MINISTÈRE
DES SOLIDARITÉS
ET DE LA SANTÉ

Liberté
Égalité
Fraternité

Recommandations de bonnes pratiques pour intégrer l'éthique dès le développement des solutions d'Intelligence Artificielle en Santé : mise en œuvre de « *l'éthique by design* »

Présentation des travaux du GT3 de la Cellule éthique du numérique en santé
de la Délégation ministérielle au Numérique en Santé

Avril 2022

Sommaire

Editorial de Jean-Gabriel Ganascia

Qu'entend-on par « éthique by design » et pourquoi conserver une locution anglaise dans le titre d'un texte en français ? La question mérite qu'on s'y appesantisse, et ce d'autant plus qu'à l'évidence la gêne à traduire ne relève ici ni de la paresse, ni d'une ignorance des ressources de la langue française, ni même — et encore moins — d'une affection, comme cela arrive parfois, mais d'une ambivalence du terme anglais difficile à rendre en français.

Pris littéralement, « éthique by design » se transpose en « éthique par conception » ou, plus précisément, en « éthique dès la conception ». Et, il en va bien de cela lorsqu'on développe des solutions d'IA pour la santé ; en effet, on doit songer à l'éthique dès les premières ébauches, qu'il s'agisse de la finalité, qu'il faut peser en regard des besoins et des contraintes des différents acteurs — soignants, assurances sociales et/ou soignés — censés les employer, ou de l'architecture informatique que l'on doit penser pour dissuader, voire pour interdire lorsque c'est envisageable, des utilisations oublieuses de la morale. Ainsi préfèrera-t-on un traitement local à une centralisation excessive des données personnelles, pour éviter leur piratage ; de même, on recueillera le consentement éclairé des patients, chaque fois que nécessaire. Dans un registre symétrique, on doit exiger la « garantie humaine » de toute décision, ce qui présuppose une compréhension claire des recommandations des machines, et donc des explications qui aident à en interpréter la signification.

On doit aussi se souvenir qu'en anglais, « by design » veut d'abord dire « intentionnellement », « de façon consciente et préméditée », ce qui s'oppose à l'inattendu et à l'incertain. En cela, l'éthique by design résulte d'une volonté opiniâtre de penser l'action, sa légitimité et ses conséquences, de toujours donner satisfaction, de ne rien abandonner au hasard, de ne pas laisser aller les choses telles qu'elles sont sans intervenir. À cette fin, on doit tester en situation les dispositifs intégrant des technologies d'intelligence artificielle pour éviter les accidents. Le caractère inductif de l'apprentissage machine rend cette validation — validation des données et des résultats de l'apprentissage — d'autant plus nécessaire ; on doit prévenir, par tous les moyens, les risques d'erreurs systématiques — ou de ce que l'on a coutume d'appeler les biais — ceux-ci provenant de la répartition et de la description des instances sur lesquelles on entraîne les systèmes.

Ce rapport recense l'ensemble des exigences éthiques requises par les solutions d'IA en santé au double sens du « by design », à savoir tant au cours de la conception, pour intégrer autant que faire se peut, des valeurs éthiques dans l'architecture matérielle et logicielle des systèmes, qu'à l'issue de la réalisation, pour valider les solutions proposées et éviter les aléas consécutifs au caractère principalement inductif de l'IA aujourd'hui.

Professeur d'informatique et philosophe
Faculté des Sciences de Sorbonne Université

Editorial des pilotes du GT3 « *Ethique by Design* »

Ce guide méthodologique représente la concrétisation d'une démarche de deux ans, fondée sur une méthodologie de concertation large des différentes parties prenantes de l'écosystème de la santé numérique, d'auditions d'experts du domaine, accompagnée d'une veille nationale et internationale sur les sujets relatifs à l'IA en santé. Elle capitalise ainsi sur les meilleures pratiques de régulation éthique positive du numérique et de l'intelligence artificielle en santé. Elle a permis aussi de capitaliser sur les contributions des représentants des patients, des professionnels de santé et des innovateurs en IA.

Cette dynamique s'est, par ailleurs, doublée, en parallèle, de contacts étroits avec les acteurs du processus législatif bioéthique et de la préparation du projet de règlement européen sur l'intelligence artificielle en cours de finalisation dans le cadre de la présidence française de l'Union européenne. Dans ces conditions, une large partie des recommandations formulées sont d'ores et déjà entrées dans le droit et dans le réel des pratiques au plan national et à l'échelle européenne. C'est le cas notamment, dans ce nouveau cadre juridique de l'IA en santé, du devoir d'information du patient sur le recours à des traitements algorithmiques dans sa prise en charge ou du nouveau principe de Garantie Humaine de l'intelligence artificielle.

Pour autant, ainsi qu'il l'était attendu au moment de l'initialisation de ce travail, ce guide ne fait pas que porter des obligations nouvelles. Il s'agit véritablement d'une démarche éthique « by design », depuis la conception de la solution d'IA jusqu'à sa mise en œuvre en vie réelle. L'un des objectifs de ce guide est donc, de façon lisible et pragmatique, d'accompagner les innovateurs en intelligence artificielle pour s'inscrire résolument dans cette approche éthique dès le début du processus de création de leurs futures solutions. Ce faisant, la régulation éthique devient véritablement positive et dynamique en ce sens qu'elle positionne la France et l'Europe au plus haut niveau d'exigence tout en leur conférant également un véritable avantage comparatif en termes de valeur ajoutée et de compétitivité au plan mondial.

Les porteurs de cette démarche ont, en effet, la conviction profonde que l'éthique et l'innovation en IA ne s'opposent pas. Tout à l'inverse, dans le nouveau cadre mondialisé de l'IA en santé, seule une approche véritablement éthique permettra de donner à la France une compétitivité durable au service de l'amélioration de la prise en charge des patients.

David Gruson
Directeur Programme Santé LUMINESS
Fondateur ETHIK-IA

Brigitte Séroussi
Directrice de Projets DNS
Responsable de la Cellule Ethique

Membres du GT3 de la Cellule éthique du numérique en santé

« Ethique By Design » pour les solutions numériques en santé embarquant de l'Intelligence Artificielle (IA) »

AMAR Nicolas (DGE, Secrétariat d'Etat chargé du numérique)
BAUDOUIN Peggy (CEA)
BERANGER Jérôme (ADELIAA)
BOURDEN Aude (APF France handicap)
BRULE Emeline (École d'ingénierie et d'informatique, Université du Sussex)
CATHERINOT Sabine (ASI)
CAVET Madeleine (Médecin radiologue)
CHALLIER Jordan (Pharmacien de santé publique)
CLATZ Olivier (DNS)
COLLIGNON Corinne (HAS)
DAVID Claudie (SIB)
DENIS Christophe (Sorbonne Université)
DEVILLERS Laurence (Sorbonne Université)
FALISE MIRAT Béatrice (Care Insight)
FRAYSSE Jean-Louis (Bot Design)
FRIJA Raphaëlle (Syntec Numérique)
Goehrs Clément (Synapse-Medicine)
GOGLIN Jean-François (Connective Santé)
GUILLOT Caroline (Health Data Hub)
GZIL Fabrice (Espace éthique Ile de France, CCNE)
HELLOCO Camille (Doctolib)
HUET Benoit (ANAP)
JAAFAR Delphine (Cabinet Vatier)
LAMOUREUX Philippe (LEEM)
LE FOL Vincent (VYV3)
LEFEVRE Xavier (Fair&Smart)
LETHEUX Corinne (Présanse)
LUCAS Jacques (ANS)
MANAUD Nathalie (CEA, LEEM)
OLLE Florence (SNITEM)
PARROT Daniela (Déléguée à la protection des données, Ministère des solidarités et de la santé)
PERSON Anaïs (Women in Legaltech)
PLOUVIER Claire
RAMIREZ Juan-Fernando (Air liquide)
ROUZO Delphine (Doctolib)
SEREIN Fréderic (Groupe Nehs)
SEVAL Frédéric (Droit de la Santé, Ministère des solidarités et de la santé)
TRANG Stéphanie (AI for Health)
VAUGELADE Cécile (SNITEM)

Copilotes du GT3 : GRUSON David (ETHIK-IA, Luminess)
SEROUSSI Brigitte (DNS, Sorbonne Université)

1. Synthèse

Dans un contexte de montée en puissance du recours à l'intelligence artificielle dans le domaine de la santé, les pouvoirs publics font face à la préoccupation croissante des professionnels de santé et des citoyens sur la difficulté à comprendre clairement les processus de fonctionnement des algorithmes, les processus de sélection et de traitement des données et l'impact médico-légal et éthique des services numériques embarquant de l'intelligence artificielle (IA).

Convaincus du bien-fondé de la mise en œuvre de l'IA dans le domaine de la santé et de la nécessaire promotion de l'innovation au bénéfice de la médecine, les membres du groupe de travail de la Cellule Ethique du numérique en santé, dédié à l'éthique de l'IA ont souhaité définir un cadre éthique des services numériques de santé embarquant l'IA pour protéger le public et circonscrire d'éventuelles dérives.

A l'issue d'auditions de personnes qualifiées, d'instructions des textes de référence et de la littérature scientifique disponible sur cette thématique, le présent guide a été produit.

Il vise à présenter les questionnements éthiques à avoir et les recommandations de bonnes pratiques pour intégrer la dimension éthique dès la construction d'une solution d'IA en santé. Bien que revenant sur les différents types d'IA, ce guide s'adresse en priorité aux acteurs souhaitant implémenter des solutions d'IA basées sur des algorithmes ayant été entraînés à partir de données massives (IA connexionniste).

A travers ce présent guide, le lecteur bénéficie d'une démarche méthodologique clé en main pour construire sa solution d'IA. Il est ainsi préconisé de suivre une démarche projet commençant par une démarche de cadrage visant à installer d'une part un comité scientifique, technique et éthique et d'autre part à définir les modalités d'implication des parties prenantes tout au long de la construction de la solution d'IA. Au même titre que l'implication des parties prenantes, il est préconisé de mener en continu, une démarche d'analyse et de sécurisation des risques. L'étape de cadrage est fondamentale dans la démarche méthodologique proposée par le groupe de travail, elle doit en effet également permettre de poser le cadre en définissant les finalités de la solution (définition qui viendra sous-tendre les étapes suivantes dédiées à la collecte et au pré-traitement des données) mais également les principes de gouvernance de la solution et les rôles et responsabilités des différentes parties prenantes.

Les quatre étapes suivantes viennent donner corps au cadre préalablement posé et contribuent à construire effectivement, au moyen de questionnements éthiques précis, la solution d'IA dans toutes ses composantes.

1. **La collecte des données**. Cette étape, pour laquelle un certain nombre de questionnements éthiques relèvent par ailleurs de la bonne application du RGPD, doit permettre, à travers les questionnements éthiques préconisés d'aboutir à la formalisation :

- De mesures visant à assurer :

    - Le consentement éclairé des patients à l'origine des données sur lesquelles les algorithmes sont entraînés à la réutilisation de leurs données au-delà de la finalité première du recueil
    - La proportionnalité des données collectées par rapport à la finalité du traitement
    - La non ré-identification directe des données
    - La qualité des données (lutte contre les biais cognitifs)
    - La représentativité de la population d'analyse/population cible/prévention des discriminations (lutte contre les biais de sélection)
    - L'implication des utilisateurs
- Des mesures de sécurité visant à assurer :
    - Le transfert sécurisé des données
    - La qualité de l'hébergement des données
    - La cyber-sécurité à l'état de l'art
- Des mesures pour garantir la non-réutilisation non éthique des données

2. **Le pré-traitement des données**. Au même titre que lors de l'étape précédente, le groupe de travail met à disposition les questionnements éthiques permettant in fine la mise en œuvre des mesures de :
    - Traitement des données manquantes (réduction des biais)
    - Rééquilibrage des populations minoritaires (réduction des biais)
    - Séparation des données (représentativité de l'échantillon d'apprentissage et d'évaluation par rapport à la population cible et la finalité du traitement)

3. **La construction de l'algorithme**. Cette étape est également clé, les recommandations données doivent accompagner le lecteur dans la construction de l'objet cœur de son projet à savoir l'algorithme en lui-même. Les questionnements éthiques proposés et bonnes pratiques associées doivent permettre aux porteurs de projets de solution d'IA de :
- Choisir l'algorithme d'apprentissage en adéquation avec la finalité
- Définir :
    - La politique qualité de l'algorithme
    - Des mesures de transparence
    - La politique de traçabilité de la démarche de construction de l'algorithme
    - La politique d'explicabilité des résultats explicables, le processus d'auditabilité des résultats non explicables
- Définir et implémenter des fonctionnalités et mécanismes visant à assurer :
    - L'identification et l'élimination des biais
    - La correction des erreurs
    - La traçabilité des traitements
    - L'adaptabilité
    - L'intégration des évolutions réglementaires et des avancées médiales
    - La maintenance et le versionning
- Définir les indicateurs de dérive du système

4. **L'évaluation de l'algorithme en amont de la mise en production de la solution numérique.** Cette ultime étape vise, à travers les questionnements posés, à accompagner les acteurs dans la :

- Mise en œuvre des principes d'évaluation externe :
    - Technique (bugs), clinique (gold standard, score de précision)
    - De l'utilisabilité (professionnels de santé, patients, usagers)
    - De la non-discrimination/équité
    - De la robustesse/reproductibilité
- Mise en œuvre des procédures en cas de cyber-attaques (analyse d'impact sur la sécurité du système d'IA)
- Mise en œuvre des mesures pour assurer l'information (juste et égalitaire) des utilisateurs (professionnels de santé, patients) relative à :
    - La finalité, gouvernance, responsabilité
    - L'architecture
    - L'origine des données et qualité (légalité de la collecte et des traitements)
    - L'explication des processus, explication du périmètre de la partie non explicable
    - La méthode d'apprentissage, d'inférence, etc.
    - Les limites de l'utilisation de l'algorithme (faux positifs, faux négatifs si classification)
    - Les modalités de recours en cas d'erreurs
    - L'implication des utilisateurs
- Mise en œuvre des mécanismes de garantie humaine (professionnels de santé, équipe de soins) pour assurer :
    - Le contrôle de l'IA par l'humain
    - L'autonomie décisionnelle des utilisateurs
    - Le maintien des compétences des utilisateurs
    - L'intervalle de confiance de l'IA / garde-fou des erreurs de l'IA
    - Les audits (désaccords IA / professionnels de santé)
- Définition de l'instance de régulation (audit, Label Ethique-IA)
- Analyse d'impact organisationnel sur le parcours de soins
- Analyse d'impact environnemental et éco-responsabilité

Ce guide doit par ailleurs permettre aux lecteurs de s'orienter dans le paysage des recommandations existantes sur la thématique et produites par d'autres institutions françaises ou internationales. Une présentation synthétique des autres sources pertinentes est ainsi proposée ainsi qu'une analyse croisée permettant au lecteur d'identifier, voire de s'alimenter, des autres recommandations.

2. Introduction

Objet du document

Le présent document a pour objectif de proposer des recommandations de bonnes pratiques pour intégrer la dimension éthique tout au long du processus de construction d'une solution numérique pour la santé intégrant une intelligence artificielle (IA). Le présent guide vise par ailleurs à s'inscrire dans un cadre de réflexion globale sur la question de l'IA et à s'articuler avec les récentes publications d'autres institutions.

Contexte général

La montée en puissance de l'IA

« L'intelligence artificielle est partout dans nos vies » disait Jean-Gabriel Ganascia, professeur d'informatique à Sorbonne Université, et l'un des principaux spécialistes français de l'intelligence artificielle, dans une interview donnée dans La Tribune en 2015. Plus de 6 ans après, c'est encore plus vrai avec des applications particulièrement intéressantes dans le domaine de la santé : aide à la décision diagnostique et thérapeutique, médecine prédictive, médecine de précision, prothèses intelligentes, chirurgie assistée par ordinateur, prévention épidémiologique, etc. Le développement des algorithmes d'apprentissage automatique (dont les récentes avancées en matière d'apprentissage profond et de réseaux de neurones), la prolifération des données numériques et biométriques, l'accélération de la puissance de calcul, et les progrès dans les domaines médical et biologique figurent parmi les ingrédients de cette révolution du système de santé accélérée par le contexte récent de la crise sanitaire.

Avec la télémédecine, les thérapies digitales et l'IA, on estime que le secteur de la santé numérique représentera un marché de 234,5 milliards de dollars au niveau mondial d'ici 2023, soit une hausse de 160 % par rapport à 2019[1]. Ces prévisions créent un appétit politique et financier pour le développement rapide de l'industrie de l'IA, notamment dans le domaine de la santé. L'IA en santé a d'ores et déjà donné des résultats significatifs (apprentissage par reconnaissance d'images en radiologie, ophtalmologie ou encore dermatologie) et la médecine algorithmique est dès à présent entrée dans les faits. Néanmoins, si la valeur clinique de l'IA a été démontrée sur des études ponctuelles, elle n'est pas encore complètement effective en routine. Par exemple, les données actuelles sur l'utilisation des systèmes d'IA dans le dépistage du cancer du sein sont d'une qualité et d'une quantité insuffisantes pour être mises en œuvre dans la pratique clinique[2]. Par ailleurs, le manque de compréhension claire des processus sous-jacents aux traitements des données (effet « boîte noire ») et les préoccupations croissantes concernant l'impact éthique et médico-légal des systèmes d'IA constituent des difficultés à résoudre en amont du déploiement de ces outils.

[1] https://store.frost.com/global-digital-health-outlook-2020.html
[2] https://www.bmj.com/content/374/bmj.n1872

Le nécessaire cadre éthique de l'IA

Promouvoir l'innovation en matière d'IA s'avère primordial pour permettre des avancées majeures dans le domaine de la santé. Mais l'IA présente toujours des zones d'ombre et nécessite un cadre éthique à même de protéger les acteurs du système de santé et les citoyens d'éventuelles dérives.

De nombreuses études ont mis en lumière le fait que nous ne sommes pas tous égaux devant les algorithmes utilisés par l'IA. Leur « partialité » peut engendrer de réelles conséquences sur nos vies. Par conséquent, si l'on souhaite faire émerger des technologies d'IA conformes à nos valeurs et normes sociales, il est essentiel de mobiliser la communauté scientifique, les pouvoirs publics, les industriels, les entrepreneurs et les organisations issues de la société civile sur le sujet d'une IA « *digne de confiance* »[3].

Le rôle de la Délégation ministérielle au Numérique en Santé

Afin de répondre à ces objectifs, la Délégation ministérielle au Numérique en Santé du Ministère des Solidarités et de la Santé (DNS) s'est dotée **d'une cellule Ethique**. Cette cellule a pour mission de faire de l'éthique un élément central du virage numérique en santé, notamment grâce à l'élaboration d'outils pratiques de sensibilisation, d'évaluation et de labellisation éthique à destination des professionnels de santé, des industriels, des usagers du système de santé et des pouvoirs publics. Des travaux sur l'éthique du numérique sont, par ailleurs, engagés dans d'autres cadres comme le Comité pilote d'éthique du numérique.

Au sein de la cellule Ethique du numérique en santé de la DNS, le GT3 a pour mission d'établir un document d'appui opérationnel utilisable par les producteurs d'algorithmes à destination des secteurs sanitaires, sociaux et médico-sociaux afin de les accompagner dans la conception éthique *by design* de leurs solutions d'IA en santé. La gestion de crise COVID-19 et les cas de recours constatés sur le pilotage par les données et l'intelligence artificielle ont encore davantage mis en lumière la nécessité d'une conception éthique dès l'origine de ces algorithmes et le besoin d'un pilotage continu selon ces principes au cours de la vie réelle de ces algorithmes.

De nombreux documents et rapports ont été publiés sur le sujet au niveau national, européen et international. Le GT3 de la cellule éthique du numérique en santé a souhaité s'inscrire dans ce mouvement avec la volonté d'être pragmatique. Aussi, il est proposé un logigramme opérationnel listant les questionnements éthiques associés à chacune des grandes étapes de conception de ces algorithmes et les cadres de référence utilisables, associé à une note de repérage synthétique sur les enjeux éthiques *by design* de la conception des algorithmes médicaux.

La création de la cellule Ethique du numérique en santé représente l'action 4 « Ethique » de l'orientation 2 « *Intensifier l'éthique, la sécurité, l'interopérabilité des systèmes d'information*

[3] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

de santé » de la Feuille de Route du numérique en santé telle que présentée lors du Conseil du Numérique en Santé du 18 Juin 2020. [4]

[4] https://esante.gouv.fr/sites/default/files/media_entity/documents/cns_18062021_v0.6-vf--6.pdf

3. L'intelligence artificielle : quelle définition ?

Domaine de recherche en pleine expansion, l'intelligence artificielle (IA) est récemment devenue partie intégrante du langage courant. Son histoire est pourtant plus ancienne et remonte aux débuts de l'ère informatique, dans les années 1950.

Le premier à avoir abordé la notion d'intelligence artificielle est le mathématicien Alan Turing en 1950. En 1956, Marvin Lee Minsky, scientifique américain, définit l'intelligence artificielle comme « *la construction de programmes informatiques qui s'adonnent à des tâches qui sont, pour l'instant, accomplies de façon plus satisfaisante par des êtres humains, car elles demandent des processus mentaux de haut niveau tels que : l'apprentissage perceptuel, l'organisation de la mémoire et le raisonnement critique.* »

Globalement, l'IA vise à développer des dispositifs, matériels et logiciels, capables de mettre en œuvre un traitement visant à produire le même résultat que celui obtenu par les mécanismes cognitifs mis en œuvre par un expert humain engagé dans une tâche de résolution de problèmes, à des fins d'assistance ou de substitution à des activités humaines.

Le champ de l'IA est donc naturellement extrêmement vaste, tant en ce qui concerne les procédures techniques utilisées que les disciplines convoquées : mathématiques, informatiques, sciences cognitives… Les méthodes d'IA sont nombreuses, diverses et souvent assez anciennes : beaucoup d'algorithmes utilisés aujourd'hui ont été développés il y a déjà plusieurs dizaines d'années.

L'IA des années 70-80 diffère quelques peu de l'IA actuelle. En effet, l'IA des années 70-80 était une **IA essentiellement symbolique** qui s'appuyait sur un raisonnement logique, en exploitant les connaissances préexistantes modélisées dans une base de connaissances (par exemple sous la forme de règles de décision). A partir de la caractérisation d'un problème à résoudre, le moteur d'inférences parcourt la base de connaissances et propose la solution adaptée. L'IA symbolique est *explicable*, la trace du raisonnement effectué par le système permet de documenter la construction de la solution proposée.

La notion d'IA a évolué depuis les années 70-80, il s'agit dorénavant d'une **IA connexionniste** incarnée par l'apprentissage automatique (ou *machine learning*, ML), illustré par des algorithmes d'apprentissage profond et des réseaux de neurones entraînés à partir de données massives. Le développement de l'IA connexionniste se fait dans un contexte technologique marqué par la « mise en données » du monde (datafication) qui touche l'ensemble des domaines et des secteurs. L'IA connexionniste qui s'appuie sur des calculs est souvent comparée à une « boîte noire » dont les résultats sont difficiles à expliquer.

Il existe plusieurs types d'apprentissage :

- L'apprentissage supervisé consiste à apprendre à une machine à catégoriser des objets à partir d'un grand nombre d'objets préalablement étiquetés (la catégorie de chaque objet est connue). Pendant la phase d'apprentissage, l'algorithme est calibré pour adapter ses paramètres de catégorisation aux données fournies. Cet algorithme est ensuite utilisé pour catégoriser de nouveaux objets.

- L'apprentissage non supervisé consiste à fournir à la machine un grand nombre d'objets non catégorisés. Le système va repérer des régularités, des proximités, des corrélations pour construire un algorithme de classification basée sur la ressemblance (les objets qui se ressemblent seront dans la même catégorie).

- Dans le cas de l'apprentissage par renforcement, le système va induire le comportement d'un « agent » (par exemple un robot) évoluant dans un « environnement » donné (par exemple dans une pièce où il doit manipuler des objets) qui apprend grâce à un système de « récompense ».

L'apprentissage par renforcement a permis au programme Alpha Zero de Google de battre le champion de jeu de Go Lee Sedol en 2016. Si l'on voulait entraîner un programme à jouer au jeu de Go avec de l'apprentissage supervisé il faudrait lui donner comme données d'entraînement un grand nombre de parties de maîtres et il apprendrait à en reproduire les coups. En apprentissage par renforcement, le programme joue des parties contre lui-même et n'a pour unique information pour apprendre que les récompenses finales des victoires (ou les « punitions » des défaites).

***Il est à noter que l'apprentissage supervisé reste le cas de figure le plus courant*.**

L'IA généraliste qui rapproche l'IA symbolique et l'IA connexionniste est le prochain enjeu stratégique de la recherche en IA (IA explicable). En même temps, les systèmes de ML devraient évoluer pour aller vers des systèmes devenant plus proactifs et axés sur la récompense, apprenant continuellement à répondre à des applications de plus en plus complexes. Ils nécessiteront de fait plus de surveillance pour s'assurer qu'ils fonctionnent comme prévu (cf. Figure 1[5]).

[5] Extrait de https://qualitysafety.bmj.com/content/28/3/231

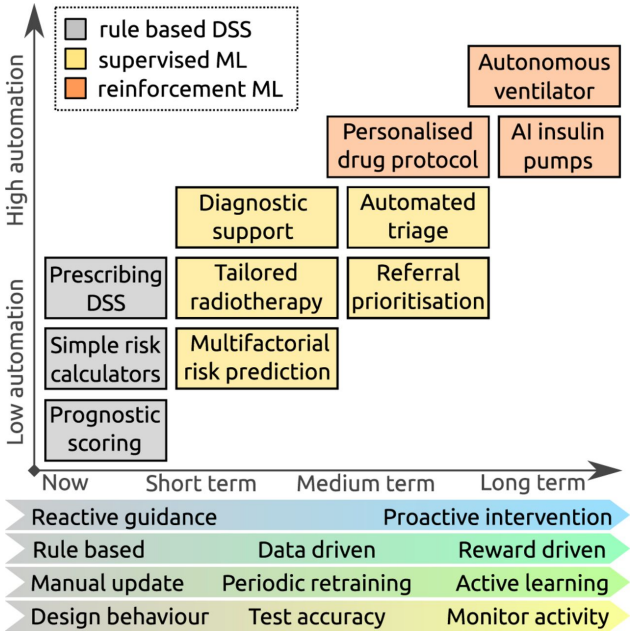


Figure 1 : Tendances attendues dans la recherche sur l'apprentissage automatique : les encadrés montrent des exemples représentatifs de l'aide à la décision offerte par des systèmes à base de règles (en gris), et les applications de ML (en jaune et orange), allant dans le sens d'une automatisation croissante.

4. La méthode de travail adoptée par le GT3

Ce guide vise à présenter les recommandations de bonnes pratiques pour intégrer la dimension éthique dès la construction d'une solution d'IA en santé. Ont été considérées les solutions d'IA ayant été entraînées à partir de données massives (IA connexionniste). Les systèmes évolutifs auto-apprenants n'ont quant à eux pas été abordés dans les différentes discussions du groupe de travail.

Il est nécessaire de rappeler par ailleurs que certaines solutions d'IA utilisées en santé ont un statut de dispositif médical (DM). La conception, la gestion des risques, la mise sur le marché et la surveillance des DM, y compris ceux embarquant de l'IA, sont régis par un cadre réglementaire européen stricte (Règlement UE 2017/745) qui prévoit un processus de certification conduisant au marquage CE médical du produit. Cette certification s'appuie sur deux éléments clés à produire par le fabricant et régulièrement contrôlés par les organismes notifiés :

- La documentation technique relative au produit qui démontre sa conformité aux exigences du règlement (dont notamment performance, bénéfice clinique et sécurité)
- Et le système de management de la qualité de l'entreprise qui démontre la capacité de l'entreprise à reproduire des produits conformes dans le temps et pour toutes les productions/versions

Le marquage CE médical est un pré-requis réglementaire indispensable à l'utilisation de ces produits. Les recommandations décrites dans ce rapport peuvent alimenter certains des éléments de démonstration attendus au regard de la conformité aux exigences de ce règlement en ce qui concernent les solutions d'IA DM.

Ce guide vise par ailleurs à mettre en lumière l'articulation (convergences et divergences) entre les principes proposés par la délégation ministérielle au numérique en santé et les principes proposés par d'autres institutions venant également de publier des recommandations sur la même thématique à savoir :

- La Commission Européenne dont le groupe d'experts indépendants de haut niveau a publié fin 2019 ses *« Lignes directrices en matière d'éthique pour une IA digne de confiance* »[6]
- L'Organisation Mondiale de la Santé (OMS) dont le guide « *Ethics and governance of artificial intelligence for health* » a été publié en juin 2021[7]
- La Haute Autorité de Santé (HAS) qui a publié fin 2019 des recommandations sur l'IA au sein de son rapport d'analyse prospective *« Numérique : quelle (R)évolution ? »*[8] et,

[6] https://op.europa.eu/fr/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1

[7] https://apps.who.int/iris/rest/bitstreams/1352854/retrieve

[8] https://www.has-sante.fr/upload/docs/application/pdf/2019-07/rapport_analyse_prospective_20191.pdf

en septembre 2020, son outil pour l'évaluation des dispositifs médicaux embarquant de l'intelligence artificielle. [9]

- La *Food and Drug Administration* (FDA) des Etats-Unis, Health Canada et the *Medicines & healthcare products regulatory agency* du Royaume-Uni [10] qui ont conjointement produit en octobre 2021 des recommandations sur le développement des produits médicaux faisant appel à l'IA.
- Le conseil de l'OCDE [11]
- L'UNESCO [12]qui a publié le 24 novembre 2021 ses recommandations sur l'éthique de l'IA. L'UNESCO, à travers ses recommandations souhaite mettre à disposition « *un instrument normatif mondial visant à doter l'IA d'une base éthique solide, qui non seulement protégera mais aussi promouvra les droits humains et la dignité humaine. Cet instrument constituera une boussole éthique et une base normative globale permettant de faire respecter l'État de droit dans le monde numérique* ».

Entre 2019 et juin 2020, les travaux menés par la cellule éthique de la délégation ministérielle du numérique en santé se sont déroulés de la manière suivante :

1. Audition de personnes qualifiées sur le sujet (voir annexe « liste des personnes qualifiées auditionnées par le groupe de travail »), d'entrepreneurs ayant développé des solutions d'IA en santé (voir annexe « liste des entrepreneurs auditionnés par le groupe de travail »), et d'organisations du secteur de la santé ayant travaillé à l'intégration de l'éthique dans leur stratégie digitale (voir annexe « liste des organisations auditionnées par le groupe de travail »).
2. Analyse de la littérature sur les démarches éthiques by design pour les solutions d'IA en santé.
3. Modélisation des étapes de développement d'une solution d'IA en santé.
4. Analyse des documents princeps issus de l'administration Trump (US), des travaux de l'Union Européenne, et des travaux nationaux conduits par la Haute Autorité de Santé (étude des versions préliminaires de la grille d'évaluation des dispositifs médicaux intégrant des systèmes d'IA), et travail sur table pour coconstruire les questionnements et les critères éthiques propres à chaque étape de développement.

Ces propositions représentent le résultat des réflexions du GT3 sur le sujet.

[9] https://www.has-sante.fr/upload/docs/application/pdf/2016-01/guide_fabricant_2016_01_11_cnedimts_vd.pdf#page=51
[10] https://www.fda.gov/media/153486/download
[11] https://oecd.ai/en/ai-principles
[12] https://fr.unesco.org/artificial-intelligence/ethics#recommandation

5. Résultats : questionnements éthiques propres à chaque étape de la construction d'une solution d'IA en santé

Préalablement au début des travaux, il est recommandé de se doter d'un conseil scientifique, technique et éthique (CSTE ou CoSTE). Le conseil scientifique devra, autant que faire se peut, inclure :

- Les développeurs de la solution d'IA ;
- Un panel d'utilisateurs (professionnels de santé, patients, grand public) ciblés par la solution d'IA.

Le groupe de travail préconise, pour la construction d'une solution d'IA en santé, d'intégrer une démarche méthodologique comprenant quatre étapes, précédées d'une étape de cadrage amont.

L'étape de cadrage amont doit permettre de définir la finalité de la solution d'IA, de déterminer le type d'apprentissage choisi pour la solution et de stabiliser les principes de gouvernance des données, et de responsabilité de la solution.

Les quatre étapes de la démarche méthodologique proposée comprennent :

- La collecte des données,
- Le pré-traitement des données,
- La construction de l'algorithme
- L'évaluation de l'algorithme en amont de la mise en production de la solution numérique.

Il convient de préciser, que pour chacune des étapes identifiées, un certain nombre de critères sont déjà partiellement couvert par le RGPD et le règlement UE 2017/745 en ce qui concerne les solutions d'IA qui sont des dispositifs médicaux (DM). Afin de produire des recommandations de bonnes pratiques qui soient « auto-suffisantes », le groupe de travail a néanmoins fait le choix d'intégrer ces critères partiellement couverts aux questionnements éthiques présentés dans ce guide. Une indication de l'article du RGPD auxquels ils sont rattachés est présente dès lors que cela s'avère nécessaire. Le schéma de la figure 2 vise à donner une vision synthétique et temporelle des principes proposés pour chacune des étapes de la méthodologie d'élaboration d'une solution d'IA.

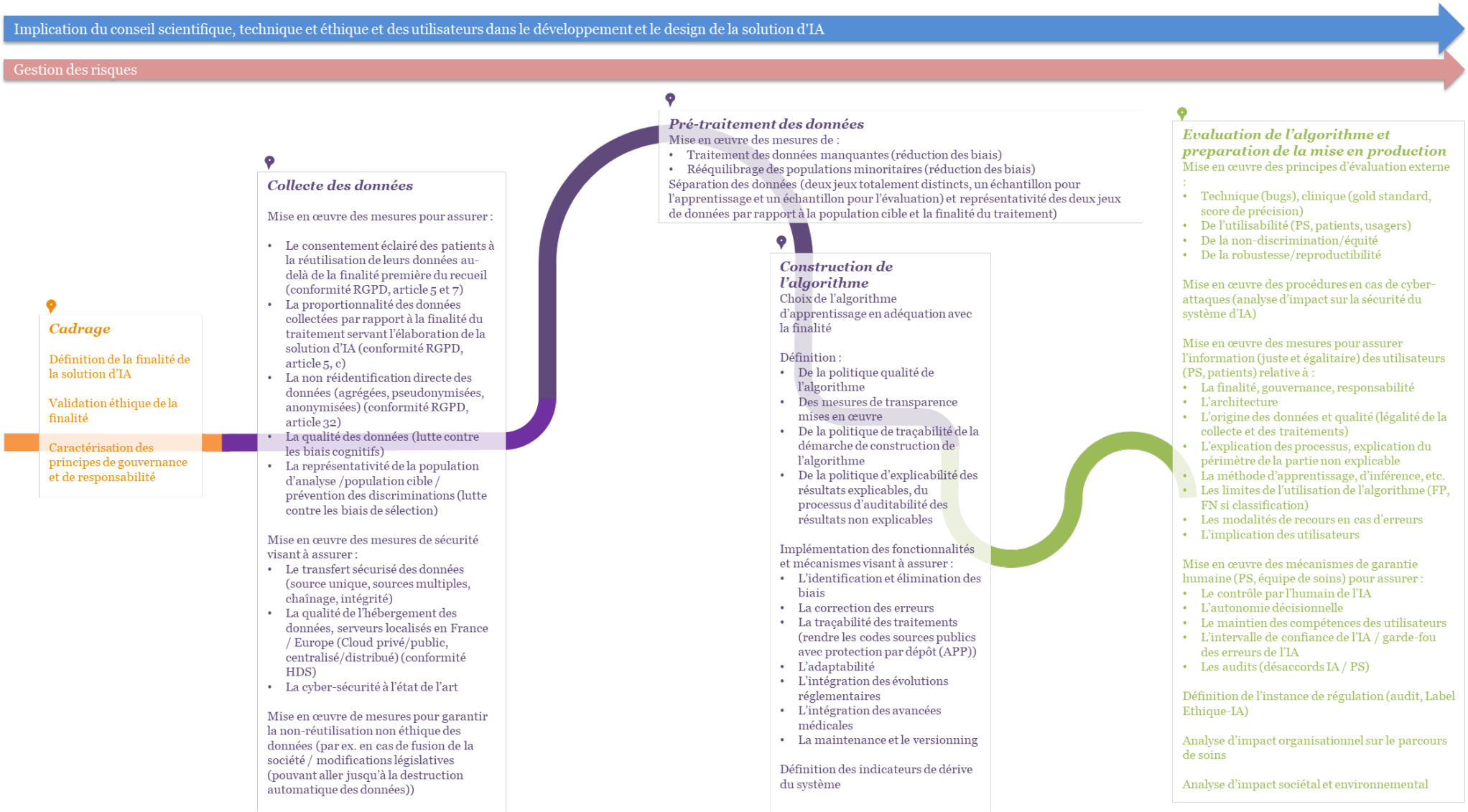


Figure 2 : vision synthétique des principes proposés pour chacune des étapes de la construction d'une solution d'IA en santé.

Etape de cadrage en amont de la collecte des données : Définir la finalité de la solution d'IA et valider l'éthique de la finalité

Cette étape préalable doit permettre de définir :

- La finalité de la solution d'IA. Pour ce faire, il est essentiel de présenter l'inscription du service d'IA à chaque étape du parcours de soins ;
- Ses utilisateurs cibles, tant professionnels de santé que patients ou grand public ;
- Le type d'apprentissage utilisé ;
- Les sujets pour lesquels elle sera utilisée (définition des critères d'inclusion et d'exclusion).

A titre d'exemple, la finalité d'une solution d'IA peut être d'apprendre à détecter des lésions mammaires à partir d'une mammographie, la solution d'IA étant destinée à être utilisée par les radiologues pour des femmes âgées de 50 à 74 ans sans signe clinique, dans le cadre du dépistage organisé du cancer du sein.

Etape 1 : Collecte de données

Principes défendus par le groupe de travail

Les algorithmes ayant recours à l'intelligence artificielle sont en mesure de déceler une fonction mathématique adéquate à la résolution d'un problème complexe, sans qu'il soit nécessaire de la coder au préalable. Pour mener à bien cette opération, l'algorithme doit avoir accès à une grande quantité de données pour répéter plusieurs fois les essais et optimiser son fonctionnement.
Le conseil national du numérique rappelle d'ailleurs que « *la valeur créée par l'intelligence artificielle provient des données nécessaires à l'apprentissage bien plus que de l'algorithme* ».[13] Dès lors, la collecte de donnée s'avère essentielle. Elle doit cependant s'accompagner d'une réflexion éthique pour éviter les écueils inhérents à l'intelligence artificielle.

Le développeur de la solution embarquant l'intelligence artificielle doit dans un premier temps s'assurer que les données utilisées pour la construction de l'algorithme ont été obtenues avec le consentement des patients à l'origine de leur production. Les données de santé répondent à des enjeux éthiques et moraux qu'il est important de préserver. En France, le cadre règlementaire relatif à la protection des données personnelles de santé est régi par le droit européen avec le règlement général sur la protection des données (RGPD), entré en vigueur le 25 mai 2018, et par le droit national français au travers de la loi n°78-17 du 6 janvier 1978 modifiée, dite loi « Informatique et libertés ». Etant considérées comme des données à caractère sensible, le traitement des données de santé relève d'une réglementation précise. Sauf exceptions, il est d'ailleurs interdit de traiter des données de santé et tout acteur qui souhaiterait procéder au traitement de données de santé doit justifier de l'une des exceptions prévues à l'article 9.2 du RGPD.

[13] Bonnet, « Le CNNum poursuivra ses travaux sur l'intelligence artificielle après la remise du rapport FranceIA ».

Pour exploiter des données personnelles de manière éthique, tout en respectant les droits et libertés des personnes, il peut être proposé aux éditeurs de logiciels embarquant de l'IA d'anonymiser ou pseudonymiser les données de santé colligées. Le processus d'anonymisation permet d'éliminer toute possibilité de réidentification du patient à partir de ses données et ainsi de s'affranchir de la législation relative à la protection des données, Le règlement général sur la protection des données ne comportant pas d'obligation générale d'anonymisation[14]. L'action de pseudonymisation consiste quant à elle à modifier les données sensibles du patient par un pseudonyme/un artefact (modifier le nom ou la date de naissance du patient). La pseudonymisation a l'avantage de ne pas masquer certaines corrélations, qui dans un contexte d'optimisation médicale peuvent être intéressantes. En ce sens, cette technique est souvent privilégiée à l'anonymisation complète lorsqu'il est question de recherche médicale. Néanmoins, la pseudonymisation expose à un risque de réidentification des patients à partir de leurs données. Il est nécessaire que les éditeurs de logiciels embarquant de l'IA actualisent et renforcent leurs méthodes en continu afin de garantir et préserver la sécurité des données des patients.

Les données de santé sont de plus en plus abondantes, il est parfois complexe pour les éditeurs de technologies recourant à l'IA d'avoir accès aux données de santé. Faciliter l'accès à ces données de santé aux industriels parait nécessaire pour encourager l'innovation mais doit s'accompagner d'un cadre et d'une réflexion éthique pour garantir la sécurité des citoyens. La France, à l'instar d'autres pays européens a décidé de se doter d'une plateforme d'hébergement des données de santé : « le Health Data Hub ».

Le Health Data Hub a vocation à mutualiser la quasi-totalité des données de santé auparavant détenues par les différents acteurs du système de santé, et à en garantir un accès sécurisé. La plateforme propose une interface sécurisée nommé « espace projet », permettant de recevoir, stocker et traiter l'ensemble des données nécessaires à l'optimisation d'un logiciel embarquant de l'IA. L'interface dispose d'une grande puissance de calcul, de logiciels performants de développement d'algorithmes, et garantit des mesures de sécurité pour réduire les risques liés à l'utilisation de données de santé. Ce service est d'ores et déjà mis à disposition des industriels et pourrait s'avérer utile pour les acteurs qui ne bénéficient pas encore d'infrastructures permettant de répondre aux exigences de sécurité liées à l'hébergement et au traitement des données de santé à caractère personnel.
Investir dans ce type de systèmes informatisés permet par ailleurs de se prémunir contre les risques de cyberattaques.

Il existe donc des outils gérés par la puissance publique notamment les entrepôts de données des établissements de santé ou des bases de données en open accès permettant aux industriels d'avoir accès à des données de santé leur permettant d'entraîner des algorithmes d'intelligence artificielle. La demande est effectuée par l'industriel qui sélectionne les données qui seront ensuite utilisées lors du processus d'optimisation.
En amont de cette demande, il est conseillé aux éditeurs de s'interroger sur la finalité de leur outil, de quantifier leur réel besoin et d'adapter leur collecte en fonction.

[14] https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles

En effet, il peut être judicieux d'évaluer au préalable le volume de données nécessaires à un apprentissage robuste afin d'adopter une certaine « sobriété numérique » et ainsi diminuer l'impact environnemental du numérique. Cette approche relève du principe de proportionnalité des données par rapport à la finalité du traitement (article 5, c du RGPD) en se limitant aux données nécessaires à l'apprentissage.
De plus, il est souhaitable de mettre en œuvre des mécanismes de destruction automatique des données pour éviter de conserver des données au-delà des besoins de la recherche dans le cadre de laquelle elles ont été recueillies et utilisées. Cela permet de garantir que ces données ne sont pas utilisées à des fins autres que celles initialement définies.

A titre d'exemple, le rachat par Google/Alphabet de DeepMind qui avait développé l'application de santé Streams (permettant de détecter les insuffisances rénales aiguës jusqu'à 48 heures avant qu'elles ne se manifestent) a permis à Google d'accéder aux données de santé de 1,6 millions de personnes sans qu'elles en aient préalablement donné leur accord[15].

Les recommandations sur l'étape 1

Il convient de se poser les questions suivantes :

- Les données servant à l'entraînement de l'algorithme ont-elles été obtenues auprès de tiers garantissant le consentement éclairé des patients qui les ont produites pour une réutilisation au-delà de la finalité première du recueil ?
- Les données servant à l'entraînement de l'algorithme ont-elles été obtenues via des modalités garantissant la sécurisation du transfert et l'intégrité des données transférées ?
- Les données ont-elles été pseudonymisées selon des modalités garantissant leur confidentialité ?
- Est-ce que l'ensemble des données collectées et utilisées pour l'entraînement de l'algorithme respecte le principe de proportionnalité (RGPD) et se réduit aux seules données nécessaires au traitement prévu compte tenu de la finalité de l'algorithme
- Est-ce que la qualité des données a été évaluée de façon à limiter le biais cognitif ? Est-ce qu'il existe par exemple une analyse de la fiabilité des données faisant la différence entre les données brutes faisant l'objet d'une mesure automatique (données issues des examens complémentaires, des capteurs, etc.) et les données ayant été interprétées (données issues de l'interrogatoire, de l'examen clinique, des questionnaires de vie réelle des patients, etc.) ? Est-ce que des mécanismes visant à réduire le poids du biais culturel ont été mis en œuvre ?
- Est-ce que des mécanismes ont été mis en œuvre pour limiter le biais de sélection et garantir que le même traitement s'applique de la même manière aux différentes sous-populations minoritaires de la population générale cible de l'utilisation de la solution d'IA ?

[15] https://techcrunch.com/2019/10/22/google-has-used-contract-swaps-to-get-bulk-access-terms-to-nhs-patient-data/?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAC3lzTcuXAamOPIwlJEGHGMB9PanEXSDrN2TWEuHCa85rSmVivfdTfKfguUVn3h4T6K-x-j8tXDHPUxmJ467Z9is6eBy20gfMArW40N_Eo0pexZ815PwqX6Ad_Rl30s3xibgzls7l-uYiax8CCqJBJcXeHOlp5vT2cjIyiaUPfhY

- Est-ce que des mécanismes permettant de sécuriser le stockage des données ont été mis en œuvre (hébergement certifié HDS, serveurs localisés en Europe, réflexion sur les *clouds* centralisés / distribués) ?
- Est-ce que des procédures à l'état de l'art ont été mises en œuvre pour garantir la protection des données contre les cyber-attaques ?
- Est-ce que des mécanismes ont été développés afin de garantir que les données ne seront pas réutilisées pour d'autres finalités que celle à l'origine du service produit par la solution d'IA, et ne seront pas réutilisées de façon non éthique dans le cas notamment d'une fusion / rachat de la société ayant développée la solution d'IA ? Ces mécanismes peuvent inclure le fait de contacter tous les patients producteurs des données afin de valider leur consentement éclairé à l'utilisation des leurs données pour ces nouvelles finalités. Ils peuvent également aller jusqu'à la possibilité d'une destruction immédiate des données.

Etape 2 : Pré-traitement (préparation) des données

Principes défendus par le groupe de travail

Une fois les données collectées, les éditeurs de technologies embarquant de l'IA doivent assurer une conception efficiente des jeux de données obtenus afin de s'affranchir des risques éthiques et d'efficacité qui en découlent.
En effet, l'utilisation de données faussées peut avoir des conséquences graves, notamment pour des technologies qui se destinent au secteur médical : mauvais diagnostic de cancer, erreur médicamenteuse iatrogène etc...

L'objectif est donc de mettre en forme les données en vue de l'étape de construction de l'algorithme. La première étape consiste à mener des traitements permettant de qualifier les données, et notamment de proposer une solution au problème des données manquantes (ces données manquantes pouvant être l'expression d'un biais qu'il soit cognitif ou de sélection). Il peut également s'agir de détecter les données aberrantes et les éliminer pour éviter qu'elles ne faussent les résultats des algorithmes. Le risque d'avoir des données manquantes et/ou aberrantes est majoré lorsque les données ont été recueillies avec un objectif initial autre que celui fixé par l'éditeur de la solution, rendant leur exploitation plus complexe. Ces étapes de prétraitement doivent donc être renseignées explicitement pour assurer un traçage et être en capacité d'améliorer les pratiques à termes.

Le risque de biais dans la construction d'un algorithme IA est majeur et doit être anticipé dès la phase de prétraitement des données. Il ne s'agit pas uniquement de déceler les données manquantes ou aberrantes mais également de veiller à la qualité des jeux de données utilisés. Depuis plusieurs mois, de nombreux articles critiquent ce qu'il est convenu de nommer le biais algorithmique pouvant conduire les solutions d'IA à des comportements discriminatoires. Parce que l'apprentissage automatique s'alimente de données produites par des êtres humains, l'IA ne ferait ainsi que reproduire et accentuer les préjugés sexistes et racistes présents chez les êtres humains. Les difficultés sont liées aux conditions d'apprentissage des

décisions, les données biaisées sont utilisées pour entraîner les algorithmes qui deviennent, à leur tour, biaisés. Ainsi, le Massachusetts Institute of Technology (MIT) est récemment tombé dans le piège d'une IA devenue raciste[16]. L'établissement a annoncé avoir mis hors ligne un jeu de données sur lequel une IA mal entraînée avait annoté un immense volume d'images par des insultes racistes et des termes péjoratifs envers les femmes. Le problème n'est pas seulement que les algorithmes reflètent pour la plupart les préjugés de notre monde, mais qu'ils le font à grande échelle et sans surveillance appropriée.

On distingue en effet de nombreux types de biais. Le premier, plus ou moins facilement identifiable, est lié à la non-représentativité du jeu de données. C'est le biais de sélection (ou encore biais statistique ou biais d'échantillonnage). C'est le biais dont souffrent par exemple les algorithmes de reconnaissance faciale entraînés sur des ensembles de données contenant plus de visages de personnes à peau claire que de visages de personnes à peau noire conduisant, comme l'a montré l'étude de Buolamwini du MIT, le système d'IA à mieux reconnaître le visage des hommes blancs que celui des femmes noires[17]. Un jeu de données peut effectivement correspondre à une population cible mais avoir des sur-représentations ou sous-représentations de certaines populations. L'apprentissage mené avec un tel jeu de données peut conduire à un résultat faussé et avoir de graves répercussions sur la santé des patients sur lesquels l'IA est utilisée. Par exemple, un apprentissage réalisé dans le cadre du dépistage organisée du cancer du sein effectuée à partir de la mammographie qui reposerait uniquement sur des données de femmes blanches sans surpoids aurait plus de difficulté à être efficient pour les femmes d'une autre ethnie et/ou avec des caractéristiques morphologiques différentes. Pour tenter d'éviter ces biais, il est nécessaire d'entraîner les algorithmes sur des clichés représentatifs de différentes couleurs de peau. Récemment, une équipe de chercheurs du MIT a mis au point une IA capable de détecter un cancer du sein jusqu'à cinq ans avant sa formation, l'IA étant capable de repérer les signes avant-coureurs d'un cancer aussi bien sur des personnes blanches que non blanches.[18]

Afin d'éviter la reproduction des disparités raciales, sociales ou de genre [19]et de mettre à mal toute considération éthique, il est primordial que les concepteurs de logiciels accordent une vigilance particulière à la construction de l'algorithme, et ce, dès l'étape de préparation des données.

Dès lors, il convient de s'interroger quant à la mise en place d'un comité consultatif regroupant les acteurs du système de santé et les développeurs des technologies embarquant l'IA. Il permettrait aux concepteurs, qui sont bien souvent extérieurs au monde de la santé, de se familiariser aux spécificités du secteur et d'adopter une dimension éthique dès la phase de préparation des données.

Ce comité peut notamment contribuer à normaliser les procédures de traitement de données quand il est question d'optimisation de logiciels destinés au secteur de la santé, et en assurer la reproductibilité par les éditeurs de logiciels.

[16] https://www.lemondeinformatique.fr/actualites/lire-le-mit-coupe-le-dataset-entrainant-une-ia-devenue-raciste-et-offensante-79628.html
[17] http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf
[18] https://pubs.rsna.org/doi/full/10.1148/radiol.2019182716
[19] https://www.arte.tv/fr/videos/102997-000-A/sante-les-femmes-sont-elles-discriminees/

Les recommandations sur l'étape 2

Il convient de se poser les questions suivantes :

- Est-ce que des mécanismes sont mis en œuvre pour traiter les données manquantes afin de limiter les biais (moyenne des données disponibles, modélisation) ?
- Est-ce qu'une stratégie de construction des populations d'apprentissage et d'évaluation (deux populations distinctes) permettant de garantir la représentativité de chacune d'elles par rapport à la population cible a été mise en œuvre (tirage au sort, stratification sur les groupes minoritaires, rééquilibrage par sur et /ou sous-échantillonnage, etc.) ?
- Est-ce qu'un comité rassemblant les designers, les *data scientists* et les utilisateurs finaux (professionnels de santé, patients, grand public) a été constitué afin d'engager les concepteurs et les utilisateurs de la solution d'IA dans un processus de co-construction ?

Etape 3 : Construction de l'algorithme

Principes défendus par le groupe de travail

Au moment de la construction de l'algorithme, il est essentiel de consolider les choix du type d'apprentissage automatique et de décider de l'algorithme à utiliser. En fonction du problème à résoudre mais aussi des données à disposition, le choix va se porter sur plusieurs types d'apprentissage (présentés supra), apprentissage supervisé, non supervisé ou par renforcement.

Outre le type d'apprentissage, il va également être question du choix de l'algorithme, selon qu'il relève des modèles d'apprentissage automatique (régression, arbres de décision, forêts aléatoires, gradient boosting, etc) ou des modèles d'apprentissage profond (réseaux de neurones). Le choix et le développement de l'algorithme sont des éléments importants permettant de perfectionner le système.

Malgré leurs importances considérables, les algorithmes n'ont pas encore de véritable cadre juridique en mesure de les protéger. En effet, un algorithme est considéré comme une suite d'opérations mathématiques et relève du cadre des « idées ». Il n'est donc pas protégé par le droit d'auteur, et ce même lorsqu'il est directement intégré à un logiciel.
De la même manière, les méthodes mathématiques » et les « programmes d'ordinateur » sont exclues du champ des propriétés brevetables, empêchant ainsi les algorithmes d'être protégés par le droit des brevets[20].

[20] ASTIER, « Quelle protection juridique pour vos algorithmes ? »

Dans un tel contexte, marqué notamment par une forte compétitivité entre les acteurs, il est dès lors compréhensible que les industriels puissent être réticents à divulguer le fonctionnement de leurs algorithmes innovants.

Pourtant, l'enjeu autour de la transparence du fonctionnement des algorithmes est de taille. Exiger davantage de transparence aux industriels apparait essentiel pour s'assurer que les algorithmes ne soient pas susceptibles de porter atteinte à leurs utilisateurs et respectent des considérations éthiques. Cette transparence repose en partie sur le traçage des étapes de construction et sur l'explicabilité des procédures mises en place.

L'algorithme embarquant de l'IA est souvent décrit comme une boite noire, peu lisible et prenant des décisions qui ne sont pas toujours explicables. Pour pallier les difficultés d'explicabilité de ces algorithmes, des critères précis issus de l'auditabilité ou encore des indicateurs de détection des dérives et d'adaptabilité (en cas d'avancées médicales remettant en question l'apprentissage, ou de modifications réglementaires) pourront être proposés.
Cet effort de transparence devra être effectué tout au long du cycle de vie du produit et porter également sur les versions incrémentales ultérieures. Il est attendu que les industriels soient en mesure de fournir davantage de documentation relative aux changements opérés entre les différentes versions et les raisons de ces changements.

Pour intégrer une dimension éthique dès la conception de l'algorithme tout en garantissant sa sécurité, la cellule éthique du numérique en santé propose également que les codes sources de l'algorithme soient publiés par l'industriel et que leurs protections soient assurées par l'agence pour la protection des programmes (APP), organisme européen chargé de la protection des logiciels et applications mobiles.

Les recommandations sur l'étape 3

Il convient de se poser les questions suivantes :

- Existe-t-il une méthode garantissant que l'algorithme d'apprentissage choisi permette d'obtenir les meilleurs résultats par rapport à la finalité de la solution d'IA ?
- Est-ce que des mécanismes ont été mis en œuvre pour corriger les erreurs de l'algorithme (sur-apprentissage, sous-apprentissage) et les biais ?
- Est-ce que des solutions permettant d'établir la traçabilité de la démarche de construction de l'algorithme ont été mises en œuvre ?
- Des traitements dans un souci de transparence ont-ils été implémentés ?

Etape 4 : Evaluation de l'algorithme

Principes défendus par le groupe de travail

L'évaluation représente une étape cruciale dans le développement d'un produit qui se destine au secteur médical, et ce d'autant plus lorsqu'il s'agit d'outils embarquant de l'IA.
Les évaluateurs auront la responsabilité de procéder aux tests permettant de s'assurer que le comportement de la technologie est en adéquation avec les enjeux sécuritaires et éthiques liés à l'intelligence artificielle.
Pourtant, le groupe de travail commandité par le Comité consultatif national d'éthique (CCNE) concluait le 19 novembre 2018 que « *les travaux de certification et de normalisation sur l'intelligence artificielle et la robotisation en santé, en dépit de leur intérêt, en restent, en l'état, à un stade très parcellaire* ».[21]

Il convient de rappeler que le thème relatif aux solutions d'IA "en apprentissage continu" (également connu sous le nom de systèmes d'IA "en adaptation constante" ou auto-apprenants), bien que longuement discuté au sein du GT, a finalement été exclu du périmètre de ce guide de bonnes pratiques.
Sont étudiées dans ce document uniquement les solutions, qui en l'absence de caractère évolutif en continu, ne nécessitent pas de mettre en place des mécanismes d'évaluation à plusieurs reprises au cours du cycle de vie de la technologie. Ce domaine devra néanmoins être suivi et réexaminé lors des prochaines itérations du guide de bonnes pratiques pour s'assurer que les technologies embarquant de l'IA en apprentissage continu conservent une efficience dans le temps.

Compte tenu des spécificités de l'IA (difficulté d'évaluation en situation réelle, existence de biais, boite noire), les processus d'évaluation diffèrent des processus d'évaluation classiques. Les éditeurs qui souhaitent évaluer leur technologie doivent être en mesure de décrire les résultats de toute analyse des erreurs de performance et, le cas échéant, expliciter comment les erreurs ont été identifiées. Ces informations et leurs implications doivent être communiquées pour anticiper et éviter les risques liés à leur utilisation en situation réelle.

La technologie embarquant de l'IA en santé doit également garantir l'autonomie décisionnelle du praticien. Tout effet de l'interaction entre l'homme et la solution d'IA sur les résultats doit être décrit en détail, y compris le niveau d'expertise requis pour comprendre les résultats et toute formation et/ou instruction qui devrait être fournie à cette fin.

Par exemple, un système de détection du cancer de la peau qui a produit un pourcentage de probabilité de cancer comme résultat doit être accompagné d'une explication de la manière dont ce résultat doit être interprété et utilisé par l'utilisateur, en précisant à la fois les voies prévues (par exemple, l'excision d'une lésion cutanée si le diagnostic est positif) et les seuils d'entrée dans ces voies (par exemple, l'excision d'une lésion cutanée si le diagnostic est positif et la probabilité est > 80 %). Les informations produites par les interventions de comparaison doivent être décrites de manière similaire, avec une explication de la manière dont ces

[21] « rapportNumériqueEtSanté-2018-16.11-version-10h20-1.pdf ».

informations ont été utilisées pour arriver à des décisions cliniques sur la prise en charge des patients, le cas échéant. Toute divergence entre la manière dont la décision a été prise et la manière dont elle devait l'être (c'est-à-dire comme spécifié dans le protocole d'essai) doit être signalée.

Les recommandations sur l'étape 4

Il convient de se poser les questions suivantes :

- Une évaluation de l'algorithme a-t-elle été réalisée sur les parties suivantes :
    - Technique (bugs), clinique (gold standard, score de précision)
    - Utilisabilité (les utilisateurs cibles de l'outil, qu'ils soient professionnels de santé, patients, ou grand public, sont-ils en mesure d'utiliser la solution d'IA tel que l'usage en a été prévu au moment de son développement ? Comment cela a-t-il été évalué ?)
    - Non-discrimination/équité (des tests de performance de l'outil ont-ils été réalisés sur les populations minoritaires ?)
    - Robustesse/reproductibilité
- Quelle est la procédure en cas de cyber-attaque ?
- Quel est le processus d'information des utilisateurs ? Est-il juste et égalitaire ? Des mécanismes ont-ils été mis en œuvre afin de s'assurer de la bonne compréhension de l'utilisateur (au-delà de la seule information) ? L'utilisateur a-t-il accès aux informations suivantes :
    - Finalité de la solution d'IA, gouvernance, responsabilité
    - Architecture
    - Origine des données et qualité (légalité de la réutilisation des données et des traitements)
    - Explication des processus, explication du périmètre de ce qui est explicable et de ce qui ne l'est pas, description des principes de la partie non explicable
    - Méthode d'apprentissage, d'inférence, etc.
    - Définition des limites de l'utilisation de l'algorithme (faux positifs, faux négatifs, et précision, rappel et score F1 dans le cadre d'une classification binaire)
    - Description des recours en cas d'erreurs
    - Implication des utilisateurs
- Quelles sont les garanties humaines ?
    - Contrôle par l'humain de l'IA
    - Autonomie décisionnelle
    - Maintien des compétences des utilisateurs
    - Intervalle de confiance de l'IA / garde-fou des erreurs de l'IA
    - Audits (désaccords IA / PS)
- Une instance de régulation est-elle en place ? Audit, Label Ethique-IA ?
- Quel est l'impact organisationnel sur le parcours de soins ?
- Quels sont les impacts sociétaux, les impacts sur les personnes, les impacts sur l'environnement ?

En synthèse, et sur l'ensemble des 4 étapes, le GT3 recommande de porter une attention particulière aux critères suivants :

- Recueil du consentement éclairé (RGPD)
- Mesures de protection de la vie privée
- Mesures de non ré-identification
- Collecte de données proportionnée au but
- Réduction des biais cognitifs : qualité des données
- Réduction des biais de sélection : représentativité des données
- Mesures pour assurer la séparation des données
- Mesures de traitement des données manquantes
- Mesures d'identification et élimination des biais dans l'algorithme
- Prise en compte du risque de cyber-attaques
- Mesures de traçabilité
- Mesures d'explicabilité
- Transparence de l'algorithme, du modèle
- Voies de recours / réparation
- Critères d'utilisabilité
- Mesures de non-discrimination/équité
- Mesures de reproductibilité
- Mesures pour l'audit
- Critère de label éthique IA
- Mesures pour le contrôle humain
- Résilience du système
- Analyse d'impact sur le parcours de soins
- Analyse d'impact environnemental
- Analyse d'impact sociétal
- Choix de l'algorithme d'apprentissage en adéquation avec la finalité
- Participation des parties prenantes
- Conduite en cas d'erreur et responsabilité
- Amélioration de la qualité et surveillance
- Adaptation aux évolutions réglementaires et médicales
- Analyse des risques

6. Articulation des recommandations du GT3 avec les recommandations des autres institutions ayant instruit cette question

Les recommandations du groupe d'experts de la Commission Européenne[22]

La commission européenne souhaite pouvoir impulser une « IA éthique, sûre et de pointe réalisée en Europe ». Pour ce faire, elle a mis sur pied un groupe d'experts de haut niveau chargés de définir les lignes directrices d'une IA, digne de confiance, qui se doit d'être :

- Licite, respectueuse des législations et réglementations applicables ;
- Ethique, assurant l'adhésion à des valeurs et principes éthiques ;
- Robuste, sur le plan technique et social (pour éviter tout préjudice involontaire).

Le groupe d'experts fonde une IA digne de confiance sur 4 principes éthiques :

- Le respect de l'autonomie humaine
- La prévention de toute atteinte
- L'équité
- L'explicabilité

Le groupe met par ailleurs en avant deux recommandations essentielles :

- *Accorder une attention particulière aux situations concernant des groupes plus vulnérables tels que les enfants, les personnes handicapées et d'autres groupes historiquement défavorisés, exposés au risque d'exclusion, et/ou aux situations caractérisées par des asymétries de pouvoir ou d'information, par exemple entre les employeurs et les travailleurs, ou entre les entreprises et les consommateurs.*
- *Reconnaître et être conscient que certaines applications d'IA sont certes susceptibles d'apporter des avantages considérables aux individus et à la société, mais qu'elles peuvent également avoir des incidences négatives, y compris des incidences pouvant s'avérer difficiles à anticiper, reconnaître ou mesurer (par exemple, en matière de démocratie, d'état de droit et de justice distributive, ou sur l'esprit humain lui-même). Adopter des mesures appropriées pour atténuer ces risques le cas échéant, de manière proportionnée à l'ampleur du risque.*

Le groupe d'experts déduit de ces principes, sept exigences essentielles (déclinées dans un second temps en liste d'évaluations pour une IA digne de confiance). Il s'agit d'exigences (techniques et non techniques) non spécifiquement centrées sur les applications de santé de l'IA : nécessité d'une action humaine et d'un contrôle humain, robustesse technique et sécurité des outils, respect de la vie privée et gouvernance des données, transparence des traitements, prise en compte de la diversité, non-discrimination et équité, bien-être sociétal et environnemental et responsabilité.

[22] https://op.europa.eu/fr/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1

Au niveau européen, les experts ont souhaité rappeler qu'un certain nombre d'acteurs ont un rôle essentiel à tenir dans le cadre d'un projet de mise en œuvre de l'IA à savoir, les développeurs, les prestataires et les utilisateurs finaux.

Les experts de la commission européenne ont adopté une démarche sensiblement différente de celle du GT3 animé par la Cellule Ethique du numérique en santé. Les experts ont ainsi pris le parti de décliner les sept exigences en critères d'évaluation d'une solution numérique intégrant de l'IA. Ces critères sont détaillés en annexes.

- Nécessité d'une action humaine et d'un contrôle humain
    - Droits fondamentaux
    - Action humaine
    - Contrôle humain
- Robustesse technique et sécurité des outils
    - Résilience aux attaques et sécurité
    - Solutions de secours et sécurité générale
    - Précision
    - Fiabilité et reproductibilité
- Respect de la vie privée et gouvernance des données
    - Respect de la vie privée et protection des données
    - Qualité et intégrité des données
    - Accès aux données
- Transparence des traitements
    - Traçabilité
    - Explicabilité
    - Communication
- Prise en compte de la diversité, non-discrimination et équité
    - Éviter les biais injustes
    - Accessibilité et conception universelle
    - Participation des parties prenantes
- Bien-être sociétal et environnemental
    - IA durable et respectueuse de l'environnement
    - Incidence sociale
    - Société et démocratie
- Responsabilité
    - Auditabilité
    - Minimisation et documentation des incidences négatives
    - Documentation des arbitrages
    - Voies de recours

Il est à noter que la Commission Européenne a proposé un outil « ALTAI »[23] permettant aux éditeurs de solutions d'IA d'évaluer si leur solution implémente une IA de confiance en accord avec les différents critères précédents. L'outil se présente sous la forme d'une plateforme où l'on répond à différentes questions permettant d'évaluer une solution d'IA. Le résultat apparaît sous la forme d'un radar quantifiant l'IA selon les sept exigences.

[23] https://altai.insight-centre.org/ puis s'enregister (« Register ») ou se connecter (« Login »)

En synthèse, le groupe d'experts de haut niveau de la commission européenne préconise de s'appuyer sur les critères suivants :

- Recueil du consentement éclairé (RGPD)
- Mesures de protection de la vie privée
- Mesures de non ré-identification
- Collecte de données proportionnée au but
- Réduction des biais cognitifs : qualité des données
- Réduction des biais de sélection : représentativité des données
- Mesures relatives à l'accessibilité des données
- Prise en compte du risque de cyber-attaques
- Résilience du système
- Mesures de traçabilité
- Mesures d'explicabilité
- Voies de recours / réparation
- Mesures de non-discrimination/équité
- Mesures de reproductibilité
- Mesures pour l'audit
- Mesures pour le contrôle humain
- Analyse d'impact sur les droits fondamentaux
- Analyse d'impact environnemental
- Analyse d'impact sociétal
- Participation des parties prenantes
- Conduite en cas d'erreur et responsabilité

Les recommandations de l'Organisation Mondiale de la Santé[24]

L'OMS met en avant cinq critères à respecter, basés sur les prérequis suivants :

- Eviter de « blesser » (non-malfaisance)
- Promouvoir le bien-être (bienfaisance)
- S'assurer que les personnes sont traitées avec équité et justice (non-discrimination, négligence d'un groupe, manipulation, abus ou domination)
- Respecter l'intérêt des personnes (autonomie)

L'OMS promeut les critères suivants :

- Critère de respect de l'autonomie :
  - Il s'agit de faire en sorte que tout transfert de prise de décision au service ne vienne in fine, entraver l'autonomie des êtres humains concernés par le service numérique. Tout service numérique de santé faisant appel à l'IA doit ainsi permettre aux personnes de rester maîtres de leurs décisions médicales, notamment en leur donnant l'information la plus éclairée possible. Ce principe implique selon l'OMS, la capacité pour un fournisseur de services, à restreindre l'autonomie de son système par une action humaine.
  - Pour respecter ce principe, tout système doit par ailleurs mettre en œuvre les mesures nécessaires permettant de protéger la vie privée et la confidentialité des données des personnes ciblées par le service. Les systèmes embarquant l'IA doivent pour ce faire, intégrer les règles du cadre légal de la protection de données à caractère personnel en vigueur sur le territoire (RGPD sur le territoire Européen). L'OMS s'inspire par ailleurs des règles contenues dans le RGPD et défendues par le groupe de travail français, en recommandant une vigilance particulière concernant le « *behavioural data surplus* » (surplus de données comportementales comparable au principe de proportionnalité des données au regard de la finalité dans la Loi Informatique et Libertés, et au principe de minimisation des données de l'article 5-c du RGPD) de manière à éviter certains biais dus aux données comportementales, et en recommandant par conséquent d'indiquer précisément le périmètre et la finalité d'utilisation des données prévues dans le dispositif. Le respect de ces règles doit permettre d'obtenir les consentements éclairés des utilisateurs cibles du service et permettre à ces mêmes utilisateurs de rester maîtres de leurs données, de leurs parcours de santé.
- Critère d'adéquation aux principes de bien-être humain, de sécurité et d'intérêt public :
  - Les services numériques embarquant l'IA doivent intégrer des mesures de contrôle et de vérification de la cohérence, ainsi que des mesures d'amélioration continue de la qualité. Pour ce faire, un dispositif de surveillance continue intégrant les responsables du service, les développeurs et les utilisateurs doit être mis en place afin de mesurer et contrôler en continue les performances de l'algorithme utilisé. Cette règle doit permettre d'éviter tout préjudice collectif ou individuel sur des patients.

[24] https://apps.who.int/iris/rest/bitstreams/1352854/retrieve

    - Dans un souci de conservation de l'universalité du soin, il est essentiel de s'assurer que l'utilisation d'un service numérique de santé embarquant l'IA ne vienne pas empêcher une partie de la population d'accéder à une prise en charge médicale. Des outils doivent ainsi être implémentés pour éviter toute discrimination ou stigmatisation par le service numérique.
- Critère de transparence, d'explicabilité, d'intelligibilité :
    - Le service numérique embarquant l'IA doit être explicité et documenté de manière compréhensible par l'ensemble de la population. L'information relative au service et à l'algorithme doit être diffusée et actualisée en continu.
    - Le service numérique doit pouvoir être régulièrement audité. Pour ce faire, il convient d'intégrer le dispositif technique et organisationnel nécessaire.
    - Le service numérique doit enfin être testé en conditions réelles de manière à vérifier sa capacité à répondre aux standards de sécurité en vigueur. Les tests et évaluations doivent être régulièrement effectués, de manière transparente, pour s'assurer qu'il n'existe aucune atteinte aux droits humains.
- Critère de responsabilité :
    - Le fournisseur d'un service numérique embarquant l'IA doit mettre en œuvre les règles techniques permettant la vérification « humaine » de l'outil et de l'algorithme. Les mécanismes visant à interroger l'algorithme doivent être implémentés. En cas de dysfonctionnements, des mesures de réparation doivent être prévues.
    - La responsabilité individuelle ou collective doit être clairement définie de manière à éviter la problématique de « le problème de tous n'est de la responsabilité de personne »
- Critère d'inclusion et d'équité :
    - Les données utilisées doivent être larges, équitables et l'accès, sans discriminations et sans biais. Des dispositifs d'identification et de correction des biais doivent être implémentés dans les algorithmes.
    - Les technologies utilisées doivent être, autant que faire se peut, largement diffusées, de manière transparente.
- Critère d'accessibilité et de soutenabilité :
    - Tout service numérique embarquant l'IA doit intégrer des règles permettant de s'adapter en permanence aux attentes et aux problématiques des utilisateurs
    - Les systèmes doivent par ailleurs s'engager dans une démarche de sobriété numérique et intégrer des règles de réduction de l'empreinte environnementale. Une attention à l'impact sociétal, notamment sur l'emploi et les potentielles suppressions de postes associées au déploiement à grande échelle d'un service numérique embarquant l'IA doit être portée.

En synthèse, l'OMS encourage à s'appuyer sur les critères suivants :

- Recueil du consentement éclairé (RGPD)
- Mesures de protection de la vie privée
- Mesures de non ré-identification
- Collecte de données proportionnée au but
- Réduction des biais cognitifs : qualité des données
- Réduction des biais de sélection : représentativité des données
- Mesures d'explicabilité

- Transparence de l’algorithme, du modèle
- Voies de recours / réparation
- Mesures de non-discrimination/équité
- Mesures pour l'audit
- Mesures pour le contrôle humain
- Analyse d’impact sur les droits fondamentaux
- Analyse d’impact environnemental
- Analyse d’impact sociétal
- Participation des parties prenantes
- Conduite en cas d’erreur et responsabilité
- Amélioration de la qualité et surveillance
- Adaptation aux évolutions réglementaires et médicales

Les principes d'évaluation de la Haute Autorité de Santé[25]

Dans le cadre des travaux sur les dispositifs médicaux, la HAS a progressivement intégré dans ses réflexions l'arrivée de l'IA et intègre dorénavant les dispositifs médicaux (DM) embarquant de l'IA.

Cette intégration de la question de l'IA aux instructions de la HAS intervient en réponse aux études de la CNIL mettant en lumière les craintes des français liées à l'essor de l'intelligence artificielle et des algorithmes dans la vie quotidienne et notamment dans le secteur de la santé et de la protection sociale. Ces craintes relatives à la sécurité, à l'accumulation de données personnelles sur « les choix, les goûts et les comportements de chacun » sont particulièrement prégnantes dans le secteur de la santé et de la protection sociale, secteur qui nécessite une protection plus élevée des données du fait de leur grande sensibilité et de leur capacité à révéler, selon la HAS, « des aspects intimes de notre vie privée, des fragilités qui pourraient être exploitées à notre désavantage ».

Pour cette raison, la HAS souhaite concourir au renforcement de la confiance dans le numérique en prônant un « usage raisonné et raisonnable » du numérique et de l'intelligence artificielle et en participant à son encadrement juridique et éthique.

La HAS concourt à cet encadrement à travers son action dans l'évaluation des dispositifs médicaux en vue de leur inscription sur la liste des produits et prestations remboursables dont sa commission spécialisée, la Commission nationale d'évaluation des dispositifs médicaux et des technologies de santé est en charge.
La HAS, à travers les critères contenus dans le dossier de demande d'inscription d'un dispositif médical à la liste des produits et prestations remboursables, permet d'évaluer les DM embarquant l'IA qui auraient vocation à être pris en charge par la sécurité sociale.
Le dossier demandé par la HAS doit permettre de :

- Démontrer le bénéfice clinique du produit et son éventuelle plus-value par rapport à l'arsenal disponible au travers des critères règlementaires (Service Attendu et Amélioration du Service Attendu)
- D'apporter des informations requises pour l'évaluation d'un acte professionnel nécessaire à son utilisation.

Aussi, dès lors qu'un dispositif médical s'appuie sur au moins un procédé d'apprentissage à partir de données, la grille d'évaluation standardisée doit être renseignée par l'industriel pour apporter aux membres de la Commission nationale d'évaluation des dispositifs médicaux et des technologies de santé les informations qui leur sont nécessaires pour évaluer le dispositif.

La grille s'appuie sur les principes suivants :

- Autonomie : accroître le contrôle des usagers sur leur vie et leur environnement
- Bienfaisance : action bénéfique et utile, et délivre une information claire et compréhensible
- Non-malfaisance : éviter la souffrance et le préjudice pour l'usager

[25] https://www.has-sante.fr/upload/docs/application/pdf/2016-01/guide_fabricant_2016_01_11_cnedimts_vd.pdf#page=51

- Justice : ni biais, ni discriminations, ni formes d'exclusion.

La grille d'évaluation matérialise ces principes à travers un certain nombre de questions évaluatives concernant :

- Finalités d'usage :
  - Présentation de l'usage et domaine d'application,
  - Description des caractéristiques populationnelles,
  - Précision sur l'intérêt des informations fournies
  - Description de l'environnement de fonctionnement
- Description des échantillons utilisés pour l'apprentissage initial ou le réapprentissage du modèle et description des données d'entrée impliquées (mode d'acquisition, pré-traitements, déploiement) :
  - Description des caractéristiques de la population ou échantillon dont les données sont utilisées
  - Présentation de la méthodologie de segmentation des échantillons
  - Description des pré-traitements et des procédures de gestion des données aberrantes
- Description du modèle :
  - Description du type d'apprentissage
  - Description de l'algorithme
  - Description des phases, de l'entrainement, de la validation et du test, avant et après le déploiement du DM
  - Description des stratégies d'entraînement
  - Description du processus d'intervention humaine pour contrôle et réapprentissage
- Caractéristiques fonctionnelles
  - Description des mesures de performance
  - Description des risques
  - Robustesse du système (surveillance du système, seuils, mesures en cas de dérives…)
  - Résilience du système (gestion des anomalies, impacts cliniques)
  - Eléments d'explicabilité du système
  - Eléments d'interprétabilité
  - Confrontation aux recommandations professionnelles

Il appartient à l'entreprise qui sollicite le remboursement de décrire le ou les systèmes experts apprenants intégrés dans le DM soumis à l''évaluation de la CNEDiMTS qui en analysera les caractéristiques.

En synthèse, les critères étudiés par la HAS sont les suivants :

- Collecte de données proportionnée au but
- Réduction des biais cognitifs : qualité des données
- Réduction des biais de sélection : représentativité des données
- Mesure pour assurer la séparation des données
- Mesures de traitement des données manquantes
- Mesures d'identification et élimination des biais dans l'algorithme

- Résilience du système
- Mesures de traçabilité
- Mesures d'explicabilité
- Analyse d’impact sur le parcours de soins
- Transparence de l’algorithme, du modèle
- Voies de recours / réparation
- Mesures de non-discrimination/équité
- Mesures de reproductibilité
- Mesures pour l'audit
- Mesures pour le contrôle humain
- Participation des parties prenantes
- Choix de l’algorithme d’apprentissage en adéquation avec la finalité
- Conduite en cas d’erreur et responsabilité
- Amélioration de la qualité et surveillance
- Adaptation aux évolutions réglementaires et médicales.

Les principes défendus par le consortium composé de la *Food and Drug Administration* (FDA – Etats-Unis), Health Canada et *Medicines and Healthcare products Regulatory Agency* (MHRA – Royaume-Uni)

Les ministères des trois pays anglo-saxons ont conjointement défini 10 principes directeurs visant à promouvoir des dispositifs médicaux recourant à l'intelligence artificielle et l'apprentissage automatique, sécurisés, efficaces et de haute qualité.
Ces acteurs défendent l'idée qu'un encadrement de l'IA avec ces principes peut permettre in fine :

- D'adopter et adapter des bonnes pratiques probantes issues d'autres secteurs, à la technologie médicale et aux soins ;
- De créer de nouvelles pratiques spécifiques à la technologie médicale et au secteur de la santé et donner aux parties prenantes les moyens de faire progresser les innovations responsables

Les 10 principes directeurs proposés sont les suivants :

- Mobiliser une expertise pluridisciplinaire tout au long du cycle de vie du projet : Ce principe devant permettre le comprendre l'intégration du modèle dans le flux de travail clinique, de disposer d'une vision des avantages souhaités et des risques associés pour le patient, et in fine de s'assurer tout au long de la vie du produit, de sa capacité à répondre des besoins cliniques significatifs.
- Implémenter les bonnes pratiques d'ingénierie logicielle et de sécurité : ce principe directeur implique par conséquent que les « fondamentaux de l'ingénierie logicielle » soit intégrés notamment en termes de gestion des données (intégrité et authenticité de la donnée et assurance qualité associée) et politique de cybersécurité. Il est par ailleurs recommandé de mener en continu, une démarche de gestion des risques et de mettre en place un dispositif assurant la transparence nécessaire sur les décisions prises, sur l'architecture, sur les choix technologiques etc…
- Assurer la représentativité des données : le système devra donner l'assurance de la représentativité des données utilisées notamment à des fins de généralisation, s'assurer de leur pertinence (en termes d'âge, genre, sexe, origine ethnique) et de biais dans l'analyse
- Assurer l'indépendance des données d'entrainement des données de test
- Recourir aux meilleures méthodes disponibles pour établir les ensembles de données de référence : s'assurer que les données sont cliniquement pertinentes et bien caractérisées pour promouvoir et démontrer la robustesse et la généralisation du modèle
- Assurer la conception du modèle en adéquation avec les données disponibles et les finalités d'utilisation. Ce principe sous-tend notamment l'atténuation active des risques connus, tels que le sur-apprentissage, la dégradation des performances et la sécurité
- Assurer le contrôle humain : ce principe d'un « humain dans la boucle » est clé pour ne pas laisser le modèle apprenant, isolé.
- Mettre en œuvre les tests dans des conditions cliniques pertinentes indépendamment de l'ensemble de données d'entraînement

- Assurer aux utilisateurs l'accès à des informations claires et pertinentes : les utilisateurs disposent d'un accès facile à des informations claires et contextuellement pertinentes (informations appropriées au public visé et comprenant notamment des précisions sur l'utilisation des données, la performances du modèle, les caractéristiques des données utilisées pour entraîner et tester le modèle, modifications et mises à jour et un moyen de communiquer les préoccupations relatives au produit au développeur
- Permettre que les modèles déployés soient surveillés en continu à l'issue de leur déploiement en mettant l'accent sur le maintien ou l'amélioration de la sécurité et des performances.

En synthèse, les critères proposés par les administrations anglo-saxonnes sont les suivants :

- Collecte de données proportionnée au but
- Réduction des biais cognitif : qualité des données
- Réduction des biais de sélection : Représentativité
- Mesure pour assurer la séparation des données
- Mesures relatives à l'accessibilité des données
- Mesures de contrôle des données manquantes
- Mesures d'identification et élimination des biais dans l'algorithme
- Prise en compte des éventuelles cyber-attaques
- Résilience du système
- Mesures de traçabilité
- Mesures d'explicabilité
- Transparence de l'algorithme, du modèle
- Critères d'utilisabilité
- Mesures de non-discrimination/équité
- Mesures de reproductibilité
- Mesures pour le contrôle humain
- Analyse d'impact sur le parcours de soins
- Choix de l'algorithme d'apprentissage en adéquation avec la finalité
- Amélioration de la qualité et surveillance
- Adaptation aux évolutions réglementaires et médicales

Les principes et recommandations de l'UNESCO[26]

L'UNESCO a souhaité produire des recommandations sur la prise en compte de l'éthique en IA en réponse aux questions soulevées par les experts indépendants de son organe consultatif scientifique multidisciplinaire qui ont pu mettre en avant la nécessité de s'attaquer aux vastes implications sociétales et culturelles des progrès de l'IA. Les questions éthiques les plus centrales pour les domaines de compétence de l'UNESCO concernent « ses implications pour la culture et la diversité culturelle, l'éducation, la connaissance scientifique, et la communication et l'information ». Mais cela va plus loin, puisque « compte tenu de l'orientation mondiale de l'UNESCO, les thèmes éthiques de portée planétaire que sont la paix, la durabilité, l'égalité des genres et les défis spécifiques de l'Afrique, méritent aussi une attention particulière ».
Le groupe d'experts de l'UNESCO a défini un ensemble de principes éthiques consensuels qui concernent les dimensions éthiques de l'IA et pourraient être inscrits dans une éventuelle recommandation sur l'éthique de l'IA. Ces principes éthiques ont été définis à partir des conventions internationales et de la littérature existantes, qui ont été classées et enrichies en termes de contenu et de pertinence.

Deux principes majeurs et nouveaux ressortent au sein des recommandations de l'UNESCO, il s'agit des recommandations relatives à l'analyse de l'impact sur les droits fondamentaux et la protection des lanceurs d'alertes.

- Recueil du Consentement éclairé (RGPD) (non explicite)
- Collecte de données proportionnée au but
- Réduction des biais cognitif : qualité des données
- Réduction des biais de sélection : Représentativité
- Mesures d'identification et élimination des biais dans l'algorithme
- Mesures de traçabilité
- Mesures d'explicabilité
- Transparence de l'algorithme, du modèle
- Voies de recours / réparation
- Critères d'utilisabilité
- Mesures de non-discrimination/équité
- Mesures pour l'audit
- Mesures pour le contrôle humain
- Analyse d'impact sur les droits fondamentaux
- Analyse d'impact environnemental
- Analyse d'impact sociétal
- Choix de l'algorithme d'apprentissage en adéquation avec la finalité
- Participation des parties prenantes
- Conduite en cas d'erreur et responsabilité
- Amélioration de la qualité et surveillance
- Protection des lanceurs d'alertes

[26] https://fr.unesco.org/artificial-intelligence/ethics#recommandation

Les principes et recommandations du conseil de l'OCDE[27]

Adoptés en mai 2019, les principes de l'OCDE ont été imaginés pour proposer un standard pratique et flexible, dans le but de promouvoir l'utilisation de l'IA dans un cadre innovant, de confiance et respectant les droits de l'Homme comme les valeurs démocratiques. Ces principes sont généraux et ne concernent pas spécifiquement l'utilisation de l'IA en santé.

Les membres de l'OCDE sont signataires des principes, auxquels il faut ajouter les adhérents de l'OCDE ainsi que les membres du G20, ayant adopté des principes largement inspirés de ceux de l'OCDE :

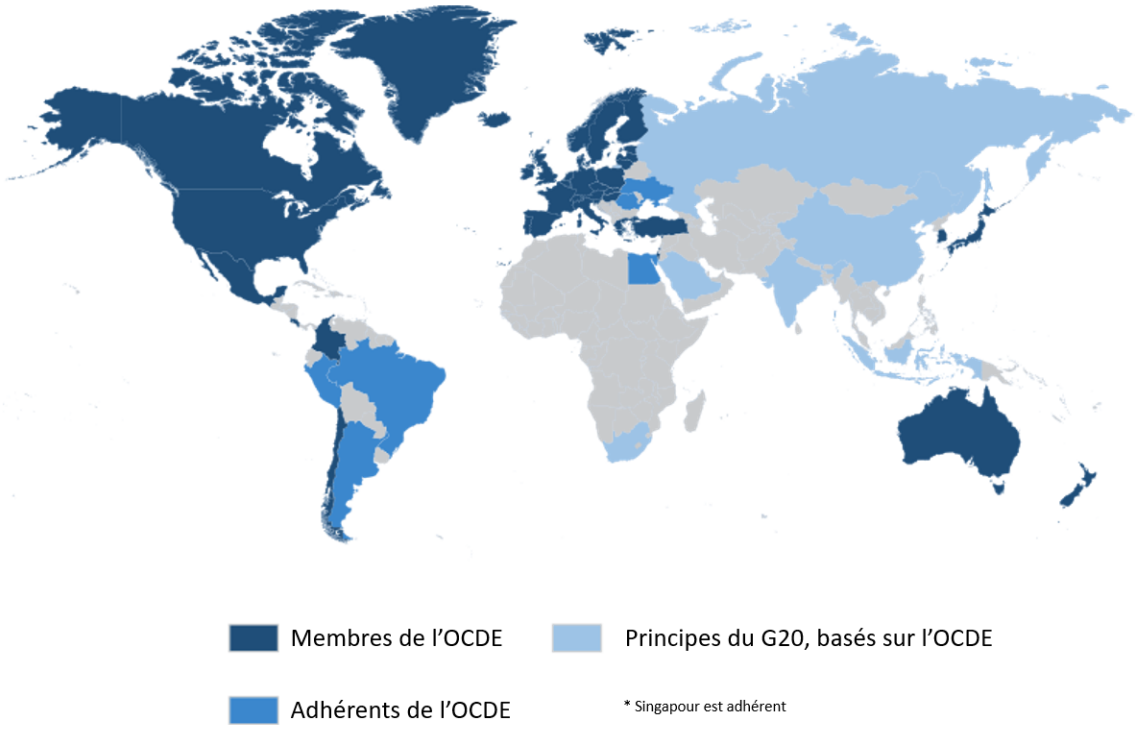


Figure 3 : Gouvernements s'étant engagés sur les principes de l'OCDE (Source : OCDE).

Les principes de l'OCDE sont les suivants :

- Croissance inclusive, développement durable et bien-être
  - Est soulignée ici la nécessité d'obtenir des résultats positifs pour les humains et la planète. La réduction des inégalités et l'inclusion des catégories de population sous-représentées sont également des objectifs mentionnés.
- Valeurs centrées sur l'humain et équité
  - Ce principe rassemble de nombreuses valeurs communes aux recommandations mentionnées précédemment, comprenant notamment la nécessité de la décision finale laissée à l'homme, le respect des valeurs démocratique et des droits de l'homme (liberté, dignité, autonomie, protection de la vie privée et des données, non-discrimination, égalité, diversité, équité, justice sociale, ainsi que les droits des travailleurs)

[27] https://oecd.ai/en/ai-principles

- Transparence et explicabilité
- Robustesse, sûreté et sécurité
- Responsabilité
    - Les acteurs de l'IA doivent être responsables du fonctionnement des systèmes d'IA et du respect des principes.

L'OCDE établit également des recommandations à l'intention des gouvernements, qui doivent :

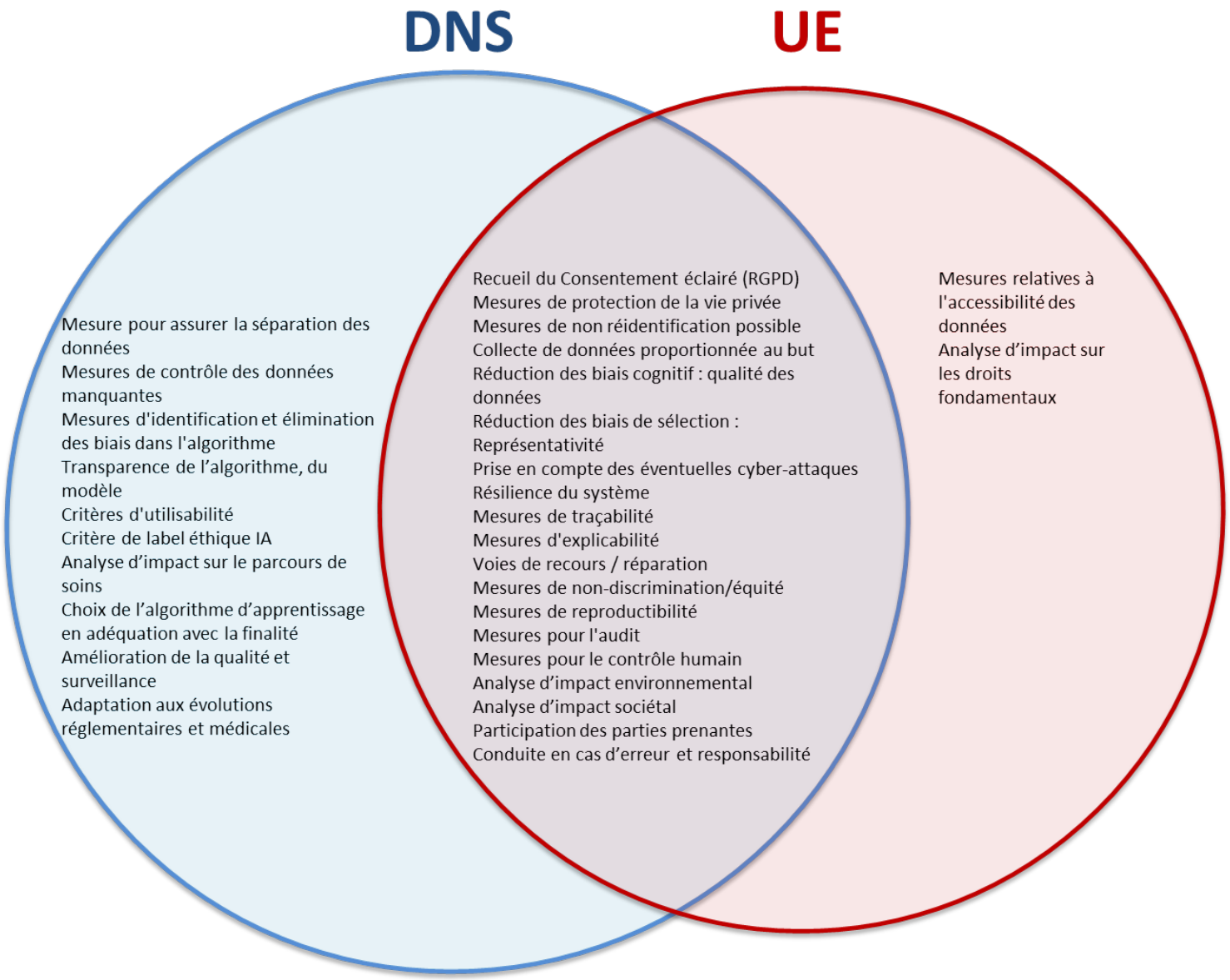
- Investir dans la recherche et le développement en matière d'IA
- Favoriser l'instauration d'un écosystème numérique pour l'IA
- Façonner un cadre d'action favorable à l'IA
- Renforcer les capacités humaines et préparer la transformation du marché du travail
- Favoriser la coopération internationale au service d'une IA digne de confiance

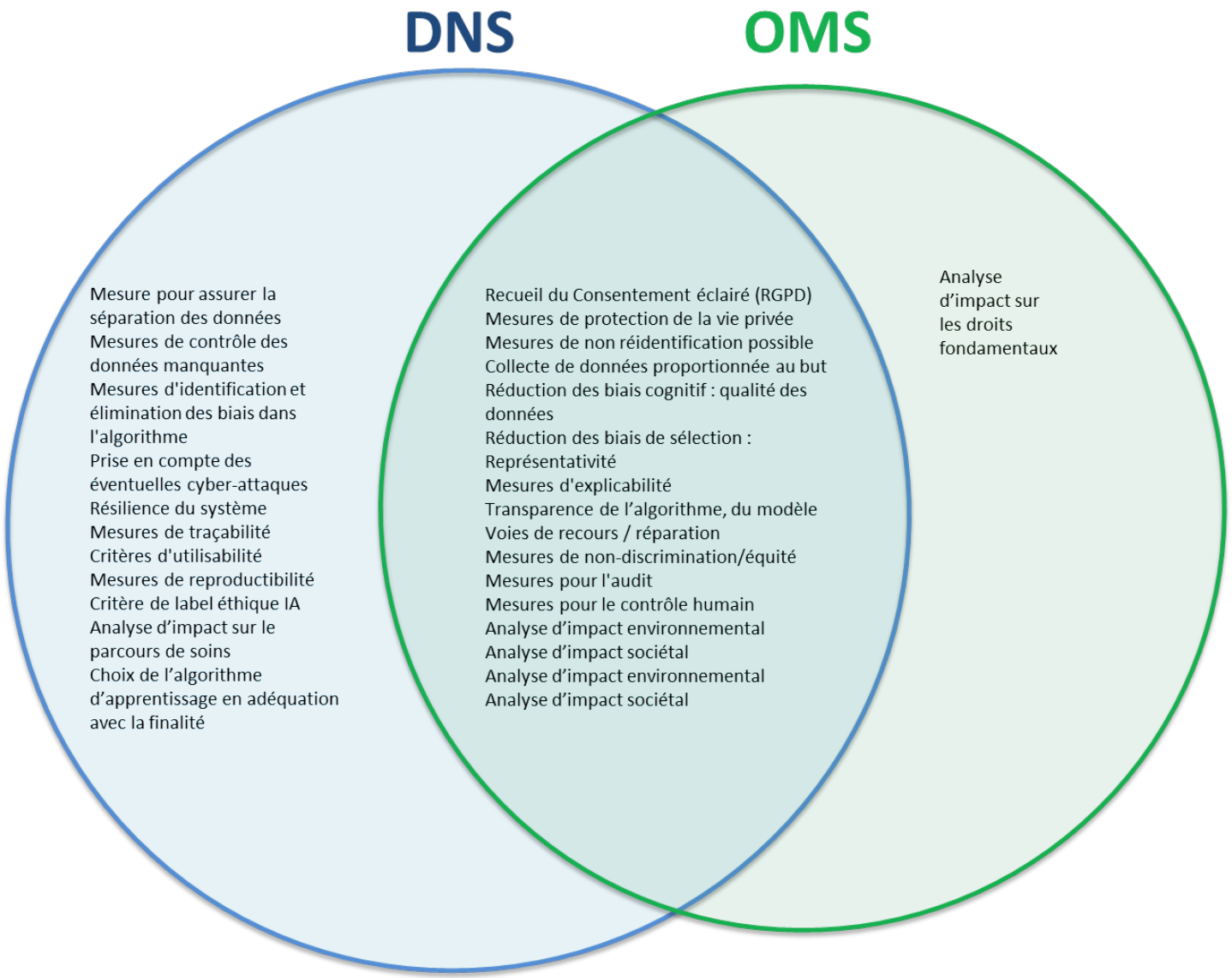
Analyse croisée

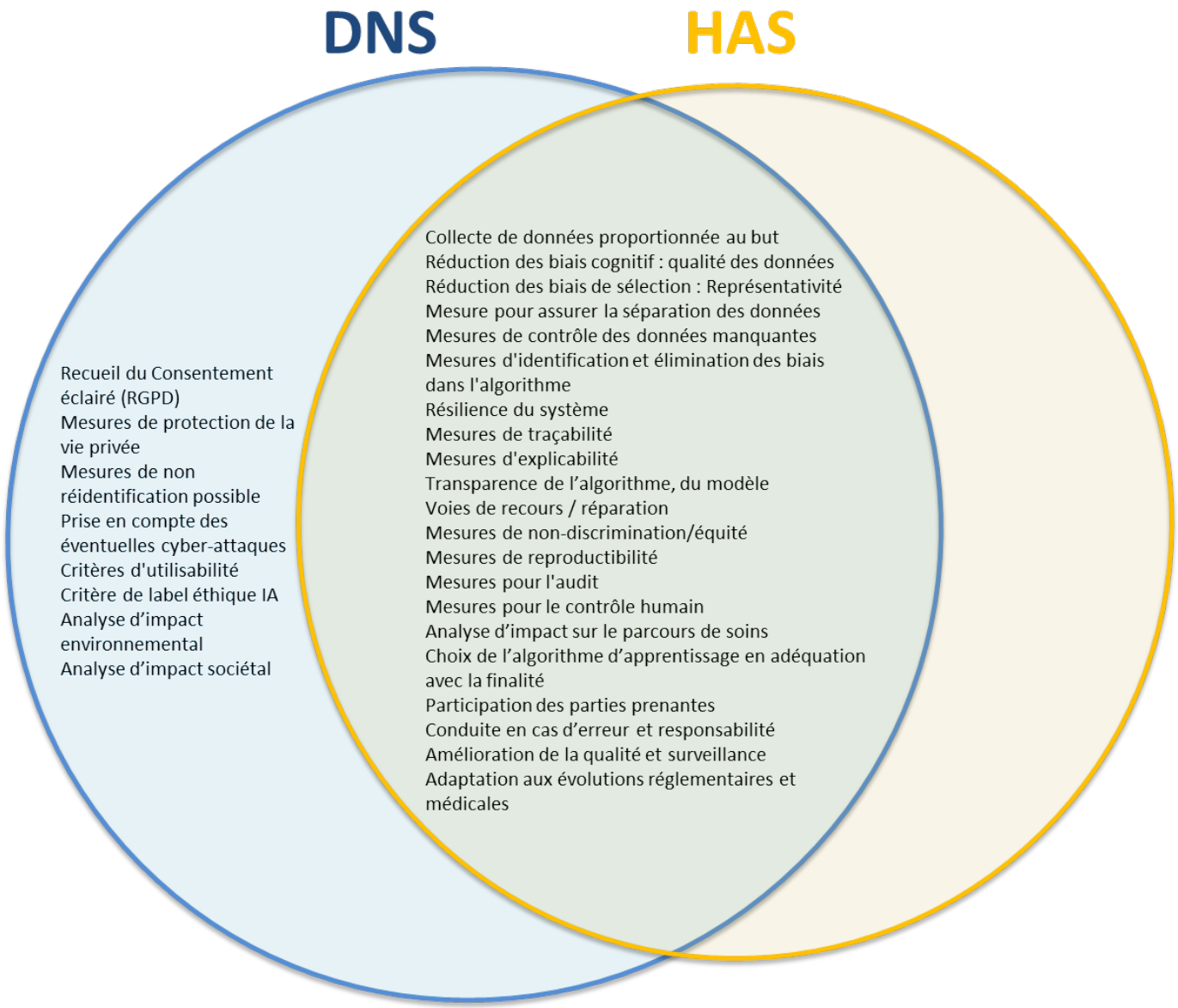
Le tableau suivant et les diagrammes de Venn associés ont vocation à montrer les chevauchements et différences avec les recommandations du GT3. Comme précisé au début du rapport, l'analyse croisée se concentre sur les principes et recommandations publiées concomitamment aux travaux du GT3. Depuis la conclusion des travaux, d'autres organismes internationaux ou nationaux ont pu avoir l'occasion de publier leurs propres recommandations en la matière, ils ne sont pas abordés et présentés dans cette étude.

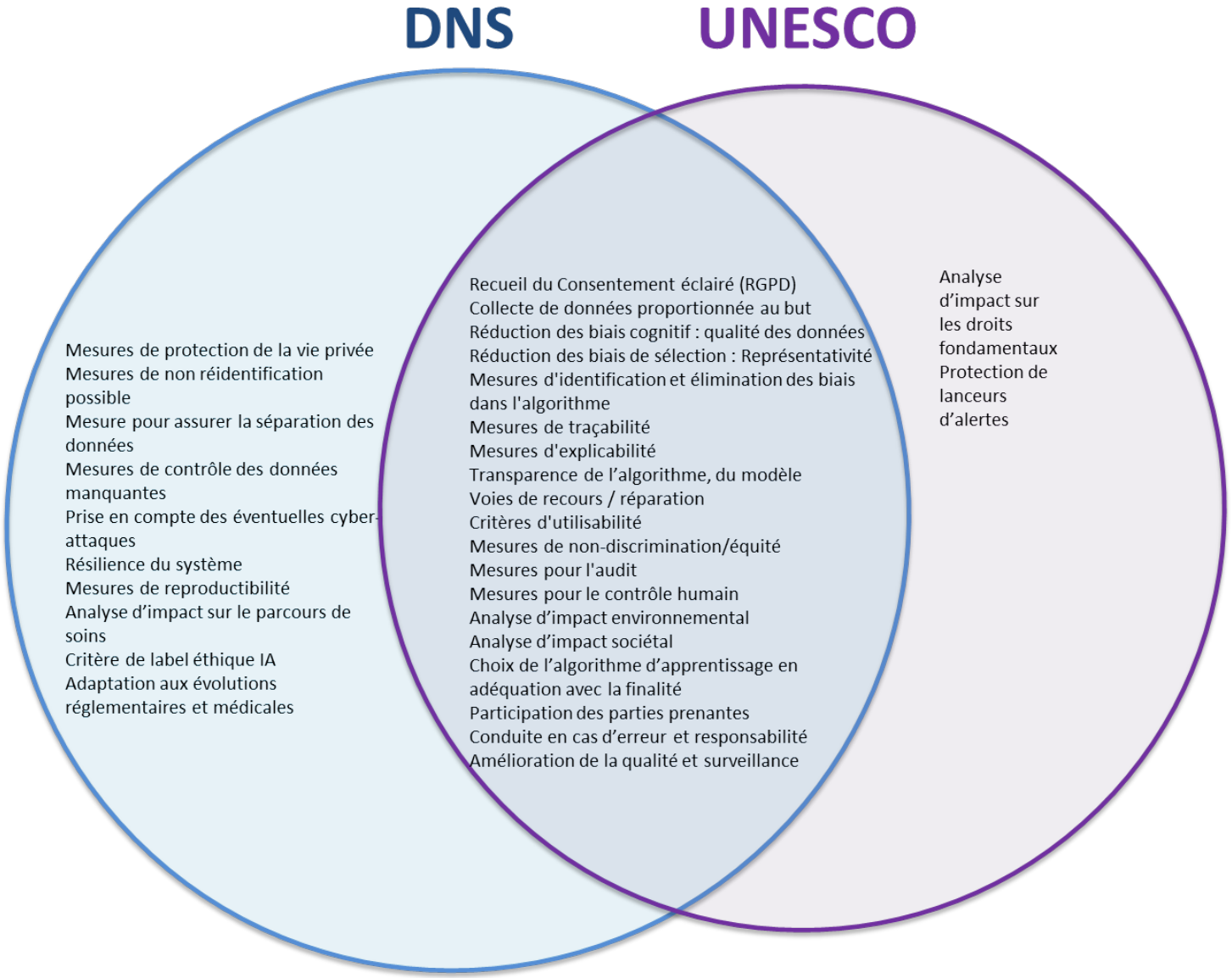
	GT3 Cellule Ethique du numérique en santé	Groupe d'experts de haut niveau de la commission européenne	OMS	HAS	UNESCO	OCDE	Ministères de la santé UK, Canada et FDA
Recueil du Consentement éclairé (RGPD)	Oui	Oui	Oui		Oui (non explicite)		
Mesures de protection de la vie privée	Oui	Oui	Oui			Oui	
Mesures de non réidentification possible	Oui	Oui	Oui				
Collecte de données proportionnée au but	Oui	Oui	Oui	Oui	Oui		Oui
Réduction des biais cognitif : qualité des données	Oui	Oui	Oui	Oui	Oui		Oui
Réduction des biais de sélection : Représentativité	Oui	Oui	Oui	Oui	Oui	Oui	Oui
Mesure pour assurer la séparation des données	Oui			Oui			Oui
Mesures relatives à l'accessibilité des données		Oui					Oui
Mesures de contrôle des données manquantes	Oui			Oui			Oui
Mesures d'identification et élimination des **biais dans l'algorithme**	Oui			Oui	Oui	Oui	Oui
Prise en compte des éventuelles **cyber-attaques**	Oui	Oui				Oui	Oui

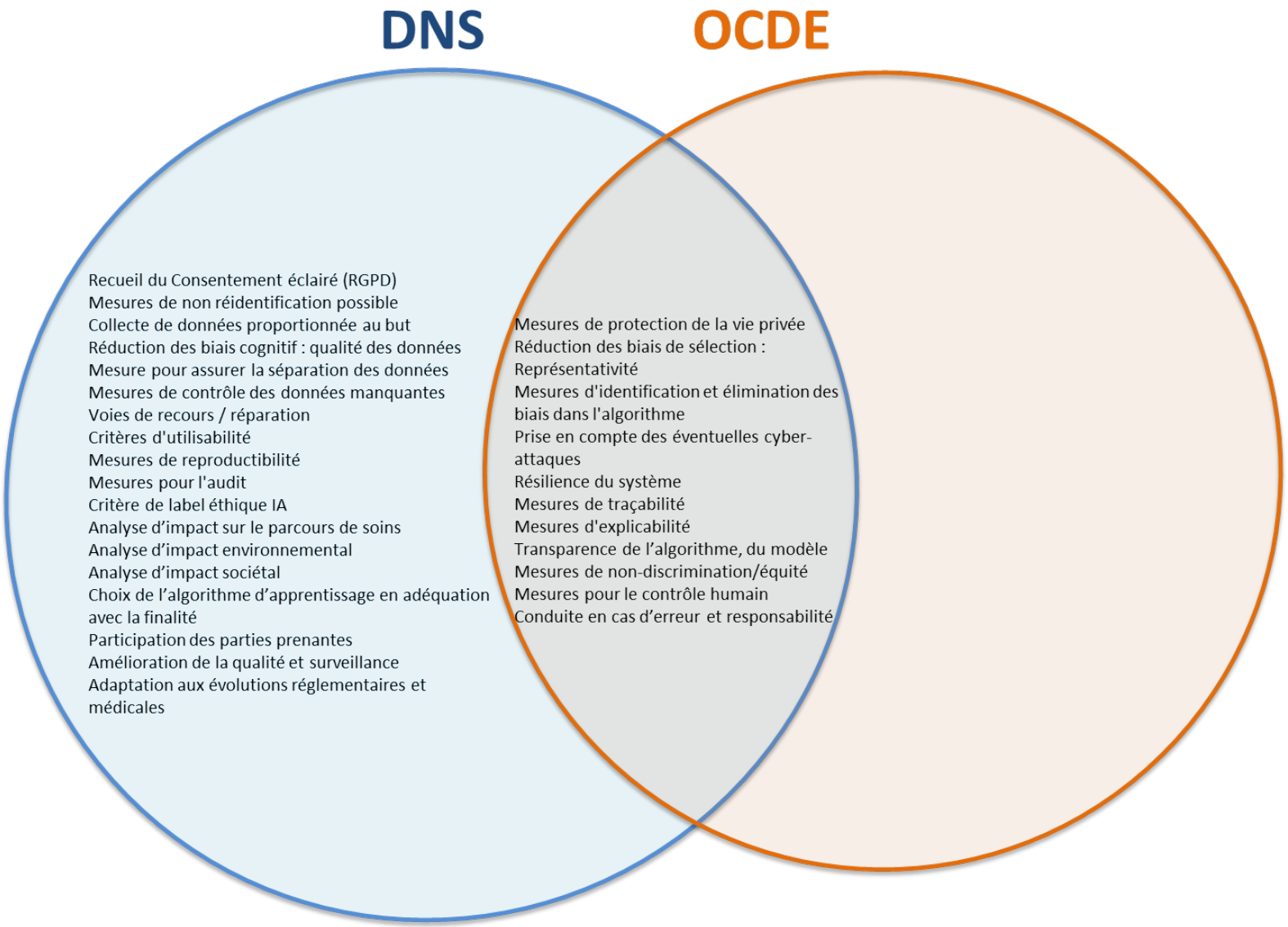
Résilience du système	Oui	Oui		Oui		Oui	Oui
Mesures de traçabilité	Oui	Oui		Oui	Oui	Oui	Oui
Mesures d'explicabilité	Oui	Oui	Oui	Oui	Oui	Oui	Oui
Transparence de l'algorithme, du modèle	Oui		Oui	Oui	Oui	Oui	Oui
Voies de **recours** / réparation	Oui	Oui	Oui	Oui	Oui		
Critères d'utilisabilité	Oui				Oui		Oui
Mesures de non-discrimination/équité	Oui	Oui	Oui	Oui	Oui	Oui	Oui
Mesures de reproductibilité	Oui	Oui		Oui			Oui
Mesures pour l'audit	Oui	Oui	Oui	Oui	Oui		
Critère de label éthique IA	Oui						
Mesures pour le contrôle humain	Oui	Oui	Oui	Oui	Oui	Oui	Oui
Analyse d'impact sur le **parcours de soins**	Oui			Oui			Oui
Analyse d'impact sur les **droits fondamentaux**		Oui	Oui		Oui		
Analyse d'impact **environnemental**	Oui	Oui	Oui		Oui		
Analyse d'impact **sociétal**	Oui	Oui	Oui		Oui		
Choix de l'algorithme d'apprentissage en adéquation avec la finalité	Oui			Oui	Oui		Oui
Participation des **parties prenantes**	Oui	Oui	Oui	Oui	Oui		
Conduite en cas d'erreur et responsabilité	Oui	Oui	Oui	Oui	Oui	Oui	
Amélioration de la **qualité** et **surveillance**	Oui		Oui	Oui	Oui		Oui
Adaptation aux **évolutions réglementaires et médicales**	Oui		Oui	Oui			Oui
Protection de lanceurs d'alertes					Oui		

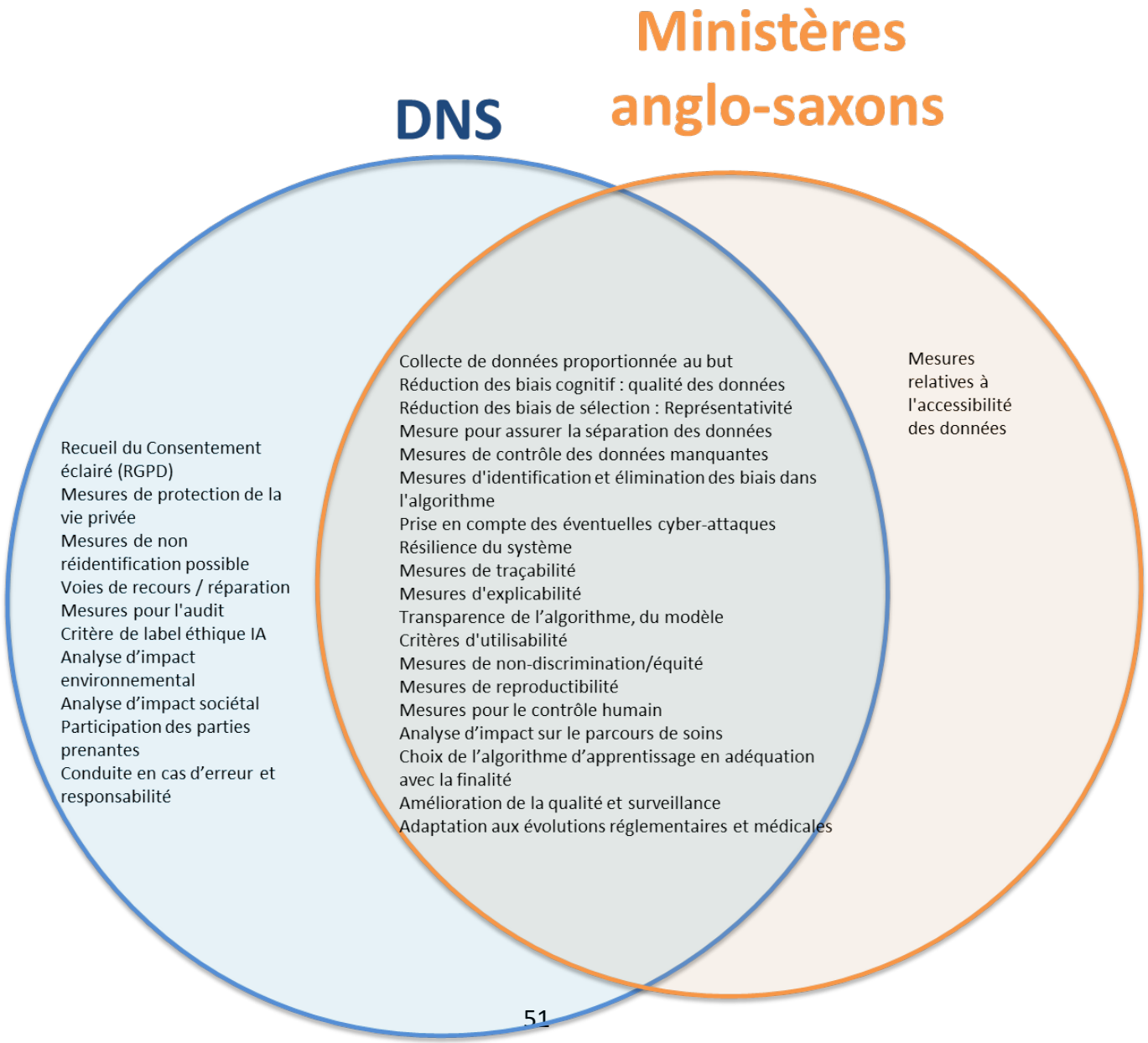

DNS
UE
Mesure pour assurer la séparation des données
Mesures de contrôle des données manquantes
Mesures d'identification et élimination des biais dans l'algorithme
Transparence de l'algorithme, du modèle
Critères d'utilisabilité
Critère de label éthique IA
Analyse d'impact sur le parcours de soins
Choix de l'algorithme d'apprentissage en adéquation avec la finalité
Amélioration de la qualité et surveillance
Adaptation aux évolutions réglementaires et médicales
Recueil du Consentement éclairé (RGPD)
Mesures de protection de la vie privée
Mesures de non réidentification possible
Collecte de données proportionnée au but
Réduction des biais cognitif : qualité des données
Réduction des biais de sélection : Représentativité
Prise en compte des éventuelles cyber-attaques
Résilience du système
Mesures de traçabilité
Mesures d'explicabilité
Voies de recours / réparation
Mesures de non-discrimination/équité
Mesures de reproductibilité
Mesures pour l'audit
Mesures pour le contrôle humain
Analyse d'impact environnemental
Analyse d'impact sociétal
Participation des parties prenantes
Conduite en cas d'erreur et responsabilité
Mesures relatives à l'accessibilité des données
Analyse d'impact sur les droits fondamentaux


DNS
OMS
Mesure pour assurer la séparation des données
Mesures de contrôle des données manquantes
Mesures d'identification et élimination des biais dans l'algorithme
Prise en compte des éventuelles cyber-attaques
Résilience du système
Mesures de traçabilité
Critères d'utilisabilité
Mesures de reproductibilité
Critère de label éthique IA
Analyse d'impact sur le parcours de soins
Choix de l'algorithme d'apprentissage en adéquation avec la finalité
Recueil du Consentement éclairé (RGPD)
Mesures de protection de la vie privée
Mesures de non réidentification possible
Collecte de données proportionnée au but
Réduction des biais cognitif : qualité des données
Réduction des biais de sélection : Représentativité
Mesures d'explicabilité
Transparence de l'algorithme, du modèle
Voies de recours / réparation
Mesures de non-discrimination/équité
Mesures pour l'audit
Mesures pour le contrôle humain
Analyse d'impact environnemental
Analyse d'impact sociétal
Analyse d'impact environnemental
Analyse d'impact sociétal
Analyse d'impact sur les droits fondamentaux


DNS
HAS
Recueil du Consentement éclairé (RGPD)
Mesures de protection de la vie privée
Mesures de non réidentification possible
Prise en compte des éventuelles cyber-attaques
Critères d'utilisabilité
Critère de label éthique IA
Analyse d'impact environnemental
Analyse d'impact sociétal
Collecte de données proportionnée au but
Réduction des biais cognitif : qualité des données
Réduction des biais de sélection : Représentativité
Mesure pour assurer la séparation des données
Mesures de contrôle des données manquantes
Mesures d'identification et élimination des biais dans l'algorithme
Résilience du système
Mesures de traçabilité
Mesures d'explicabilité
Transparence de l'algorithme, du modèle
Voies de recours / réparation
Mesures de non-discrimination/équité
Mesures de reproductibilité
Mesures pour l'audit
Mesures pour le contrôle humain
Analyse d'impact sur le parcours de soins
Choix de l'algorithme d'apprentissage en adéquation avec la finalité
Participation des parties prenantes
Conduite en cas d'erreur et responsabilité
Amélioration de la qualité et surveillance
Adaptation aux évolutions réglementaires et médicales


DNS
UNESCO
Mesures de protection de la vie privée
Mesures de non réidentification possible
Mesure pour assurer la séparation des données
Mesures de contrôle des données manquantes
Prise en compte des éventuelles cyber-attaques
Résilience du système
Mesures de reproductibilité
Analyse d'impact sur le parcours de soins
Critère de label éthique IA
Adaptation aux évolutions réglementaires et médicales
Recueil du Consentement éclairé (RGPD)
Collecte de données proportionnée au but
Réduction des biais cognitif : qualité des données
Réduction des biais de sélection : Représentativité
Mesures d'identification et élimination des biais dans l'algorithme
Mesures de traçabilité
Mesures d'explicabilité
Transparence de l'algorithme, du modèle
Voies de recours / réparation
Critères d'utilisabilité
Mesures de non-discrimination/équité
Mesures pour l'audit
Mesures pour le contrôle humain
Analyse d'impact environnemental
Analyse d'impact sociétal
Choix de l'algorithme d'apprentissage en adéquation avec la finalité
Participation des parties prenantes
Conduite en cas d'erreur et responsabilité
Amélioration de la qualité et surveillance
Analyse d'impact sur les droits fondamentaux
Protection de lanceurs d'alertes


DNS
OCDE
Recueil du Consentement éclairé (RGPD)
Mesures de non réidentification possible
Collecte de données proportionnée au but
Réduction des biais cognitif : qualité des données
Mesure pour assurer la séparation des données
Mesures de contrôle des données manquantes
Voies de recours / réparation
Critères d'utilisabilité
Mesures de reproductibilité
Mesures pour l'audit
Critère de label éthique IA
Analyse d'impact sur le parcours de soins
Analyse d'impact environnemental
Analyse d'impact sociétal
Choix de l'algorithme d'apprentissage en adéquation avec la finalité
Participation des parties prenantes
Amélioration de la qualité et surveillance
Adaptation aux évolutions réglementaires et médicales
Mesures de protection de la vie privée
Réduction des biais de sélection : Représentativité
Mesures d'identification et élimination des biais dans l'algorithme
Prise en compte des éventuelles cyber-attaques
Résilience du système
Mesures de traçabilité
Mesures d'explicabilité
Transparence de l'algorithme, du modèle
Mesures de non-discrimination/équité
Mesures pour le contrôle humain
Conduite en cas d'erreur et responsabilité


DNS
Ministères anglo-saxons
Recueil du Consentement éclairé (RGPD)
Mesures de protection de la vie privée
Mesures de non réidentification possible
Voies de recours / réparation
Mesures pour l'audit
Critère de label éthique IA
Analyse d'impact environnemental
Analyse d'impact sociétal
Participation des parties prenantes
Conduite en cas d'erreur et responsabilité
Collecte de données proportionnée au but
Réduction des biais cognitif : qualité des données
Réduction des biais de sélection : Représentativité
Mesure pour assurer la séparation des données
Mesures de contrôle des données manquantes
Mesures d'identification et élimination des biais dans l'algorithme
Prise en compte des éventuelles cyber-attaques
Résilience du système
Mesures de traçabilité
Mesures d'explicabilité
Transparence de l'algorithme, du modèle
Critères d'utilisabilité
Mesures de non-discrimination/équité
Mesures de reproductibilité
Mesures pour le contrôle humain
Analyse d'impact sur le parcours de soins
Choix de l'algorithme d'apprentissage en adéquation avec la finalité
Amélioration de la qualité et surveillance
Adaptation aux évolutions réglementaires et médicales
Mesures relatives à l'accessibilité des données

7. Discussion : quelle régulation éthique de l'IA en santé ?

La diffusion d'une vague d'innovations sur l'intelligence artificielle en santé modifie très profondément les pratiques médicales. Elle soulève également des enjeux majeurs de régulation des enjeux éthiques associés. La crise COVID19 a montré tout à la fois l'ampleur de ces opportunités et de ces risques. Une voie de passage pragmatique doit donc être aménagée pour permettre la mise en œuvre pratique de ces innovations au service des patients. L'objectif est donc celui d'une régulation positive pour préserver les valeurs essentielles de notre système de santé et plus particulièrement les fondements d'une médecine personnalisée. C'est le sens de la Garantie Humaine de l'IA, principe reconnu par l'article 17 de la loi n° 2021-1017 du 2 août 2021 relative à la bioéthique. L'intégration de ce principe dans les *Guidelines* de l'OMS sur la régulation de l'IA en santé lui donne une dimension encore plus large.

La nécessité d'une régulation éthique positive de l'IA en santé

Le rapport du Conseil national de l'ordre des médecins de janvier 2018[28] montrait à quel point notre système de santé était en train de connaître une transformation profonde sous l'effet de l'émergence progressive d'un véritable pilotage par les données de santé.

Le déploiement de l'intelligence artificielle est, en effet, une source d'améliorations potentielles littéralement extraordinaires pour notre santé. Ce saut qualitatif possible pour notre système de soins s'accompagne d'un effet de levier potentiel majeur pour la croissance de la France. Il induit également des transformations possiblement radicales des métiers du champ sanitaire et médico-social. Nous vivons actuellement une véritable révolution des cas d'usage de recours à l'intelligence artificielle. La technique la plus mature est la reconnaissance d'image par apprentissage machine qui trouve d'ores et déjà à s'appliquer largement en radiologie, dermatologie, ophtalmologie ou encore en oncologie. Dans toutes ces disciplines, les performances de diagnostic des algorithmes dépassent désormais fréquemment celles de l'Humain.

Les avis 129 et 130 du Comité consultatif national d'éthique[29], émis dans le cadre de la préparation de la révision bioéthique, a identifié des risques éthiques associés à la diffusion de cette nouvelle vague d'innovations technologiques en santé et notamment ceux d'une délégation pratique de la décision du médecin et du consentement du patient à l'IA. Comme nous le ressentons tous, la matérialisation de ces risques est devenue palpable dans le contexte épidémique COVID-19 qui affecte si durement les patients et les professionnels. Les

[28] Médecins et patients dans le monde des data, des algorithmes et de l'intelligence artificielle (https://www.conseil-national.medecin.fr/sites/default/files/external-package/edition/od6gnt/cnomdata_algorithmes_ia_0.pdf)

[29] https://www.ccne-ethique.fr/fr/actualites/lavis-129-contribution-du-ccne-la-revision-de-la-loi-de-bioethique-est-en-ligne

principes au fondement de notre médecine personnalisée en France et en Europe peuvent entrer en confrontation avec la logique collectiviste intrinsèque aux algorithmes d'IA. Cette logique est, d'une certaine manière, hyperbolisée en contexte épidémique.

La collecte massive de données – ce que l'on a pris l'habitude d'appeler le « *big data* » – constitue une condition *sine qua non* au déploiement de ce *data management* et de l'intelligence artificielle. Cette dernière se fonde, en effet, sur des algorithmes qui nécessitent la mobilisation de données fiables et en nombre suffisant pour dégager des calculs robustes de probabilités permettant d'appuyer les orientations de l'intelligence artificielle. Cette donnée-requérance ne constitue pas un principe nouveau : les *data* constituent la matière première, la base de l'alimentation de tout programme informatique. Le déploiement actuel du pilotage par les données et de l'IA sur un large spectre fait simplement changer l'enjeu d'échelle.

Une « course aux données de santé » s'est donc engagée au niveau mondial avec les signaux de plus en plus visibles d'une compétition exacerbée. Pour pouvoir approvisionner les algorithmes, ces données doivent être médicalement et techniquement fiables mais également en volume suffisant pour permettre à l'IA de s'appuyer sur des régularités statistiques robustes. Cette compétition internationale pour les données de santé est naturellement marquée par un facteur temps. Le premier fabricant de solutions de *data management* et d'IA qui sera parvenu à élaborer une solution en santé opérationnelle alimentée par une masse de données fiables et en nombre suffisant aura acquis un avantage sans doute décisif sur ses concurrents. Avec naturellement des perspectives financières colossales à la clé.

Pour autant, cette diffusion rapide de l'intelligence artificielle en santé est aussi génératrice d'un besoin de régulation éthique. La crise sanitaire que nous traversons l'a, à nouveau, fortement montré. Dans la réponse au COVID19, certains pays – en particulier en Asie – ont eu plus largement recours que d'autres à l'IA, au pilotage par les données et aux technologies numériques. Des dispositifs de reconnaissance faciale ainsi que l'utilisation de thermomètres connectés ont permis la surveillance de la température et l'identification de personnes à risque d'être positive au COVID19. Les données de géolocalisation ont été largement utilisées pour connaitre les flux des personnes et bloquer certains déplacements. Des robots ont été introduits dans certains hôpitaux, pour accompagner voire renforcer les équipes médicales, assurer une présence auprès des patients et répondre à leurs besoins, décontaminer certains services… Le *Data Tracking* a été mis en œuvre sans restriction et un choix d'efficacité a été fait au détriment de la protection des données de santé.

Mais au-delà des recours à ces dispositifs fortement visibles et dont les usages ont été pour certains contestables au regard des principes de nos sociétés démocratiques, le recours à l'intelligence artificielle en temps de crise épidémique porte aussi la promesse d'accélération de certains procédés relevant aussi bien du diagnostic que de la recherche. Des méthodes diagnostiques reposant sur la reconnaissance d'images par apprentissage machine permettraient un diagnostic beaucoup plus rapide et efficace sur la base de clichés de tomodensitométrie. L'IA induit aussi un potentiel d'apport majeur concernant l'identification d'éventuels traitements efficaces.

Les options radicales, choisies par certains pays, pour répondre au risque collectif au détriment de la protection des libertés individuelles, semblent très éloignés des principes fondateurs du règlement général sur la protection des données (RGPD) en Europe et, plus largement, des valeurs essentielles de notre médecine personnalisée. Cette gestion de crise illustre l'importance et l'intérêt de la Garantie Humaine de l'IA, consistant dans la mise en place d'une supervision humaine lors du recours à un algorithme d'intelligence artificielle.

La garantie humaine de l'IA en santé : reconnaissance d'un vecteur de régulation positive

Les principes au fondement de notre médecine personnalisée en France et en Europe peuvent entrer en confrontation avec un certain nombre de principes éthiques. Le principal danger est sans doute celui de la perte d'un recul critique des soignants et des soignés avec, en arrière-plan, une mécanique algorithmique fondée sur la loi du plus grand nombre, cette dernière pouvant aller à l'encontre d'intérêts d'individus ou de groupes d'individus.

Pour éviter une perspective aussi sinistre, le principe de Garantie Humaine de l'IA est issu d'un mouvement de propositions académiques, citoyennes mais aussi de professionnels de santé. La reconnaissance de ce principe est portée dans l'article 11 du projet de loi bioéthique devenu l'article 17 de la loi du 2 août 2021 relative à la bioéthique[30]. Il comprend deux normes nouvelles : l'information du patient sur le recours à l'IA dans sa prise en charge, d'une part, et le principe de Garantie Humaine de l'IA lui-même, d'autre part.

Le concept de « Garantie Humaine » peut paraître abstrait mais il est, en réalité, très concret. Dans le cas de l'IA, l'idée est d'appliquer les principes de régulation de l'intelligence artificielle en amont et en aval de l'algorithme lui-même en établissant des points de supervision humaine. Non pas à chaque étape, sinon l'innovation serait bloquée. Mais sur des points critiques identifiés dans un dialogue partagé entre les professionnels, les patients et les

[30] « Le chapitre Ier du titre préliminaire du livre préliminaire de la quatrième partie du code de la santé publique est complété par un article L. 4001-3 ainsi rédigé :

« Art. L. 4001-3.-I.-Le professionnel de santé qui décide d'utiliser, pour un acte de prévention, de diagnostic ou de soin, un dispositif médical comportant un traitement de données algorithmique dont l'apprentissage a été réalisé à partir de données massives s'assure que la personne concernée en a été informée et qu'elle est, le cas échéant, avertie de l'interprétation qui en résulte.
« II.-Les professionnels de santé concernés sont informés du recours à ce traitement de données. Les données du patient utilisées dans ce traitement et les résultats qui en sont issus leur sont accessibles.
« III.-Les concepteurs d'un traitement algorithmique mentionné au I s'assurent de l'explicabilité de son fonctionnement pour les utilisateurs.
« IV.-Un arrêté du ministre chargé de la santé établit, après avis de la Haute Autorité de santé et de la Commission nationale de l'informatique et des libertés, la nature des dispositifs médicaux mentionnés au I et leurs modalités d'utilisation. »

concepteurs d'innovation. La supervision peut s'exercer avec le déploiement de « collèges de garantie humaine » associant médecins, professionnels paramédicaux et représentants des usagers. Leur vocation serait d'assurer *a posteriori* une révision de dossiers médicaux pour porter un regard humain sur les options thérapeutiques conseillées ou prises par l'algorithme. L'objectif consiste à s'assurer « au fil de l'eau » que l'algorithme reste sur un développement de *Machine Learning* à la fois efficace médicalement et responsable éthiquement. Les dossiers à auditer pourraient être définis à partir d'événements indésirables constatés, de critères prédéterminés ou d'une sélection aléatoire. La première session pilote de collège de garantie humaine a été organisée, en décembre 2020, sous l'égide de l'Union française pour la santé bucco-dentaire (UFSBD) dans le cas d'une solution d'IA appliquée dans le domaine des soins bucco-dentaires (protocole innovant de l'article 51 de la loi de financement de la Sécurité sociale).

Il est aussi à relever que le principe de garantie humaine avait reçu, en 2020 et 2021, des concrétisations dans des cadres très significatifs :

- La garantie humaine a été incorporée dans la grille d'auto-évaluation des dispositifs médicaux intégrant de l'IA publiée par la Haute Autorité de Santé au mois d'octobre dernier ;
- Le principe de garantie humaine a été incorporé par l'OMS à la régulation de l'IA en santé [31];
- Le principe a été repris dans le Livre Blanc sur l'IA publié par la Commission européenne le 19 février 2020.

Cette dernière reconnaissance a été prolongée par le projet de règlement établi par la Commission européenne sur la base de cette consultation. L'article 14 de ce texte qui institue ce principe se situe dans le droit fil de ces démarches pilotes et de ces recommandations.

Le paragraphe 1 de cet article énonce ainsi que les solutions d'intelligence artificielle doivent être conçues et développées de façon à pouvoir être supervisé par des Humains. Le paragraphe suivant précise que la supervision humaine permettra de prévenir ou de minimiser les risques pour la santé, la sécurité ou les droits fondamentaux pouvant émerger d'un système d'IA susceptible de présenter un niveau de risque élevé. Par cet énoncé, **le projet de règlement consacre ainsi la nécessité d'une garantie humaine pour un déploiement éthique de l'IA**. Le paragraphe 3 donne ainsi des indications sur la mise en application de la supervision humaine de l'IA. En effet, **la garantie humaine doit être identifiée et construite par le fournisseur avant sa mise sur le marché ou sa mise en service et/ou identifiée par le fournisseur et pouvant être mise en œuvre par l'utilisateur**, et ce, toujours en amont de sa mise sur le marché ou de sa mise en service. Cette garantie humaine doit pouvoir faire l'objet d'un **suivi en vie réelle de l'intelligence artificielle**. **Les mesures prévues au paragraphe 3 de**

[31] A relever la définition précise de la Garantie Humaine donnée par l'OMS selon laquelle la régulation de l'IA « *peut être assurée par l'application de la "garantie humaine", qui implique une évaluation par les patients et les cliniciens lors du développement et du déploiement des technologies d'IA. La garantie humaine nécessite l'application de principes réglementaires en amont et en aval de l'algorithme en établissant des points de supervision humaine. Si quelque chose ne va pas avec une technologie d'IA, il devrait y avoir une responsabilité. Des mécanismes appropriés devraient être disponibles pour la remise en question et pour la réparation des individus et des groupes qui sont affectés négativement par des décisions basées sur des algorithmes. » / Recommandations de l'OMS, Éthique et gouvernance de l'intelligence artificielle en santé, 28 juin 2021, p. 13*

cet article fixent un certain nombre d'objectifs d'information autour de cette Garantie Humaine : comprendre entièrement les capacités et limites du système d'IA et être capable de surveiller l'opération de façon à ce que les risques d'anomalies, de dysfonctionnements et de performance inattendus puissent être détectés ; être conscient des risques liés aux IA d'aide à la décision ; être capable d'interpréter correctement le résultat de l'IA à haut risque et, si nécessaire, ne pas tenir compte de ce résultat, le remplacer ou l'ignorer ; et enfin pouvoir interrompre l'IA à tout moment.

On retrouve donc dans l'article 14 du règlement européen les deux axes essentiels de l'article 17 et des méthodologies construites depuis 2017 dans le champ de la santé dans le cadre de la loi de bioéthique :

- L'information des utilisateurs de la solution d'IA ;
- La supervision humaine de l'IA dans sa phase de conception et, dans une logique d'amélioration continue de la qualité, dans son utilisation en vie réelle.

8. Conclusion

L'intelligence artificielle s'invite progressivement dans plusieurs pans de notre société. Elle laisse entrevoir des perspectives prometteuses dans le domaine de la santé où elle pourrait contribuer à améliorer significativement la pratique clinique et la recherche médicale. La crise sanitaire liée au COVID-19 a vu émerger de nouvelles applications ou technologies embarquant l'IA, permettant ainsi de mettre en exergue des problématiques éthiques qui jusqu'à présent n'avaient pas ou peu été abordées par les acteurs : surveillance, atteinte au droit à la vie privée et à l'autonomie, etc. En effet, l'IA présente des spécificités qui rendent son développement et son explicabilité parfois complexe. La construction de tels outils nécessite donc une vigilance éthique particulière de la part des éditeurs, dès la conception de ces outils.

Le questionnement éthique passe par ailleurs par la connaissance de la réglementation en vigueur relative aux données de santé pour en assurer la collecte, le prétraitement et le stockage dans des conditions qui garantissent le respect des droits des usagers. Cela passe également par plus de transparence tout au long du processus d'optimisation de l'algorithme.

En effet, la transparence et l'explicabilité de l'algorithme sont des conditions essentielles pour renforcer la confiance des usagers et des professionnels de santé qui bénéficieront de ces technologies.

La sensibilisation des acteurs aux considérations éthiques accompagnée d'un modèle d'évaluation efficient à même de prendre en compte les spécificités de l'IA permettra à termes de placer l'éthique et le respect des droits des usagers au cœur de la conception de toute solution de santé.

9. Annexes

Liste des personnes qualifiées auditionnées par le groupe de travail

Nom – prénom	Titre	Organisme de rattachement
Anaïs PERSON	Juriste numérique	Institut Droit et Santé
Marie-Pauline TALABARD	CEO et Co-fondatrice	NetR
Clément GOEHRS	CEO et Co-fondateur	Synapse Medicine
Dr Jean-Louis FRAYSSE	Co-fondateur	BOTdesign
Jérôme BERANGER	CEO	ADELIAA
Dr Madeleine CAVET	Radiologue et directrice médicale	Groupe CTM
David GIBLAS	Directeur général délégué, en charge de la direction des opérations assurance et relation client, la direction de l'innovation, des Partenariats santé, du digital et de la data, des Achats et de MH Innov	Malakoff médéric humanis
Stéphane BARDE	Directrice Data	Malakoff médéric humanis
Corinne COLLIGNON	Cheffe de service adjointe évaluation des dispositifs (SED) au sein de la direction de l'évaluation médicale, économique et de santé publique	HAS
Caroline GUILLOT	Adjointe à la direction citoyenne	Health Data Hub
Claire PATOUILLET	Conseillère technique	GESMS

Documentation étudiée par le groupe de travail (liste non exhaustive)

Titre du document	Références
Guidance for Regulation of Artificial Intelligence Applications (7 Janvier 2020)	https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf
Ethics guidelines for trustworthy AI (Juin 2018)	https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai
Grille HAS (Novembre 2019)	https://www.has-sante.fr/upload/docs/application/pdf/2019-11/notice_consultation_algorithmes.pdf
Avis 130 du CCNE	https://www.ccne-ethique.fr/sites/default/files/avis_130.pdf

WHITE PAPER On Artificial Intelligence - A European approach to excellence and trust (Février 2020)	https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en
ESC e-Cardiology Working Group Position Paper: Overcoming challenges in digital health implementation in cardiovascular medicine (Marx 2019)	https://www.researchgate.net/publication/332046115_ESC_e-Cardiology_Working_Group_Position_Paper_Overcoming_challenges_in_digital_health_implementation_in_cardiovascular_medicine
Lignes directrices en matière d'éthique pour une IA digne de confiance (Avril 2019)	https://op.europa.eu/fr/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1
ETHICS AND GOVERNANCE OF ARTIFICIAL INTELLIGENCE FOR HEALTH WHO guidance (2021)	https://apps.who.int/iris/rest/bitstreams/1352854/retrieve
Un nouvel outil pour l'évaluation des dispositifs médicaux embarquant de l'intelligence artificiel (Octobre 2020)	https://www.has-sante.fr/upload/docs/application/pdf/2016-01/guide_fabricant_2016_01_11_cnedimts_vd.pdf#page=43
Good Machine Learning Practice for Medical Device Development: Guiding Principles (Octobre 2021)	https://www.fda.gov/media/153486/download

Liste des critères d'évaluation du GT3 de la Cellule Ethique du numérique en santé

Etape de cadrage amont : Définir la finalité de la solution d'IA et valider l'éthique de la finalité (CSTE ou CoSTE) Caractériser les principes de gouvernance et de responsabilité			
1. Collecte des données	**2. Pré-traitement des données**	**3. Construction de l'algorithme**	**4. Evaluation de l'algorithme et préparation de la mise en production**
▪ Données obtenues auprès de tiers garantissant le consentement éclairé des patients à la réutilisation de leurs données au-delà de la finalité première du recueil (conformité RGPD, article 5 et 7) ▪ Données non directement ré-identifiables (agrégées, pseudonymisées, anonymisées) (conformité RGPD, article 32) ▪ Proportionnalité des données collectées par rapport à la finalité du traitement servant l'élaboration de la solution d'IA (conformité RGPD, article 5, c) ▪ Transfert sécurisé des données (source unique, sources multiples, chaînage, intégrité) ▪ Qualité des données (biais cognitif) ▪ Représentativité de la population d'analyse /population cible / prévention des discriminations (biais de sélection) ▪ Qualité de l'hébergement des données, serveurs localisés en France / Europe (Cloud privé/public, centralisé/distribué) (conformité HDS) ▪ Cybersécurité à l'état de l'art	▪ Traitement des données manquantes (réduction des biais) ▪ Stratégie de rééquilibrage des populations minoritaires (réduction des biais) ▪ Ethique de la séparation des données (deux jeux totalement distincts, un échantillon pour l'apprentissage et un échantillon pour l'évaluation) et représentativité des deux jeux de données par rapport à la population cible et la finalité du traitement)	▪ Choix de l'algorithme d'apprentissage en adéquation avec la finalité ▪ Qualité de l'algorithme ○ Identification et élimination des biais ○ Correction des erreurs ▪ Transparence ○ Traçabilité de la démarche de construction de l'algorithme ○ Traçabilité des traitements (rendre les codes sources publics avec protection par dépôt (APP)) ○ Explicabilité des résultats explicables, processus d'auditabilité des résultats non explicables ▪ Définition d'indicateurs de dérive du système ▪ Mise en œuvre de mécanismes d'adaptabilité ○ Évolution réglementaire ○ Avancées médicales ▪ Versionning/maintenance	▪ Evaluation (externe) ○ Technique (bugs), clinique (gold standard, score de précision) ○ Utilisabilité (PS, patients, usagers) ○ Non-discrimination/équité ○ Robustesse/reproductibilité ▪ Procédures en cas de cyber-attaques (analyse d'impact sur la sécurité du système d'IA) ▪ Information (juste et égalitaire) des utilisateurs (PS, patients) ○ Finalité, gouvernance, responsabilité ○ Architecture ○ Origine des données et qualité (légalité de la collecte et des traitements) ○ Explication des processus, explication du périmètre de la partie non explicable ○ Méthode d'apprentissage, d'inférence, etc. ○ Définition des limites de l'utilisation de l'algorithme (FP, FN si classification) ○ Recours en cas d'erreurs ○ Implication des utilisateurs ▪ Garantie humaine (PS, équipe de soins)

▪ Garantie de non-réutilisation non éthique des données (par ex. en cas de fusion de la société / modifications législatives (pouvant aller jusqu'à la destruction automatique des données)) ▪			○ Contrôle par l'humain de l'IA ○ Autonomie décisionnelle ○ Maintien des compétences des utilisateurs ○ Intervalle de confiance de l'IA / garde-fou des erreurs de l'IA ○ Audits (désaccords IA / PS) ▪ Instance de régulation ○ Audit, Label Ethique-IA ▪ Impact organisationnel sur le parcours de soins ▪ Analyse d'impact environnemental ▪ Analyse d'impact sociétal
Analyse des risques (en continu)			
Implication des utilisateurs dans le développement et le design de la solution d'IA			

Liste des critères d'évaluation du groupe d'experts de haut niveau de la commission européenne (outil ALTAI)

- Action humaine et contrôle humain
    - Droits fondamentaux :
        - Dans les cas d'utilisation susceptibles d'entraîner des effets négatifs sur les droits fondamentaux, avez-vous réalisé une analyse d'impact sur les droits fondamentaux ? Avez-vous déterminé et documenté le recours potentiel à des arbitrages entre les différents principes et droits ?
        - Le système d'IA interagit-il avec la prise de décision par un utilisateur final humain (par exemple, en recommandant des mesures ou décisions à prendre, ou en présentant des choix possibles)?
            - Dans de tels cas, existe-t-il un risque que le système d'IA affecte l'autonomie humaine en interférant de manière involontaire avec le processus décisionnel de l'utilisateur final?
            - Estimez-vous qu'un système d'IA devrait communiquer aux utilisateurs qu'une décision, un contenu, un conseil ou un résultat découlent d'une décision algorithmique?
            - Lorsque le système d'IA comporte un robot ou système conversationnel, les utilisateurs humains sont-ils informés du fait qu'ils interagissent avec un agent virtuel?
    - Action humaine :
        - Lorsque le système d'IA est intégré dans un processus de travail, avez-vous réfléchi à la répartition des tâches entre le système d'IA et les travailleurs humains pour permettre des interactions constructives ainsi qu'une supervision et un contrôle humains appropriés ?
        - Le système d'IA renforce-t-il ou augmente-t-il les capacités humaines ?
        - Avez-vous prévu des garanties pour empêcher toute confiance ou dépendance excessives envers le système d'IA dans les processus de travail ?
    - Contrôle humain :
        - Avez-vous réfléchi au niveau approprié de contrôle humain pour le système d'IA et le cas d'utilisation en question ?
        - Pouvez-vous décrire le niveau de contrôle ou de participation humains, le cas échéant ? Qui est «l'humain aux manettes» et à quel moment y a-t-il intervention humaine, ou avec quels outils ?
        - Avez-vous mis en place des mécanismes et des mesures pour garantir un contrôle ou une supervision humains potentiels de cette nature, ou pour veiller à ce que les décisions soient prises sous la responsabilité globale d'êtres humains ?
        - Avez-vous pris des mesures pour permettre la réalisation d'audits et résoudre des questions liées à la gouvernance de l'autonomie de l'IA ?
    - Dans le cas d'un système d'IA ou d'une utilisation capables d'auto-apprentissage ou autonomes, avez-vous mis en place des mécanismes plus spécifiques de contrôle et de supervision ?
        - Quel type de mécanismes de détection et de réponse avez-vous mis sur pied pour évaluer le risque que des problèmes surviennent ?

        - Avez-vous veillé à la présence d'un «bouton d'arrêt» ou à l'existence d'une procédure pour suspendre, en cas de besoin, une opération en toute sécurité? Cette procédure suspend-elle le processus dans sa totalité, en partie, ou délègue-t-elle le contrôle à un être humain ?
- Robustesse technique et sécurité
    - Résilience aux attaques et sécurité :
        - Avez-vous évalué des formes d'attaques potentielles auxquelles le système d'IA pourrait être vulnérable ?
            - Avez-vous en particulier envisagé différents types et différentes natures de vulnérabilités, comme la pollution des données, l'infrastructure physique, les cyber attaques ?
        - Avez-vous prévu des mesures ou systèmes pour veiller à l'intégrité et à la résilience du système d'IA face à de potentielles attaques ?
        - Avez-vous évalué le comportement de votre système dans des situations ou des environnements imprévus ?
        - Avez-vous envisagé si, et dans quelle mesure, votre système pourrait avoir un double usage ? Le cas échéant, avez-vous pris des mesures préventives appropriées contre un tel cas de figure (y compris, par exemple, ne pas publier la recherche ou ne pas déployer le système)?
    - Solutions de secours et sécurité générale :
        - Avez-vous veillé à ce que votre système dispose de suffisamment de solutions de secours pour faire face à d'éventuelles attaques antagonistes ou autres situations imprévues (par exemple, procédures de relais technique ou demande de communication avec un opérateur humain avant d'agir) ?
        - Avez-vous envisagé le niveau de risque posé par le système d'IA dans ce cas d'utilisation spécifique ?
            - Avez-vous mis en place un processus pour mesurer et évaluer les risques et la sécurité ?
            - Avez-vous fourni les informations nécessaires en cas de risque pour l'intégrité physique humaine ?
            - Avez-vous réfléchi à une politique d'assurance pour couvrir les dégâts potentiels provoqués par le système d'IA ?
            - Avez-vous recensé les risques potentiels en matière de sécurité d'(autres) utilisations prévisibles de la technologie, y compris d'utilisation abusive accidentelle ou malveillante ? Existe-t-il un plan pour atténuer ou gérer ces risques ?
        - Avez-vous évalué s'il est probable que le système d'IA cause des dommages ou préjudices aux utilisateurs ou à des tiers ? Le cas échéant, avez-vous évalué la probabilité, les dommages potentiels, le public concerné et la gravité ?
            - En cas de risque qu'un système d'IA cause des dommages, avez-vous réfléchi à des règles de responsabilité et de protection des consommateurs, et de quelle manière en avez-vous tenu compte ?
            - Avez-vous réfléchi à l'incidence potentielle ou au risque en matière de sécurité sur l'environnement ou les animaux ?

- Vous êtes-vous demandé, dans le cadre de votre analyse des risques, si des problèmes de sécurité ou de réseau (par exemple, des menaces pesant sur la cyber sécurité) pourraient mettre en péril la sécurité ou entraîner des préjudices du fait d'un comportement involontaire du système d'IA ?

- Avez-vous évalué l'incidence probable d'une défaillance de votre système d'IA entraînant la production de résultats erronés, l'indisponibilité de votre système, ou la production de résultats inacceptables pour la société (par exemple, pratiques discriminatoires) ?
  - Avez-vous mis en place des seuils et une gouvernance pour les scénarios ci-dessus afin de déclencher d'autres plans/solutions de secours ?
  - Avez-vous défini et testé des solutions de secours ?

- Précision
  - Avez-vous évalué le niveau de précision et la définition de la précision nécessaires dans le contexte du système d'IA et du cas d'utilisation concerné ?
    - Avez-vous réfléchi à la manière dont la précision est mesurée et assurée ?
    - Avez-vous mis en place des mesures pour veiller à ce que les données utilisées soient exhaustives et à jour ?
    - Avez-vous mis en place des mesures pour évaluer si des données supplémentaires sont nécessaires, par exemple pour améliorer la précision et éliminer les biais ?
  - Avez-vous évalué le préjudice que causeraient des prédictions inexactes du système d'IA ?
  - Avez-vous prévu des moyens de mesurer si votre système produit un nombre inacceptable de prédictions inexactes ?
  - En cas de prédictions inexactes, avez-vous mis en place une série d'étapes pour résoudre le problème ?

- Fiabilité et reproductibilité :
  - Avez-vous mis en place une stratégie afin de contrôler le système d'IA et de vous assurer qu'il répond aux objectifs, aux finalités et aux applications prévues ?
    - Avez-vous vérifié si des contextes spécifiques ou conditions particulières doivent être pris en compte pour garantir la reproductibilité ?
    - Avez-vous mis en place des processus ou méthodes de vérification pour mesurer et garantir les différents aspects de la fiabilité et de la reproductibilité ?
    - Avez-vous mis en place des processus visant à décrire certains réglages susceptibles d'entraîner une défaillance du système d'IA ?
    - Avez-vous clairement documenté et appliqué ces processus aux fins des essais et de la vérification de la fiabilité du système d'IA ?
    - Avez-vous mis en place un mécanisme ou une communication pour garantir aux utilisateurs (finaux) la fiabilité du système d'IA ?

- Respect de la vie privée et gouvernance des données

- Respect de la vie privée et protection des données :
  - En fonction du cas d'utilisation, avez-vous mis sur pied un mécanisme permettant à autrui de signaler des problèmes en rapport avec le respect de la vie privée et la protection des données durant les processus suivis par le système d'IA pour la collecte des données (aux fins de l'entraînement et du fonctionnement) et le traitement des données ?
  - Avez-vous évalué le type et la portée des données constituant vos ensembles de données (par exemple, si elles contiennent des données à caractère personnel) ?
  - Avez-vous réfléchi à des manières de mettre au point le système d'IA ou d'entraîner le modèle sans utiliser (ou en utilisant de manière limitée) des données potentiellement sensibles ou à caractère personnel ?
  - Avez-vous intégré des mécanismes de notification et de contrôle concernant les données à caractère personnel en fonction du cas d'utilisation (comme un consentement valable et la possibilité de révoquer le consentement, le cas échéant) ?
  - Avez-vous pris des mesures pour renforcer le respect de la vie privée, par exemple des mesures de cryptage, d'anonymisation et d'agrégation ?
  - Lorsqu'il existe un responsable de la protection des données, avez-vous mobilisé cette personne à un stade précoce dans le processus ?
- Qualité et intégrité des données :
  - Avez-vous aligné votre système sur d'éventuelles normes pertinentes (par exemple, ISO, IEEE) ou des protocoles largement adoptés dans le cadre de votre gestion et de votre gouvernance quotidiennes des données ?
  - Avez-vous mis sur pied des mécanismes de contrôle pour la collecte, le stockage, le traitement et l'utilisation des données ?
  - Avez-vous évalué la mesure dans laquelle vous contrôlez la qualité des sources externes des données utilisées ?
  - Avez-vous mis en place des processus pour garantir la qualité et l'intégrité de vos données ? Avez-vous envisagé d'autres processus ? De quelle manière vérifiez-vous que vos ensembles de données n'ont pas été compromis ou piratés ?
- Accès aux données :
  - Quels protocoles, processus et procédures ont été suivis pour gérer et garantir la gouvernance appropriée des données ?
    - Avez-vous analysé qui peut accéder aux données des utilisateurs et dans quelles circonstances ?
    - Avez-vous veillé à ce que ces personnes soient qualifiées, qu'elles aient effectivement besoin d'accéder aux données et à ce qu'elles disposent des compétences nécessaires pour comprendre précisément la politique de protection des données ?
    - Avez-vous prévu un mécanisme de contrôle pour consigner quand, où, comment, par qui et dans quel but les données ont été consultées ?

- Transparence
  - Traçabilité :

- Avez-vous mis des mesures en place susceptibles de garantir la traçabilité ? Cela pourrait consister à documenter :
    - Les méthodes appliquées aux fins de la conception et de la mise au point du système algorithmique :
        - Dans le cas d'un système d'IA fondé sur des règles, la méthode de programmation ou la manière dont le modèle a été mis au point devraient être documentées;
        - Dans le cas d'un système d'IA fondé sur l'apprentissage, la méthode d'entraînement de l'algorithme, y compris quelles données d'entrée ont été collectées et sélectionnées, et dans quelles conditions, devrait être documentée.
    - Les méthodes appliquées pour tester et valider le système algorithmique :
        - Dans le cas d'un système d'IA fondé sur des règles, les scénarios ou cas utilisés pour tester et valider devraient être documentés ;
        - Dans le cas d'un système d'IA fondé sur l'apprentissage, les informations relatives aux données utilisées pour tester et valider devraient être documentées.
    - Les résultats du système algorithmique :
        - Les résultats d'un algorithme ou les décisions qu'il prend, ainsi que les éventuelles autres décisions qui résulteraient de différents cas (par exemple, pour d'autres sous-groupes d'utilisateurs) devraient être documentés.

- Explicabilité :
    - Avez-vous évalué la mesure dans laquelle les décisions prises, et donc les résultats obtenus, par le système d'IA peuvent être compris
    - Avez-vous veillé à ce qu'une explication de la raison pour laquelle un système a procédé à un certain choix entraînant un certain résultat puisse être rendue compréhensible pour l'ensemble des utilisateurs qui pourraient souhaiter obtenir une explication ?
    - Avez-vous évalué la mesure dans laquelle la décision du système influence les processus décisionnels de l'organisation ?
    - Avez-vous évalué pourquoi ce système particulier a été déployé dans ce domaine spécifique ?
    - Avez-vous évalué le modèle économique concernant ce système (par exemple, en quoi crée-t-il de la valeur pour l'organisation) ?
    - Avez-vous conçu le système d'IA en ayant dès le départ l'interprétation à l'esprit ?
        - Avez-vous cherché à utiliser le modèle le plus simple et le plus facile à interpréter pour l'application en question ?
        - Avez-vous évalué si vous êtes en mesure d'analyser les données que vous avez utilisées aux fins de l'entraînement et des essais ? Cela peut-il être modifié et actualisé au fil du temps ?
    - Avez-vous évalué si des solutions s'offrent à vous suite à l'entraînement et à la mise au point du modèle pour examiner l'interprétation ou si vous avez accès à la séquence des opérations du modèle ?

- Communication :
  - Avez-vous informé les utilisateurs (finaux) – au moyen d'une clause de non-responsabilité ou de tout autre moyen – qu'ils interagissent avec un système d'IA et pas avec un autre être humain ? Avez-vous indiqué clairement que votre système est doté de l'IA ?
  - Avez-vous mis en place des mécanismes pour informer les utilisateurs des raisons et critères expliquant les résultats du système d'IA ?
    - Les utilisateurs visés en sont-ils informés de manière claire et intelligible ?
    - Avez-vous établi des processus pour tenir compte des commentaires des utilisateurs et utiliser ces commentaires pour adapter le système ?
    - Avez-vous également communiqué les risques potentiels ou perçus, tels que les biais ?
    - En fonction du cas d'utilisation, avez-vous également réfléchi à la communication et à la transparence envers d'autres publics, des tiers ou le grand public ?
  - Avez-vous clairement indiqué la finalité du système d'IA et qui ou ce qui pourrait bénéficier du produit/service ?
    - Les scénarios d'utilisation du produit ont-ils été définis et clairement expliqués, en envisageant également d'autres moyens de communication pour veiller à ce qu'ils soient compréhensibles et appropriés pour le destinataire ?
    - En fonction du cas d'utilisation, avez-vous réfléchi à la psychologie humaine et aux potentielles limites humaines, comme le risque de confusion, les biais de confirmation ou la fatigue cognitive ?
  - Avez-vous clairement expliqué les caractéristiques, les limites et les éventuelles lacunes du système d'IA :
    - S'agissant de la mise au point : à toute personne chargée de son déploiement pour en faire un produit ou service?
    - S'agissant du déploiement : à l'utilisateur final ou au consommateur?
- Diversité, non-discrimination et équité
  - Éviter les biais injustes :
    - Avez-vous prévu une stratégie ou un ensemble de procédures pour éviter de créer ou de renforcer des biais injustes dans le système d'IA, en ce qui concerne tant l'utilisation des données d'entrée que la conception de l'algorithme ?
      - Avez-vous évalué et reconnu les éventuelles limites provenant de la composition des ensembles de données utilisés ?
      - Avez-vous réfléchi à la diversité et à la représentativité des utilisateurs dans les données ? Avez-vous procédé à des essais portant sur des populations spécifiques ou des cas d'utilisation problématiques ?
      - Avez-vous recherché et utilisé les outils techniques disponibles pour améliorer votre compréhension des données, du modèle et de la performance ?

        - Avez-vous mis en place des processus pour tester et contrôler les biais éventuels au cours de la phase de mise au point, de déploiement et d'utilisation du système ?
    - En fonction du cas d'utilisation, avez-vous prévu un mécanisme permettant à autrui de signaler des problèmes liés aux biais, à la discrimination ou aux mauvaises performances du système d'IA ?
        - Avez-vous envisagé des mesures et des moyens de communication clairs pour savoir comment et à qui ces problèmes peuvent être signalés ?
        - Avez-vous tenu compte non seulement des utilisateurs (finaux) mais également des autres personnes susceptibles d'être indirectement affectées par le système d'IA ?
    - Avez-vous évalué si, dans des conditions identiques, une éventuelle variabilité des décisions est possible ?
        - Le cas échéant, avez-vous réfléchi aux causes probables ?
        - Concernant la variabilité, avez-vous mis sur pied un mécanisme de mesure ou d'évaluation de l'incidence potentielle de cette variabilité sur les droits fondamentaux ?
    - Avez-vous prévu une définition appropriée de l'«équité» que vous appliquez dans la conception des systèmes d'IA ?
        - Votre définition est-elle couramment utilisée ? Avez-vous envisagé d'autres définitions avant de choisir celle-ci ?
        - Avez-vous prévu une analyse quantitative ou des indicateurs pour mesurer et tester la définition appliquée de l'équité ?
        - Avez-vous mis sur pied des mécanismes visant à garantir l'équité dans vos systèmes d'IA ? Avez-vous envisagé d'autres mécanismes potentiels ?
- Accessibilité et conception universelle :
    - Avez-vous veillé à ce que le système d'IA réponde aux besoins d'un large ensemble de préférences et de capacités individuelles ?
        - Avez-vous évalué si le système d'IA peut être utilisé par les personnes présentant des besoins spécifiques ou un handicap ou qui sont exposées au risque d'exclusion ? Comment cet aspect a-t-il été intégré à la conception du système et comment est-il vérifié ?
        - Avez-vous veillé à ce que les informations relatives au système d'IA soient également accessibles aux utilisateurs de technologies d'assistance ?
        - Avez-vous mobilisé ou consulté cette communauté d'utilisateurs au cours de la phase de mise au point du système d'IA ?
    - Avez-vous tenu compte de l'incidence de votre système d'IA sur le groupe d'utilisateurs potentiels ?
        - L'équipe participant à la mise au point du système d'IA est-elle représentative de votre groupe cible d'utilisateurs ? Est-elle représentative de la population au sens large, compte tenu également d'autres groupes susceptibles d'être indirectement concernés ?

- Avez-vous évalué si certaines personnes ou certains groupes pourraient subir de manière disproportionnée des effets négatifs ?
- D'autres équipes ou groupes présentant différents parcours professionnels et expériences vous ont-ils fait parvenir des réactions ?

- Participation des parties prenantes :
  - Avez-vous réfléchi à un mécanisme pour inclure la participation de différentes parties prenantes dans la mise au point et l'utilisation du système d'IA ?
  - Avez-vous préparé la voie à l'introduction du système d'IA au sein de votre organisation en informant et en mobilisant au préalable les travailleurs concernés et leurs représentants ?

- Bien-être sociétal et environnemental
  - IA durable et respectueuse de l'environnement :
    - Avez-vous mis en place des mécanismes pour mesurer l'impact environnemental de la mise au point, du déploiement et de l'utilisation du système d'IA (par exemple, énergie consommée par les centres de données, type d'énergie consommée par les centres de données, etc.)?
    - Avez-vous prévu des mesures pour réduire l'impact environnemental du cycle de vie de votre système d'IA ?
  - Incidence sociale :
    - Lorsque le système d'IA interagit directement avec des êtres humains:
      - Avez-vous évalué si le système d'IA encourage les êtres humains à développer de l'attachement et de l'empathie pour le système ?
      - Avez-vous veillé à ce que le système d'IA indique clairement que son interaction sociale est simulée et qu'il n'a nullement la capacité de «comprendre» et de «ressentir»?
    - Avez-vous veillé à ce que les incidences sociales du système d'IA soient bien comprises ? Par exemple, vous êtes-vous demandé s'il existe un risque de perte d'emplois et de perte de compétences de la main-d'œuvre ? Quelles mesures ont été prises pour contrer ces risques ?
  - Société et démocratie :
    - Avez-vous évalué l'incidence plus large de l'utilisation du système d'IA sur la société, au-delà de l'utilisateur (final) individuel, par exemple les parties prenantes susceptibles d'être indirectement concernées ?

- Responsabilité
  - Auditabilité :
    - Avez-vous mis en place des mécanismes pour faciliter l'auditabilité du système par des acteurs internes et/ou indépendants, en veillant pas exemple à la traçabilité et à la journalisation des processus et des résultats du système d'IA ?
  - Minimisation et documentation des incidences négatives :
    - Avez-vous réalisé une analyse des risques ou de l'impact du système d'IA qui tienne compte de différentes parties prenantes qui sont directement et indirectement concernées ?
    - Avez-vous mis en place des cadres de formation et d'éducation pour définir des pratiques en matière de responsabilité ?

- Quels travailleurs ou branches de travailleurs sont concernés ? Le sont-ils au-delà de la phase de mise au point ?
- Ces formations portent-elles également sur le cadre juridique potentiellement applicable au système d'IA ?
- Avez-vous envisagé la mise sur pied d'un « comité d'examen pour l'IA éthique» ou d'un mécanisme similaire pour discuter des pratiques globales en matière de responsabilité et d'éthique, y compris des zones grises potentiellement floues?

▪ Outre les initiatives ou cadres internes destinés à contrôler l'éthique et la responsabilité, des orientations externes ou des processus d'audit ont-ils également été mis en place ?
▪ Existe-t-il des processus permettant aux tiers (par exemple, fournisseurs, consommateurs, distributeurs/vendeurs) ou aux travailleurs de signaler de possibles vulnérabilités, risques ou biais dans le système/l'application d'IA ?

o Documentation des arbitrages :
▪ Avez-vous mis sur pied un mécanisme permettant de recenser les intérêts et les valeurs pertinents concernés par le système d'IA et les éventuels arbitrages entre eux ?
▪ Quel processus utilisez-vous pour prendre des décisions relatives à ces arbitrages ? Avez-vous veillé à ce que les décisions d'arbitrage soient documentées ?

o Voies de recours :
▪ Avez-vous mis en place un ensemble approprié de mécanismes permettant un recours en cas de préjudice ou d'effet néfaste ?
▪ Avez-vous mis en place des mécanismes pour fournir des informations aux utilisateurs (finaux)/tiers à propos des possibilités de recours ?