Fieke Jansen, « Dutch Comfort: The limits of AI governance through municipal registers », arXiv preprint arXiv:2109.02944 (2021) ; Luciano Floridi, « Artificial Intelligence as a Public Service: Learning from Amsterdam and Helsinki », *Philosophy & Technology* n° 33 p. 541-546 (2020) ; Timnit Gebru *et al.*, « Datasheets for Datasets » (2018), https://arxiv.org/abs/1803.09010 (consulté le 1er oct. 2018) ; Margaret Mitchell *et al.*, « Model Cards for Model Reporting », *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19* p. 220-229 (2019) ; Sarah Holland *et al.*, « The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards », arXiv:1805.03677 [cs] (2018), http://arxiv.org/abs/1805.03677 (consulté le 1er oct. 2018).

[226] « Amsterdam and Helsinki launch algorithm registries to bring transparency to public deployments of AI », *VentureBeat* (2020), https://venturebeat.com/2020/09/28/amsterdam-and-helsinki-launch-algorithm-registries-to-bring-transparency-to-public-deployments-of-ai/ (consulté le 1er déc. 2021).

[227] *Id.*

transparence algorithmique qui fonctionnera réellement comme une sorte de registre public.[228] Au niveau international, la récente proposition de législation sur l'intelligence artificielle (Artificial Intelligence Act) contient une disposition pour la création d'une base de données publique européenne dans laquelle les applications d'une IA autonome à haut risque doivent être enregistrées.[229] Le Conseil du l'Europe a l'opportunité de compléter ces normes émergentes en matière de transparence en instaurant un registre public de l'IA médicale dans les états membres destiné aux patients, afin de les sensibiliser aux systèmes d'IA actuellement utilisés par leurs services de santé publics.

Collecte de données sensibles pour la vérification des biais et de l'équité

Il est à prévoir que les biais présents dans les systèmes d'IA associés à des lacunes dans les données d'entraînement et d'essai encouragent une collecte plus importante de données sensibles sur des groupes juridiquement protégés dans le but de tester les biais et l'équité. Il est généralement admis qu'afin de prévenir les résultats discriminatoires ou biaisés, des données concernant les groupes sensibles doivent être collectées. Sans cela, la discrimination contre les groupes protégés ne pourra pas être évitée et sera vraisemblablement plus difficile à détecter.[230] Les données sensibles sont nécessaires pour vérifier si la prise de décision automatisée a désavantagé certains groupes en se basant sur des caractéristiques protégées (par ex. les données sur la « race », le handicap ou l'orientation sexuelle).[231] D'un autre côté, la collecte de telles données a des implications non négligeables en termes de vie privée. Il s'agit d'une préoccupation légitime et étroitement liée à des expériences historiques inquiétantes qui ont gravement porté préjudice à certains groupes sociaux spécifiques.[232] Les données collectées à des fins publiques ou de recherche ont par

[228] « UK government publishes pioneering standard for algorithmic transparency », gov.uk, https://www.gov.uk/government/news/uk-government-publishes-pioneering-standard-for-algorithmic-transparency (consulté le 1er déc. 2021).

[229] Commission européenne, cf. *supra* note 16, art. 51 et 60.

[230] Sandra Wachter, Brent Mittelstadt et Chris Russell, « Why Fairness Cannot Be Automated : Bridging the Gap Between EU Non-Discrimination Law and AI » p. 34-35 (2020), https://papers.ssrn.com/abstract=3547922 (consulté le 19 avr. 2020) ; Cynthia Dwork et Deirdre K. Mulligan, « It's not privacy, and it's not fair », n° 66 *Stan. L. Rev. Online* p. 35 (2013) ; Cynthia Dwork *et al.*, « Fairness Through Awareness », arXiv:1104.3913 [cs] (2011), http://arxiv.org/abs/1104.3913 (consulté le 15 fév. 2016) ; Anupam Datta *et al.*, « Proxy Non-Discrimination in Data-Driven Systems », arXiv:1707.08120 [cs] (2017), http://arxiv.org/abs/1707.08120 (consulté le 9 janv. 2021) ; Kusner *et al.*, cf. *supra* note 189.

[231] Kusner *et al.*, cf. *supra* note 189 ; Chris Russell *et al.*, « When worlds collide: integrating different counterfactual assumptions in fairness », dans *Advances in Neural Information Processing Systems* p. 6396-6405 (2017).

[232] Mayer-Schönberger et Cukier, cf. *supra* note 31 ; Pour une comparaison des États-Unis et du Royaume-Uni, voir Joris Van Hoboken, « From collection to use in privacy regulation? A forward looking comparison of European and US frameworks for personal data processing », p. 231 *Exploring the Boundaries of Big Data* (2016) ; Pour une perspective internationale, voir p. 63 Lee A. Bygrave, « Data privacy law: an international perspective », (2014) ; Pour une perspective européenne, voir Sandra Wachter, « Normative challenges of identification in the Internet of Things: Privacy, profiling, discrimination, and the GDPR », *Computer Law & Security Review* n° 34, p. 436-449 (2018) ; Sandra Wachter, « The GDPR and the Internet of Things: a three-step transparency model », *Law, Innovation and Technology* n° 10, p. 266-294 (2018) ; Pour une perspective européenne et allemande, voir Mario

exemple contribué à l'eugénisme en Europe, au Royaume-Uni[233] et aux États-Unis,[234] au génocide de la Seconde guerre mondiale, à des pratiques racistes en matière d'immigration et à un non-respect des droits de l'homme fondamentaux aux États-Unis,[235] à la justification de l'esclavage,[236] à la stérilisation forcée au Royaume-Uni,[237] aux États-Unis, en Allemagne et à Porto Rico dans la première moitié du XXe siècle,[238] à la répression, la castration et l'emprisonnement des membres de la communauté LGBT,[239] et au refus de donner aux femmes l'égalité des droits et de les protéger (par ex. contre les violences sexuelles).[240] La protection de la vie privée doit clairement être prise au sérieux lorsqu'on envisage la collecte de données personnelles sensibles à des fins de vérification de l'existence de biais.[241]

En mettant un instant de côté ces préoccupations, on serait tenté de penser que les problèmes de biais seront naturellement résolus en recueillant davantage de données (sensibles) et en comblant les écarts de représentation dans les ensembles de données d'entraînement et d'essai. Cependant, la résolution des écarts de représentation et autres biais des données n'entraînera pas automatiquement des résultats justes et équitables. Être conscient des inégalités ne revient pas à les rectifier.[242] Au contraire, la persistance de biais sociaux dans les sociétés occidentales suggère que des efforts politiques, sociaux et juridiques importants sont nécessaires pour les éliminer, plutôt que de simplement recueillir et tester davantage de données.

La lutte contre les inégalités exige des changements délibérés et souvent coûteux dans les processus décisionnels, les modèles commerciaux et les politiques. Pour justifier la collecte et l'utilisation de données sensibles supplémentaires, il faut avant tout démontrer une volonté politique et un engagement sérieux à remédier aux inégalités. Du point de vue de l'élaboration des normes, ces observations suggèrent que toute proposition de collecte de données sensibles afin de contrôler la présence de biais dans les systèmes d'IA médicale doit comprendre une limitation claire des finalités ainsi que des garanties de confidentialité, en plus d'un engagement à remédier aux inégalités sociales qui sous-tendent les biais découverts grâce aux tests. La

---

Martini, Wiebke Fröhlich et Saskia Fritzsche, « Algorithmen als Herausforderung für die Rechtsordnung » (2017) ; Pour une preuve empirique de la collecte mobile des données, voir Reuben Binns *et al.*, « Third party tracking in the mobile ecosystem » dans *Proceedings of the 10th ACM Conference on Web Science* p. 23-31 (2018) ; Sur les préjudices en ligne, voir Woods Lorna et Perrin William, « An updated proposal by Professor Lorna Woods and William Perrin », https://d1ssu070pg2v9i.cloudfront.net/pex/carnegie_uk_trust/2019/01/29121025/Internet-Harm-Reduction-final.pdf (consulté le 11 mai 2019).

[233] Cela s'est passé jusqu'aux années 30, voir Reni Eddo-Lodge, *Why I'm no longer talking to white people about race* p. 20-21 (2020).

[234] Jean Halley, Amy Eshleman et Ramya Mahadevan Vijaya, *Seeing white: An introduction to white privilege and race* p. 36 (2011).

[235] *Id.* p. 25.

[236] *Id.* p. 36-37.

[237] Eddo-Lodge, cf. *supra* note 232, p. 20-21.

[238] Halley, Eshleman et Vijaya, cf. *supra* note 233, p. 36-38.

[239] Jean Halley et Amy Eshleman, *Seeing straight: An introduction to gender and sexual privilege* p. 15-17 (2016).

[240] Saini, cf. *supra* note 178, p. 233-235.

[241] Sur la surveillance et les effets dissuasifs, voir Jon Penney, « Chilling Effects: Online Surveillance and Wikipedia Use » (2016), https://papers.ssrn.com/abstract=2769645 (consulté le 27 déc. 2017).

[242] Eddo-Lodge, cf. *supra* note 232, p. 208.

concrétisation de ces engagements n'est pas simple. L'UE Artificial Intelligence Act propose par exemple la création de « sandbox » (bacs à sable) réglementaires dans lesquels les fournisseurs d'IA pourraient vérifier la présence de biais dans leurs systèmes en utilisant une catégorie particulière de données spécifiquement recueillies à des fins d'essai.[243] Il manque à cette proposition l'élément essentiel d'un engagement à remédier aux inégalités découvertes.

[243] Commission européenne, cf. *supra* note 16, art. 53.

8. OBSERVATIONS FINALES

Les soins médicaux sont de plus en plus disséminés et concernent aujourd'hui une très grande diversité d'établissements, de personnels et de technologies. Aux fils des ans, la relation médecin-patient n'a cessé de s'adapter aux progrès de la médecine et de la recherche biomédicale ainsi qu'aux nouvelles pratiques en matière de soins. Parallèlement, l'IA, qui a la capacité d'étoffer ou de remplacer l'expertise clinique humaine par des analyses très complexes sur des données d'une ampleur et d'une diversité sans précédent, pourrait bien modifier la relation médecin-patient dans des proportions jamais atteintes auparavant.

L'adoption de l'IA n'est pas nécessairement un obstacle insurmontable à une bonne relation entre le médecin et son patient. L'IA peut certes modifier les relations de soins et déplacer les responsabilités qui incombent depuis toujours aux professionnels de la santé, mais cette issue n'est pas inévitable. C'est en effet le choix du modèle de service qui détermine dans quelle mesure un système d'IA fait obstacle à la « bonne » pratique de la médecine. Si l'IA vient seulement compléter l'expertise des professionnels de la santé, lesquels sont liés par le devoir de loyauté vis-à-vis à du patient, ses effets sur la fiabilité et la qualité humaine des entretiens cliniques peuvent se révéler minimes.

D'un autre côté, dans le cas où l'IA est utilisée pour étoffer largement l'expertise clinique humaine ou pour la remplacer, son impact sur la relation de soins est plus difficile à prévoir. Avec le recours croissant aux systèmes d'IA, de nouvelles normes, largement admises, en matière de « bons » soins verront sans doute le jour, les cliniciens passant plus de temps en face à face avec leurs patients tout en s'appuyant largement sur des recommandations issues de systèmes automatiques.

L'impact de l'IA sur la relation médecin-patient reste très incertain. Il est peu probable que nous assistions dans les cinq prochaines années à une reconfiguration radicale des soins, c'est-à-dire à un remplacement de l'expertise humaine par l'intelligence artificielle. Cela dit, certains événements comme la pandémie de Covid-19 et les pressions accrues qu'elle exerce sur les services de santé pourraient bien transformer le mode d'administration des soins, voire l'expertise qui les sous-tend. La télémédecine, par exemple, pourrait se banaliser, même si le diagnostic et le traitement restent l'apanage des professionnels de santé humains.

Une reconfiguration radicale de la relation médecin-patient telle que certains l'imaginent, où des systèmes artificiels diagnostiqueraient et traiteraient les patients directement, les cliniciens humains intervenant a minima, reste, semble-t-il, une perspective lointaine. Pour s'engager dans cette voie, il faut encore passer l'épreuve de l'efficacité clinique, ce qui, comme on l'a vu plus haut, reste un obstacle à la commercialisation et à l'adoption généralisée de ces systèmes[244]. De même, il faudrait définir de nouveaux modes de soins cliniques exploitant les meilleurs atouts des cliniciens humains et des systèmes artificiels, corriger autant que possible les

[244] Liu et al., *supra* note 112 ; Robbins et Brodwin, *supra* note 5.

insuffisances et les biais implicites de ces deux intervenants et mettre en place des contrôles de sécurité et de résilience adaptés. Si l'on ne prend pas dûment en considération les incidences de l'IA sur la pratique médicale, l'« intégrité morale de la relation médecin-patient » risque d'être largement dictée par des intérêts institutionnels et externes, le vécu des soins par le patient en faisant les frais[245].

À mesure que l'IA est adoptée par différents systèmes de santé et diverses juridictions, il importe de rappeler que les obligations morales liées à la relation médecin-patient sont toujours affectées voire déplacées par l'arrivée de nouveaux prestataires de soins. Alors que la technologie continue de progresser à grands pas, la façon dont le patient vit la maladie (vulnérabilité, dépendance, etc.) et les attentes que suscite la relation thérapeutique ne changent ni rapidement ni de manière radicale. La relation médecin-patient est une pierre angulaire de la « bonne » pratique médicale, et pourtant, elle semble évoluer vers une relation médecin-patient-IA. Le défi auquel sont confrontés les fournisseurs d'intelligence artificielle, les autorités de réglementation et les décideurs est de définir des normes et des exigences solides pour ce nouveau type de relation thérapeutique, afin que les intérêts des patients et l'intégrité morale de la médecine en tant que profession ne soient pas fondamentalement entamés par la mise en œuvre de technologies émergentes, profondément déstabilisatrices.

[245] Bauer, *supra* note 136, p. 90.

ANNEXE : VERTUS MEDICALES

Les vertus sont définies par rapport aux finalités de la pratique qu'elles sont censées servir. Dans le cas de la médecine, ces finalités consistent à apporter des soins d'un niveau suffisant afin d'améliorer la santé physique et mentale et le bien-être, au bénéfice d'une société composée de patients individuels. Elles sont réalisées au travers de la relation thérapeutique, dont la nature crée certaines obligations morales.

À l'instar des autres pratiques, la *phronesis* ou prudence est une vertu capitale en médecine, sans laquelle les autres vertus ne peuvent être intégrées dans les pratiques au moyen d'actes vertueux[246]. La justice, l'honnêteté et le courage sont aussi nécessaires pour protéger la médecine du pouvoir corrupteur des institutions médicales, notamment des hôpitaux, des organismes payeurs et des ministères[247]. Ces trois vertus essentielles sont nécessaires à la révision continue, par les praticiens, des normes d'excellence et des biens internes, ce qui exige une analyse personnelle et critique de la relation entre les actes que l'on accomplit et les normes de la pratique, ou de l'incidence des institutions sur la définition et la mise en œuvre des normes[248].

La justice est définie au sens large comme « la ferme volonté de donner à autrui ce qui lui est dû »[249] ou « la vertu d'accorder un dédommagement et de réparer ainsi des manquements au sein d'une communauté déjà constituée »[250]. Pour être justes, les normes relatives au traitement des personnes au sein d'une communauté doivent être « uniformes et impersonnelles », ce qui veut dire qu'il est injuste de privilégier les personnes de son entourage. Dans les systèmes de santé sociaux ou nationaux, la justice peut s'appliquer à la répartition parfaitement équitable des ressources médicales (par exemple, médicaments, traitements, consultations, etc.). La justice n'est pas qu'un concept quantitatif en vertu duquel toutes les parties prenantes reçoivent une part égale ; elle suppose également de mettre en adéquation les ressources avec les besoins du patient et de statuer selon l'importance relative des différents besoins.

La fidélité à la confiance et la bienveillance peuvent aussi être vues comme des valeurs essentielles, propres à la médecine, en raison du besoin de confiance qui caractérise les relations thérapeutiques[251]. Entre le médecin vertueux et le patient doit s'instaurer, au fil du temps, une relation de confiance, dans laquelle les valeurs, les attentes et les questionnements sur la maladie et les soins médicaux adaptés sont mis en commun. Pour que la confiance existe, il faut, au minimum, que le patient soit convaincu que le

---

[246] MACINTYRE, *supra* note 122, p. 154 ; G. Widdershoven & Lieke Van der Scheer, « Theory and methodology of empirical ethics: a pragmatic hermeneutic perspective », dans *EMPIRICAL ETHICS IN PSYCHIATRY* *23–36* (2008), http://books.google.co.uk/books?hl=en&lr=&id=Lvq0lkDyEBQC&oi=fnd&pg=PA23&dq=Theory+and+methodology+of+empirical+ethics:+a+pragmatic+hermeneutic+perspective&ots=IXt3OC6Obh&sig=EU-idi92-6EzBI6uTp8UNReq4AY#v=onepage&q&f=false; PELLEGRINO ET THOMASMA, *supra* note 120.

[247] MACINTYRE, *supra* note 122, p. 192.

[248] *Id.*, p. 191.

[249] PELLEGRINO ET THOMASMA, *supra* note 120, p. 92.

[250] MACINTYRE, *supra* note 122, p. 156.

[251] PELLEGRINO ET THOMASMA, *supra* note 120, pp. 71 et 156.

médecin agit, dans une certaine mesure, avec bienveillance ou dans son intérêt et pour son bien-être[252].

Parmi les autres vertus figurent notamment la compassion, la force morale, l'intégrité et la tempérance. La compassion est le trait de caractère d'un médecin qui « se met à la place » de son patient afin, d'une part, de comprendre comment les valeurs de ce dernier ainsi que ses attentes en matière de soins et de bien-être social, émotionnel et physique jouent sur son expérience de la maladie, et, d'autre part, d'adapter ses soins et recommandations aux besoins de chaque patient en tant qu'individu unique[253]. La compassion nécessite aussi parfois de promouvoir certaines valeurs liées à la santé et de se concerter avec le patient pour le convaincre de la meilleure intervention au sens de l'adéquation entre ses valeurs et les résultats sur la santé tels qu'ils sont perçus par le médecin[254].

La force morale est une forme de courage par laquelle un individu est prêt à « subir un préjudice personnel au nom d'un bien moral », à l'image d'un médecin refusant d'agir selon des règles institutionnelles qui nuiraient au bien-être de son patient, quitte à mettre en danger sa carrière et son affiliation professionnelle[255]. La force morale peut créer chez le médecin une obligation de dénoncer les préjudices que pourraient causer à ses patients de nouvelles politiques institutionnelles, de nouvelles technologies ou de nouveaux traitements. La tempérance désigne, quant à elle, le fait de restreindre son comportement dans le cadre d'une pratique afin de satisfaire aux obligations de cette dernière. La tempérance peut être employée comme un synonyme de la vertu, mais elle s'en distingue en tant qu'elle est un trait de caractère du médecin vertueux qui renonce à son intérêt personnel dans le traitement de ses patients. Sans cette retenue, les autres vertus ne peuvent pas être exercées[256].

Enfin, l'intégrité s'entend de la possession de toutes les vertus, couplée à la capacité de faire la distinction entre les différents principes moraux, au moment de choisir les actions susceptibles d'apporter les bienfaits de la médecine dans différentes situations[257]. C'est la vertu fondamentale de la quête narrative de la vie bonne et elle est visible dans une vie de comportement vertueux[258]. L'intégrité peut être exercée lorsqu'un médecin défend les intérêts et le bien-être de son patient face aux pressions institutionnelles, par exemple lorsqu'il décide de ne pas renvoyer, trop tôt, un patient hospitalisé[259]. Edgar et Pattison définissent l'intégrité comme « la capacité de débattre et de réfléchir utilement, en tenant compte de la situation, des connaissances, de l'expérience et de l'information (celles que l'on possède et celles des autres), au sujet de facteurs complexes et antagoniques qui ont une incidence sur l'action ou l'action

[252] *Id.*, p. 156.
[253] *Id.* pp. 79 et 81.
[254] Emanuel et Emanuel, *supra* note 123, p. 2226.
[255] PELLEGRINO ET THOMASMA, *supra* note 120, p. 109.
[256] *Id.*, p. 117.
[257] *Id.*, p. 127 ; Edgar et Pattison, *supra* note 145, p. 102.
[258] MACINTYRE, *supra* note 122.
[259] Edgar et Pattison, *supra* note 145, p. 94.

potentielle[260]. » L'intégrité est donc peut-être indissociable de la *phronesis*, de la tempérance et de la force morale.

[260] *Id.*, p. 102.