Recherches en Intelligence Artificielle et Santé: État de l'art et perspectives

Malik Ghallab

Recherches en Intelligence Artificielle & Santé : État de l'art et perspectives

Malik Ghallab
LAAS–CNRS, Université de Toulouse

Résumé

Cet article propose une synthèse de l'état de l'art en Intelligence Artificielle (IA) pour la santé, et évoque des perspectives de recherche sur des questions interdisciplinaires touchant aux deux domaines. Il ne prétend pas répondre aux besoins légitimes du praticien de connaître quels outils seraient directement utilisables pour quelles tâches médicales.

Dans ce qui suit, on introduit brièvement l'IA, ses méthodes et développements récents, illustrés dans le domaine de la médecine. On fait le point sur l'état de l'art et les problèmes ouverts pour l'IA dans l'aide au diagnostic et à l'action thérapeutique. On évoque enfin les risques et questions éthiques relatives à l'IA pour la santé.

1 Introduction

La médecine est une des premières sciences de l'humanité. Les techniques médicales, comme celles d'autres domaines, ont précédé de longtemps la science médicale. Elles remontent à la préhistoire [1], bien avant l'écriture, sans laquelle il n'y a pas de science. Aujourd'hui la santé s'appuie sur un très vaste champ interdisciplinaire couvrant, entre autres :

- les sciences de la vie : biologie, anatomie, physiologie, pharmacologie, génétique, épidémiologie ;
- les sciences de la matière : biochimie, biomécanique, physique, chimie ;
- les sciences humaines et sociales : psychologie, sociologie, économie ;
- les sciences de la modélisation et de l'information, qui restent parfois moins présentes dans les formations médicales.

Pour notre propos, la médecine se caractérise en particulier par :

1. L'archéologie documente des instruments chirurgicaux de suture de l'âge de pierre, et des interventions complexes suivies de guérisons, telles que des trépanations, qui remontent au proto-néolithique, il y a plus de 12.000 ans.

- le rôle critique des relations humaines dans l'action thérapeutique ;
- les interdépendances fortes entre divers domaines de savoir, et le caractère essentiellement intégratif de la médecine, peu compatible avec une approche réductrice ;
- la très grande complexité des données et de l'imagerie médicales, et la technicité croissante de la médecine ;
- les exigences éthiques élevées de la recherche et de l'action médicale.

Les apports possible de l'intelligence artificielle (IA) dans le domaine de la santé doivent prendre en compte ce qui précède et s'inscrire dans le cadre de défis sociaux et techniques, dont principalement :

- une médecine pour toute la société, accessible à tous, qui vise autant sinon davantage à prévenir et anticiper qu'à guérir ;
- une médecine personnalisée, factuelle et de précision.

Les contributions de l'IA pour la santé devraient s'attacher, selon nous, à conjuguer trois finalités essentielles :

- répondre aux besoins de la population, des patients et des praticiens pour améliorer l'état de la santé publique ;
- contribuer aux connaissances médicales à tous les niveaux, depuis celles fondamentales, jusqu'à la caractérisation et la qualification de nouvelles méthodes et techniques ;
- mieux répondre aux besoins de formation initiale et continue des médecins, du personnel de santé, mais aussi du grand public.

Ces deux derniers points nous semblent aussi importants que le premier. On constate que l'IA est avant tout un amplificateur du développement des savoirs dans tous les domaines. Pour la santé, l'IA est un atout pour le développement des sciences médicales autant qu'un instrument amplifiant les techniques médicales. L'IA ouvre également un accès très large aux savoirs, via de riches modalités d'interrogation et d'interaction avec des bases de connaissances, des moyens de simulation, de "jeux sérieux", et de réalité virtuelle.

Les trois sections suivantes présentent brièvement l'intelligence artificielle, les fonctions cognitives qu'elle étudie, ainsi que ses méthodes, à base de connaissances et à base de données. La section 5 présente des travaux très récents et des perspectives de recherche en IA & Santé dans l'aide au diagnostic et à l'action thérapeutique. On évoque ensuite les risques et problèmes éthiques d'un tel programme, avant de conclure.

2 Qu'est ce que l'Intelligence Artificielle

L'IA est une discipline scientifique qui cherche à comprendre et à modéliser l'intelligence naturelle des êtres vivants par des approches com-

putationnelles ou algorithmiques [2]. L'objet de l'IA est donc l'intelligence, laquelle est mal définie et multiforme. Cependant on peut constater des manifestations de l'intelligence – animale ou humaine – dans la réalisation de tâches. La démarche de l'IA consiste alors à mécaniser des tâches de plus en plus complexes qui requièrent de l'intelligence. Il s'agit, par exemple, d'interpréter une scène, de manipuler des outils, de planifier des actions (comme y excellent de nombreuses espèces), ou d'apprendre à compter, lire, écrire et comprendre des textes (les fondamentaux de nos écoles). Pour ce faire, on modélise la tâche en question, on cherche des méthodes et des algorithmes pour la résoudre, et on développe des implémentations logicielles et matérielles de ces algorithmes. Ensuite, on évalue empiriquement les capacités des modèles et algorithmes et leurs performances pour cette tâche, et on s'efforce de les améliorer.

L'IA relève des sciences de la modélisation et de l'information, avec des recouvrements vers les sciences cognitives et les sciences de la décision (recherche opérationnelle, économie). C'est un domaine relativement récent, même si ses racines sont très anciennes [3]. Depuis le milieu du 20^e siècle, l'IA bénéficie et contribue significativement aux développements des matériels et logiciels informatiques et de télécommunication. Elle a transformé considérablement nos capacités de modélisation, d'analyse, d'organisation et de recherche d'informations. Elle a d'ores et déjà un impact important sur pratiquement tous les champs d'investigation scientifiques et techniques, dont ceux relatifs à la santé.

L'IA, comme toute *technoscience* aujourd'hui, est aussi un champ technologique, difficilement séparable du volet scientifique. Le savoir y est motivé par l'action. L'IA alimente de nombreuses industries en techniques et déploiements, que l'on espère socialement utiles. Les investissements de R&D en IA sont aujourd'hui considérables [4]. Ils sont très massivement dominés par quelques multinationales, lesquelles contrôlent ainsi l'évolution du domaine selon des logiques capitalistes plutôt sociales (cf. section 6).

Les recherches en IA & Santé visent à contribuer aux sciences médicales et à la pratique clinique, pour une meilleure maîtrise de la complexité du domaine, vers une médecine intégrative, personnalisée, préventive et socialement efficiente. Cette recherche interdisciplinaire se traduit par une activité de publication dans pratiquement toutes les revues scientifiques des deux domaines, et dans des revues spécialisées [5].

En efforts de R&D, la santé est un des premiers champs d'application

2. Privilégier des approches algorithmiques n'est pas un biais. L'humanité tente de comprendre et donc de modéliser le monde d'abord en nommant et décrivant, puis en représentant mathématiquement ce qui peut l'être, et aujourd'hui par modèles algorithmiques.

3. Pour une histoire de l'IA voir par exemple [73].

4. Ils sont estimés entre 100 et 200 milliards de dollars par an.

5. En particulier, la revue *Artificial Intelligence in Medicine* qui existe depuis 1989.

de l'IA. À côté de quelques succès, on note moult prédictions fortement exagérées [6]. On sait en effet qu'il existe un long parcours depuis les premiers prototypes jusqu'aux déploiements médicaux effectifs [7]. Ce parcours passe par des procédures de certification exigeantes. Des preuves de fiabilité sont indispensables : l'erreur humaine est inacceptable si elle est le fait d'une machine. Par ailleurs, il ne suffit pas qu'un système excelle dans une tâche étroite pour qu'il apporte une valeur ajoutée à la santé publique.

Cependant, l'accélération du nombre de systèmes d'IA certifiés par des agences internationales de santé démontre une maturité croissante de certaines applications [8]. Elle confirme l'importance de ce champ interdisciplinaire pour le développement d'une santé publique efficiente.

3 Fonctions cognitives étudiées en IA

L'IA étudie l'intelligence en mécanisant des tâches. On peut, en simplifiant, classer ces tâches à mécaniser en les cinq fonctions cognitives suivantes : *percevoir, agir, interagir, raisonner*, et *apprendre*.

Ces tâches sont plus riches et plus complexes pour une machine dotée de capacités sensori-motrices. C'est le cas de robots contrôlant des capteurs pour percevoir, et des actionneurs pour agir et interagir, avec des boucles de rétroaction sur la perception et l'action. La rétroaction est également essentielle pour raisonner et apprendre lorsque la machine doit faire face à des tâches et des environnements variables. Ces tâches sont présentes dans des applications d'aide chirurgicale, de réadatation physiologique, ou de dispositifs d'assistance physique (exosquelettes) pour personnes handicapées. Focalisons nous ici principalement sur des problèmes de "machines logicielles", sans capacités sensori-motrices propres et sans autonomie. Ces problèmes sont plus simples et davantage étudiés en IA & Santé. Illustrons brièvement les cinq fonctions cognitives étudiées en IA ; plusieurs points seront détaillés ultérieurement en sections 4 et 5.

3.1 Percevoir

La perception en IA consiste à analyser des données et à leur associer un sens. Dans les cas les plus simples, ceci se ramène à la classification de données selon des catégories prédéfinies. Par exemple, un enregistrement ECG particulier est classé "fibrillation ventriculaire" [67] ; une image cutanée est classée "kératose actinique" [33]. Il s'agit généralement d'un simple

6. Par exemple, il n'y aurait plus besoin de radiologues, selon un pionnier des réseaux de neurones.

7. Par exemple, MYCIN, un prototype d'aide au diagnostic en infectiologie, remonte à 1976 [101], mais le déploiement de tels systèmes est très récent

8. Plus de 700 systèmes d'IA certifiés par la FDA, l'organisme américain de certification, dont 170 pour la seule année 2023.

étiquetage, sans compréhension du sens des étiquettes. Cette classification de données d'un état médical est aujourd'hui relativement bien résolue. Plus compliquée est la perception d'une série spatio-temporelle de signaux à corréler pour en extraire une interprétation quantitative. C'est par exemple l'analyse de plusieurs scanners crâniens d'un patient, étalés dans le temps, pour y localiser, détecter les évolutions, et mesurer en 3D toute trace d'anomalie pathologique.

Par ailleurs, les dossiers médicaux comportent naturellement des données très hétérogènes : textes, sons, images, observations et signaux de toutes sortes. La perception du praticien prend en compte l'apparence d'un patient, le timbre de sa voix, ses paroles et son comportement, ainsi que moult autres informations et signaux de l'examen clinique. Peut-on intégrer ces informations très riches dans un système d'IA ?

La perception des données hétérogènes en IA soulève des problèmes difficiles de fusion multisensorielle nécessitant de conjuguer diverses méthodes. Elle soulève également des difficultés pour la gestion de l'incertitude présente dans tous les problèmes de perception. L'incertitude est due au bruit des capteurs pour acquérir l'information à interpréter. Elle est également due aux modèles anatomiques et physiologiques sur lesquels se base implicitement ou explicitement la perception : ces modèles ne sont pas déterministes. S'il est peu fréquent qu'un rapport de radiologie fournisse la marge d'erreur ou le degré de confiance à accorder à ses conclusions, la perception avec des techniques d'IA se doit de qualifier l'incertitude des interprétations proposées. Enfin, ces interprétations nécessitent des justifications et des explications. Interpréter une imagerie cérébrale comme annonçant une évolution dans quelques années vers une maladie Alzheimer serait peu crédible sans justification [80, 108].

3.2 Agir

L'action en IA, dans le contexte de la médecine de précision personnalisée, est généralement une aide à la décision sur les actions thérapeutiques les plus indiquées et leur planification. Dans de nombreux cas, par exemple pour des pathologies virales, en chimiothérapie, voire en hypertension artérielle, plusieurs molécules ou principes actifs sont potentiellement pertinents. Appliquer le même protocole pour tous, ou tâtonner par essai et erreur pour trouver la bonne combinaison adaptée à un cas spécifique, n'est pas satisfaisant. Parfois, un modèle non déterministe des effets possibles de chaque action est disponible. Ce modèle permet de planifier à *horizon glissant*. Ceci signifie que, périodiquement, une politique optimale pour l'état courant d'un patient particulier est planifiée ; la première étape de cette politique est appliquée ; après quoi l'état du patient est de nouveau observé (par analyse biologique et autres examens) ; un nouveau plan optimal est calculé ; on en applique de nouveau la première étape, etc. Cette approche,

avec mise à jour hebdomadaire, a été mise en oeuvre par exemple pour le HIV [15, 2].

La planification d'actions chirurgicales complexes (e.g., traumatologie, cancérologie, neurochirurgie) est un autre champ d'investigation très actif. Cette planification est en particulier associée à des actions de positionnements anatomiques pré-opératoires précis, et de localisation de ces positions prédéterminées en cours d'intervention (indispensable pour les tissus non rigides : une tumeur cérébrale, peu visible à l'oeil nu, peut se décaler relativement au crâne de plus de 1 cm [49]). Les mêmes techniques peuvent être mises en oeuvre pour simuler l'intervention et entrainer le chirurgien sur les phases critiques.

Notons également l'action continue avec rétroaction en boucle fermée sur un monitoring médical, par exemple en anesthésie.

3.3 Interagir

L'interaction en IA fait souvent référence à la participation d'un système à une activité collaborative, pour la coordination et la résolution commune d'un problème. Elle nécessite la communication avec des humains. Un premier problème est la communication en langage naturel. C'est par exemple l'interrogation orale d'un système de gestion de données médicales sur le dossier d'un patient particulier, ou sur des analyses factorielles pour une cohorte. C'est la communication orale avec traduction simultanée, e.g., pour un soignant qui ne parle pas la langue du patient. C'est également les conseils téléphoniques à un patient pour le suivi de son ordonnance, ou les conseils donnés à un soignant isolé sur un cas médical qui lui est peu familier. Dans tous ces cas, on cherche à interpréter les questions posées, à comprendre, via le dialogue, des requêtes complexes, similaires à celles adressées à un médecin spécialiste, et obtenir des réponses, des justifications et des explications adaptées à l'interrogateur. Les développements récents sur ces tâches, bien qu'encore très couteux et complexes, sont prometteurs (cf. sous-section 5.4).

L'interaction avec un robot, par exemple en chirurgie ou physiothérapie, soulève d'autres problèmes que ceux de communication. Il s'agit de perception et retours sensoriels haptiques, pour le mouvement, le touché et les sensations d'efforts [107, 113]. Enfin l'interaction pour la résolution partagée d'un problème, par exemple de gestion de ressources médicales, soulève des questions de raisonnement et de décision distribuées, où l'IA doit modéliser les autres intervenant et prendre en compte leur spécificités et contraintes.

3.4 Raisonner

Le raisonnement est naturellement présent dans toutes les tâches qui précédent, pour l'interprétation des données, la planification d'actions,

la communication, et l'interaction. Le raisonnement est particulièrement présent dans l'aide au diagnostic médical à base de modèles numériques. On souhaite passer des symptômes aux causes probables ayant pu les produire, avec une quantification de l'incertitude, et des explications étayant le diagnostic proposé. On souhaite également la proposition d'examens complémentaires, si besoin, avec une appréciation de l'amélioration attendue du diagnostic relativement au coût des tests. Ces sujets ont donné lieu à beaucoup de travaux, mais à relativement peu de déploiements [55, 103]. Les systèmes généralement déployés, et les plus simples à mettre en oeuvre, sont des *classifieurs* relativement opaques, qui associent sans raisonnement explicite un état médical à un diagnostic. Ils sont difficilement acceptables pour le praticien. Plus intéressants sont les systèmes à base de connaissances causales capables de raisonnement contrefactuel [95, 81].

3.5 Apprendre

L'apprentissage est essentiel pour toutes les tâches qui précèdent. Il s'agit de faire acquérir par une machine les modèles nécessaires à la perception, à l'action, à l'interaction, et au raisonnement. L'apprentissage *supervisé* synthétise un modèle à partir d'exemples et d'instructions fournies par un tuteur. L'apprentissage par *renforcement* obtient le modèle en répétant la tâche un grand nombre de fois avec des actions différentes et, en fonction des résultats, retenir une méthode qui maximise un critère de performance (métaphoriquement, la *récompense* de l'apprenant).

Idéalement, on aimerait pouvoir apprendre des modèles causaux explicites, capables de fournir des explications, et de quantifier l'incertitude des conclusions obtenues à l'aide des modèles appris. Pour une IA de confiance, on souhaite que ces modèles soient vérifiables et certifiables.

En pratique, on sait apprendre facilement des modèles statistiques qui ne traitent que de corrélation. De ce fait, l'IA est parfois perçue comme synonyme des méthodes d'apprentissage supervisé par réseaux de neurones, dits profonds, qui ne fournissent que des modèles statistiques opaques. Ces techniques, relativement robustes et faciles à déployer, sont aujourd'hui très répandues. Elles gagnent cependant à être intégrées à des méthodes de raisonnement à base de connaissances et à des méthodes d'apprentissage par renforcement. Sur ce dernier point, par exemple, un système de planification pour le HIV peut être amélioré par renforcement [57, 30]. L'apprentissage par renforcement est exploré, e.g., en diagnostic médical [56], ou pour la planification d'essais cliniques en oncologie [129].

4 Méthodes de l'IA

En simplifiant, on peut distinguer deux catégories de méthodes en IA : celles à base de connaissances et celles à base de données, présentées ci-après.

4.1 Méthodes à base de connaissances

Il s'agit de traduire le savoir humain sur un sujet donné en des formalismes mathématiques et algorithmiques exploitables par une machine, et de les utiliser par des méthodes adaptés pour réaliser des tâches. Ces formalismes sont qualifiés de *représentations des connaissances*. Un exemple bien connu de représentation qui fut d'un apport majeur en mathématique est la notation décimale. En IA, on cherche des représentations très expressives, économes en efforts humains pour exprimer des connaissances, mais efficacement calculables, ce qui nécessite des compromis (les plus expressives étant souvent incalculables). L'IA a développé de nombreuses représentations des connaissances et des algorithmes utilisant en particulier la logique, l'algèbre, la géométrie, les probabilités, la combinatoire ou les graphes. Leur présentation détaillée dépasse le cadre de cet article (pour une introduction générale, voir [97]). Illustrons cependant quelques exemples.

Représentations relationnelles ou ontologiques. Une *ontologie* généralise la notion de *taxonomie* hiérarchique, très populaire dans les sciences de la vie. C'est un graphe dont les noeuds sont les concepts clés d'un domaine et les arcs sont des relations entre ces concepts. Il s'agit par exemple de relations d'inclusion et d'appartenance ensemblistes, avec des mécanismes de transitivité (e.g., $x \in A$ et $A \subset B \subset C$, donc $x \in C$), des relations de similitude et d'incompatibilité, ou des relations spécifiques donnant des symptômes, des tests, etc. (cf. figure 1). Une ontologie de plusieurs milliers de concepts serait illisible pour nous, mais très efficacement exploitable par des algorithmes permettant de répondre à des requêtes multiples (qui est quoi, en relation avec qui). Les ontologies permettent de consolider les concepts de plusieurs spécialités de façon cohérente et facilitent l'interopérabilité sémantique de traitements variés sur des dossiers médicaux [69].

Tout médecin connait (implicitement) une fraction de l'ontologie de son domaine, qui sera généralement distincte de celle d'un autre médecin, et non exploitée systématiquement. Developper une ontologie formelle à partir d'une page blanche demande un effort considérable. Fort heureusement, il existe de nombreuses ontologies médicales dans le domaine public [9], couvrant l'anatomie, la physiologie, la génétique, l'infectiologie, les vaccins, l'oncologie, etc. Les ontologies sont un outil essentiel pour la recherche médicale [16], mais aussi pour la mise en oeuvre et le suivi d'une politique de santé publique.

Représentations logiques. Elles combinent des propositions, des prédicats portant sur des ensembles d'objets, et des formules avec des quantificateurs et des opérateurs logiques (conjonction, implication, etc.)

9. cf par exemple //guides.lib.umich.edu/ontology/ontologies

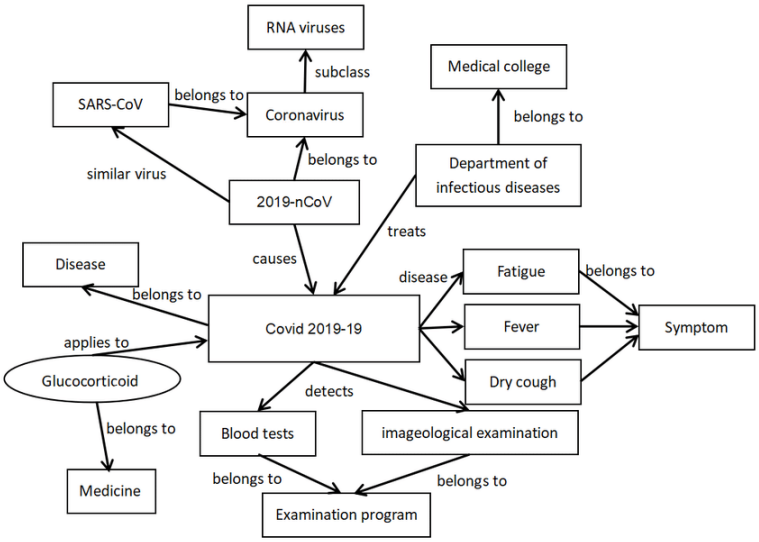


FIGURE 1 – Un exemple simple d'ontologie [122].

pour exprimer formellement des assertions complexes. Les représentations logiques se prêtent à diverses figures de raisonnement, dont la déduction, l'abduction, l'induction, l'analogie, le raisonnement hypothétique, etc.

Le cas le plus simple est celui de la *déduction*, qui s'appuie sur le général pour expliciter le particulier (cf. Exemple 1). L'*abduction* est à la base du diagnostic : partant de relations causales et d'observation de symptômes, on remonte aux hypothèses de leurs causes possibles. L'*induction* logique permet de généraliser, donc d'apprendre. On peut également raisonner sur des changements d'états, spontanés ou résultant d'actions dont on décrit les conditions et les effets possibles, en particulier pour planifier. C'est également pertinent pour faire un raisonnement contrefactuel. On peut enrichir la représentation de diverses modalités pour distinguer, par exemple, ce qui est toujours vrai, ce qui peut l'être possiblement, ce qui l'a été dans le passé ou le sera dans le futur. Cependant les logiques modales ne permettent pas une gestion quantitative des incertitudes. On utilise pour cela des logiques floues ou les probabilités.

Exemple 1. On connait le célèbre syllogisme que tous les hommes sont mortels et que Socrate est un homme : $\forall x\ Homme(x) \rightarrow Mortel(x)$, et $Homme(Socrate)$; on en déduit : $Mortel(Socrate)$.

Ce même mécanisme répond à des questions moins triviales. Soit 3 personnes i, j et k ; i travaille quotidiennement avec j, et j avec k ; k a une infection virale z ; i ne l'a pas. Y-a-il parmi ces trois une personne non infectée qui travaille avec une personne qui l'est ? La réponse à ce problème simple (il n'y a que 2 prédicats) n'est pas toujours intuitive pour un humain

et donne parfois lieu à des erreurs. Un algorithme de déduction automatique y répond, preuve à l'appui, même s'il y a des millions de prédicats et d'objets.

Autre illustration plus réaliste d'interactions moléculaires avec des réactions d'activation et d'inhibition entre diverses substances. Un substance x peut être *présente, activée* ou *inhibée*, dénotée respectivement $P(x), A(x)$ et $I(x)$. Toute substance présente dans un milieu est soit activée soit inhibée, mais pas les deux, ce qui se traduit en : $\forall x\ P(x) \leftrightarrow A(x) \vee I(x)$ et $\neg\exists x\ (A(x) \wedge I(x))$. On exprime ensuite des relations d'activation–inhibition entre deux substances x et y ; puis entre trois substances, e.g., z inhibant la capacité d'activation de x sur y. Des formules permettent d'exprimer des chaines d'activation-inhibition, ou d'introduire des mécanismes d'interaction protéinique spécifiques. Une telle base de connaissances permet de traiter des requêtes sur les substances, présentes ou à introduire dans un milieu, pour obtenir ou inhiber une activité particulière. □

Représentations probabilistes. L'abduction logique permet de remonter des observations vers leurs causes possibles, mais ne quantifie pas leur degré de possibilité. Ceci peut être fait avec des modèles probabilistes. Partant de la probabilité *à priori* de la véracité d'un fait, on veut déterminer sa probabilité *à postériori* compte tenu des symptômes observés. Le calcul repose sur la règle de Bayes, à partir de statistiques sur les liens symptômes-causes (cf. Exemple 2).

La représentation par *réseaux bayésiens* relie en un graphe acyclique un ensemble de variables et leurs distributions de probabilités. Ce graphe exprime les relations de dépendance conditionnelle entre les variables (cf. Exemple 3). Au-delà du calcul des probabilités à postériori, des algorithmes performants permettent de fournir des explications ou de faire des analyses de sensibilités sur de très grands graphes de plusieurs dizaines de milliers de variables. On dispose également de méthodes pour apprendre les paramètres d'un réseau bayésien ou aider à en définir la structure. Le point délicat est lié aux relations de dépendance et aux facteurs de confusion : une variable de confusion influence à la fois la variable dépendante et les variables explicatives ; elle est à la source de la différence entre corrélation et causalité.

Pour l'aide à la décision, un *diagramme d'influence* étend un réseau bayésien à l'aide de variables de décision, ou des actions possibles, et de variables d'utilité ou de coût des actions. Une mise en oeuvre dans l'aide à la prise en charge de maladie pulmonaire en soins intensifs est illustrée dans [70].

Par ailleurs, la prise en compte le temps, par exemple pour l'évolution de l'état d'un patient, peut se faire à l'aide d'un *réseau bayésien dynamique*. Enfin, soulignons que l'extension particulièrement intéressante des *réseaux bayésiens causaux* qui distinguent les dépendances probabilistes des dépendances causales (la probabilité de A *sachant* B n'y est pas synonyme

de la probabilité de A *faisant* B). Les réseaux causaux permettent la mise en oeuvre d'un raisonnement contrefactuel (que se passerait-il si on fait ceci ou cela) permettant l'action délibérée et la planification [88].

Exemple 2. La cause C provoque le symptôme S. La probabilité d'avoir les deux est : $P(C \text{ et } S) = P(C) \times P(S|C) = P(S) \times P(C|S)$, ce qui donne la règle de Bayes pour la probabilité conditionnelle de C sachant S :

$$P(C|S) = P(S|C) \times P(C)/P(S)$$

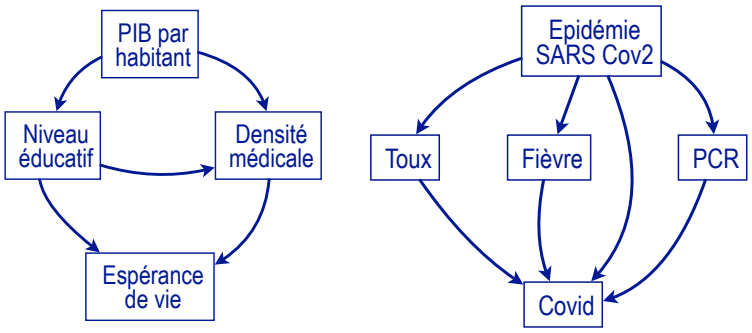
Appliquons ceci pour déterminer la probabilité d'être atteint du Covid sachant comme symptôme un test PCR positif. Supposons une probabilité à priori de 2% (i.e., une période et une zone de forte épidémie), et un test PCR de bonne qualité ayant une sensibilité (ou $P(S|C)$) de 96% et une spécificité (ou $1 - P(S|\text{non } C)$) de 7%. Le calcul simple donne $P(C|S) = 21,8\%$, illustrant que le seul test PCR est loin d'indiquer une grande probabilité, et encore moins la certitude de la maladie. Le même calcul sur d'autres types d'épidémies, par exemple le VIH, donnerait des résultats similaires. □

Exemple 3. Considérons quatre variables (qui interviennent par exemple dans le calcul de l'indice de développement humain) : l'espérance de vie, la richesse moyenne (PIB par habitant), le niveau éducatif moyen et la densité médicale (nombre de médecins par habitant) d'une région. On peut compiler des statistiques sur diverses régions et vérifier que ces variables ne sont pas indépendantes. Mais à elles seules, ces statistiques n'informent pas sur la nature des dépendances. On ne sait pas sur quoi on pourrait agir, par exemple pour augmenter l'espérance de vie, et quelle est l'importance de chaque lien. Le réseau en Figure 2(a), à compléter par de nombreuses autres variables qui influent sur l'espérance de vie, donne quelques réponses. Chaque arc est associé aux tables des probabilités conditionnelles permettant les calculs des probabilités à postériori. À noter que les rétroactions vers le PIB des trois autres variables (qui rendraient le graphe cyclique) ne sont pas prises en compte directement dans ce réseau, mais peuvent l'être par d'autres moyens.

Le calcul sur deux variables C et S pour la Covid donne un résultat faible (cf. Exemple 2). Un calcul plus informatif prend en compte d'autres symptômes, par exemple toux ou fièvre, et divers niveaux d'épidémie. Le réseau bayésien en Figure 2(b) peut être le support de ces calculs. □

Les méthodes à base de connaissances utilisent les représentations mentionnées, conjointement avec d'autres représentations, dont par exemple :

- les représentations à base de *contraintes*, relationnelles ou algébriques ;
- les représentations à base de programmes, dont la *programmation mathématique* (programmation dynamique, programmation linéaire



(a) (b)

FIGURE 2 – (a) Un réseau bayésien sur quatre variables de l'indice de développement humain. Noter que l'espérance de vie est conditionnellement indépendante du PIB *sachant* les deux autres variables, et que la densité médicale dépend du niveau éducatif, mais pas l'inverse. (b) Un réseau bayésien sur quelques variables liées au Covid. Noter que la présence de la maladie dépend dans ce cas de toutes les autres variables.

et en nombres entiers, etc.), la *programmation logique*, et la *programmation probabiliste*.

La programmation probabiliste est plus récente. Elle étend les relations des réseaux bayésiens, qui ne portent que sur des variables, à des relations plus générales sur des prédicats [39, 18]. Elle est très expressive et offre un fort potentiel pour traiter des connaissances médicales.

Les méthodes à base de connaissances ont donné lieu à de nombreux succès. Elles ont des propriétés particulièrement désirables pour une IA de confiance dans le domaine de la santé : elles sont vérifiables et intelligibles à l'utilisateur ; elles peuvent être justifiées sur des bases scientifiques, empiriques et/ou rationnelles ; elles sont capables de fournir des explications convaincantes, et de quantifier l'incertitude des réponses. Mais elles se heurtent cependant à un goulet d'étranglement limitatif : celui de devoir exprimer formellement nos connaissances sur un domaine. Ceci nécessite des efforts conséquents de développement, de modélisation et d'adaptation à chaque tâche. Les systèmes résultants ont tendance à être étroits et spécialisés, difficiles à étendre à de nouvelles tâches [10]. En effet, l'apprentissage automatique de connaissances formelles, ou l'aide à leur acquisition et formalisation (par exemple par la seule lecture des manuels de référence) restent insuffisants. Ceci réduit les capacités de généralisation et de transfert d'une tâche à une autre.

10. Un exemple est le système Watson, champion impressionnant du jeu de questions/réponses "Jeopardy", qui n'a pas pu être transposé avec succès au domaine médical malgré des investissements considérables [36].

Des capacités d'apprentissage et d'adaptation ont pu être atteintes efficacement grâce aux progrès des méthodes à base de données.

4.2 Méthodes à base de données

Ces méthodes s'appuient sur les principes et les hypothèses restrictives de l'induction statistique. Expliquons ce dont il s'agit avant d'en présenter les apports pour l'interprétation et la génération de données à l'aide des réseaux de neurones.

Induction statistique. Considérons un problème de prédiction classique : ayant une séquence de termes $\langle x_1, \ldots, x_{n-1} \rangle$, quel est le terme suivant x_n ? Lorsqu'il s'agit d'une séquence temporelle des états d'un système dont l'évolution est bien modélisée, on aborde ce problème de prédiction grâce aux connaissances sur le système considéré. C'est ce que l'on fait par exemple pour l'évolution d'un processus biochimique, physique, ou météorologique.

Lorsqu'aucun modèle n'est disponible mais que le domaine n'est pas trop erratique, on a recours à la prédiction statistique. Cette approche produit un modèle "superficiel", qui fait des prédictions basées sur des statistiques, et non sur des relations causales. Un tel modèle cherche à prédire le comportement d'un système sans avoir de relations de cause à effet qui engendrent ce comportement et permettent de l'expliquer. Par exemple, on chercherait à prédire l'effet de telle dose d'aspirine sur la tension artérielle d'une personne en utilisant des statistiques sur des cas similaires. Une modélisation des mécanismes d'acétylation et des effets de l'acide acétylsalicylique sur l'inhibition irréversible de plusieurs enzymes nous apporte une compréhension de ce qui se passe, ainsi qu'une validation et une confiance dans une prédiction exploitant les statistiques de façon plus précise.

Ces remarques sur des problèmes de prédiction s'appliquent aux problèmes de classification ou d'interprétation : ayant un état x, quelle est l'interprétation $y = f(x)$ associée. Par exemple x est le vecteur des paramètres d'état d'un patient, y est un élément de diagnostic correspondant. La fonction f est estimée sur des statistiques de couples (x, y) connus. Dans les cas très simples, f est une régression linéaire classique, par exemple x est la quantité d'insuline injectée à un patient diabétique, et y la concentration d'acétoacétate dans son sang.

En science, la prédiction statistique à elle seule est une solution de repli en l'absence de modèle à base de connaissances. Elle suppose que le domaine est régulier. Elle s'appuie sur l'*hypothèse de l'induction statistique*. Cette hypothèse admet que ce qui est vrai sur les données observées reste vrai pour celles non observées. Elle exige réserve et prudence, en particulier en médecine et dans tout domaine où la variabilité est grande, la "normalité" faible, et où peuvent se produire des événements rares aux effets importants.

L'induction statistique, lorsqu'elle est applicable, nécessite l'acquisition d'un nombre important de données pour estimer les distributions servant à prédire. Mais les observations seront toujours finies, alors que l'ensemble des possibles est potentiellement infini.

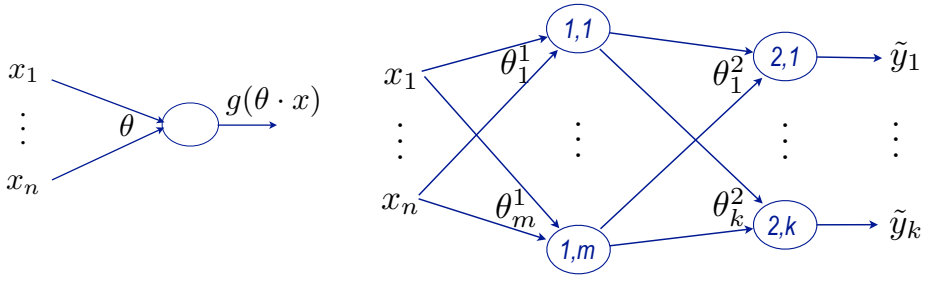
Interprétation et génération de données. Les méthodes à base de données ont conduit à des progrès très significatifs en IA dans deux tâches importantes : l'interprétation et la génération de données de toutes sortes – signaux, textes, images, sons, vidéos, etc.

L'interprétation des données consiste à trouver quelle interprétation $y = f(x)$ associer à une donnée x. C'est le problème classique de la *reconnaissance des formes*. Pour simplifier l'apprentissage de la fonction f, pendant longtemps on s'est efforcé de représenter intelligemment les données plutôt que de les traiter à l'état brut. On a utilisé pour cela des fonctions caractéristiques des données, par exemple des transformées de Fourrier, des splines ou des ondelettes calculées sur un électrocardiogramme. On a consacré des efforts considérables à la conception de ces fonctions caractéristiques pour chaque type de données particulières. Aujourd'hui, grâce aux réseaux de neurones, on se dispense de cette phase de caractérisation des données. Les réseaux permettent d'estimer des fonctions d'interprétation robustes pour toutes sortes de données. Ils fournissent automatiquement, après apprentissage sur des données brutes, les caractéristiques adaptées à ces données. Les méthodes d'interprétation des données ne sont plus coûteuses et spécialisées pour chaque application. Elles sont désormais largement déployées pour l'analyse de données multimodales (signaux physiques, images, sons, vidéos, textes) dans de nombreuses applications exigeantes, en particulier en imagerie médicale.

La génération de données correspond au problème de la prédiction d'un terme x_n associé à une séquence $\langle x_1, \ldots, x_{n-1} \rangle$. Ici aussi, les principes sont connus depuis longtemps. Il s'agit d'estimer une distribution de probabilité qui représente adéquatement les données d'intérêt, puis d'échantillonner dans cette distribution, dans le contexte de la séquence considérée, pour générer des instances de suites probables. L'échantillonnage génératif a également bénéficié des progrès des réseaux de neurones en termes de performances matérielles, algorithmiques, et d'architectures. Les développements d'IA pour la génération de sons, d'images, et de vidéos sont de plus en plus performants. Les méthodes de traitement automatique du langage naturel rentrent dans le cadre de cette évolution. Ces méthodes sont utilisées pour la compréhension de textes, la traduction, la synthèse ou le dialogue. Les développements des méthodes de génération de données textuelles ont données lieu à changement radical avec des logiciels dits "grands modèles de langage". Les versions dites "modèles de fondement" intègrent plusieurs modalités de données : textes, paroles, et images. Ils sont polyvalents, mais

peuvent être adaptés à des domaines spécifiques. Au-delà de la génération de données, ces modèles démontrent des capacités de raisonnement et de résolution de problèmes, certes imparfaites, mais surprenantes, car imprévues dans leur conception, et non encore bien comprises. Les systèmes ont été testés sur de nombreux examens universitaires. Ils ont démontré des performances bonnes, voire excellentes [14], y compris pour des examens réputés difficiles, tels que ceux de médecine [96].

Réseaux de neurones artificiels. Un neurone artificiel[11] est une fonction multivariable relativement simple : $f(x_1, \ldots, x_n) = g(\sum_{i=0,n} \theta_i x_i)$, où les θ_i sont des paramètres, g est une fonction non linéaire, et le terme θ_0 (pour $x_0 = 1$) est un biais. En notation vectorielle, cette fonction est : $f(x) = g(\theta \cdot x)$ (cf. Figure 3(a)). L'apprentissage consiste à estimer les valeurs des paramètres θ_i qui permettent de se rapprocher le plus des valeurs souhaitées de f. Il s'agit d'un problème d'optimisation, similaire dans son principe à ce qui est fait classiquement dans une régression.



(a) (b)

FIGURE 3 – (a) Un neurone artificiel : θ est un vecteur de paramètres, x est le vecteur d'entrée, $\theta \cdot x$ est leur produit scalaire ; (b) Un réseau perceptron à deux couches : la sortie de chaque couche est un vecteur, Θ_i est la matrice des paramètres de la couche i, la sortie est $\tilde{\mathbf{y}} = g_2(\Theta_2 \times g_1(\Theta_1 \times \mathbf{x}))$.

À lui seul, un neurone artificiel présente peu d'intérêt. C'est l'interconnection d'un très grand nombre de telles fonctions qui permet de modéliser des tâches complexes. Un réseau de neurones connecte des entrées et des sorties de plusieurs de ces fonctions. C'est par exemple un *perceptron multicouches*, qui organise les neurones en couches régulières : un neurone de la couche j prend comme entrées les sorties de tous les neurones de la couche $j-1$ et fournit sa sortie à tous ceux de la couche $j+1$ (Figure 3(b)).

L'apprentissage dans un réseau se fait sur une collection de couples : entrée–sortie souhaitée. Soit (x, y) un tel couple d'apprentissage, et $\tilde{y} = f(x)$

11. L'analogie avec des neurones naturels est ténue : il s'agit davantage d'une métaphore que d'un modèle biologiquement plausible.

la valeur calculée par le réseau sur l'entrée x. On modifie les paramètres du réseau pour minimiser l'écart entre $\tilde{y}$ et y. On utilise pour cela un algorithme de minimisation par descente de gradient, dit de *rétro-propagation*. À chaque exemple d'entrainement (x, y) on adapte les paramètres, couche par couche, de la dernière à la première, pour rapprocher $\tilde{y}$ de la sortie désirée y.

Ce même principe d'adaptation des paramètres pour minimiser l'écart entre ce qui est souhaité et ce qui est calculé est utilisé aussi bien pour l'interprétation que pour la génération de données. C'est de l'apprentissage supervisé pour l'interprétation, en ce sens qu'un humain dit qu'elle est l'interprétation y à associer à x pour chaque exemple d'apprentissage. On parle d'apprentissage *auto-supervisé* pour la génération quand on entraine un réseau sur la prédiction du terme suivant dans des séquences connues. Pour les modèles du langage, ce sont des séquences de mots recueillies automatiquement à partir de moult textes sur internet.

Un réseau de neurones peut approximer n'importe quelle fonction continue à condition (i) d'avoir suffisamment de neurones, de couches et de paramètres, (ii) de disposer de suffisamment d'exemples sur ce que calcule cette fonction, et (iii) de pouvoir calculer les paramètres du réseau en un temps raisonnable. Plus la fonction à approximer est complexe, plus la taille du réseau nécessaire est grande, et plus ces conditions sont difficiles à remplir. Les réseaux de neurones, bien que connus depuis très longtemps [74], n'ont que récemment pu remplir ces conditions pour des fonctions non triviales avec des réseaux dits "profonds", i.e., ayant de nombreuses couches. Ceci a pu être fait grâce à la loi de Moore pour la puissance des processeurs numériques, aux progrès algorithmiques pour les méthodes d'estimation des paramètres, aux architectures de réseaux multi-couches performants, et à internet pour la quantité de données disponibles,

4.2.1 Données *versus* connaissances

Les succès des réseaux de neurones sur des tâches d'interprétation et de génération de données, voire de façon émergente sur des tâches de raisonnement, peuvent conduire à estimer que ces réseaux appréhendent correctement l'intelligence en générale [14], et qu'en pratique, d'autres approches à base de connaissances, plus difficiles à mettre en oeuvre, seraient désormais superflues. Or, la prédiction statistique, à la base des réseaux de neurones, a été qualifiée ici de "superficielle". Elle ne fournit à elle seule qu'une solution de repli en l'absence de modèle à base de connaissances. Qu'en est-il ?

- Les méthodes à base de données produisent, après apprentissage, des *modèles*. Par exemple, un réseau de neurones modélise la relation entre des enregistrement ECG et des interprétations possibles de ces enregistrements, sans appréhender ce que ces interprétations signifient. Il en va de même pour des réseaux très complexes, tel

que celui de GPT, qui modéliserait le langage naturel, ou ceux des "modèles de fondement". La nature de ces modèles statistiques est particulière. Il s'agit de modèles uniquement *prédictifs*. Ils sont opaques (de type "boite noire"), et non intelligibles. Ils sont validés uniquement sur une base statistique, modulo l'hypothèse de l'induction. Retenons qu'apprendre à étiqueter un ECG n'est pas la même chose que de comprendre ce dont il s'agit et d'en apprendre un modèle intelligible et explicatif [12].

- Les méthodes à base de connaissances élaborent ou utilisent des modèles qui s'appuient sur des principes fondamentaux et les relations causales que donnent leurs connaissances. Il s'agit de modèles *justifiés, intelligibles, et explicatifs*. Les raisonnements mis en oeuvre sont traçables. Ils fournissent des explications de leurs prédictions ou conclusions, avec des arguments rationnels et factuels. Ils sont intelligibles car ils composent des connaissances validées, que nous leur avons fournies et que nous comprenons. Ces modèles peuvent aussi mettre en oeuvre des connaissances probabilistes, mais sur des relations élémentaires locales (e.g., sensibilité et spécificité d'un test), plus facilement testables et acceptables, qu'un modèle statistique global. Enfin, ces modèles peuvent aussi comporter des paramètres, pour prendre en compte des éléments non modélisés et appréhender plus précisément la réalité. Mais ces paramètres sont en petit nombre et jouent un rôle différent de ceux des réseaux de neurones [13].

Le tableau 1 résume les propriétés respectives des méthodes à base de données et celles à base de connaissances. Soulignons que l'apprentissage automatique des réseaux de neurones à partir de données brutes permet de couvrir un spectre large de domaines. Il donne des méthodes généralisables et extensibles. Les méthodes à base de connaissances sont par contre étroites, plus difficiles à formaliser, mais bien plus fiables. Ainsi, les approches à base de données ont pu passer avec succès des examens de médecine mais sont incapables de justifier leurs réponses, parfois farfelues ; les approches à base de connaissances excellent de façon fiable sur des sujets étroits, mais n'ont pas passé des examens généralistes (ce n'est pas infaisable mais les efforts de saisie et formalisation des connaissances trop couteux).

Les capacités d'apprentissage autonome, d'adaptation et d'interaction en langage naturel avec un utilisateur sont particulièrement désirables. Mais l'intelligibilité, l'explicabilité et la fiabilité des modèles sont essentielles, en particulier dans les applications médicales. La recherche interdisciplinaire

12. Dans le cas d'habilités sensori-motrices, l'apprentissage d'un modèle "boite-noire" réactif peut-être suffisant, e.g., nul besoin de connaitre la biomécanique pour apprendre à faire du vélo, mais la biomécanique est essentielle à l'orthopédie traumatologique.

13. La théorie de la relativité générale comporte un seul paramètre qui continue de faire l'objet de moult débats scientifiques ; GPT4 comporte plus de mille milliards de paramètres.

TABLE 1 – Méthodes à base de données *versus* celles à base de connaissances.

Méthodes à base de données	*Méthodes à base de connaissances*
Large répertoire de données brutes, difficilement traçables	Base étroite de connaissances formelles, individuellement identifiées
Apprentissage supervisé ou auto-supervisé	Spécification humaine des connaissances
Raisonnements superficiels	Raisonnements profonds
Processus opaque	Processus intelligible
Modèles uniquement prédictifs	Modèles prédictifs et explicatifs
Non fiables	Correctes et prouvables
Généralisables	Difficilement généralisables
Extensibles	Difficilement extensibles
Interface naturelle	Interface formelle

en IA & Santé devrait s'efforcer de combiner les approches d'induction neuronale, à des systèmes de raisonnement intégrant nos connaissances et capables de développer des argumentaires rationnels, étayés et convaincants. De nombreux travaux s'inscrivent dans cette perspective. Parmi les pistes explorées, citons les approches neuro-symboliques [38, 98, 50], ou celles conjuguant des réseaux de neurones à la programmation probabiliste [119]. D'autres recherches plus fondamentales situent l'intelligence naturelle à partir de connaissances intuitives et innées sur le monde, intégrées aux données des expériences vécues et d'observations d'autrui, qui complètent et étayent ces connaissances [61]. Les problèmes en IA dans cette direction sont largement ouverts.

5 Aide au diagnostic et à l'action thérapeutique

Cette section discute de possibles perspectives de recherche en IA & Santé, pour l'aide au diagnostic et à l'action thérapeutique. On ne considère que des systèmes d'aide qui étendent la palette des instruments du praticien et des personnels de santé, voire aident le patient, à l'exclusion de systèmes qui se voudraient autonomes. On abordera successivement (par ordre de difficulté croissante) l'aide à l'interprétation de données, au développement de modèles biophysiques, à la planification d'actions thérapeutiques, et à la synthèse de données par des méthodes génératives.

5.1 Interprétation des données médicales

Données homogènes. L'aide à l'interprétation de données homogènes, de même nature, a donné lieu à de nombreux développements et déploiements industriels. Les applications les plus simples portent sur les analyses biologiques (avec des logiciels commercialisés il y a plus de 25 ans). Les plus récentes portent par exemple sur l'interprétation de données telles que :

- des signaux ECG [67] ; à noter que les signaux EEG et EMG, plus bruités et peu utilisés par les praticiens, sont moins étudiés [22], e.g., pour des interfaces cerveaux-machines (BCI), ou pour le contrôle de prothèses et l'aide à la réhabilitation médicale [58] ;
- des images photographiques, e.g., en dermatologie [33] ou ophtalmologie [47, 112] ;
- des images microscopiques, e.g., en biologie et oncologie [118] ;
- des images radiographiques [17] et échographiques [115, 65], e.g., en oncologie gynécologique [79] ;
- des tomographies de divers types [117, 93].

Tous ces travaux s'appuient sur des méthodes d'apprentissage neuronal.

Des développements d'applications, similaires en fonctionnalités et en performances à celles citées, peuvent utiliser des méthodologies désormais matures et de nombreux logiciels dans le domaine public. Les recherches en cours visent des apports en performances et/ou en fonctionnalités complémentaires, par exemple pour la justification et l'explication des interprétations.

La valeur ajoutée pour le praticien de ces systèmes est à questionner finement. Ainsi, identifier une fracture osseuse dans une radio est de peu d'intérêt. Aider à identifier une pathologie rare sur des images photographiques, prises avec un smartphone par un personnel soignant non spécialiste, soumises à un site de traitement en ligne [14], pourrait permettre de diligenter une prise en charge adéquate. Pour des images plus complexes, qui passent généralement par un radiologue ou un spécialiste, la valeur ajoutée pourrait être par exemple :

- de localiser sur un grand nombre d'images (plusieurs milliers de coupes dans un scanner) celles présentant de possibles anomalies sur lesquelles attirer l'attention du praticien ;
- d'identifier des signes précurseurs d'anomalies non encore manifestes, par exemple des risques d'Alzheimer avant les symptômes [48, 29] ;
- de fournir un modèle anatomique précis par une reconstruction volumique à partir d'un imageur échographique 3D [90, 51] ou tomographique de l'organe d'intérêt pour le praticien ;

14. Similaire aux sites d'identification des plantes (e.g., //plantnet.org) ou des sommets de montagnes, en bien plus fiable.)

- de positionner précisément en 3D relativement aux tissus environnants une tumeur ou le lieu d'intérêt du chirurgien pour lui permettre de planifier et de réaliser une intervention, e.g., en neurochirurgie [49].

On ne dispose pas souvent d'une qualification précise de l'utilisation d'une application d'interprétation de données médicales. Il s'agit d'analyser, dans le contexte hospitalier ou d'un centre de soin, les conditions qui font qu'une application est effectivement utilisée ou ne l'est pas, de savoir pourquoi, comment, et comprendre quelles fonctionnalités complémentaires (ou propriétés non fonctionnelles) permettent ou améliorent son utilisation.

Un autre sujet essentiel, pertinent pour ce qui précède, mais qui n'a pas nécessairement besoin d'expérimentation sur le terrain, porte sur la fiabilité d'un système particulier d'interprétation de données médicales. Souvent, ces systèmes sont validés (en sensibilité, spécificité, précision, et taux d'erreur) sur la même cohorte que celle utilisée pour leur entrainement (une fraction des données sert à la validation). C'est une source de fragilité, documentées dans plusieurs études, e.g., [66]. Il est très important de qualifier, dans le contexte des données locales contextualisées, la confiance qui peut être accordée à l'interprétation d'un système d'induction statistique.

Données hétérogènes. L'interprétation de données combinant signaux, images de divers types, textes, et sons soulève des problèmes ouverts. Elle répond davantage à la nature des informations médicales et apporte potentiellement beaucoup [94]. Ainsi, on obtient de meilleurs résultats par l'intégration de données génétiques, d'images IRM, et d'observations cliniques pour l'aide au diagnostic des pathologies cognitives [116]. L'utilisation des grands "modèles de fondement" multimodaux, bien que très lourde pour le moment, ouvre des perspectives prometteuses. De telles études pourraient privilégier des sujets tels que :

- l'aide au praticien (généraliste, interniste, pédiatre, gynécologue) devant prendre en compte une vision globale et intégrative d'un patient et son évolution dans le temps, par opposition aux aides focalisées sur un organe particulier pouvant s'appuyer davantage sur des données homogènes ;
- l'aide aux chercheurs, par exemple en épidémiologie, pour le suivi et l'analyse spatio-temporelle de pathologies dans la population ;
- l'aide à l'analyse du système de santé globalement, selon des indicateurs pour un suivi géographique précis reliant les ressources et les mesures spécifiques déployées, aux résultats constatés.

La formation, initiale et continue, de spécialistes et de personnels soignants est également une finalité pertinente des systèmes d'interprétation de données médicales. Les étudiants auront à utiliser ces systèmes ; ils doivent en maîtriser l'usage critique. Par ailleurs, des logiciels didactiques intégrant de tels systèmes sont une ressource d'apprentissage puissante.

Enfin, les reconstructions 3D de modèles anatomiques spécifiques d'un organe, conjuguées à des outils de simulation, de réalité virtuelle, voire de "jeux sérieux", apportent des possibilités d'entrainement peu couteuses et particulièrement formatrices, e.g., [10].

5.2 Intégration de modèles biophysiques

Plusieurs arguments militent pour augmenter les approches qui précèdent par des connaissances physiques, biologiques et médicales. D'une part, il est difficile d'accorder confiance à un système "boite noire" de prédiction statistique, peu capable d'expliquer et justifier ses réponses. Par ailleurs, l'hétérogénéité des données médicales et le besoin de leur intégration ont déjà été soulignés ; ne considérer que des données brutes, sans associer à chaque composante ce qu'elle représente, est une perte d'information considérable [15]. Enfin, la complémentarité des méthodes à base de données et celles à bases de connaissances ouvre des perspectives essentielles en IA et médecine. Il s'agit d'assoir les aides au praticien sur le savoir scientifique, mais également aider le chercheur à étendre ce savoir grâce à l'apport des données spécifiques.

À un niveau fondamental, de nombreux travaux dans cette direction portent sur la modélisation biologique. Ils sont à l'intersection entre IA et bio-informatique (cf. [35, 84]), avec des applications en particulier en pharmacologie. Ils alimentent également des recherches qui se focalisent sur un ou quelques organes, et s'efforcent de les aborder par des connaissances générales et des données spécifiques à un patient, sous toutes les facettes pertinentes pour la médecine. C'est par exemple le système cardio-vasculaire d'une personne que l'on va appréhender par l'intégration de ses modèles anatomique, mécanique, physiologique, électrophysiologique, métabolique et de dynamique des fluides [25, 26]. On peut interroger ces systèmes via des simulations précises du fonctionnement et des dysfonctionnements éventuels de l'organe d'intérêt [6]. On peut également les utiliser pour planifier une intervention chirurgicale, concevoir et dimensionner très précisément une prothèse, e.g., un stint ou un manchon périvasculaire, et étudier son impact et évolution à moyen et long terme [77, 7].

Dans ces perspectives d'intégration ambitieuse, les recherches peuvent désormais bénéficier de modèles et simulateurs disponibles, parfois dans le domaine publique [16], et/ou développés par des consortiums de R&D internationaux ouverts [17].

15. "*It is rather unlikely that a machine learning algorithm that knows nothing of anatomy, physiology, infection, pharmacokinetics, or instrumentation can make head or tail of such data—not least because of the lack of uniformity in what data are available concerning any particular patient at any particular time* " [71].
16. e.g., //www.imagwiki.nibib.nih.gov/physiome, ou //physionet.org/.
17. e.g., //3dexperiencelab.3ds.com/en/projects/life/living-heart

5.3 Planification d'actions thérapeutiques

La planification en IA est synonyme de raisonnement sur l'action : quelles actions entreprendre, comment les organiser et les réaliser dans un contexte et pour un objectif donnés [42]. Dans le domaine médical, les actions conjuguent par exemple des actes pharmacologiques, physiques (radiothérapie, kinésithérapie, sport), chirurgicaux, voire verbaux, comportementaux et liés à l'environnement du patient. La planification doit aussi prendre en compte des actions dites épistémiques : quels tests faire et quand, pour informer la poursuite du plan thérapeutique. L'aide à la planification est peu utile au praticien si un diagnostic est associé au même protocole thérapeutique pour tous : ce protocole inclut généralement un plan, avec une marge de paramétrage relativement focalisée. La médecine personnalisée entraine un besoin d'aide à la planification, laquelle nécessite des modèles, généralement probabilistes, des effets possibles des actions.

Considérons les cas (comme ceux mentionnés en section 3.2), où plusieurs thérapies peuvent être combinées. On aimerait connaitre à chaque étape la combinaison *optimale* à mettre en oeuvre, compte tenu de l'évolution de l'état d'un patient particulier, et des contraintes pharmacologiques et médicales. Ce problème peut être modélisé par la représentation probabiliste des *processus décisionnels de Markov*. Si un modèle des effets possibles de chaque thérapie est connu (sous forme d'une distribution de probabilité des états résultants), on dispose de bons algorithmes pour résoudre le problème [43, chap. 8 et 9].

Dans le domaine médical, l'état d'un patient peut être décrit par un vecteur à hautes dimensions (i.e., dans R^n pour $n \geq 10^6$) de paramètres biologiques et d'observations cliniques ; l'espace des états possibles est infini. Les algorithmes les plus performants procèdent par tirages aléatoires adaptatifs. Ils présentent l'avantage de permettre une optimisation à horizon glissant : à chaque étape on observe l'état du patient, on recalcule une politique optimale dont on n'applique que le premier pas, et on recommence à l'étape suivante. Des approches de ce type ont été mises en oeuvre avec succès [15, 2].

Cependant, on doit fréquemment aborder le problème de planification de thérapies sans avoir de modèle à priori des effets possibles des actions. On utilise alors des approches d'*apprentissage par renforcement*. Ces approches, inspirées de travaux anciens de neuropsychologues ([111, 106]), sont homothétiques aux techniques algorithmiques précédentes. Elles conduisent à apprendre implicitement les modèles des effets des actions via l'estimation de fonctions de qualité relative, et convergent également vers des politiques quasi-optimales. Elles se conjuguent efficacement avec l'utilisation d'estimateurs des fonctions de qualité par réseaux de neurones profonds, mais avec des propriétés théoriques et explicatives différentes (cf. [43, chap. 10 et 13]).

L'apprentissage par renforcement ne converge généralement qu'après de

nombreux tests et essais. Ceci n'est pas compatible avec des applications médicales. Pour y remédier, deux approches sont envisageables :

- L'utilisation de simulateurs sur la base de modèles biophysiques ; c'est un argument important en faveur de ces modèles (cf. section 5.2). Par exemple, [89] est une illustration pour le traitement des sepsis.
- L'utilisation d'historiques sur de nombreux cas médicaux des effets observés des thérapies à planifier. L'apprentissage par renforcement sur données est particulièrement prometteur en médecine. Il a été étudié par exemple en hémodialyse [31], en oncologie pulmonaire [130], en transplantation de moelle osseuse [68], et dans d'autres applications, dont on trouvera des revues dans [21, 126, 63]. Cette approche pourrait également être pertinente pour l'aide à la prise en charge précise du diabète [109, 125].

L'apprentissage par renforcement est également pertinent pour la planification d'une politique de santé publique, par exemple pour le dépistage précoce de certaines pathologies, telles que les cancers du sein [123]. Notons également son utilisation pour le contrôle optimal en continu du niveau de sédation anesthésique par rétroaction sur les signaux EEG fournis par les appareils classiques de monitoring ; les mises en oeuvre, sous supervision de l'anesthésiste, donnent d'excellents résultats [99, 127].

La planification d'actes médicaux physiques et chirurgicaux est un autre volet important. Citons à titre d'exemples les utilisations suivantes :

- planification de radiothérapies [5, 78] et son acceptabilité clinique [9] ;
- planification de chirurgies, e.g., en hépatectomie [86], ou en chirurgie orthognathique avec l'aide d'une plateforme de modélisation, de simulation et de planification d'interventions à partir de données tomographiques [121] ;
- planification de physiothérapies et protocoles de réhabilitations [45, 62, 91].

Sur ces trois points, l'utilisation de capteurs précis de position et de mouvement, voire d'actionneurs, de robots et d'exoskelettes, accroit les besoins en planification. Elle peut être très bénéfique, mais reste généralement couteuse.

La planification en pharmacologie joue également un rôle important, par exemple pour l'aide à la synthèse de protéines ayant des propriétés de conformation et de docking moléculaires précises [32, 76, 85]. Dans ce sens, l'apport récent du système d'IA AlphaFold [54], qualifié parfois de révolution en biologie, est considérable. AlphaFold permet de passer de la séquence d'acides aminés d'une protéine à sa structure, et, pour AlphaFold.3, à ses interactions biomoléculaires et appariements possibles [1]. Ce système, aujourd'hui très largement utilisé par les biologistes, a permis d'explorer les fonctions de plus de 200 millions de protéines [18].

18. Ceci justifie le prix Nobel de Chimie accordé en 2024 à ces deux concepteurs D. Hassabis et J. Jumper.

Enfin, rappelons pour mémoire, que l'aide à la planification de l'usage optimal de ressources hospitalières est un sujet mature, pour lequel divers outils sont commercialement disponibles. L'extension de ces outils pour la mise en réseau géographique et l'optimisation des développements de moyens de santé publique peut poser des problèmes appliqués intéressants.

5.4 IA générative et synthèse de données

Les développements récents les plus visibles de l'IA sont produits par les grands modèles du langage et les "modèles de fondement" multimodaux, désignés ici brièvement par LLM (*Large Language Models*). Sur l'état de l'art des LLM voir par exemple [41, 83, 128], ou [12, 64] pour leurs extensions multimodales. Il s'agit de systèmes polyvalents qui démontrent une excellente maîtrise du langage et de l'interactivité conversationnelle. Leurs performances dans des tâches cognitives, de raisonnement et de sens commun sur un très large spectre de domaines rendent les LLM extrêmement attractifs pour de nombreuses applications. Qu'en est-il pour la médecine ?

Un premier indicateur de réponse est le succès des LLM aux examens finaux de médecine dans plusieurs systèmes académiques, par exemple en Pologne [96], au Japon [124], ou aux USA [44]. Mais ceci qualifie probablement davantage la nature des examens que les capacités médicales de la machine qui y réussit. On sait que la pratique clinique requière bien plus de compétences que des réponses correctes à un QCM.

Rappelons que les LLM font de l'induction statistique sur d'énormes bases de données. Ils ne manipulent pas de connaissances médicales formelles ; ils n'ont pas les modèles biophysiques explicites. De ce fait, leurs limitations, dans l'état de l'art actuel, sont nombreuses. Relativement aux exigences d'applications médicales, on relève en particulier :

- le caractère non factuel des réponses, même quand les bonnes réponses sont présentes dans les données d'apprentissage, ce qui donne lieu parfois à des erreurs grossières, qualifiées d"hallucinations" [19] ;
- l'absence de justification, d'explication des réponses et de transparence sur les sources ;
- l'absence de qualification de la fiabilité des propos ; ainsi réagir à une réponse d'un LLM en lui demandant "*ès-tu sur ?* " donne généralement lieu à des appréciations ni fiables ni étayées ;
- l'absence de vérification de la validité des propos ; une telle vérification ne peut se faire que relativement à des connaissances explicites, qui sont absentes des LLM.

19. Ce qualificatif est trompeur, car un réseau de neurones calcule une d'approximation statistique, et non une requête dans une base de données.

Beaucoup de travaux en cours s'efforcent de pallier ces limitations. Par exemple, par des méthodes d'interrogation particulières pour des réponses plus précises, dites *Chain-of-Thought* [100, 53], ou par des méthodes de liens aux sources pour des justifications, dites *Retrieval Augmented Generation* [19, 37, 52]. Il est également envisageable de connecter un LLM à des logiciels encodant des modèles biophysiques, ce qui reste à faire à notre connaissance.

Les recherches multidisciplinaires dans ces directions sont prometteuses. Il s'agit de mettre à profit les capacités langagières et multimodales des LLM pour l'aide au diagnostic à partir d'images médicales, d'analyses de symptômes et de dossiers d'antécédents ; [87] en illustre un exemple. L'aide au dépistage et à la prévision de risques est également une perspective importante. Ainsi, le système Foresight [59], un LLM entrainé sur des concepts biomédicaux et les dossiers de près d'un million de patients, serait capable de prévoir, à partir de dossiers textuels et de données médicales, des risques de pathologies, des effets de procédures thérapeutiques, ou des progressions de troubles. Foresight vise à fournir un outil de recherche clinique pour faire des essais virtuels, simuler des thérapies et des explorations contrefactuelles. Mentionnons aussi Med-PaLM2, un système affiné sur des données médicales à partir du LLM polyvalent PaLM2, qui a été évalué positivement par des praticiens sur plusieurs bases de tests cliniques (MedQA, MedMCQA, PubMedQA, et MMLU), y compris sur des questions malicieuses (*adversarial tests*) destinées à le prendre en défaut [104, 92]. D'autres projets de recherche sur ces sujets sont en cours, e.g., [110].

Enfin, et au-delà des LLM polyvalents, soulignons que les capacités génératives sont particulièrement importantes dans d'autres applications médicales, par exemple en pharmacologie. Ainsi, plusieurs systèmes utilisent des grands modèles du langage spécialisés sur des données biologiques, e.g., ProGen pour l'aide à la synthèse de protéines ayant des fonctions biologiques désirées [72], ou ZymCTRL pour la synthèse enzymatique [82].

À plus court terme, des développements peuvent d'ores et déjà mettre à profit les capacités langagières et d'interaction des LLM pour assister les praticiens et personnels soignants dans des tâches plus administratives, par exemple dans l'élaboration de comptes rendus, de synthèses de dossiers médicaux, de rapports consolidés [3]. Les méthodes de traitement automatique du langage, déjà utilisées en santé [120], sont largement améliorées par les LLM. Ces systèmes peuvent également apporter des compléments de réponse détaillés aux patients, expliquer un diagnostic et préciser des éléments dans une ordonnance (bien mieux que les notices des médicaments, souvent peu lisibles). Des tests (évalués par des médecins) indiquent que les réponses des LLM aux patients seraient appréciées en termes de qualité et d'empathie [4]. La maîtrise du langage naturel est également pertinente pour l'aide au chercheur, e.g., dans la fouille et l'analyse de publications médicales, le résumé ou la traduction d'articles.

Les bonnes performances des LLM dans les examens de médecine les

rendent possiblement utiles à des fins éducatives [60]. Au-delà de tâches relativement simples d'aide à la préparation d'examens, on peut envisager le développement de tuteurs virtuels personnalisés, capables de suivre et d'assister un étudiant tout au long de sa formation par des entrainements renforçant ses faiblesses. Enfin, des actions de formation des personnels soignants et aides soignants, par exemple dans une spécialité particulière telle qu'en maïeutique [46], ou pour l'utilisation d'appareils complexes à maîtriser (e.g., échographie obstétrique), peuvent être particulièrement bénéfiques en termes de santé publique dans des zones rurales à relativement faible couverture médicale [20].

Un projet mettant en oeuvre un LLM peut, soit utiliser un système existant et l'adapter à ses besoins, soit développer un nouvel LLM. La première option devrait préférablement s'appuyer sur un système ouvert à code public pour permettre une réelle maîtrise de son contenu [21]. Des systèmes ouverts aussi performants que les meilleurs LLM commencent à être disponibles, par exemple Molmo [23]. La deuxième option est confrontée à deux difficultés : la disponibilité de bases d'entraînement larges, et les coûts de calcul pour faire cet entrainement. Quelques bases de pré-entrainement génériques commencent à être disponibles. Le recueil de bases spécifiques à divers contextes reste à faire. Les coûts de calcul sont exorbitants. Il a été estimé que le pré-entrainement de Gemini-Ultra a nécessité 5×10^{25} opérations de calcul et couté 200M$, celui de ChatGPT a consommé 1,3 GWh [22]. Même après entrainement, les coûts d'utilisation des grands LLM restent élevés.

En résumé sur l'utilisation des méthodes génératives et des LLM polyvalents en médecine, il faut retenir que ces systèmes ouvrent des perspectives importantes pour la recherche, la formation et l'aide clinique. Cependant, l'absence d'explication, de justification et de mesure d'incertitude réduit aujourd'hui leurs possibilités d'utilisation dans des applications médicales. À notre connaissance, aucun système utilisant des LLM polyvalents n'a encore été certifié par des organismes publics de santé, alors que plusieurs centaines de systèmes spécialisés, à base de connaissances ou de données, l'ont été. Il faut donc privilégier des projets de recherche visant à qualifier sur de vastes cohortes, la confiance qui peut être accordée à de tels systèmes. En attendant de telles qualifications précises, il faut aborder tout déploiement éventuel avec beaucoup de prudence et le circonscrire à des tâches non critiques et bien encadrées.

En conclusion de cette section sur les quatre volets abordés d'interprétation de données, d'intégration de modèles biophysique, de planification, et des méthodes génératives, soulignons que plusieurs "feuilles de

20. Voir par exemple au Kenya le projet //jacarandahealth.org/
21. Voir //github.com/eugeneyan/open-llms ou //llmmodels.org/
22. Soit les besoins mensuels moyens d'une ville de plus de 1000 personnes.

route" pour des développements en IA et Santé ont été proposées, e.g., [8]. Elles méritent d'être prises en compte. Par ailleurs une analyse par spécialités médicales seraient également utile ; elle apporterait un éclairage complémentaire à ce qui a été dit ici. Cette analyse (disséminée dans de nombreuses publications et non consolidée à notre connaissance) révélerait sans doute que la cardiologie et la cancérologie, les deux principales causes de mortalité, attirent le plus de travaux. Elle révélerait aussi que l'IA est explorée pour toutes les spécialités médicales, sans exceptions, le plus souvent avec de nombreux prototypes mais très peu de validations cliniques. Ainsi, [27] fait une analyse de 66 publications (sur près 600) de l'IA en gynécologie-obstétrique, y compris en médecine reproductive et médecine foetale. On y souligne le large spectre des méthodes utilisées, mais le plus souvent, "la validation clinique reste une condition préalable non remplie". Il est important que des recherches en IA & Santé prennent en compte cette dimension essentielle, prioritairement à des développements exploratoires de "preuves de concept", formateurs mais moins pertinents.

6 Risques et problèmes éthiques

La recherche sur IA & Santé se doit de consacrer des efforts spécifiques à l'analyse des risques. Elle devra également contribuer à des recommandations, issues de réflexions collectives, pour la prise en compte d'impératifs éthiques. Discutons ces deux points.

Analyse des risques. Les taux d'erreur des praticiens sont loin d'être négligeables [28] : l'erreur est humaine et la décision médicale peut devoir être prise dans un contexte cognitif stressant, sans délibération [23]. Cependant, certifier un système d'IA uniquement sur un taux d'erreur plus faible que celui du praticien moyen n'est pas suffisant. Il faudrait que l'aide au praticien apporte globalement une amélioration avérée au processus clinique, en termes de fiabilité, de précision, d'intégration de l'ensemble des faits sur chaque cas, de délibération, de vérification de la décision et de qualité de l'action thérapeutique.

L'importance de projets de recherche sur la fiabilité de systèmes d'IA en médecine a déjà été soulignée. Les systèmes à base de connaissances sont en principe prouvables, encore faut-il développer et mettre en oeuvre des algorithmes de vérification et de validation associés aux connaissances fournies. Les systèmes d'interprétation par induction statistique peuvent être testés empiriquement sur des cohortes indépendantes de celles utilisées pour leur apprentissage. Ceci est plus facile pour les systèmes focalisés, et se complique pour ceux plus larges à données hétérogènes.

23. Ces erreurs ont aussi un impact sur la santé des praticiens [105].

Le test des grands modèles polyvalents du type LLM est encore plus difficile à réaliser ; il exige des recherches substantielles. Dans tous les cas, il faut estimer la fiabilité d'un système dans son environnement clinique opérationnel, ce qui inclut la prise en compte des utilisateurs et des usages (lacunes de formation, mauvaises utilisations, etc.).

Aux difficultés de tester une "boite noire" opaque, se rajoutent celles de prendre en compte des équipements intégrant des fonctionnalités IA de façon ubiquitaire, peu visible, par exemple dans de nouvelles générations d'équipements d'imagerie. Mentionnons également la consultation de systèmes en ligne sur internet. Ces systèmes utilisent de plus en plus des LLM, avec les risques mentionnés d'apparence de compétence polyvalente, risques aggravés pour des utilisateurs non avertis [24]. Ce point doit faire l'objet d'information, de formations, et d'une vigilance du régulateur de santé publique.

Aux risques fortuits, dus aux imperfections, se rajoutent malheureusement ceux liés à des actes intentionnels, voire à des malveillances. On connait la catastrophe sanitaire des opiacés aux USA [75] [25], qui a fait près d'un million de victimes avant que le régulateur n'intervienne. La justice à postériori (avec un compromis des industriels à 26 milliards pour "solder les litiges" restants) est peu satisfaisante. Il faut prévenir à priori de tels risques de catastrophes sanitaires, dont les exemples ne sont pas rares (e.g., Mediator, avec condamnation pour "tromperie aggravée", Levothyrox, Dépakine, etc.). Ces risques sont dûs en particulier au développement capitaliste, lequel, par définition, est régi par une logique de profit davantage que par une logique d'amélioration de la santé publique. Les mécanismes de régulation, dont le critère principal est la santé publique, doivent pouvoir s'appuyer autant que faire se peut sur l'*anticipation des risques*. Les systèmes d'IA en santé rajoutent à ce qui précède tous les risques liés à la manipulation des données (en particulier par l'IA générative), au piratage, à la cybersécurité, voire à la manipulation intentionnelle du personnel médical et des patients.

En conclusion sur ce point, il serait pertinent de promouvoir quelques actions de recherche visant à établir une *taxonomie des risques* de l'IA pour la santé, comment les prévenir, comment y pallier et quel "code de pratique" exiger des développeurs de systèmes d'IA pour la santé.

Impératifs éthiques. Plusieurs considérations éthiques sont à prendre en compte dans un programme de recherche en IA & Santé, dont celles de l'éthique clinique, l'éthique de la recherche médicale, et l'éthique en IA.

Le premier point est connu depuis Hippocrate ; il fait l'objet de serments

24. Selon un sondage récent par l'association KFF, un adulte sur 6 aux USA consulte au moins une fois par mois un chatbot pour obtenir des conseils et informations de santé.

25. cf. aussi //en.wikipedia.org/wiki/Opioid_epidemic_in_the_United_States

formels sur des principes régulièrement débattus et mis à jour.

L'éthique de la recherche médicale est également une question bien traitée, depuis la convention de Nuremberg (1947) et la déclaration de Helsinki (1964), jusqu'aux nombreux textes juridiques récents (e.g., 5 lois en France entre 1988 et 2016) sur la recherche biomédicale qui instituent une hiérarchie de comités d'éthiques et d'autorisations pour encadrer les travaux. Concentrons nous ici sur des éléments nouveaux apportés par l'IA.

Les préoccupations sur les risques et problèmes éthiques que soulève l'IA donnent lieu à de nombreuses publications et recommandations (e.g., [13, 34, 114], ou [20, 102] plus spécifiquement en IA et Santé). La plupart des travaux portent sur des questions éthiques centrées sur l'usage des données, telles que les biais, la confidentialité, la protection de la vie privée, l'équité, la transparence, la fiabilité, la propriété des données et les droits d'accès. Ces questions sont particulièrement importantes dans le contexte des données médicales. Les résultats de ces travaux doivent être traduits en des recommandations de mise en oeuvre dans des réglementations, des institutions (par exemple, des fiduciaires de données [24]) et des processus de surveillance active.

Toutefois, les préoccupations éthiques relatives aux données sont pour la plupart "individualistes". Elles ne prennent pas suffisamment en compte les responsabilités sociales sur de possibles effets néfastes, plus larges et plus profonds. Ce point soulève une difficulté : l'acceptabilité individuelle d'une technologie, même très répandue dans un marché lucratif, n'est pas équivalente à l'acceptabilité sociale. Cette dernière doit prendre en compte le long terme, les incidences sur l'environnement, et les effets sur la cohésion et les valeurs sociales [40]. Considérons trois points : à court terme, les risques de tromperie et de manipulation sociale, et à long terme, ceux liés aux systèmes autonomes et à l'augmentation humaine.

Les questions de tromperie en santé remontent aux charlatans, évoqués dès Hippocrate. Les outils d'IA, en particulier les méthodes génératives et les LLM, augmentent considérablement les possibilités de manipulation des faits et des personnes. À petite échelle, c'est le déploiement de services d'IA & Santé non fiables, incapables de faire ce qu'ils prétendent. À grande échelle, ce sont les "vérités alternatives", illustrées par les réseaux sociaux lors de l'épidémie du SARS-CoV2, qui peuvent être amplifiées par l'IA. Des agents conversationnels, qui parlent notre langue et qui semblent apparemment bien informés sur nous et notre environnement, peuvent être dotés de connaissances détaillées de chacun de leurs interlocuteurs (il suffit d'observer leurs clics sur internet [11]), et donc capables de les manipuler en vue de finalités douteuses. Ceci est particulièrement vrai en médecine, la santé étant un sujet de grande fragilité psychologique où le croire l'emporte souvent sur le savoir.

À plus long terme, le déploiement de systèmes autonomes en santé soulève des questions éthiques moins étudiées. La prise en charge clinique

entièrement mécanisée n'est heureusement pas à l'ordre du jour. Mais des machines autonomes sont sérieusement considérées pour certaines applications, par exemple pour l'aide à des personnes handicapées ou âgées ayant perdu leur autonomie sensorimotrice ou cognitive. On peut certes argumenter que des machines empathiques, serviables et dignes de confiance apporteraient un soutien désirable dans certains cas. Il reste néanmoins nécessaire de s'interroger sur les limites des fonctions acceptables pour de telles machines autonomes et l'encadrement humain exigé.

Un autre problème à long terme est la frontière parfois ténue entre ce qui relève de l'aide pour pallier une déficience, et ce qui relève de l'amplification des capacités d'une personne bien portante. Le pas entre ces deux finalités est vite franchi. Les mêmes travaux pour l'aide à un tétraplégique peuvent être détournés pour amplifier les capacités de mouvement d'un combattant. De même, les interfaces invasives cerveau-machine peuvent aller d'objectifs médicaux légitimes, e.g., Alzheimer ou prothèses visuelles, à des objectifs d'augmentation humaine [26]. Le *transhumanisme* s'affirme comme une "ingénierie de l'espèce humaine" visant à dépasser nos limitations physiques et mentales, à abolir le vieillissement, et à éloigner la mort. Le mouvement transhumaniste est défendu, ouvertement ou de façon plus insidieuse, par exemple sur des questions de procréation pour lesquelles on peut passer du criblage génétique pour des objectifs médicaux moralement acceptables, à la recherche de croisements "optimaux" pour certains traits. On imagine comment des outils tels que AlphaFold et ProGen pourraient être prolongés vers des mises en oeuvre qui s'apparentent à l'eugénisme. Le transhumanisme ne peut qu'accentuer les inégalités sociales à un niveau essentiel, et pousser vers une *spéciation de l'espèce humaine*.

Des recherches spécifiques sur ces questions, en particulier sur les deux derniers points évoqués, ne sont peut-être pas prioritaires. Par contre, il est important d'y sensibiliser tous les acteurs de la recherche en IA et Santé, par exemple sous forme de rencontres focalisées impliquants des chercheurs des sciences humaines et sociales, de rédaction de codes de conduite, et d'adhésion à ceux existant [27].

7 Conclusion

On a présenté ici une introduction générale à l'IA pour la Santé. On s'est efforcé de situer les fonctions cognitives étudiées par l'IA pour des applications médicales. On a discuté et comparé des méthodes d'IA, à base de connaissances et à base de données. Des perspectives de recherches ont été évoquées successivement pour :

26. L'acquisition de Neuralink par Elon Musk, très présent en IA et adepte du transhumanisme, n'a sans doute pas qu'une finalité industrielle et médicale.

27. Cf. e.g., , le code pour la conception de protéines : //responsiblebiodesign.ai/

- l'interprétation de données médicales, homogènes ou hétérogènes ;
- l'integration de modèles biophysiques ;
- la planification d'actions thérapeutiques ; et
- la synthèse de données par des grands modèles polyvalents.

Bien entendu, les méthodes et approches mentionnées sont illustratives de ce qui nous semble pertinent sur le sujet, mais n'ont aucune prétention à l'exhaustivité. Dans tous les cas, on s'est efforcé de mettre en avant les limitations de l'état de l'art actuel en IA. Les avantages et les inconvénients de diverses approches ont été discutés. Schématiquement, on peut les résumer ainsi :

- Les méthodes d'interprétation de données homogènes par réseaux de neurones sont relativement matures, mais les modèles résultants sont opaques et non intelligibles ; la fiabilité de ces méthodes dans des utilisations cliniques effectives reste à qualifier.
- Les méthodes de raisonnement à base de connaissances sont correctes et prouvables, elles donnent lieu à des modèles justifié, explicatifs et intelligibles, mais elles sont étroites et difficiles à mettre en oeuvre.

Des recherches ambitieuses devraient pouvoir conjuguer avantageusement ces deux types d'approches.

L'IA peut considérablement contribuer à la recherche médicale et à la pratique clinique, pour une meilleure maîtrise de la complexité des problèmes de santé, vers une médecine intégrative, personnalisée, préventive et socialement efficiente. Elle peut également être d'un apport important pour la formation du personnel de santé, pour l'aide au patient et l'information du grand public. Cependant, l'IA, comme toute technologie, est ambivalente. Elle présente des risques, qui doivent faire l'objet d'études, et soulève des problèmes éthiques qui méritent d'être débattus et donner lieu à des recommandations de régulations sociales.

Références

[1] J. Abramson, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 2024.

[2] B. M. Adams, et al. Dynamic multidrug therapies for HIV : Optimal and STI control approaches. *Mathematical Biosciences & Engineering*, 2004.

[3] A. Arora and A. Arora. The promise of large language models in health care. *The Lancet*, 2023.

[4] J. W. Ayers, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine*, 2023.

[5] A. Babier, et al. Knowledge-based automated planning with three-dimensional generative adversarial networks. *Medical Physics*, 2020.

[6] B. Baillargeon, et al. The living heart project : a robust and integrative simulator for human heart function. *European Journal of Mechanics*, 2014.

[7] B. Baillargeon, et al. Human cardiac function simulator for the optimal design of a

novel annuloplasty ring with a sub-valvular element for correction of ischemic mitral regurgitation. *Cardiovascular engineering and technology*, 2015.

[8] J. Bajwa, et al. Artificial Intelligence in healthcare : transforming the practice of medicine. *Future healthcare journal*, 2021.

[9] H. Baroudi, et al. Automated contouring and planning in radiation therapy : what is 'clinically acceptable' ? *Diagnostics*, 2023.

[10] A. Bauer, et al. Living book of anatomy (LBA) project : see your insides in motion ! In *SIGGRAPH Asia 2015 Emerging Technologies*. 2015.

[11] E. L. Boldyreva, et al. Cambridge analytica : Ethics and online manipulation with decision-making process. *European Proceedings of Social and Behavioural Sciences*, 2018.

[12] R. Bommasani, et al. On the opportunities and risks of foundation models. *arXiv :2108.07258*, 2022.

[13] B. Braunschweig and M. Ghallab, editors. *Reflections on Artificial Intelligence for Humanity*. Springer, 2021.

[14] S. Bubeck, et al. Sparks of artificial general intelligence : Early experiments with GPT-4. *arXiv :2303.12712*, 2023.

[15] L. Busoniu, et al. Optimistic planning for sparsely stochastic systems. In *IEEE Symp. Adaptive Dynamic Programming And Reinforcement Learning*. 2011.

[16] T. J. Callahan, et al. Ontologizing health systems data at scale : making translational discovery a reality. *Digital Medicine*, 2023.

[17] E. Çallı, et al. Deep learning for chest x-ray analysis : A survey. *Medical Image Analysis*, 2021.

[18] B. Carpenter, et al. Stan : A probabilistic programming language. *Journal of statistical software*, 2017.

[19] J. Chen, et al. Benchmarking large language models in retrieval-augmented generation. In *AAAI*, 2024.

[20] Y. Chino. AI in medicine : Creating a safe and equitable future. *Lancet*, 2023.

[21] A. Coronato, et al. Reinforcement learning for intelligent healthcare applications : A survey. *Artificial intelligence in medicine*, 2020.

[22] A. Craik, et al. Deep learning for electroencephalogram (EEG) classification tasks : a review. *Journal of neural engineering*, 2019.

[23] M. Deitke, et al. Molmo and pixmo : Open weights and open data for state-of-the-art multimodal models. *arXiv*, 2024.

[24] S. Delacroix, et al. *Democratising the digital revolution : The role of data governance*, chapter 3. In Braunschweig and Ghallab [13], 2021.

[25] H. Delingette, et al. Cardiosense3D : Patient-Specific Cardiac Simulation INRIA Large Wingspan Project. Technical report, INRIA, 2005.

[26] H. Delingette, et al. Cardiosense3d : patient-specific cardiac simulation. In *2007 4th IEEE International Symposium on Biomedical Imaging : From Nano to Macro*. IEEE, 2007.

[27] F. Dhombres, et al. Contributions of Artificial Intelligence reported in obstetrics and gynecology journals : systematic review. *Journal of medical Internet research*, 2022.

[28] M. S. Donaldson, et al. *To err is human : building a safer health system*. National Academies Press, 2000.

[29] M. A. Ebrahimighahnavieh, et al. Deep learning to detect Alzheimer's disease from neuroimaging : A systematic literature review. *Computer methods and programs in biomedicine*, 2020.

[30] D. Ernst, et al. Clinical data based optimal STI strategies for HIV : a reinforcement learning approach. In *Proceedings of the 45th IEEE Conference on Decision and Control*. IEEE, 2006.

[31] P. Escandell-Montero, et al. Optimization of anemia treatment in hemodialysis patients via reinforcement learning. *Artificial Intelligence in Medicine*, 2014.

[32] A. Estaña, et al. Investigating the formation of structural elements in proteins using local sequence-dependent information and a heuristic search algorithm. *Molecules*, 2019.
[33] A. Esteva, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 2017.
[34] EU High-Level Expert Group on AI. Ethics Guidelines for Trustworthy AI, 2019.
[35] F. Fages. Artificial intelligence in biological modelling. In *A Guided Tour of Artificial Intelligence Research : Volume III : Interfaces and Applications of Artificial Intelligence*. Springer, 2020.
[36] D. Ferrucci, et al. Watson : beyond Jeopardy ! *Artificial Intelligence*, 2013.
[37] Y. Gao, et al. Retrieval-augmented generation for large language models : A survey. *arXiv :2312.10997*, 2023.
[38] A. d. Garcez and L. C. Lamb. Neurosymbolic AI : The 3 rd wave. *Artificial Intelligence Review*, 2023.
[39] Z. Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 2015.
[40] M. Ghallab. Responsible AI : requirements and challenges. *AI Perspectives*, 2019.
[41] M. Ghallab. GPT et les grands modèles du langage en intelligence artificielle : principes et défis, 2024. https://laas.hal.science/hal-04742089.
[42] M. Ghallab, et al. *Automated Planning : Theory and Practice*. Morgann Kaufmann, Oct. 2004.
[43] M. Ghallab, et al. *Acting, Planning and Learning*. Cambridge University Press (in press), 2024. https://projects.laas.fr/planning/.
[44] A. Gilson, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE) ? The implications of large language models for medical education and knowledge assessment. *JMIR medical education*, 2023.
[45] J. C. González, et al. A three-layer planning architecture for the autonomous control of rehabilitation therapies based on social robots. *Cognitive Systems Research*, 2017.
[46] A. Grünebaum, et al. The exciting potential for ChatGPT in obstetrics and gynecology. *American Journal of Obstetrics and Gynecology*, 2023.
[47] V. Gulshan, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 2016.
[48] H. A. Helaly, et al. Deep learning approach for early detection of Alzheimer's disease. *Cognitive computation*, 2022.
[49] P. Hellier, et al. Retrospective evaluation of intersubject brain registration. *IEEE transactions on medical imaging*, 2003.
[50] P. Hitzler and M. K. Sarker. *Neuro-symbolic artificial intelligence : The state of the art*. IOS press, 2022.
[51] J. P. Howard, et al. Improving ultrasound video classification : an evaluation of novel deep learning methods in echocardiography. *Journal of medical artificial intelligence*, 2020.
[52] Z. Jiang, et al. Active retrieval augmented generation. *arXiv :2305.06983*, 2023.
[53] M. Jovanović and P. Voss. Towards incremental learning in large language models : A critical review, 2024. Online report.
[54] J. Jumper, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 2021.
[55] S. Kaur, et al. Medical diagnostic systems using artificial intelligence algorithms : Principles and perspectives. *IEEE Access*, 2020.
[56] R. Khajuria and A. Sarwar. Reinforcement learning in medical diagnosis : An overview. In *Proceedings of ICRIC Recent Innovations in Computing*. 2022.
[57] S. N. Khatami and C. Gopalappa. A reinforcement learning model to inform optimal decision paths for hiv elimination. *Mathematical biosciences and engineering : MBE*, 2021.
[58] E. A. Kirchner and J. Bütefür. Towards bidirectional and coadaptive robotic

exoskeletons for neuromotor rehabilitation and assisted daily living : a review. *Current Robotics Reports*, 2022.

[59] Z. Kraljevic, et al. Foresight–generative pretrained transformer (GPT) for modelling of patient timelines using Ehrs. *arXiv :2212.08072*, 2022.

[60] T. H. Kung, et al. Performance of ChatGPT on USMLE : potential for AI-assisted medical education using large language models. *PLoS digital health*, 2023.

[61] B. M. Lake, et al. Building machines that learn and think like people. *Behavioral and brain sciences*, 2017.

[62] M. H. Lee, et al. Enabling AI and robotic coaches for physical rehabilitation therapy : iterative design and evaluation with therapists and post-stroke survivors. *International Journal of Social Robotics*, 2024.

[63] S. Levine, et al. Offline reinforcement learning : Tutorial, review, and perspectives on open problems. *arXiv :2005.01643*, 2020.

[64] C. Li, et al. Multimodal foundation models : From specialists to general-purpose assistants. *arXiv :2309.10020*, 2023.

[65] S. Liu, et al. Deep learning in medical ultrasound analysis : a review. *Engineering*, 2019.

[66] X. Liu, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging : a systematic review and meta-analysis. *The lancet digital health*, 2019.

[67] X. Liu, et al. Deep learning in ECG diagnosis : A review. *Knowledge-Based Systems*, 2021.

[68] Y. Liu, et al. Deep reinforcement learning for dynamic treatment regimes on medical registry data. In *IEEE International Conference on Healthcare Informatics*, 2017.

[69] H. Liyanage, et al. Using ontologies to improve semantic interoperability in health data. *BMJ Health & Care Informatics*, 2015.

[70] P. J. Lucas, et al. A probabilistic and decision-theoretic approach to the management of infectious disease at the ICU. *Artificial Intelligence in medicine*, 2000.

[71] J. Maclure and S. Russell. *AI for humanity : The global challenges*, chapter 8. In Braunschweig and Ghallab [13], 2021.

[72] A. Madani, et al. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 2023.

[73] P. Marquis, et al. Elements for a history of artificial intelligence. In *A Guided Tour of Artificial Intelligence Research : Volume I : Knowledge Representation, Reasoning and Learning*. Springer, 2020.

[74] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 1943.

[75] C. McGreal. *American overdose : The opioid tragedy in three acts*. PublicAffairs, 2018.

[76] V. Milia, et al. Exploring Molecular Energy Landscapes by Coupling the DFTB Potential with a Tree-Based Stochastic Algorithm : Investigation of the Conformational Diversity of Phthalates. *Journal of Chemical Information and Modeling*, 2024.

[77] P. Moireau, et al. External tissue support and fluid–structure simulation in blood flows. *Biomechanics and modeling in mechanobiology*, 2012.

[78] K. L. Moore. Automated radiotherapy treatment planning. In *Seminars in radiation oncology*, volume 29. Elsevier, 2019.

[79] F. Moro, et al. Role of artificial intelligence applied to ultrasound in gynecology oncology : A systematic review. *International Journal of Cancer*, 2024.

[80] L. Mosconi, et al. Early detection of Alzheimer's disease using neuroimaging. *Experimental gerontology*, 2007.

[81] C. S. Muñoz-Valencia, et al. Employing bayesian networks for the diagnosis and prognosis of diseases : A comprehensive review. *arXiv :2304.06400*, 2023.

[82] G. Munsamy, et al. Conditional language models enable the efficient design of

proficient enzymes. *bioRxiv*, 2024.
[83] H. Naveed, et al. A comprehensive overview of large language models. *arXiv :2307.06435*, 2023.
[84] J. Nicolas. Artificial intelligence and bioinformatics. In *A Guided Tour of Artificial Intelligence Research : Volume III : Interfaces and Applications of Artificial Intelligence*. Springer, 2020.
[85] P. Notin, et al. Machine learning for functional protein design. *Nature biotechnology*, 2024.
[86] Y. Oshiro and N. Ohkohchi. Three-dimensional liver surgery simulation : computer-assisted surgical planning with three-dimensional simulation software and three-dimensional printing. *Tissue engineering part A*, 2017.
[87] D. P. Panagoulias, et al. Evaluating llm–generated multimodal diagnosis from medical images and symptom analysis. *arXiv :2402.01730*, 2024.
[88] J. Pearl. *Causality*. Cambridge university press, 2009.
[89] B. K. Petersen, et al. Deep reinforcement learning and simulation as a path toward precision medicine. *Journal of Computational Biology*, 2019.
[90] R. Prevost, et al. 3D freehand ultrasound without external tracking using deep learning. *Medical image analysis*, 2018.
[91] J. C. Pulido, et al. Goal-directed generation of exercise sets for upper-limb rehabilitation assisted by humanoid robots. In *Knowledge Engineering for Planning and Scheduling Workshop*, 2014.
[92] J. Qian, et al. A liver cancer question-answering system based on next-generation intelligence and the large model Med-PaLM 2. *International Journal of Computer Science and Information Technology*, 2024.
[93] H. Rahman, et al. A systematic literature review of 3D deep learning techniques in computed tomography reconstruction. *Tomography*, 2023.
[94] P. Rajpurkar, et al. AI in health and medicine. *Nature medicine*, 2022.
[95] J. G. Richens, et al. Improving the accuracy of medical diagnosis with causal machine learning. *Nature communications*, 2020.
[96] M. Rosol, et al. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination. *Nature Scientific Reports*, 2023.
[97] S. Russell and P. Norvig. *AI : a modern approach (4th Edition)*. Pearson, 2020.
[98] M. K. Sarker, et al. Neuro-symbolic Artificial Intelligence. *AI Communications*, 2021.
[99] G. Schamberg, et al. Continuous action deep reinforcement learning for propofol dosing during general anesthesia. *Artificial Intelligence in Medicine*, 2022.
[100] H. Shi, et al. Continual learning of large language models : A comprehensive survey. *arXiv :2404.16789*, 2024.
[101] E. Shortliffe. *Computer-based medical consultations : MYCIN*. Elsevier, 1976.
[102] I. Sim and C. Cassel. The Ethics of Relational AI - Expanding and Implementing the Belmont Principles. *The New England journal of medicine*, 2024.
[103] D. Singh, et al. A comprehensive review of intelligent medical diagnostic systems. In *2020 4th International conference on trends in electronics and informatics (ICOEI)(48184)*. IEEE, 2020.
[104] K. Singhal, et al. Large language models encode clinical knowledge. *Nature*, 2023.
[105] R. Sirriyeh, et al. Coping with medical error : a systematic review of papers to assess the effects of involvement in medical errors on healthcare professionals' psychological well-being. *Quality and safety in health care*, 2010.
[106] B. F. Skinner. *The Behavior of Organisms : An Experimental Analysis*. Appleton, 1938.
[107] R. H. Taylor. A perspective on medical robotics. *Proceedings of the IEEE*, 2006.
[108] S. Teipel, et al. Multimodal imaging in Alzheimer's disease : validity and usefulness for early detection. *The Lancet Neurology*, 2015.
[109] M. Tejedor, et al. Reinforcement learning application in diabetes blood glucose

control : A systematic review. *Artificial intelligence in medicine*, 2020.
[110] A. J. Thirunavukarasu, et al. Large language models in medicine. *Nature medicine*, 2023.
[111] E. L. Thorndike. *Animal Intelligence*. Macmillan, 1911.
[112] D. S. W. Ting, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*, 2017.
[113] J. Troccaz, et al. Frontiers of medical robotics : from concept to systems to clinical translation. *Annual review of biomedical engineering*, 2019.
[114] UN AI Advisory Board. Governing AI for Humanity, Interim Report, 2023.
[115] R. J. Van Sloun, et al. Deep learning in ultrasound imaging. *Proceedings of the IEEE*, 2019.
[116] J. Venugopalan, et al. Multimodal deep learning models for early detection of Alzheimer's disease stage. *Scientific reports*, 2021.
[117] G. Wang, et al. Deep learning for tomographic image reconstruction. *Nature machine intelligence*, 2020.
[118] S. Weng, et al. Combining deep learning and coherent anti-stokes raman scattering imaging for automated differential diagnosis of lung cancer. *Journal of biomedical optics*, 2017.
[119] L. Wong, et al. From word models to world models : Translating from natural language to the probabilistic language of thought. *arXiv :2306.12672*, 2023.
[120] H. Wu, et al. A survey on clinical natural language processing in the United Kingdom from 2007 to 2022. *NPJ digital medicine*, 2022.
[121] J. Xia, et al. Three-dimensional virtual reality surgical planning and simulation workbench for orthognathic surgery. *The International journal of adult orthodontics and orthognathic surgery*, 2000.
[122] X. Xue, et al. Efficient ontology meta-matching based on interpolation model assisted evolutionary algorithm. *Mathematics*, 09 2022.
[123] A. Yala, et al. Optimizing risk-based breast cancer screening policies with reinforcement learning. *Nature medicine*, 2022.
[124] Y. Yanagita, et al. Accuracy of ChatGPT on medical questions in the national medical licensing examination in Japan : evaluation study. *JMIR Formative Research*, 2023.
[125] K.-L. A. Yau, et al. Reinforcement learning models and algorithms for diabetes management. *IEEE Access*, 2023.
[126] C. Yu, et al. Reinforcement learning in healthcare : A survey. *ACM Computing Surveys (CSUR)*, 2021.
[127] W. J. Yun, et al. Hierarchical deep reinforcement learning-based propofol infusion assistant framework in anesthesia. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
[128] W. X. Zhao, et al. A survey of large language models. *arXiv :2303.18223*, 2023.
[129] Y. Zhao, et al. Reinforcement learning design for cancer clinical trials. *Statistics in medicine*, 2009.
[130] Y. Zhao, et al. Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics*, 2011.