Statistique et société
Décembre 2020

Volume 8, Numéro 3

Varia

Sommaire
Statistique et société

Volume 8, Numéro 3

Sommaire
Statistique et société

Volume 8, Numéro 3

Statistique et société

Magazine quadrimestriel publié par la Société Française de Statistique.
Le but de Statistique et société est de présenter, d'une manière attrayante et qui invite à la réflexion, l'utilisation pratique de la statistique dans tous les domaines de la vie. Il s'agit de montrer comment l'usage de la statistique intervient dans la société pour y jouer un rôle souvent inaperçu de transformation, et est en retour influencé par elle. Un autre dessein de Statistique et société est d'informer ses lecteurs avec un souci pédagogique à propos d'applications innovantes, de développements théoriques importants, de problèmes actuels affectant les statisticiens, et d'évolutions dans les rôles joués par les statisticiens et l'usage de statistiques dans la vie de la société.

ISSN 2269-0271

Éditorial

Emmanuel DIDIER
Rédacteur en chef de Statistique et société

Chère lectrice, cher lecteur,

Un élément très important de ce numéro n'est pas visible au premier coup d'œil car il ne figure pas dans la table des matières mais dans l'ours : le comité éditorial de notre revue a été renouvelé. Ce comité joue pour nous un rôle capital. Il regroupe des représentants des groupes de la SFdS et des personnalités extérieures qui nous aident à diffuser la revue vers un plus large public et à vous proposer des articles ou des dossiers riches et variés. En plus de ces échanges directs, nous nous rencontrons collégialement une fois l'an pour faire le point sur l'avancement de la revue. Un grand merci donc, d'abord aux sortants pour le travail accompli et, ensuite, aux entrants et à ceux qui restent de nous accorder leur confiance et leur intérêt.

Ce numéro est un numéro varia regroupant quatre contributions. Loup Wolff, directeur du service statistique du Ministère de la culture, présente la dernière édition de l'enquête sur les pratiques culturelles des Français, qui fait suite à la précédente édition de 2008. Son article insiste sur la croissance impressionnante du numérique parmi les pratiques culturelles. Ensuite, deux articles abordent les impacts sociaux de l'usage des algorithmes d'apprentissage automatique et d'intelligence artificielle : Philippe Besse et ses coauteurs réfléchissent au cadre juridique et éthique du numérique en santé ; Nicolas Bousquet et ses coauteurs de l'entreprise Quantmetry présentent certains usages de l'apprentissage profond. Enfin, Cyprien Rousset et Ariane Sessego reviennent sur l'histoire d'Amira, un ensemble de chercheurs de l'ORSTOM et administrateurs de l'INSEE très intéressant quoique trop méconnu et qui a longuement réfléchi aux spécificités de la statistique africaine et aux transferts technologiques entre la France et les anciennes colonies.

Trois compte-rendu d'ouvrage closent le numéro : le premier concerne le livre intitulé *La Nationalité, une histoire de chiffres*, de Morgane Labbé, présenté par Jean-Jacques Droesbeke, le second, *Le Théorème d'hypocrite. Histoire de la manipulation par les chiffres de Pythagore au Covid-19* de Antoine Houlou-Garcia et Thierry Maugenest, présenté par Antoine Rolland et enfin le troisième, *L'Empire des chiffres*, d'Olivier Martin par Camille Beaurepaire.

Bonne lecture !

Emmanuel Didier

Observer les pratiques culturelles à l'ère du numérique

Loup WOLFF[1]
Chef du Département des études, de la prospective et des statistiques, Ministère de la culture

TITLE
Surveying cultural participation in the digital age

RÉSUMÉ
Rééditée à six occasions de 1973 à 2018, l'enquête sur les pratiques culturelles est un dispositif structurant pour la connaissance tant dans le champ administratif que scientifique, voire médiatique. Cet article retrace les questionnements ayant accompagné la reconduction de ce dispositif pour sa sixième édition, amorcée dès 2015, jusqu'à son administration sur le terrain en 2018 et début 2019. Outil de connaissance, mais également perçu comme outil d'évaluation, donnant une image souvent jugée défavorable de l'action des politiques publiques, cette dernière édition a pourtant bénéficié de moyens renforcés, étendant les potentialités d'analyse de cette enquête. Au-delà des enjeux d'image, l'originalité de ce dispositif, son ancienneté, sa légitimité scientifique et statistique ont assuré la poursuite de cette initiative exceptionnelle.

Mots-clés **:** *enquête, pratiques culturelles.*

ABSTRACT
Continued on six occasions from 1973 to 2018, the cultural participation survey is a structuring device for knowledge in the administrative, scientific and even media fields. This article traces the questions that accompanied the renewal of this device for its sixth edition, which began in 2015, until the fieldwork held in 2018 and early 2019. Knowledge tool, but also perceived as an evaluation tool, giving an image often considered unfavorable for the action of public policies, this latest edition nevertheless benefited from increased resources, extending the analysis potential of this survey. The originality of this device, its age, its scientific and statistical legitimacy ensured the continuation of this exceptional initiative.

Keywords**:** *cultural practices, survey.*

1. Loup.wolff@culture.gouv.fr

1. Introduction

Continuant une série ayant débuté en 1973 pour sa première édition, puis poursuivie en 1981, 1989, 1998 et 2008, l'*enquête sur les pratiques culturelles* a été reconduite en France pour une sixième édition en 2018. Soutenue et financée comme les précédentes par le ministère de la culture et portée par son service d'études et de statistique – le Département des études, de la prospective et des statistiques (DEPS) –, cette sixième édition bénéficie de moyens étendus : un doublement de la taille d'échantillon financée (qui passera entre 2008 et 2018 de 5 000 à 9 300 répondants pour la France métropolitaine), ainsi qu'une extension aux départements et régions d'Outre-Mer (La Réunion, Mayotte, Martinique, Guadeloupe et Guyane) – rejoignant ainsi pour la première fois ce dispositif qui devient véritablement national.

Décider de reconduire en 2018, à l'ère de la généralisation des pratiques numériques, un tel dispositif d'enquête (interrogation d'un échantillon aléatoire d'individus, en face à face, avec l'intervention d'enquêteurs) pose question à plusieurs titres. Ce choix, lourd de conséquences à la fois pour la communauté scientifique, pour le service qui le porte, ainsi que pour les finances publiques a fait l'objet de discussions serrées. Pour sa reconduction en 2018, l'enquête sur les pratiques culturelles a ainsi dû faire face à trois défis, mêlant de façon indissociable enjeux institutionnels et méthodologiques, sur lesquels ce texte souhaite revenir : réaffirmer la pertinence scientifique et administrative d'une nouvelle édition en 2018 et ainsi obtenir l'engagement du ministère de la culture dans le portage administratif et financier de cette opération de grande ampleur ; assurer la continuité avec les éditions précédentes et ainsi préserver la continuité historique des analyses ; opérer malgré tout les évolutions nécessaires permettant d'adapter le questionnement aux nouvelles formes de pratiques.

2. Centralité de l'enquête sur les pratiques culturelles dans le champ de la connaissance

Pour comprendre ce choix, il faut commencer par rappeler l'importance de la place occupée en France par les éditions successives de cette enquête tant dans le champ administratif, que scientifique[2]. Le contexte initial de sa conception « cherchait à concilier deux traditions sociologiques qui s'opposent sur bien des points, celle des travaux pionniers de J. Dumazedier dans le domaine des loisirs et celle des premières investigations de P. Bourdieu dans celui de la culture » (Donnat, 2003). La première édition de cette série fait suite à la parution en 1962 de l'ouvrage de J. Dumazedier sur la « société du loisir » (Dumazedier, 1962) et surtout à la parution en 1966 de *L'Amour de l'art* (Bourdieu et Darbel, 1966) – ouvrage qui mobilisait de nombreux résultats issus d'enquêtes menées avec le soutien du service d'études et de statistique du ministère de la culture, alors nommé *Service des études et recherches* (dirigé par Augustin Girard), et qui toutes préfiguraient ce que deviendra l'enquête sur les pratiques culturelles (Girard, 1994). La poursuite de ces travaux conduira Pierre Bourdieu, alors membre du *Centre de sociologie de l'éducation et de la culture*, à énoncer en 1979 dans *La Distinction* son analyse novatrice de la structure sociale, portant au jour les conditions sociales de la production du goût (Bourdieu, 1979). Décrivant le rôle joué par la culture – sédimentée sous la forme d'un capital symbolique inégalement détenu par les individus – dans la structuration des rapports sociaux, cette analyse a eu un impact majeur sur les sciences sociales, en France comme à l'étranger, et a retenti bien au-delà du monde académique (Coulangeon, 2011).

Seul dispositif d'observation abordant en France les pratiques culturelles de façon transversale sur l'ensemble du champ (spectacle vivant, industries culturelles, patrimoine et au-delà), au

2. Pour s'en faire une idée, contradictoire, le lecteur pourra à la fois se référer à l'ouvrage de Ph. Coulangeon poursuivant les analyses de P. Bourdieu dans *La Distinction* et faisant un usage intensif de cette série d'enquêtes (Coulangeon, 2011) ou encore l'article d'H. Glévarec déplorant les effets de la centralité et des usages théoriques de ce dispositif dans le champ scientifique (Glévarec, 2016).

niveau national et avec une telle profondeur historique, l'enquête sur les pratiques culturelles occupe une place tout aussi centrale au sein des dispositifs mobilisés par l'administration française pour objectiver et penser son action. Depuis ses débuts, elle est en effet restée fidèle aux quatre objectifs aux origines de sa conception :

- Observer les comportements et pratiques culturels de la population résidant en France, en conservant une acception large de ce qui fait la culture, pour mieux appréhender la diversité des rapports à la culture ;
- Fournir des analyses détaillées sur l'évolution de ces comportements et pratiques ;
- Adapter le questionnement aux comportements et pratiques émergents (notamment liés aux nouvelles technologies et nouveaux modes d'accès à la culture) ;
- Mieux identifier les facteurs d'accès ou de distanciation à la culture.

Le dispositif, bien qu'ayant connu quelques évolutions notables (liées principalement à l'élargissement progressif du champ de l'enquête à des pratiques plus populaires), est resté relativement stable depuis ses débuts aussi bien dans sa méthodologie que dans ses objectifs et dans la formulation des questions. Les éditions successives constituent ainsi un corpus de données cohérent, qui a pu donner lieu à des exploitations longitudinales (en coupes répétées et quasi-panels). Ce corpus permet aujourd'hui d'actualiser la connaissance des transformations structurelles qui touchent depuis près d'un demi-siècle les comportements pouvant être qualifiés de « culturels » au sens large – intégrant aussi bien les formes les plus légitimes de la culture (au sens de [Bourdieu, 1979]) que des pratiques connexes (jardinage, tricot, spectacles sportifs, …). Fidèle là aussi à ses origines, l'enquête continue de jouer « de l'ambiguïté du terme 'pratiques culturelles' qui permettait à la fois de mener une sociographie de la fréquentation des équipements culturels dans la perspective de la planification culturelle et de revendiquer une approche large des usages du temps libre, ce qui était la meilleure manière – la suite allait en apporter la preuve – de tomber sous la double critique contradictoire d'imposer une vision légitimiste de la culture et de participer activement au triomphe du relativisme culturel » (Donnat, 2003).

3. Une édition 2018 entre continuité et adaptation

L'édition 2018 a été pensée à la fois comme un prolongement de cette série, tout en intégrant une réflexion de fond sur un nécessaire renouvellement des problématiques et des moyens mis en œuvre par l'enquête pour y répondre. Identifiée en France comme un instrument incontournable pour le suivi des comportements culturels, cette enquête a en effet fait l'objet d'une forte demande de renouvellement de la part des décideurs en charge de la politique culturelle, aussi bien au sein des services centraux et déconcentrés du ministère ainsi que de la part des responsables d'établissements et des acteurs de la vie culturelle.

L'édition 2018 poursuit les objectifs des éditions précédentes : décrire l'évolution des pratiques culturelles et analyser les relations entre les différentes formes d'accès à l'art et à la culture, dans un contexte de généralisation de l'usage du numérique (via les ordinateurs, les tablettes, les téléphones, …). Tout au long d'un questionnaire dont la passation a duré en moyenne près d'une heure sur le terrain, l'enquête a ainsi pour vocation d'apporter des éléments de réponse aux interrogations historiques allouées au dispositif : décrire les publics du théâtre, du concert, du cinéma, des bibliothèques et de la lecture, les usages culturels des médias, les pratiques amateurs (notamment pratique d'un instrument de musique, mais aussi la photographie, la danse, etc.). Du fait de l'existence des cinq éditions précédentes, il s'agit aussi de mesurer l'évolution de la diffusion des différentes pratiques culturelles et celle du profil des publics concernés : la fréquentation des musées, théâtres, salles de cinéma, etc. a-t-elle augmenté ou baissé ? Et la lecture de livres, l'écoute de musique ou la pratique en amateur d'activités artistiques ? Dans quelle mesure le profil des personnes concernées par ces diverses activités a-t-il changé : féminisation du lectorat de livres, vieillissement du public des théâtres, concerts

classiques, expositions, etc. ?

Compte tenu de l'importance croissante prise depuis 2008 par les équipements et les contenus numériques, du taux élevé de pénétration d'Internet dans les ménages (selon l'enquête sur les « conditions de vie et les aspirations » commandée par l'ARCEP au CREDOC, 89% des Français ont utilisé internet en 2018, contre 52% en 2005) et de l'évolution conjointe des caractéristiques de l'offre culturelle (développement des contenus trans-médias, fragilisation des filières traditionnelles), l'édition 2018 de l'enquête sur les pratiques culturelles a dû mieux tenir compte des nouveaux usages culturels et a été confrontée au moment de sa conception à une double contrainte : garder un protocole et une méthodologie les plus proches possibles de ceux des éditions précédentes pour pouvoir comparer les résultats sur longue période ; mais également, tenir compte de l'émergence du numérique et ne plus aborder les pratiques culturelles uniquement par le médium (le musée, le livre, le journal, la télévision, le cinéma, le disque, la scène, etc.), mais également par le contenu, dans la mesure où les œuvres peuvent désormais être de plus en plus consommées chez soi (y compris les spectacles, les expositions), et sans supports physiques. Aussi la nouvelle édition intègre-t-elle ces nouvelles formes d'accès à l'art et à la culture dans la perspective d'offrir une description complète de leurs usages et d'apporter des éléments de réponse aux interrogations que suscite leur développement : les usages culturels du numérique sont-ils plutôt le fait de personnes ayant un fort niveau d'engagement dans la culture ou concernent-ils des personnes peu ou pas habituées des équipements culturels ? Plus généralement, quelles relations existent entre ces usages et les pratiques culturelles traditionnelles ?

Les résultats issus des éditions successives de l'enquête sont régulièrement utilisés tant par les médias que par les professionnels des différents domaines relevant du champ de compétence du ministère de la culture. Pour répondre aux questionnements fréquemment formulés par ces différents acteurs, le questionnaire aborde des thématiques comme les loisirs et les vacances, les pratiques en amateur, les jeux vidéo, les films, les séries et les émissions, l'information, l'écoute de la musique, la fréquentation des bibliothèques et la lecture, le cinéma, le spectacle vivant comprenant le théâtre, la danse et les festivals ainsi qu'un bloc de questions portant sur les musées, le patrimoine et les expositions.

Des questions concernant la situation familiale, la situation du ménage, la situation vis-à-vis du travail, l'activité professionnelle, la situation dans l'enfance et les ressources culturelles sont aussi présentes dans le questionnaire. Celles-ci permettent de contextualiser les pratiques des personnes enquêtées et mettre en avant des facteurs explicatifs.

4. Répéter le constat d'échec de la démocratisation ?

La reconduction de l'enquête en 2018 ne s'est pas pour autant faite sans questionnements pressants. L'un des premiers fronts ouverts le fut sur la question de reconduire un dispositif qui risquerait de réitérer un constat déjà ancien et bien établi : celui de l'incapacité de la politique de l'offre, vigoureusement menée en France, à contredire une forme de déterminisme social, constaté empiriquement, dans la formation des comportements culturels. Malgré une offre objectivement croissante et foisonnante, la physionomie des publics n'a pas radicalement évolué et – plus grave encore ! – cette multiplication des propositions culturelles a largement profité aux catégories supérieures qui ont intensifié leurs pratiques (Donnat, 2009).

Ce constat, déjà établi à l'occasion des deux premières éditions de l'enquête en 1973 et 1981, n'a pas été à l'époque perçu comme problématique : il fixait le point de départ que des politiques culturelles volontaristes allaient pouvoir permettre de dépasser bientôt ! Si les pratiques culturelles observées à la fin des années 1970 étaient encore empruntes d'une certaine pesanteur sociale, les décideurs comptaient sur le nouvel élan insufflé à partir des années 1980

avec l'arrivée de François Mitterrand au pouvoir, secondé par son célèbre ministre de la culture, Jack Lang, pour permettre de changer la donne et d'ouvrir une nouvelle ère culturelle (Martigny, 2016). Or, les résultats qui se dégagèrent des éditions successives de l'enquête confirmèrent les uns après les autres le rôle prédominant et puissamment enraciné joué par les structures sociales dans la formation des goûts et des dégoûts, ainsi que dans les pratiques culturelles en France. Les analyses historiques nous apprennent que Jack Lang « fut très désagréablement surpris par les chiffres publiés dans l'Enquête sur les pratiques culturelles des Français en 1990, qui montraient que la consommation culturelle n'avait guère augmenté depuis une dizaine d'années, malgré l'augmentation du budget et la multiplication de l'offre culturelle qui en avait résulté » (Martin, 2012). Cette désagréable découverte de 1990 s'est ensuite répétée et confirmée à chacune des éditions suivantes, si bien que les analyses historiques d'évolution des pratiques culturelles conduisent avant tout à révéler la dimension structurelle des comportements culturels et l'importance des « pesanteurs sociales » dans ce domaine, nuancées tout de même par des « dynamiques générationnelles » amenant l'émergence progressive de pratiques nouvelles (Donnat, 2011). D'autres travaux ont pu nuancer le constat d'un accaparement des politiques culturelles par les classes supérieures : Glévarec (2016) montre ainsi qu'une intensification des pratiques est observable dans l'ensemble des classes sociales, certes à des degrés divers selon les formes culturelles et à un rythme décroissant à mesure que l'on descend dans l'échelle sociale. Mais malgré ces arguments partiellement contradictoires, c'est malgré tout le prisme d'un certain « échec des politiques de démocratisation culturelle » qui s'impose dans la lecture des résultats issus de l'enquête sur les pratiques culturelles.

Dans ce contexte, structurellement déceptif pour son commanditaire, le ministère de la culture, la reconduction de ce dispositif aurait pu être questionnée. De fait, des observateurs ont pu avoir le sentiment que cette enquête, ainsi que le service qui la porte depuis ses débuts au sein du ministère, relevaient d'une forme d'anomalie administrative. À propos de la personnalité qui, au sein du DEPS, a conduit et exploité les éditions 1990, 1998 et 2008 de la série, Michel Guerrin, rédacteur en chef du quotidien national Le Monde, a pu ainsi écrire dans sa chronique du 26 octobre 2018 : « Olivier Donnat est sociologue au ministère de la culture. Il est un loup dans la bergerie, l'ennemi de l'intérieur, le gars qui casse le moral, fait tomber les illusions » – montrant ainsi combien est forte la perception d'un antagonisme entre ce dispositif de connaissance et les enjeux de l'administration culturelle. Ce constat montre également toute l'ambiguïté du rôle attribué à ce dispositif d'enquête, aussi bien dans l'administration, que plus largement auprès des observateurs de la vie culturelle. Malgré les nombreux avertissements de leurs concepteurs (Girard, 1994 ; Donnat, 2003), les résultats issus des éditions successives de l'enquête ont systématiquement été interprétés – à tort – comme une évaluation des politiques culturelles, alors même qu'il ne peut s'agir que d'un point d'observation, à un moment donné, de l'état de pratiques multiples – sans possibilité de les relier causalement à l'historicité des politiques mises en œuvre, ou encore à d'autres facteurs purement exogènes à l'action publique. Les corrélations observées et leurs évolutions entremêlent en effet de façon radicalement endogène les effets possibles attribuables aux politiques publiques culturelles, portées aussi bien au niveau national que local par les collectivités territoriales, les dynamiques démographiques transformant la population française, la diffusion de technologies nouvelles, de nouvelles modalités de consommation portées par les industries. Dans la perception commune de l'outil, cette forme de schizophrénie prêtée au dispositif, ainsi que l'ambiguïté de son caractère évaluatif sont sources de tensions permanentes, qui auraient pu conduire à un désengagement de l'administration dans son soutien, à la fois humain et financier.

Cela n'a pas été le cas : non seulement l'enquête a été reconduite en 2018, mais le ministère a même consenti à doubler son financement afin de la doter d'un échantillon plus conséquent, comme précisé en introduction, permettant ainsi d'approfondir les analyses.

Ce soutien réaffirmé, et même renforcé, au dispositif s'explique par plusieurs facteurs. D'abord

probablement, en raison de la dimension patrimoniale de cette enquête, qui appelait une édition supplémentaire dix ans après 2008. Les différentes éditions, ainsi que la parution des résultats issus de leur exploitation statistique, ont scandé la vie de l'administration culturelle et des débats sur cette politique, il était manifestement impensable pour les décideurs de cette nouvelle décennie qu'une nouvelle édition de l'enquête ne voie pas le jour sous leur mandat. Dès 2015, le ministère de la culture, via son secrétaire général, Christopher Miles, a explicitement réaffirmé son soutien à la reconduction de l'enquête, inscrivant même explicitement cet objectif au cœur du programme de travail du DEPS, son service statistique. Sa réédition faisait de plus l'objet de demandes régulières de la part d'opérateurs, publics et privés, des organisations professionnelles et plus généralement de nombreux membres de la société civile, pour qui ce dispositif reste une boussole incontournable.

Enfin, la justification d'une nouvelle édition a été renforcée par un certain déplacement des enjeux de connaissances autour de ce que sont les pratiques culturelles en France. Les cinq premières éditions ont fourni des éléments de connaissance indispensables sur la structuration de ces pratiques en fonction des facteurs sociaux et démographiques. L'édition 2018 propose d'approfondir ces constats, désormais bien renseignés et dont le caractère profondément structurel ne fait plus de doute, en les décomposant selon plusieurs axes d'analyse, nouveaux pour ce dispositif : en fonction des caractéristiques territoriales définissant les lieux d'habitation et de circulation des répondants, en fonction de leurs groupes sociaux (et approfondissant donc les constats établis au niveau national, pour aller observer de plus petites différences observables entre sous-groupes), et enfin en fonction des multiples facteurs qui définissent les répondants au moment de leur interrogation (sexe et âge, mais également origine géographique, parcours migratoire, état de santé).

L'enjeu ne devient donc plus seulement de reconduire les grands constats généraux décrivant la structuration des pratiques culturelles en France, il devient également de réussir à interpréter les différences relatives existantes entre de multiples configurations de pratiques, selon les multiples facteurs qui influent sur le parcours des personnes. Apparu d'abord à l'occasion des échanges scientifiques et méthodologiques organisés au sein du DEPS et ensuite avec le comité scientifique[3] formé pour accompagner ces réflexions, cet enjeu supplémentaire a ensuite trouvé une résonnance particulière auprès des décideurs publics et a joué un rôle important, non pas tant dans la décision de reconduire ce dispositif, intervenue finalement précocement, mais plus significativement dans son renforcement, via notamment le quasi-doublement de la taille de son échantillon, et donc de son coût, près de cinquante ans après sa première édition.

5. Une enquête ménage à l'ère du numérique ?

Un autre questionnement, plus méthodologique, aurait pu conduire à l'abandon du dispositif, y compris au sein du Département des études, de la prospective et des statistiques (DEPS). Les travaux du DEPS sont en effet à la fois anciens et nombreux à renseigner le développement puissant des pratiques numériques au sein de la population française (Donnat, 2009, 2011). Les pratiques audiovisuelles, l'écoute quotidienne de musique… ont toutes progressé à un rythme soutenu, et cela même avant l'avènement d'Internet dans les foyers. Mais, en se généralisant, ces pratiques, devenues à la fois toujours plus fréquentes et plus numériques, sont de plus en

3. Ce Comité scientifique se compose de chercheurs ayant été identifiés par des travaux publiés exploitant les éditions précédentes de la série, ainsi que des agents du DEPS ayant travaillé sur les éditions anciennes ou missionnés pour contribuer à cette nouvelle édition : Philippe COULANGEON, Hervé GLEVAREC, Karim HAMMOU, Olivier ROUEFF (CNRS), Gaël DE PERETTI (Insee), Christine DETREZ (ENS Lyon), Stéphane DORIN (Université de Limoges) et Nicolas ROBETTE (CREST-Laboratoire de Sociologie Quantitative), ainsi que Nathalie BERTHOMIER, Jean-Michel GUY, Amandine LOUGUET, Sylvie OCTOBRE et Loup WOLFF (pour le DEPS). Il s'est réuni dix fois, à un rythme mensuel, à partir de septembre 2016. Ces réunions mensuelles ont permis de rythmer les travaux de rénovation du questionnaire et de discuter dans le détail, à chaque étape, les décisions qui ont été prises concernant cette sixième édition. Aux côtés du Comité scientifique, s'ajoute également le rôle joué par les deux comités d'utilisateurs réunis coup sur coup en 2017 – le premier au sein de l'administration culture (ministère et structures partenaires), le second avec la société civile (associations culturelles, professionnelles, territoriales).

plus difficiles à interroger dans le cadre canonique de l'entretien en face-à-face au domicile, avec un.e enquêteur.e dépêché.e par le commanditaire public. Elles ont en effet intégré tous les moments quotidiens de la vie, se déclenchent parfois par le simple recours à un dispositif technologique ou à une application, délivrant un flux parfois structuré algorithmiquement en fonction des préférences des consommateurs – sans que ces derniers portent une attention particulière à ce qu'ils reçoivent (Cardon, 2015). Que peuvent-ils donc répondre à un.e enquêteur.e venu.e les interroger sur leurs pratiques, leurs fréquences, leurs goûts ?

Cette dernière question est cruciale : est-il encore possible de renseigner aujourd'hui les pratiques culturelles d'une personne sur une base déclarative, dans le cadre d'une interaction avec un.e enquêteur.e chargé.e de les passer en revue ? Certaines pratiques ont à ce point intégré le quotidien des personnes qu'il devient difficile de se les remémorer et de les objectiver face à un.e enquêteur.e.

Ce n'est bien évidemment pas le cas de toutes les pratiques : les sorties ou visites (musées, monuments, spectacles vivants, cinémas) restent inscrites dans un moment et un espace qui en font des événements objectivables dans ces conditions d'enquête. Il n'en reste pas moins que cette incapacité croissante à identifier les productions culturelles fréquentées, liée à la fois à une intensification de ces pratiques et au recours de plus en plus fréquent à des dispositifs tiers pour construire le flux, touche les pratiques aujourd'hui les plus courantes : le visionnage de contenus audiovisuels (télévision, films, séries, vidéos Web, …), l'écoute de musique, de la radio.

Pour répondre à ces enjeux, évidemment majeurs pour la compréhension des évolutions des comportements culturels, chercheurs et administrations tendent désormais à développer le recours à l'analyse des traces numériques. Plutôt qu'une information collectée à l'occasion d'une interaction avec un.e enquêteur.e, ces approches privilégient l'analyse directe des actions exécutées par les personnes enregistrées par les dispositifs technologiques. Les personnes n'ont ainsi plus à déclarer ce qu'elles font, ce sont directement leurs actions qui sont observées et collectées (Cardon, 2015).

A la lumière de ces réflexions, la question s'est donc posée de la pertinence de la reconduction d'une méthodologie de type « enquête ménage » (i.e. enquête par questionnaires administrés directement auprès des personnes, et donc sur une base déclarative), avec recours à un réseau d'enquêteur.e.s, pour continuer à observer les comportements et pratiques culturels en France. Cette question a fait l'objet de vifs débats, tant au sein du DEPS, qu'au sein du Comité scientifique qui a été réuni auprès du DEPS pour mener à bien le travail de conception de cette sixième édition de l'enquête sur les pratiques culturelles.

A l'issue de ces débats, deux arguments se sont imposés pour la reconduction de la méthodologie des précédentes éditions de l'enquête sur les pratiques culturelles. D'abord la volonté de poursuivre les séries établies sur les pratiques culturelles par les précédentes éditions, avec une attention portée aux conditions méthodologiques assurant une certaine continuité des séries historiques (formulation des questions, effets de l'échantillonnage). Et ensuite, la nécessité de conserver et même de développer le lien établi entre comportements culturels et caractéristiques sociales des individus. Car, si les big data et l'analyse des traces numériques ouvrent de larges possibilités, elles trouvent rapidement leurs limites quand il s'agit de rapporter les comportements observés dans un très grand détail avec les caractéristiques des personnes agissantes : bien souvent, le chercheur ne dispose que d'informations très parcellaires concernant les personnes – sauf dans le cas des panels, mais qui posent quant à eux d'autres difficultés méthodologiques (biais de composition, attrition, complexité des dispositifs techniques à prévoir pour l'identification des personnes en action). Outre le fait que les équipements numériques, et plus encore les compétences nécessaires pour les manipuler

ne sont pas encore exhaustivement répartis au sein de la population française, les enquêtes ménage restent une méthodologie incontournable, et encore difficilement dépassable pour qui souhaite analyser les liens entre les comportements et les caractéristiques démographiques et sociales des individus, les trajectoires géographiques et biographiques, ainsi que le croisement des résultats avec des variables territoriales.

Enfin, cette approche reste également centrale si, comme c'est l'une de ses fonctions principales, il s'agit de croiser les comportements culturels entre eux : articulation de la culture d'écran, avec les pratiques de sortie, les pratiques en amateur, l'écoute de la radio, etc. Pour une herméneutique des configurations dans lesquelles l'ensemble de ces pratiques s'articulent pour un même individu, la reconduction de l'approche par une enquête par questionnaires administrés par des enquêteur.e.s au sein des ménages s'est finalement imposée. Enfin, le recours à un réseau d'enquêteur.e.s, chargé de couvrir un échantillon d'adresses tirées aléatoirement dans un registre national de référence (ici l'« échantillon maître » de l'Insee, issu du recensement de la population), reste la meilleure garantie pour limiter les biais de composition[4].

6. Un questionnaire à reconstruire : abandon de l'approche par les dispositifs, pour celle par les contenus

Une fois la question de la méthodologie tranchée, d'autres débats ont continué de rythmer la conception de cette sixième édition. L'un des plus cruciaux, et des plus difficiles à surmonter a concerné les principes mêmes à l'origine de la construction du questionnaire historique. Les versions les plus anciennes du questionnaire s'appuyaient en effet sur les dispositifs physiques pour approcher la question des contenus artistiques et culturels, profitant d'une certaine homologie de ces deux espaces dans le monde pré-numérique : ainsi la lecture pouvait-elle être approchée par le fait d'ouvrir un livre, le cinéma par le fait de se rendre dans une salle, les productions audiovisuelles par le fait d'allumer un poste de télévision. Ces questionnaires pouvaient ainsi faire le choix d'interroger les dispositifs (les objets : tels les livres, les postes de télévision, de radio – ou les lieux : les salles de cinéma, de théâtre, les musées) pour observer la façon dont les comportements des répondants se structurent par domaines artistiques et culturels.

Poursuivant des réflexions qui ont déjà été à l'origine d'évolutions importantes dans le questionnaire de la cinquième édition (2008), les débats ont fait apparaître la nécessité de parachever la reconstruction du questionnaire en abandonnant l'entrée autrefois privilégiée par les dispositifs. Pour la cinquième édition en 2008 déjà, la partie du questionnaire concernant l'écoute de musique avait fait le choix de rompre avec l'édition précédente, en interrogeant la pratique d'écoute de musique elle-même, plus que l'allumage d'un dispositif Hi-fi ou d'un poste de radio.

Mais la transformation ne fut que partielle à l'époque : la télévision restait arrimée au poste de télévision, la radio au poste de radio, le cinéma aux salles de cinéma, etc. Malgré une volonté clairement affirmée de conserver la transversalité historique du dispositif sur l'ensemble des champs culturels, il est rapidement apparu impossible de poursuivre cette entrée par les dispositifs : en 2018, le numérique a en effet largement contribué à brouiller l'homologie qui a pu exister (même imparfaitement) entre dispositifs et contenus. La télévision se regarde désormais sur de très nombreux supports, la radio et/ou la musique s'écoutent sur des dispositifs divers. Pour l'édition 2018, l'ensemble du questionnaire a ainsi dû être retravaillé pour finaliser cette mue et repenser la formulation de ces contenus en catégories, parlantes

4. Là où les méthodologies Web restent beaucoup plus tributaires de l'inégal équipement des ménages en information, en couverture Internet, ainsi qu'en compétences pour manier ces technologies.

pour les enquêtés et pertinentes pour la recherche, et qui sont devenues les nouvelles conventions mobilisées par le questionnaire pour décrire les pratiques et comportements. Ainsi, à l'issue de ce travail de rénovation, a-t-il été notamment décidé de continuer à interroger les répondants sur l'activité « regarder la télévision », en abandonnant la mention au dispositif « poste de télévision ». L'hypothèse est en effet faite qu'il continue d'exister une pratique sociale pouvant être approchée par cette formulation (et dont on cherchera ensuite à savoir si elle a été mise en œuvre à l'aide d'un poste de télévision, une tablette, un smartphone, ou tout autre support technologique) et que l'enquête se doit de la mesurer (notamment pour permettre de continuer à qualifier l'évolution de cette pratique au cours des dernières décennies). Cette approche renouvelée nous a également conduits à identifier, et donc traiter séparément dans le questionnaire, des contenus audiovisuels spécifiques, liés à des pratiques que nous voulions mesurer : le visionnage de films (on interroge là aussi dans un second temps les supports techniques utilisés pour cette pratique), de séries, de vidéos Web. La même mue a été appliquée à l'écoute de la radio. Enfin, les parties du questionnaire concernant les pratiques de sortie ou de visite (spectacle vivant et patrimoine) n'ont pas été concernées par ces réflexions.

7. L'hybridation croissante des formes et des genres culturels

Ces choix méthodologiques (et pratiques) ont été réalisés avec la claire conscience des difficultés posées par une autre évolution concernant les productions artistiques et culturelles : alors même qu'était réaffirmée la nécessité de mettre au premier plan du questionnaire les contenus artistiques et culturels (et de reléguer au second plan la question, malgré tout cruciale, des dispositifs), les discussions ont fait apparaître les difficultés croissantes que nous rencontrions, en tant qu'observateurs du monde social et culturel, à recourir à des conventions catégorielles stabilisées, à partir desquelles construire notre questionnaire. Le développement des moyens techniques pour la diffusion d'œuvres s'est accompagné d'une plus grande plasticité des formats dans lesquels ces œuvres sont conçues et perçues, avec une tendance manifeste à l'hybridation. Les espaces commerciaux dans le monde physique peuvent en effet inciter les producteurs et diffuseurs à donner une labellisation claire à leurs produits, afin de permettre de les retrouver efficacement : on peut penser ici aux rayonnages des librairies, organisés par genres littéraires, ou aux disquaires, séparant consciencieusement le rock, à l'écart des musiques du monde ou du rap et du hip hop. L'avènement du numérique a là aussi radicalement transformé les logiques de classement et d'identification : les rayonnages numériques n'ont nul besoin de s'organiser comme une partition stricte des productions, bien au contraire ! Un interprète aura plus de chance d'être écouté s'il peut émarger à plusieurs styles musicaux et donc être identifié par plusieurs tags et finalement toucher plusieurs types de publics.

Les secteurs culturels dont la diffusion s'est organisée le plus rapidement au sein des espaces numériques, comme le jeu vidéo (Benghozi et Chantepie, 2017), ont été les plus précocement touchés par cet affaiblissement, voire par la disparition de catégories de genres clairement identifiés et partagés au sein de la population[5]. Les autres secteurs (livre notamment) prennent très clairement cette direction.

Les chercheurs en sciences sociales sont des victimes collatérales de cette situation : ils voient fondre des instruments qui leur ont pour autant été précieux pour qualifier les contenus culturels. Si le travail social d'énonciation des genres et styles culturels s'émousse, au profit de l'usage de tags, paramétrés au plus près des préférences individuelles et donc sans valeur de généralité, les enquêtes par questionnaires courent le risque de devoir renoncer à recourir à des catégories génériques pour interroger et décrire les pratiques des répondants. Et les

5. Il suffit de penser à ces nouvelles catégorisations apparues sur les plateformes de streaming musical : « musiques pour le sport », « pour se sentir bien le soir », etc. Prouvant leur efficacité en termes d'usage, ces nouveaux genres sont désormais plébiscités par les plateformes, qui, semble-t-il, donnent un moindre crédit aux styles musicaux eux-mêmes.

chercheurs ne pourront plus s'intéresser aux genres et aux échelles de légitimité qui les structurent.

La rénovation du questionnaire 2018 a malgré tout fait l'hypothèse qu'un bon nombre de ces nomenclatures de genres (genres de films, musicaux, de séries, d'émissions de radio) continue de faire foi et sens dans la population. Cette sixième édition poursuit donc dans beaucoup de domaines les catégories qui ont pu être mobilisées dans les éditions précédentes, *modulo* quelques réaménagements marginaux. Elle a fait également le choix de prévoir des questions ouvertes à certains endroits stratégiques du questionnaire (notamment le genre de musique préféré) afin de tester la validité de cette hypothèse. Il sera en effet possible d'examiner si ces réponses spontanées gardent une forme de cohérence avec les questions structurées par les genres.

8. Une actualisation très attendue des constats avec l'édition 2018

In fine et conformément à l'un des objectifs assignés au dispositif, le questionnaire 2018 conserve de nombreuses caractéristiques des éditions précédentes : l'approche choisie continue de privilégier l'observation des pratiques et comportements, et se démarque d'interrogations orientées autour des représentations ou opinions. Cette continuité permet de comparer les pratiques dans le temps, avec notamment le maintien d'un noyau dur de questions tout au long de la période, à partir desquels un pseudo panel (c'est-à-dire la reconstitution d'analyses générationnelles par le cumul des enquêtes) a été constitué à partir de 1973. L'édition 2018, en reproduisant les formulations passées, permettra en particulier de prolonger les séries portant sur le visionnage de la télévision, la radio, les livres lus, l'écoute de musique enregistrée, les sorties au cinéma, au concert, au théâtre, à la danse, au musée (ou exposition), à la visite de monuments, aux jeux vidéos (depuis 2008).

Aux côtés de ces invariants, le questionnaire 2018 présente des évolutions notables, permettant de mieux prendre en compte les enjeux contemporains de la culture. Pour commencer, un plus grand degré de précision a été apporté dans la description des pratiques en sous-domaines. Ainsi, le questionnaire n'interroge plus seulement la sortie au théâtre, mais permet ensuite de distinguer « théâtre classique », « théâtre contemporain », « pièce de boulevard, vaudeville » et « one man show, café-théâtre, spectacle d'improvisation » ; la danse se décline en « classique », « traditionnelle ou folklorique », « modern jazz », « contemporaine » ou « autre » ; les catégories de jeux-vidéos joués sont détaillées, ainsi que les mémoriaux visités. Cette recherche de précision supplémentaire s'explique par les possibilités ouvertes par l'augmentation de la taille d'échantillon, qui permet de décliner différentes variantes au sein de pratiques qui étaient auparavant appréhendées par plus gros blocs. Ces déclinaisons permettront de raffiner l'analyse de l'appropriation de ces pratiques, en distinguant plusieurs niveaux de légitimité culturelle.

Deuxième innovation, un module de questions a été décliné tout au long du questionnaire afin de décrire les contextes dans lesquels se vivent les pratiques observées : autour des enjeux de sociabilité (identification des personnes avec lesquelles l'activité est pratiquée en général), de temporalité (repérage des moments dans lesquels l'activité est le plus souvent pratiquée, si ces moments existent : congés, week-ends, semaines), de mobilité (possibilité d'une pratique s'expérimentant plus souvent en déplacement qu'à domicile).

Des questions supplémentaires ont été introduites pour détailler les usages numériques liés aux pratiques culturelles observées (usage du streaming, des liseuses, des ressources numériques et des réseaux sociaux). La dimension linguistique a été également développée, en interrogeant plus systématiquement qu'auparavant la consommation de produits culturels en langues étrangères.

Enfin, un plus grand raffinement a été apporté dans les modules permettant d'observer les propriétés socio-démographiques de la personne enquêtée et de son ménage. La codification des professions (aussi bien de l'enquêté, que de son ou sa conjoint.e et de ses parents) a été retravaillée avec la division Emploi de l'Insee, afin d'en améliorer la qualité. La codification du diplôme bénéficie également des apports du pôle Diplômes et spécialités de formation de l'Insee.

Ces changements divers ont conduit à une légère réorganisation de la structure du questionnaire, par rapport à 2008, avec notamment le regroupement en début de l'ensemble des questions portant sur les pratiques en amateurs (qui ont pu également être développées, en intégrant une dimension rétrospective sur le contexte du développement de ces pratiques).
Cette édition 2018, fidèle à la lignée des éditions précédentes, a donc relevé le double défi de la continuité et de l'adaptation aux enjeux contemporains. Faut-il faire le pari qu'elle sera la dernière de la série, bientôt remplacée par d'autres méthodologies, potentiellement plus adaptées à ce que sont devenues nos pratiques ? Ce sont les exploitations scientifiques de ce matériau qui diront sa valeur empirique (largement prouvée par le passé) et les évolutions nécessaires à lui apporter à l'avenir.

Remerciements

L'auteur tient à remercier Hervé Glévarec, Jean-Michel Guy, Amandine Louguet et Sylvie Octobre pour leur relecture attentive de cet article.

Références

Benghozi P.-J. et Ph. Chantepie (2017), *Jeux vidéo : l'industrie culturelle du XXIe siècle ?*, Paris, DEPS, Ministère de la Culture, Les Presses de Sciences Po, coll. « Questions de culture ».

Bourdieu P. (1979), *La Distinction. Critique sociale du jugement*, Paris, Éditions de Minuit, coll. « Le sens commun ».

Bourdieu P. et A. Darbel (1966), *L'amour de l'art*, Paris, Éditions de Minuit, coll. « Le sens commun ».

Cardon D. (2015), *À quoi rêvent les algorithmes. Nos vies à l'heure des big data*, Paris, Le Seuil.

Coulangeon Ph. (2011), *Les métamorphoses de la distinction. Inégalités culturelles dans la France d'aujourd'hui*, Paris, Grasset, coll. « Mondes vécus ».

Donnat O. (éd.) (2003), *Regards croisés sur les pratiques culturelles*, Paris, Ministère de la culture - DEPS, « Questions de culture ».

Donnat O. (2009), *Les pratiques culturelles des Français à l'ère numérique : enquête 2008*, Paris, Découverte, Ministère de la culture et de la communication.

Donnat O. (2011), « Pratiques culturelles, 1973-2008 : dynamiques générationnelles et pesanteurs sociales », *Culture Études*, n° 7, https://doi.org/10.3917/cule.117.0001.

Dumazedier J. (1962), *Vers une civilisation du loisir ?*, Paris, Éditions du Seuil.

Girard A. (1994), « Les enquêtes sur les pratiques culturelles », *in* J.-P. Rioux et J.-Fr. Sirinelli (éds.), *Pour une histoire culturelle*, Paris, Le Seuil.

Glévarec H. (2016), « Le discours de l'échec de la démocratisation culturelle en France », *Revue européenne des sciences sociales*, vol. 54, n° 2, pp. 147-93.

Martigny V. (2016), *Dire la France : culture(s) et identités nationales (1981-1995)*, Paris, Les Presses de Sciences Po, coll. « Références académiques ».

Martin L. (2012), « Du SER au DEP, ou la constitution d'une socio-économie de la culture et d'une prospective culturelle au service de l'action (1959-1993) », *Revue historique*, vol. 663, n° 3, pp. 683-704.

de Saint-Martin M. (2013), « Les tentatives de construction de l'espace social, d'« Anatomie du goût » à *La Distinction*. Quelques repères pour l'histoire d'une recherche », *in* Ph. Coulangeon (éd.), *Trente ans après La Distinction, de Pierre Bourdieu*, Paris, La Découverte, coll. « Recherches », p. 29-44.

Wolff L. et Ph. Lombardo (2020), « Cinquante ans de pratiques culturelles en France », *Culture Études*, vol. 2020, n° 2, DEPS Ministère de la Culture.

Implications juridiques et éthiques des algorithmes d'intelligence artificielle dans le domaine de la santé

Philippe BESSE[1]

Université de Toulouse – INSA, Institut de Mathématiques de Toulouse – UMR CNRS 5219 – Chercheur régulier à l'ObvIA[2]

Aurèle BESSE-PATIN

McGill University & Montreal Neurological Institute

Céline CASTETS-RENARD

Université d'Ottawa – Titulaire de la chaire Law, Accountability and Social trust in AI, ANITI[3] – Chercheure régulière à l'ObvIA

TITLE

Legal and Ethical Implications of Artificial Intelligence Algorithms in the Health Field

RÉSUMÉ

L'Intelligence Artificielle (IA) envahit nos quotidiens et le domaine de la santé notamment pour aider au diagnostic, faire des choix thérapeutiques ou encore viser une médecine prédictive de précision. Absente de la loi française de bioéthique du 7 juillet 2011, l'IA fut très présente lors des États Généraux accompagnant la révision de la loi en 2018. La profusion de guides ou recommandations éthiques sur l'IA (*soft law*), motivés par la nécessité de conquérir la confiance des usagers, incite préalablement à se préoccuper de leur vigueur normative, en lien avec les textes juridiques promulgués depuis l'entrée en vigueur le 25 mai 2018 du RGPD (règlement 2016/679/UE – règlement général de protection des données personnelles). Une analyse conjointe de ces textes, des algorithmes d'IA déployés et d'applications concrètes en santé permet de poser les principales questions éthiques et légales soulevées dans ce domaine : principe du *consentement libre et éclairé* du patient face à l'opacité des algorithmes, risques potentiels de *discrimination* dans l'accès au soin, *intérêt public* ou *bien commun* attendu de la recherche en comparaison des *risques* encourus par l'ouverture de l'accès aux données personnelles. Les réponses conduisent à des recommandations déontologiques ou règlementaires indispensables à la transparence de ces outils : *protection* drastique des données de santé, notamment génétiques, et de leurs utilisations, rigueur des pratiques de recherche pour produire des *résultats reproductibles* donc scientifiques, *détection des biais* avant certification des dispositifs de santé et explicitation du *protocole d'information* des patients.

Mots-clés **:** *intelligence artificielle, apprentissage automatique, statistique, RGPD, code de santé publique, éthique, bioéthique, discrimination, droit de l'IA.*

ABSTRACT

Artificial Intelligence (AI) is invading our daily lives and the health field, notably to help with diagnosis, to make therapeutic choices or even to aim for precise predictive medicine. Absent from the French bioethics law of July 7, 2011, AI was very present during the "États Généraux" accompanying the revision of the law in 2018. The profusion of ethical guides or recommendations on AI (soft law), motivated by the need to win the trust of users,

1. philippe.besse@math.univ-toulouse.fr

2. ObvIA : Observatoire international sur les impacts sociétaux de l'IA et du numérique, https://observatoire-ia.ulaval.ca/
3. ANITI : *Artificial and Natural Intelligence Toulouse Institute*, https://aniti.univ-toulouse.fr/

encourages us to be concerned about their normative force, in connection with the legal texts promulgated since the entry into action on 25 May 2018 of the GDPR (regulation 2016/679/EU – general regulation on the protection of personal data). A joint analysis of these texts, of the AI algorithms deployed and of concrete applications in health, enables us to consider the main ethical and legal questions raised in this field: the principle of *free and informed consent* of the patient faced to the opacity of algorithms, potential risks of *discrimination* in access to care, *public interest* or *common good* expected from research in comparison with *risks* incurred by opening access to personal data. The responses lead to ethical or regulatory recommendations that are essential for the transparency of these tools: drastic *protection* of health data, particularly genetic data, and their uses, rigorous research practices to produce *reproducible* and therefore scientific *results*, *detection of biases* before certification of health devices and clarification of the patient *information protocol*.

Keywords**:** *artificial intelligence, machine learning, statistics, GDPR, public health code, ethics, bioethics, discrimination, AI law.*

1. Introduction

1.1 Battage médiatique

L'intelligence artificielle (IA) dite *faible*, opposée à une IA *forte* supposée disposer d'une conscience de soi et que nous laisserons à la science-fiction, recouvre une grande variété d'objets, méthodes *et al*gorithmes susceptibles d'imiter des comportements humains « intelligents » : robots, véhicules autonomes, systèmes experts, algorithmes d'apprentissage automatique…

Depuis 2012, nous sommes soumis à une déferlante médiatique sans précédent sur les applications des algorithmes d'IA associées à des succès retentissants : reconnaissance d'images et diagnostic automatique, véhicules autonomes, victoire au go, traduction automatique… Ce battage médiatique fait suite à celui sur l'avènement du stockage tous azimuts de données massives ou *big data* et leur utilisation pour alimenter les nouveaux algorithmes d'IA exécutés dans des environnements technologiques en constante progression. Cette convergence entre données massives, algorithmes performants et puissance de calcul est à l'origine de l'expansion exceptionnelle des usages de l'IA dans tous les domaines de nos quotidiens. Les principaux acteurs technologiques comme *Google*, *Facebook*, *Amazon* ou *Microsoft*, ont tout intérêt à sur-médiatiser ces succès puisque leurs considérables revenus proviennent de la vente de l'application de ces technologies à notre profilage publicitaire. Ils se doivent donc d'en promouvoir l'efficacité, même si ses succès diffèrent en fonction du domaine d'application et si elle peut s'avérer anxiogène dans ses conséquences sociétales, tant sur la destruction d'emplois même qualifiés, que sur la déresponsabilisation des acteurs humains ou encore l'exposition des données de la vie privée.

1.2 Confiance et acceptabilité

Une composante importante de la publicité excessive autour de l'IA concerne son *acceptabilité*, comme celle de toute nouvelle technologie pénétrant ou plutôt envahissant nos quotidiens. Le principal enjeu est de cultiver ou conquérir la confiance des utilisateurs, qu'ils soient consommateurs, clients, patients, contribuables, justiciables ou citoyens, pour une IA acceptable. En première ligne, les entreprises privées spécialistes des réseaux sociaux et technologies numériques, rejointes ensuite par plus de 90 partenaires, se sont empressées, dès 2015, de signer une Charte de partenariat[4] pour une *IA au bénéfice du peuple et de la société*. Dès lors, tous les acteurs publics institutionnels ont rejoint le mouvement ; citons parmi les plus récents la partie 5 du rapport Villani pour *donner un sens à l'IA* (Villani *et al.*, 2018), les lignes directrices pour une *IA digne de confiance* des hauts experts désignés par la Commission Européenne (High Level Expert Group, 2019), ou encore la *déclaration de Montréal pour le développement d'une IA responsable* (Université de Montréal, 2018). Notons la création (12/2019) du Comité Pilote d'Éthique du Numérique[5] sous l'égide du Comité Consultatif National d'Éthique pour les Sciences de la Vie et de la Santé et dont le rapport attendu début 2021 doit s'intéresser aux liens entre diagnostic médical et IA. C'est plus largement une avalanche de recommandations pour une IA éthique au service de l'humanité dont Fjeld *et al.* (2019) proposent une analyse graphique et sémantique tandis que Jobin *et al.* (2019) en explore le paysage. Les enjeux sont considérables car, en l'absence de confiance, les utilisateurs n'accepteront pas l'IA. Sans acceptation sociale, les entreprises technologiques ne pourront plus collecter toutes les données nécessaires et ne pourront pas développer une IA pertinente, source de profits. Les conséquences de l'affaire *Cambridge Analytica* sur l'encours boursier de *Facebook*, en mars 2018, en furent une démonstration éclatante (Guichard, 2018).

4. https://www.partnershiponai.org/tenets/
5. https://www.ccne-ethique.fr/sites/default/files/communique_lancement_comite_numerique.pdf

1.3 Éthique et protection juridique

Cette affaire peut être citée parmi d'autres : condamnations successives de *Google* pour entrave à la concurrence, fuites massives et répétées de données personnelles, utilisations abusives de celles-ci… nous rappellent que le but premier des entreprises commerciales ou de leurs dirigeants n'est pas l'altruisme ou la philanthropie mais des encours boursiers et le montant des dividendes à distribuer à leurs actionnaires. Ces profits nécessitent des pratiques éthiques pour être acceptables mais la confiance des usagers sera nettement plus franche et massive si elle repose sur une protection juridique, plutôt que sur de bonnes intentions éthiques (*ethical washing*), aussi louables soient-elles. En France, la première version de la loi Informatique et Liberté date de 1978. Ce texte précurseur marqua une réelle anticipation des problèmes à venir. En revanche, à l'heure actuelle, la loi peine à suivre les évolutions ou disruptions technologiques. Ce sont bien entre autres quelques-uns de ces retards que vise à combler une révision de la loi de Bioéthique.

L'entrée en vigueur du RGPD (Commission Européenne, 2018), puis son intégration dans les textes nationaux des États membres, signent une avancée majeure pour la protection des données personnelles en Europe. Le principe de sécurité et confidentialité, au cœur de l'action de la Commission Nationale de l'Informatique et des Libertés (CNIL) en France, est en effet une priorité mais d'autres aspects, tant juridiques qu'éthiques, sont à considérer pour instaurer ou restaurer la confiance des usagers envers ces nouvelles technologies. Ainsi, l'article 22§1 du RGPD (Commission Européenne, 2018) accorde aux personnes concernées le droit de ne pas faire l'objet d'une décision fondée exclusivement sur un traitement automatisé, produisant des effets juridiques la concernant ou l'affectant de manière significative. Repris dans les lois nationales des États membres, cet article a pu servir de fondement en droit français pour reconnaître un droit à l'*explicabilité* des décisions algorithmiques, dans le souci de lutter contre les risques de *discrimination*. Ces préoccupations rejoignent les exigences publiques exprimées dans un sondage réalisé au Royaume-Uni (Vayena *et al.*, 2018) au sujet des applications de l'IA en médecine.

Un large consensus est donc établi sur la nécessité de pratiques en IA respectueuses de l'éthique. Néanmoins, compte tenu des pressions financières, un cadre juridique s'avère indispensable. Il est un préalable à des pratiques vertueuses génératrices de confiance.

Tel est bien l'objectif de la Commission Européenne (CE) qui propose *les éléments clefs d'un futur cadre réglementaire* dans le livre blanc (Commission Européenne, 2020) *pour une IA basée sur l'excellence et la confiance fondée sur les droits fondamentaux de la dignité humaine et la protection de la vie privée*. La rédaction de ce livre blanc s'appuie sur les lignes directrices pour une IA de confiance (High Level Expert Group, 2019) rédigées par un groupe d'experts et dont il est important d'anticiper l'impact à venir. En résumé, les technologies de l'IA se développent à grande vitesse dans un contexte juridique très complexe mais insuffisant à encadrer les risques sociaux susceptibles de se produire. Ce cadre légal est appelé à évoluer, au moins en Europe, afin de minimiser les risques et créer les conditions d'une acceptabilité sociale de l'IA. Les limites de la norme légale étant identifiées, il est alors possible, dans un deuxième temps, de compléter la norme légale par des chartes de déontologie, telle celle des professionnels de la statistique publique européenne (Eurostat, 2017). Rappelons que l'objectif de ces chartes est tout autant de constituer une obligation réglementaire envers les employés qu'une protection de leur rigueur professionnelle contre les pressions extérieures politiques ou financières.

1.4 IA et bioéthique

Le champ d'étude ainsi esquissé est trop vaste. Nous proposons d'en limiter le périmètre en considérant le **domaine d'application restreint mais très sensible de la santé**, donc de la

bioéthique. Les questions de sécurité et confidentialité des données personnelles sont déjà abondamment traitées par des systèmes d'analyse (*Privacy Impact Assessment*[6] ou études d'impact des données personnelles) mis en place par la CNIL et rendus obligatoires par le RGPD en présence de risques qui peuvent notamment résulter du traitement de données sensibles, telles les données de santé (art. 35). Nous nous focaliserons sur *les questions touchant au risque de discrimination* et à la nécessité d'une explication intelligible des décisions, en lien avec les réglementations et textes juridiques concernés. D'autres auteurs (Racine *et al.*, 2019 ; Wiens *et al.*, 2019) ont récemment abordé ce sujet mais en privilégiant le point de vue médical, ainsi que ses interactions avec les nouvelles technologies et certaines questions éthiques émergentes. Aborder la question à partir du cadre juridique ouvre une autre perspective. Des questions essentielles sont alors soulevées, tenant en particulier à la notion de *consentement libre et éclairé* des personnes et aux *risques de discriminations*. D'autres questions émergent, qui nécessitent des réponses plus délicates à formuler, sur l'équilibre entre, d'une part, le développement de la recherche en santé et *l'intérêt public* ou *bien commun* attendu et, d'autre part, sur les risques afférents à l'accessibilité des données personnelles de santé.

Aborder la bioéthique des applications d'intelligence artificielle dans le domaine de la santé mobilise les compétences de nombreuses disciplines. L'objectif est pour le moins ambitieux. Aussi, pour en faciliter la lecture, cet article introduit de façon pédagogique les prérequis de chaque discipline, dans une volonté de faciliter les échanges réciproques. La section 2 définit plus précisément l'IA considérée dans cet article. La section 3 décrit brièvement les règles légales applicables à l'IA en santé. Les principaux domaines de santé concernés par ces modèles *et al*gorithmes sont cernés en section 4. La section 5 met en relation les champs disciplinaires du droit, de la santé, des sciences du numérique et de la statistique pour en tirer les conséquences : quelles sont les protections juridiques existantes ? Où sont leurs limites ou insuffisances qui nécessiteraient plus de réglementation ou un code déontologique des professionnels concernés ?

Les questions de bioéthique sont très culturellement marquées ainsi donc que les lois qui sont très spécifiques à un pays, notamment en France où elles s'avèrent plus restrictives par exemple sur l'accès aux tests génétiques par rapport à d'autres pays européens. Cet article, axé sur le corpus juridique européen et plus particulièrement français, ne peut évidemment prétendre à l'exhaustivité. En revanche, il peut constituer une première étape pour aborder de façon comparative les différentes situations.

Cette réflexion conduit, en conclusion, à proposer un ensemble de recommandations résumées dans le Tableau 1 permettant de préciser les responsabilités et devoirs de chaque partie : développeur, médecin, chercheur, patient, dans les actes médicaux aidés par l'IA ou pour la poursuite de recherches en santé sur des grandes bases de données personnelles.

6. https://www.cnil.fr/fr/outil-pia-telechargez-et-installez-le-logiciel-de-la-cnil

Tableau 1 – *Proposition de recommandations à trois niveaux : 1. accès aux données ; 2. déontologie de la recherche ; 3. réglementation des dispositifs de santé intégrant de l'IA*

1. L'accès aux données nationales de santé peut être ouvert (INDS, CEIP) sans consentement explicite des personnes concernées lorsque les résultats attendus le justifient : cohortes épidémiologiques, diagnostic par imagerie médicale ou protéomique, étude des maladies rares ou monogéniques… En revanche, au regard des risques encourus de ré-identification, l'accès aux données de santé publique n'est pas justifié pour des projets de recherche pangénomiques sur les maladies multifactorielles.

2. Compte tenu des enjeux et des risques encourus, les équipes de recherche doivent se soumettre à un audit externe effectif et pas seulement déclaratif de sécurité de la chaîne d'archivage et de traitements des données personnelles de santé même pseudonymisées. Elles doivent s'astreindre à une rigueur d'analyse (détection, correction des biais) et d'évaluation des erreurs, afin de publier des résultats reproductibles, première exigence vers une certification. Elles doivent donner accès aux séquences de traitement des algorithmes lors d'une soumission avant publication.

3. Les autorités de santé (*e.g.* HAS, FDA…) ont la responsabilité de la certification ou du remboursement des dispositifs de santé intégrant de l'IA. Il importe d'*harmoniser* leurs protocoles en anticipant la stratégie en cours d'élaboration de la Commission Européenne. Obligation au responsable d'un système d'IA de produire une *documentation exhaustive* décrivant comment sont : (i) validées qualité, robustesse, résilience, des décisions ; (ii) traqués les biais des données ; (iii) intégré un suivi qualité adaptatif par enrichissement des bases d'apprentissage ; (iv) identifiés les responsables à chaque étape des traitements (recueil des données, entraînement des algorithmes, validation, certification, exploitation) pour la mise en place d'une boucle vertueuse de rétroaction. Ces autorités doivent formaliser les protocoles d'explicitation auprès des patients du rôle des algorithmes dans leur prise en charge, des risques d'erreur dans l'aide à la décision et des risques encourus de leur ré-identification.

2. De quelle IA est-il question ?

2.1 Historique

L'IA est apparue sous cette appellation dès 1955 à la suite du développement des premiers ordinateurs et a commencé à être formalisée par les travaux pionniers d'Alan Turing. La notion de neurone formel est due à McCulloch (neurophysiologiste) et Pitts (logicien) en 1943 tandis que le premier réseau de neurones est proposé par Rosenblatt (1958) avec un *perceptron* censé simuler le fonctionnement de la rétine. Faute de méthodes et capacités de calcul suffisantes, cette approche de l'IA a été mise en veilleuse au profit des *systèmes experts* dans les années 70. Ces systèmes associent une base de règles logiques, explicitées par des experts humains du domaine d'application, une base de faits et un moteur d'inférence. Ce dernier met itérativement en relation faits et prémices des règles pour en déduire de nouveaux faits jusqu'à atteindre le ou les faits correspondant à la décision recherchée ou l'objectif visé. Un tel prototype de système expert (Mycin) a été développé par Buchanan et Shortliffe (1984) pour la sélection d'un antibiotique adapté aux paramètres biologiques du diagnostic d'une infection bactérienne.

Malgré les très grandes difficultés de construction des bases de règles expertes, leur manque de flexibilité, ainsi que la complexité algorithmique exponentielle de leur exécution, cette approche n'a pas été complètement abandonnée (Darlington, 2011). Néanmoins la recherche sur les systèmes experts passa en arrière-plan à la fin des années soixante-dix au profit d'un retour des réseaux de neurones bénéficiant de moyens de calculs suffisants et de résultats théoriques (Rumelhart *et al.*, 1986) sur la convergence (locale) de l'algorithme de *rétropropagation du gradient* permettant d'entraîner itérativement un réseau multicouches. Dans un réseau, la connaissance est dite répartie, dans les valeurs des poids des entrées des neurones appris sur les données, par opposition à la connaissance localisée des bases de règles construites par les experts ; une IA *empirique* opaque s'oppose à une IA *symbolique* explicable. Les années quatre-vingts ont connu un développement massif de différents types de réseaux de neurones *et al*gorithmes d'apprentissage parallèlement à l'extension des méthodes et modèles statistiques appliqués à des objets complexes de grande dimension (courbes et fonctions).

Dans les années quatre-vingt-dix, ces réseaux se trouvèrent en concurrence avec bien d'autres algorithmes : modèles statistiques avec pénalisation, *k* plus proches voisins, arbres binaires de décision, machine à vecteurs supports, *boosting*, forêts aléatoires… (James *et al.*, 2013), poursuivant les mêmes objectifs prédictifs au sein d'une très large communauté scientifique réunie autour de l'apprentissage automatique (*machine learning* ou *ML*) à l'interface entre Sciences du Numérique, Mathématiques et Statistique. Les recherches sur les réseaux de neurones ont toujours progressé jusqu'à leur succès retentissant en 2012 sous l'appellation très médiatisée d'apprentissage profond (*deep learning*). Ces réseaux associent des dizaines de couches de neurones dont celles dites *convolutionnelles* ou d'autres *récurrentes* (*LSTM*) qui firent franchirent des étapes décisives, par exemple en reconnaissance d'images ou en traduction automatique. Ces avancées ont valu à leurs promoteurs, Yoshua Bengio, Georges Hinton et Yan Le Cun, l'attribution du prix Turing en 2019.

2.2 Algorithmes d'apprentissage

Compte tenu des besoins dans le domaine de la santé, **cet article est focalisé sur cette catégorie d'algorithmes dits d'apprentissage automatique (*machine learning, ML*)** qui représente les utilisations très majoritaires de l'IA du quotidien. Schématiquement, le ML se divise principalement en quatre classes d'algorithmes répondant à quatre objectifs :

- apprentissage ou classification *non supervisé*, lorsqu'aucun objectif à atteindre n'est a *priori* connu : reconnaissance de classes ou mesure quantitative. Il peut être question de débruiter ou déflouter une image, de rechercher des groupes homogènes (taxinomie, segmentation ou *clustering*) dans une population décrite par un ensemble de variables ou caractéristiques, comme par exemple segmenter en marketing des comportements pour la gestion de la relation client, définir des classes homogènes de patients au regard de leurs analyses biologiques ;

- apprentissage par *renforcement*, lorsque l'algorithme, disposant de règles de base, apprend en optimisant une fonction sous forme d'objectif ou récompense par des successions d'essais / erreurs au cours de la réception d'un flux de données ou d'expérimentations séquentielles. Tel est par exemple le cas de l'algorithme AlphaZero (Silver *et al.,* 2017) pour jouer au go ou aux échecs, ou encore celui de bandits manchots pour les systèmes de recommandation des sites de vente en ligne ;

- détection d'anomalie ou classification à une classe ou découverte de nouveautés ;

- apprentissage *supervisé* ou *statistique (statistical learning)* (James *et al.*, 2013), lorsqu'il est question de modéliser, expliquer et principalement prévoir la valeur d'une quantité ou celle d'une classe.

2.3 Apprentissage statistique ou IA empirique

C'est principalement ce dernier type d'apprentissage qui envahit notre quotidien, lorsqu'il s'agit d'attribution ou risque d'un crédit, d'analyse automatique de textes (CV ou *tweets*), d'évaluation du risque de récidive d'un accusé ou détenu, de la gestion des patrouilles de police en prévoyant les zones les plus probables de délits, d'aides au diagnostic médical… Les applications en sont innombrables, corrélatives à une production académique considérable.

De façon générale, un modèle est estimé ou un algorithme entraîné pour rendre visibles des relations entre une variable Y cible (le risque, le diagnostic…) et un ensemble de variables ou caractéristiques (*features*) dites explicatives X^j $(j=1,\ldots,p)$: caractéristiques socio-économiques, biologiques… Toutes ces variables (Y,X^j) sont mesurées, observées, sur un ensemble $i=1,\ldots,n$ d'individus ou *instances* appelé échantillon d'*apprentissage* ou d'entraînement. Une fois un modèle estimé ou un algorithme entraîné sur ces données, la connaissance d'un vecteur x_0, contenant les observations des variables X^j pour un nouvel individu, permet d'en déduire une prévision de la valeur ou de la classe y_0 le concernant. Le modèle ou l'algorithme calcule automatiquement cette valeur y_0 en combinant, en fonction de l'algorithme utilisé, celles y_i observées sur les individus présents dans la base d'apprentissage et proches de x_0, en un certain sens, au regard des valeurs x_i^j. Autrement dit, la prévision d'une nouvelle situation et donc la décision qui en découle, est construite automatiquement à partir des situations lui ressemblant le plus dans la base d'apprentissage et dont les décisions sont déjà connues. Le principe repose sur la stationnarité des données : la loi apprise sur l'échantillon d'apprentissage est la même que celle des données que l'on veut tester. En conséquence, l'apprentissage statistique n'*invente rien*, il reproduit un modèle connu et le généralise aux nouvelles données, *au mieux* selon un critère spécifique d'ordre statistique à optimiser. Plus on possède de données, meilleure sera la connaissance fournie par ce modèle. Ceci souligne le rôle fondamental joué par les données et donc le succès des grands acteurs d'internet et des réseaux sociaux qui bénéficient d'une situation de monopole sur des masses considérables de données comportementales des internautes pour les traduire en profilage et donc en recettes publicitaires. Transposé au domaine de la santé, où l'objectif est une prise en compte toujours plus fine de la complexité du vivant, le premier enjeu est l'accès à de grandes masses de données personnelles excessivement sensibles, objet de toutes les convoitises.

2.4 Statistique inférentielle versus apprentissage statistique

Deux objectifs doivent être clairement distingués dans les applications, tant de la statistique que de l'IA en santé pour lever une ambiguïté trop répandue.

Le premier objectif est celui *explicatif* de la statistique inférentielle, poursuivi par la mise en œuvre de tests, afin de montrer l'*influence d'un facteur* en contrôlant le *risque d'erreur*, soit le risque de rejeter à tort une hypothèse dite H_0 et donc de considérer que le facteur à un impact, alors qu'il n'en a pas. C'est le cas typique des essais cliniques de phase III, durant lesquels une molécule est prescrite en double aveugle à un groupe témoin, tandis que le groupe contrôle reçoit un *placebo*. Pour beaucoup de disciplines académiques, le test statistique constitue un outil de preuve scientifique même si son usage, parfois abusif, est mis en cause voire controversé à cause du manque de reproductibilité de trop nombreuses publications scientifiques (Ioannidis, 2016).

Le deuxième objectif est *prédictif*, en utilisant des modèles statistiques classiques ou les

algorithmes d'apprentissage automatique plus récents et sophistiqués. Deux sous-objectifs sont à considérer ; le premier est une prévision avec explication des résultats, de la façon dont les variables X^j influent sur la cible ou variable réponse Y. Le deuxième est une prévision brute sans recherche ou possibilité d'explication. Mais dans les deux cas, le *data scientist* sélectionne le modèle ou algorithme minimisant une estimation ou mesure d'une *erreur de prévision* qui contrôle le *risque d'erreur* de la décision qui en découle. *In fine* l'erreur de prévision de l'algorithme sélectionné est estimée sur un *échantillon test indépendant*, différent de l'*échantillon d'apprentissage* sur lequel il a été entraîné ; c'est aussi à la base de toute procédure de certification précédant sa mise en exploitation.

Il y a donc, selon les objectifs, deux types de risque ou d'erreur. Celui de se tromper en affirmant qu'un facteur est influent et celui de se tromper de décision à cause d'une erreur de prévision. Laissons la question, largement débattue par ailleurs (Ioannidis, 2016), de la pertinence des tests statistiques pour nous focaliser sur celle de la qualité de prévision plus spécifique à l'IA.

Il existe de très nombreux critères ou métriques pour évaluer une erreur de prévision. Ce peut être un simple *taux d'erreur* pour la prévision d'une variable binaire : tissus pathologiques ou sains, une *erreur quadratique moyenne* pour une variable Y quantitative. Dans beaucoup de publications du domaine de la santé, il est fait référence à l'aire sous la courbe ROC (*Area Under the Curve, AUC*) pour évaluer la qualité d'un algorithme pour une prévision binaire. Ce critère issu du traitement du signal nécessite quelques explications rappelées en annexe1.

2.5 Facteurs de qualité d'une prévision

Plus précisément, quels sont les composants d'un modèle statistique ou algorithme d'apprentissage qui sont déterminants pour la qualité de prévision et donc pour les risques d'erreur de la décision qui en découle ?

Le point fondamental pour la qualité ou robustesse, voire la certification d'un algorithme d'apprentissage statistique, est, en tout premier lieu, la *qualité des données* disponibles, ainsi que leur *représentativité* du domaine d'étude ou d'application concerné. Les données d'entraînement de l'algorithme sont-elles bien représentatives de l'ensemble des situations ou cas de figure susceptibles d'être, par la suite, rencontrés lors de l'exploitation de l'algorithme ? Il s'agit d'anticiper une capacité de généralisation de son usage. En effet, si des groupes ou des situations sont absents ou simplement sous-représentés, c'est-à-dire si les données sont, d'une façon ou d'une autre, *biaisées*, le modèle ou l'algorithme qui en découle ne fait que reproduire les biais ou s'avère incapable de produire des prévisions correctes de situations qu'il n'a pas suffisamment apprises lors de son entraînement. Ce problème est très bien référencé dans la littérature et souligné dans les rapports et guides éthiques. C'est même un vieux problème déjà formalisé en statistique pour la constitution d'un échantillon relativement à une population de référence en planification d'expérience ou en théorie des sondages. Ce n'est pas parce que les données sont volumineuses, déjà acquises, qu'il faut pour autant tout prendre en compte ou ne pas se préoccuper d'en acquérir d'autres. Considérons l'exemple typique de la prévision d'événements rares mais catastrophiques. Un algorithme naïf, pour ne pas dire trivial, conduit à un très faible taux d'erreur s'il ne prévoit aucune occurrence de l'événement rare, mais est inutile voire dangereux. L'expérience du *data scientist* le conduit alors à sur-représenter (sur-échantillonnage) les événements rares, ou sous-échantillonner ceux très fréquents ou encore à introduire des pondérations dans le choix de la fonction objectif à optimiser. Ces pondérations dépendent de l'asymétrie des coûts, à évaluer par des *experts métier*, d'un faux positif ou prévision à tort d'un événement exceptionnel, relativement au coût induit par un faux négatif qui n'anticipe pas la catastrophe.

La précédente question concerne la représentativité des individus ou situations présentes

dans la base d'entraînement relativement à une *population théorique de référence*. La deuxième soulève celle du choix ou de la disponibilité des caractéristiques ou variables observées sur ces individus. Elle peut se formuler de la façon suivante : les *causes effectives* de la cible ou variable Y à modéliser, ou les variables qui lui seraient très corrélées, sont-elles bien prises en compte dans les observations ? Dans le même ordre d'idée et avec les mêmes conséquences, des mesures peuvent être erronées, soumises à du bruit. Ces questions ne sont pas plus faciles à résoudre que celles de représentativité précédentes, car il n'est pas possible de pallier une absence d'information ou rectifier des erreurs de mesures ou de labellisations, mais il est plus simple d'en circonscrire les conséquences en estimant précisément les erreurs d'ajustement du modèle ou d'entraînement de l'algorithme puis celles de prévision ; elles resteront plus ou moins importantes mais évaluables, quel que soit le nombre de variables prises en compte ou le volume des données accumulées.

Plus précisément, la taille de l'échantillon ou le nombre d'instances de la base d'apprentissage intervient à deux niveaux sur la qualité de prévision. La taille nécessaire dépend, d'une part, de la complexité de l'algorithme, du nombre de paramètres ou de poids qui en définit la structure et, d'autre part, de la variance du bruit résiduel ou erreur de mesure. Un algorithme est entraîné, en moyenne, et la taille de l'échantillon doit être d'autant plus grande que la variance de l'erreur de mesure est importante. Les réseaux de neurones profonds appliqués à des images de plusieurs millions de pixels sont composés de dizaines de couches pouvant comporter des millions de paramètres ou poids à estimer ; ils nécessitent des bases de données considérables.

Attention, lorsque n est très grand (*big data*), le modèle peut être bien estimé car c'est une *estimation en moyenne* dont la précision s'améliore proportionnellement avec la racine de n. En revanche, une prévision individuelle est toujours impactée par le bruit résiduel du modèle, sa variance, quelle que soit la taille de l'échantillon. Aussi, même avec de très grands échantillons, la prudence est de mise quant à la précision de la prévision d'un *comportement individuel* surtout s'il est mal ou peu représenté dans la base : acte d'achat, acte violent, défaut de paiement, occurrence d'une pathologie.

En résumé, les applications quotidiennes de l'IA sont l'exploitation d'algorithmes d'apprentissage statistique, particulièrement sensibles à la qualité des données d'entraînement. Leur quantité est importante mais ne suffit pas à garantir la précision de prévisions individuelles qui doit être évaluée avec soin, afin de garantir, certifier, l'usage d'un algorithme. Malgré les abus de communication, l'IA ne se résume pas à l'utilisation de l'apprentissage profond. Le succès très médiatisé de certaines de ses applications ne doit pas laisser croire que ces relativement bons résultats en reconnaissance d'images ou traduction automatique sont transposables à tout type de problème.

Enfin, à l'exception des modèles statistiques élémentaires car linéaires ou à celle des arbres binaires de décision, les algorithmes d'apprentissage statistique sont opaques à une interprétation fine et directe de l'influence des caractéristiques d'entrée ou variables explicatives sur la prévision de la variable cible Y. Ce point soulève des problèmes délicats lorsqu'il s'agit de fournir l'*explication intelligible* d'une décision automatique. Des pistes de solutions ou d'aide à des solutions existent (Barredo Arrieta *et al.*, 2020) mais nous verrons ci-dessous que les applications de l'IA en santé appréhendent cette question de façon spécifique.

3. Cadre juridique de l'IA en Santé

Schématiquement, trois questions sont à prendre en considération pour préciser les frontières de l'action juridique :

- **Comment rendre compte de décisions** et en préciser les responsabilités, lorsqu'elles sont issues d'algorithmes souvent caractérisés par leur opacité ?

- **Quels sont les risques de discrimination** envers des personnes protégées ou groupes sensibles ?
- **Comment évaluer l'équilibre bénéfice / risque** entre l'intérêt public, d'une part, et le risque pour la vie privée des personnes touchées par l'utilisation de leurs données personnelles, d'autre part ?

3.1 Redevabilité et information versus opacité

L'article L.1111-4 du code de la santé publique précise que : « *Aucun acte médical ni aucun traitement ne peut être pratiqué sans le consentement libre et éclairé de la personne et ce consentement peut être retiré à tout moment* ». Le projet de loi bioéthique de 2019 intègre un article 11 spécifique sur l'utilisation de l'IA dans un cadre médical. Une fois voté, il devrait être intégré au chapitre I^{er} du titre préliminaire du livre préliminaire de la quatrième partie du code de la santé publique, et complété par un article L. 4001-3 ainsi rédigé :

« *I. – Lorsque pour des actes à visée préventive, diagnostique ou thérapeutique est utilisé un traitement algorithmique de données massives, le professionnel de santé qui communique les résultats de ces actes informe la personne de cette utilisation et des modalités d'action de ce traitement.*

II. – L'adaptation des paramètres d'un traitement mentionné au I pour des actions à visée préventive, diagnostique ou thérapeutique concernant une personne est réalisée avec l'intervention d'un professionnel de santé et peut être modifiée par celui-ci.

III. – La traçabilité des actions d'un traitement mentionné au I et des données ayant été utilisées par celui-ci est assurée et les informations qui en résultent sont accessibles aux professionnels de santé concernés. »

Seraient ainsi consacrés un droit à l'information sur l'utilisation d'un dispositif d'IA et un droit à une intervention humaine, celle du professionnel de santé, pour garantir le respect du droit à l'information sur les actes médicaux réalisés (à visée préventive, diagnostique ou thérapeutique) qui fonde le consentement libre et éclairé du patient.

En outre, le rôle du médecin n'est pas seulement d'informer le patient sur le recours à l'IA mais ce dernier doit aussi avoir la capacité d'intervenir sur l'utilisation du traitement algorithmique en modifiant les paramètres. Pour que le médecin puisse prendre des décisions en connaissance de cause, la traçabilité des actions est prévue. Cette interaction homme-machine pose naturellement la question de la responsabilité du médecin. Mais dès lors que le médecin reste au centre de la relation de confiance avec le patient, qu'il continue d'assumer une obligation d'information et qu'il reste maître des choix et décisions prises, la machine doit être considérée comme une simple aide à la décision qui ne remplace pas le médecin et ne modifie en rien les règles de responsabilité. En l'état actuel, les règles de responsabilité médicale applicables au médecin ne sont donc pas modifiées par le recours à un traitement algorithmique. En principe, le médecin assume une obligation de soin qui est une obligation de moyen, et non de résultat. Il n'engage sa responsabilité qu'en cas d'erreur fautive ayant entraîné un dommage.

Si l'obligation d'information concernant le recours à un dispositif d'IA ne cause pas de problème particulier, il sera sans doute plus difficile de garantir que le médecin pourra informer le patient des « modalités d'action de ce traitement » algorithmique. Encore faudra-t-il que le médecin le comprenne lui-même, ce qui pourra s'avérer difficile voire impossible dans certaines situations. Les conditions de déploiement des algorithmes doivent donc tenir compte de ces exigences et les entreprises privées qui proposeront leur système d'IA devront par conséquent expliquer voire former les médecins à la bonne utilisation de ces outils pour que ces derniers puissent à leur tour informer les patients, au moins sommairement, de la façon dont ils fonctionnent.

Au-delà de l'obligation d'information, il paraît de toute façon pertinent que les médecins puissent maîtriser un minimum ces outils pour que le médecin ait lui-même confiance et qu'ils deviennent de véritables aides à la décision.

Au demeurant, il paraît nécessaire d'interdire l'utilisation des méthodes algorithmiques opaques en matière de santé. Sur le modèle de ce que prévoit la loi s'agissant des décisions administratives automatiques, il pourrait être imposé que le responsable de traitement doive s'assurer de la maîtrise du traitement algorithmique et de ses évolutions, afin de pouvoir expliquer, en détail et sous une forme intelligible à la personne concernée, la manière dont le traitement a été mis en œuvre à son égard (art. L.311-3-1 du code des relations entre le public et l'administration). Ne peuvent être alors utilisés des algorithmes susceptibles de réviser eux-mêmes les règles qu'ils appliquent, sans le contrôle et la validation du responsable du traitement (voir l'interprétation du Conseil constitutionnel dans sa décision n° 2018-765 DC du 12 juin 2018 (pt 71)). De telles dispositions ne supprimeraient pas mais réduiraient sensiblement les risques d'opacité.

3.2 Risque de discrimination

Les discriminations sont pénalement sanctionnées à l'article 225-1 al. 1er du code pénal qui définit la discrimination directe comme étant « *toute distinction opérée entre les personnes physiques sur le fondement de leur origine, de leur sexe, de leur situation de famille, de leur grossesse, de leur apparence physique, de la particulière vulnérabilité résultant de leur situation économique, apparente ou connue de son auteur, de leur patronyme, de leur lieu de résidence, de leur état de santé, de leur perte d'autonomie, de leur handicap, de leurs caractéristiques génétiques, de leurs mœurs, de leur orientation sexuelle, de leur identité de genre, de leur âge, de leurs opinions politiques, de leurs activités syndicales, de leur capacité à s'exprimer dans une langue autre que le français, de leur appartenance ou de leur non-appartenance, vraie ou supposée, à une ethnie, une Nation, une prétendue race ou une religion déterminée* ». La loi énumère ainsi très largement les critères exhaustifs à prendre en compte pour rechercher si une discrimination directe a été commise. Une telle discrimination est intentionnelle et sera probablement plus facile à prouver.

L'alinéa 2 vise la discrimination indirecte qui est « *une disposition, un critère ou une pratique neutre en apparence, mais susceptible d'entraîner, pour l'un des motifs mentionnés au premier alinéa, un désavantage particulier pour des personnes par rapport à d'autres personnes, à moins que cette disposition, ce critère ou cette pratique ne soit objectivement justifié par un but légitime et que les moyens pour réaliser ce but ne soient nécessaires et appropriés* ». La discrimination indirecte est plus difficile à prouver car elle est non intentionnelle. Les traitements algorithmiques sont susceptibles d'être ainsi qualifiés car les discriminations peuvent être systémiques, par exemple en raison des données d'apprentissage biaisées utilisées pour entraîner le système d'IA, et avoir des répercussions sur des individus ou groupes d'individus. La preuve risque d'être difficile à rapporter, aussi est-il particulièrement fondamental en matière de santé d'être exigeant sur les conditions de constitution et utilisation de ces outils.

La notion de discrimination individuelle est reprise à l'article L.1110-3 du code de santé publique, selon lequel « *aucune personne ne peut faire l'objet de discriminations dans l'accès à la prévention ou aux soins* » (al. 1er). En outre, « *un professionnel de santé ne peut refuser de soigner une personne pour l'un des motifs visés au premier alinéa de l'article 225-1 du code pénal ou à l'article 225-1-1 du code pénal ou au motif qu'elle est bénéficiaire de la protection complémentaire en matière de santé prévue à l'article L.861-1 du code de la sécurité sociale, ou du droit à l'aide prévue à l'article L.251-1 du code de l'action sociale et des familles ». On peut alors s'interroger sur la place que prendront les dispositifs d'IA et l'impérieuse nécessité qu'il y aura à ce que les données utilisées ne soient pas biaisées, au risque de constituer une discrimination « dans l'accès à la prévention ou aux soins* ». Surtout, on peut se demander comment prouver la discrimination.

3.3 Bénéfice d'intérêt public versus Risque individuel

Alors que les données personnelles sont protégées dans l'Union européenne par le RGPD et en droit français par la loi informatique et libertés (LIL3), l'article L.1461-3 du code de santé publique organise un régime national d'accès aux données à caractère personnel du système national des données de santé (SNDS) pour permettre des traitements suivant une finalité mentionnée au III de l'article L.1461-1 et répondant à un **motif d'intérêt public**. La mise à disposition des données peut se faire pour contribuer : (i) à l'information sur la santé ainsi que sur l'offre de soins, la prise en charge médico-sociale et leur qualité ; (ii) à la définition, à la mise en œuvre et à l'évaluation des politiques de santé et de protection sociale ; (iii) à la connaissance des dépenses de santé, des dépenses d'assurance maladie et des dépenses médico-sociales ; (iv) à l'information des professionnels, des structures et des établissements de santé ou médico-sociaux sur leur activité ; (v) à la surveillance, à la veille et à la sécurité sanitaires ; (vi) à la recherche, aux études, à l'évaluation et à l'innovation dans les domaines de la santé et de la prise en charge médico-sociale.

Le décret n° 2016-1871 du 26 décembre 2016 relatif au traitement de données à caractère personnel dénommé « système national des données de santé » fixe les règles de gouvernance et désigne les organismes autorisés à accéder de manière permanente aux données du SNDS, en fonction des missions de service public qu'ils remplissent. Tel est, entre autres, le cas de la Direction générale de la santé, des Agences régionales de santé, de l'Agence nationale de santé publique, de l'Agence nationale de sécurité du médicament et des produits de santé, l'Institut national du cancer, de l'INSERM, des équipes de recherche des CHU et des centres de lutte contre le cancer (CSP, art. R. 1461-12). Le décret définit l'étendue de cette autorisation par différents critères, tels que la profondeur historique, l'aire géographique, les caractéristiques d'une population, ainsi que la possibilité ou non d'utiliser dans un même traitement des identifiants potentiels qui permettraient d'accroître le risque de ré-identification.

Le décret organise aussi un accès aux données du SNDS soumis à autorisation de la CNIL à des fins de recherche, étude ou évaluation dans le domaine de la santé par les organismes non listés dans le décret (notamment organismes privés) et les organismes habilités à accéder de façon permanente au SNDS qui dépasseraient les limites fixées par le décret. Un autre décret n° 2016-1872 du 26 décembre 2016 précise les modalités de fonctionnement de l'Institut National des Données de Santé (INDS) et du Comité d'Expertise pour les Recherches, les Études et les Évaluations dans le domaine de la Santé (CEREES). Ce comité reprend une partie des missions du CCTIRS (Comité Consultatif sur le Traitement de l'Information en matière de Recherche dans le domaine de la Santé) et se prononce sur la mise en œuvre de tout traitement de données à caractère personnel ayant pour finalité la recherche, l'étude ou l'évaluation dans le domaine de la santé et n'impliquant pas la personne humaine. L'INDS est en lien direct avec le CEREES, afin de fournir un avis à la CNIL sur la cohérence entre la finalité de l'étude proposée, la méthodologie présentée et le périmètre des données auxquelles il est demandé accès.

En résumé, la loi met en place un Système National des Données de Santé, devenu ensuite le *Health Data Hub*, qui peut, sous réserve d'assurer la confidentialité des données et donc des personnes concernées, donner l'accès aux données pour différents objectifs dont celui de recherches scientifiques présentant un intérêt public substantiel. En plus de devoir préciser les conditions de sécurité et confidentialité des données pour éviter les risques de ré-identification, la question qui en découle directement est de savoir ce qu'est un « **intérêt public substantiel** » de la recherche en santé. Une balance des intérêts doit ici se mettre en place entre la protection des données et l'intérêt de la recherche. Le standard juridique de l'« intérêt public substantiel » est une notion floue peu claire mais qui permet une souplesse de mise en œuvre par une interprétation au cas par cas des bénéfices et des risques.

Un raisonnement comparable est intégré au sein même de la Loi Informatique et Liberté n° 2018-493 du 20 juin 2018 s'agissant du traitement des données à caractère personnel dans le domaine de la santé. L'article 66.I. prévoit que de tels traitements « *ne peuvent être mis en œuvre qu'en considération de la finalité d'intérêt public qu'ils présentent. La garantie de normes élevées de qualité et de sécurité des soins de santé et des médicaments ou des dispositifs médicaux constitue une finalité d'intérêt public* ».

3.4 Quelle évolution du cadre réglementaire ?

Le livre blanc (Commission Européenne, 2020) annonce un changement de paradigme sur le fondement des lignes directrices (High Level Expert Group, 2019) qui s'achèvent par une liste d'évaluation (pilote) *ex ante* qui constituera le dossier obligatoire et indispensable à une expertise ou audit d'un système d'IA. Cette liste de questions couvre les 7 points éthiques fondamentaux identifiés : action humaine et contrôle humain, robustesse technique et sécurité, respect de la vie privée et gouvernance des données, transparence, non-discrimination et équité, bien-être sociétal et environnemental, utilité, responsabilité. Ce n'est pas le lieu de discuter la pertinence des 10 pages de questions auxquelles le responsable d'un système d'IA devra répondre mais bien celui de souligner le renversement de la charge de preuve. Alors qu'il serait très difficile pour ne pas dire impossible à un usager d'apporter la preuve qu'il est victime par exemple d'une discrimination algorithmique, ce sera au responsable du traitement de montrer qu'il a pris les mesures nécessaires afin d'éviter des biais sources de discrimination (obligation de moyens).

La Commission européenne n'a pas encore proposé un cadre légal, mais il est important de noter qu'en matière de santé, des organismes de certification en anticipent le principe. Outre aux États-Unis où la FDA (Health, 2019) a posé un cadre pour l'autorisation de commercialisation de systèmes d'IA d'aide au diagnostic, des réflexions sont aussi menées en France au travers du guide de la CNEDiMTS (commission nationale d'évaluation des dispositifs médicaux et des technologies de santé) (Haute Autorité de Santé, 2020) pour le dépôt d'un dossier de remboursement des DSC (dispositifs de santé connectés) embarquant de l'IA.

4. Domaines de santé concernés par L'IA

Le projet de loi bioéthique de 2011 ne fait pas mention d'IA mais rend nécessaire, pour sa révision périodique, la réunion d'États Généraux de la bioéthique qui ont produit un rapport (France, 2018). Ce rapport aborde neuf points dont six ont des implications sociétales fondamentales : la procréation assistée, la recherche sur l'embryon, les dons d'organes, la fin de vie, les neurosciences, l'environnement ; trois autres concernent indirectement ou directement les applications de l'IA : les bases de données de santé, la médecine génomique, la robotisation de la médecine. Cette section a pour but de préciser les quelques domaines de santé pour lesquels il semble le plus pertinent de s'intéresser aux impacts du déploiement de l'IA.

4.1 Bases de données

L'accès aux données est un préalable indispensable. La mise en place du Système National des Données de Santé (SNDS) (art. L.1461-3 du code de santé publique) est le résultat de la volonté politique d'ouvrir un *hub des données de santé* respectant par construction l'anonymat des patients. Il est composé du regroupement de la base SNIIRAM de l'assurance maladie, de celles des hôpitaux (PMSI), de la base INSERM des causes de décès, des données relatives au handicap et de celles détenues par les caisses d'assurance maladie complémentaire. L'accès à ces données est contrôlé par l'Institut National des Données de Santé (INDS) après avis de la CNIL.

Par ailleurs, d'autres sites régionaux se mettent en place pour regrouper les données hospitalières comme celui de la *clinique des données de santé* pilotée par le CHU de Nantes pour le grand ouest et qui fait appel à une société privée (*Wedata*[7]) pour la phase d'anonymisation. De plus, le Plan Investissement d'Avenir (PIA) (plan Médecine France Génomique 2025[8]) prévoit la mise en place de plateformes de séquençage à haut débit. Deux ont été sélectionnées à la suite de l'appel d'offre : SeqOIA (Paris) et AURAGEN (Lyon). Celles-ci ont pour mission de séquencer des dizaines de milliers de génomes chaque année.

Toutes ces bases et bien d'autres s'intègrent au projet national de *Health Data Hub*[9] (HDH) qui met en place une *pseudonymisation* des données : noms et adresses des patients sont supprimés et le code national d'inscription au registre des personnes physiques (NIRPP) est crypté par une fonction de hachage non réversible. Ce code devient une clef d'appariement des données de chaque patient pour la fusion des différentes bases mais ne permet pas de revenir au NIRPP initial.

Arrêt de la Cour de Justice de l'Union Européenne, projet de décret gouvernemental, avis de la CNIL, du Conseil d'État, opposition du Conseil de la Caisse nationale d'Assurance Maladie, le choix d'une société de droit américain (*Microsoft Azure*) pour l'hébergement du HDH soulève des problèmes même avec une localisation géographique française des données. Il est inutile de tenter d'intervenir dans ce débat à notre niveau mais il semble important de souligner que son ampleur obère d'autres questions qu'il serait dommageable de laisser dans l'ombre.

4.2 Médecine génomique

L'un des principaux battages médiatiques en santé concerne les médecines dites *translationnelles* et *4p* pour médecine *prédictive* d'un risque pathologique, *préventive* de ce risque, *participative* incluant le patient à la prévention et *personnalisée* ou de *précision* avec un traitement thérapeutique spécifique au patient. Cette précision ou personnalisation peut faire appel aux caractéristiques génétiques du patient et donc à la médecine dite *génomique*. La médecine *translationnelle* a pour objectif d'accélérer les applications de la recherche, donc des médecines précédentes, pour raccourcir le cycle de mise sur le marché d'un médicament. Elle nécessite de faciliter les échanges pluridisciplinaires ainsi qu'évidemment l'accès aux données médicales personnelles.

Schématiquement, deux types de bases de données génomiques sont constitués. Certaines, les plus récentes, enregistrent des séquences complètes de chaque génome ; 3,4 milliards de paires de base soit au minimum 3,5 GO par génome. Un génome complet comprend 1,5 % de parties codantes dans 26517 gènes protéiques. Les deuxièmes bases, de mises en place plus anciennes (Klein, 2005), se limitent à enregistrer pour chaque individu les présences/absences de variants génétiques ou mutations spécifiques appelées *single nucleotide polymorphism* (SNP). Jusqu'à 165 millions de SNP sont pris en compte pour chacun des milliers, ou millions d'individus de la base, auxquels sont associés un ensemble de phénotypes, c'est-à-dire la présence ou non de pathologies, des constantes biologiques. Ces bases permettent des études dites *pangénomiques* (*genomic wide association studies, GWAS*) en cherchant à mettre en relation variants génétiques ou mutations avec l'occurrence d'une pathologie.

Deux objectifs sont principalement poursuivis avec l'analyse de ces données. Le premier vise l'identification d'un élément potentiellement causal dans la survenue d'une maladie rare ou monogénique. Une mutation, éventuellement sur un gène, est associée à une fonction biologique défaillante et donc une pathologie. La mucoviscidose est un exemple type d'une telle

7. https://octopize-md.com/
8. https://solidarites-sante.gouv.fr/systeme-de-sante-et-medico-social/recherche-et-innovation/france-genomique
9. https://www.health-data-hub.fr/

maladie parmi plus de 8000 répertoriées dont beaucoup ne touchent que quelques familles dans le monde. Point important, la détection de la mutation responsable est obtenue par un test statistique qui détecte le facteur influent mais la prévision de la maladie concernée est validée par l'interprétation biologique ; elle n'est pas le résultat d'un algorithme d'apprentissage statistique. Un exemple spectaculaire d'une démarche de médecine personnalisée génomique translationnelle est fourni par le cas clinique (Kim *et al.*, 2019) d'une petite fille atteinte d'une maladie génétique dégénérative rare (Batten) et même exceptionnellement rare, unique dans le monde pour cette fillette, car la conséquence de deux mutations génétiques. Traitée à l'Hôpital de Boston, il a fallu un an pour déterminer et lui appliquer une thérapie génique qui n'est pas susceptible de la guérir mais au moins de réduire l'impact de la maladie dont le nombre de crises d'épilepsie par jour. Le coût global de cette démarche thérapeutique est resté confidentiel.

Un deuxième objectif vise à déterminer des facteurs génétiques de maladies multigéniques ou multifactorielles et souvent chroniques affectant une grande partie de la population. Cette démarche est basée sur des seuls éléments statistiques (tests) et pas sur l'analyse biologique des fonctions mises en cause car beaucoup trop de variants génétiques sont détectés. Elle occulte complètement les influences d'autres facteurs, environnementaux, épigénétiques, qui peuvent être largement prépondérants pour certaines pathologies. Ces insuffisances soulèvent de nombreuses critiques.

4.3 IA et robotisation de la médecine

Les États Généraux de la bioéthique font état des robots de microchirurgie, mais il n'en sera pas question ici. Nous allons nous focaliser sur d'autres types d'automatisation :

- aide au diagnostic par
- magerie médicale, électroencéphalogrammes (EEG), électrocardiogramme (ECG) et reconnaissance de formes par apprentissage profond ou deep learning,
- identification de biomarqueurs préventifs par études « omiques » ;
- aide aux choix thérapeutiques : e.g. IBM Watson ;
- surveillance des effets secondaires de médicaments à partir de la base SNIIRAM (Morel *et al*., 2019) ;
- suivi épidémiologique de grandes cohortes, telles que Constances10 (Zins *et al*., 2010).

Topol (2019) propose une revue assez exhaustive et enthousiaste des applications de l'IA en médecine mais nous nous limiterons aux quelques exemples illustrant les questions émergentes, juridiques ou éthiques. Ainsi, l'analyse automatique d'ECG relève des mêmes techniques d'apprentissage que l'analyse des images obtenues en radiologie ; seule cette dernière au développement viral est évoquée. La recherche de biomarqueurs transcriptomiques, protéomiques… d'une pathologie rejoint, d'un point de vue méthodologique, le débat sur la médecine génomique ; il n'est pas nécessaire de compléter. Suite à ce qui peut être considéré comme un coûteux échec (Ross and Swetlitz, 2018), *IBM* ne communique plus sur les applications de l'algorithme *Watson* en santé. C*et al*gorithme apprenait à partir de la littérature scientifique mais pas à partir de données personnelles sensibles. Son usage est principalement commercialisé dans le tertiaire, banque, assurance, c'est pourquoi nous le laisserons également de côté.

10. https://www.constances.fr/actualites/2019/js2019.php

5. Questions juridiques / éthiques de l'IA en santé

Comme évoqué précédemment, trois questions essentielles seront ici illustrées :

- biais et discrimination dans l'accès au soin ;
- consentement éclairé face à des algorithmes opaques ;
- balance bénéfice / risque entre intérêt de santé publique et ouverture des données.

5.1 Risques de discrimination des algorithmes d'apprentissage

La littérature académique propose (Žliobaitė, 2017) une très grande variété de critères ou définitions de la notion de biais et donc de discrimination des algorithmes d'apprentissage. Mais, comme le font remarquer Friedler *et al.* (2019), beaucoup sont très corrélés voire redondants et même pour certains incompatibles (Chouldechova, 2017). Nous nous limiterons à trois niveaux possibles, donc trois indicateurs de types de biais les plus régulièrement évoqués dans la littérature. Ils sont faciles à estimer par des intervalles de confiance afin d'en intégrer la précision (Besse *et al.*, 2018) si les données sont disponibles et fournissent un premier tableau synthétique suffisamment exhaustif des risques encourus de discrimination.
Le premier, nommé *demographic equality* dans la littérature, concerne la reproductibilité et aussi le risque d'amplification ou d'exacerbation de certains biais présents dans les données d'apprentissage. Comme dans beaucoup d'autres domaines, comme l'emploi, le crédit, le logement, les données thérapeutiques sont empreintes de biais de société. Lee *et al.* (2019) mettent ainsi en évidence dans une étude portant sur 85 millions de patients, que la gestion de la douleur par des antalgiques dépend de l'origine ethnique des patients. En toute logique, des algorithmes entraînés à partir de telles données reproduisent les biais voire les renforcent en se comportant donc de façon discriminatoire. Obermayer et Mullainathan (2019) dissèquent ainsi les biais ethniques produits par un algorithme qui guide les choix thérapeutiques de 70 millions de patients aux USA. Lorsque le système prévoit qu'un patient aura des besoins de soins futurs de santé particulièrement complexes et intensifs, il est inscrit à un programme qui fournit des ressources supplémentaires et une plus grande attention de la part de prestataires qualifiés ainsi qu'une aide à la coordination de ses soins. Les auteurs mettent en évidence un biais raciste en montrant comment les patients d'origine caucasienne, ayant le même état de santé que les patients d'origine afro-américaines, sont beaucoup plus susceptibles d'être inscrits dans le programme de gestion des soins et de bénéficier de ses ressources. Il s'agit là d'un *cas de prévision auto-réalisatrice* comme ceux dénoncés par O'Neil (2016). En synthétisant différents cas de sources de discrimination en santé, Vyas *et al.* (2020) ouvrent le difficile débat sur la pertinence ou non d'inclure l'origine ethnique dans les algorithmes cliniques.

Le deuxième indicateur est nommé *overall error equality*. C'est souvent la conséquence d'un autre type de biais initial lié à la mauvaise représentativité des données. Si un sous-groupe est sous-représenté, la prévision le concernant sera de moins bonne qualité. Ce biais est bien connu pour les applications de reconnaissance faciale (Buolamwini and Gebru, 2018). Particulièrement présent dans les données pangénomiques, ce biais fait que nous ne sommes pas tous égaux devant une médecine de précision qui personnaliserait les traitements à partir de considérations génétiques. En effet, la population d'ascendance blanche européenne (Popejoy and Fullerton, 2016) est présente à 96 % dans les bases génomiques en 2009 et encore à 81 % en 2016. Cette récente évolution est très majoritairement due au développement massif de campagnes de séquençage en Chine et donc sur des populations d'origine ethnique très spécifique. D'autres sources de problèmes sont aussi relevées dans ces données conduisant à d'autres risques de biais. Alors que beaucoup de pathologies (Pulit *et al.*, 2017) dépendent largement du sexe, cet aspect est négligé : les possibles mutations du chromosome X sont très rares (Chang *et al.*, 2014) dans ces bases et le chromosome Y en est absent. Enfin, la grande abondance de personnes relativement âgées et de leurs pathologies afférentes, ainsi que l'absence de prise en compte des facteurs environnementaux, biaisent l'étude des risques pathologiques des

patients jeunes.

Les algorithmes d'IA en santé ne semblent pas ou pas encore concernés, à première vue, par un troisième niveau de discrimination : *equality of odds*. Cet indicateur est à la base d'une vive controverse aux USA sur le caractère discriminatoire du logiciel *Compas* de prévision du risque de récidive. La société diffusant ce logiciel affirme qu'il ne discrimine pas au regard des deux précédents indicateurs tout en minimisant le rôle d'un taux d'erreur élevé de l'ordre de 30 à 40%. Larson et Angwin (2016), du site d'investigation *Propublica*[11], ont suivi une cohorte de près de 7000 détenus libérés dont ils connaissaient le score *Compas* de récidive à leur libération ainsi que l'occurrence ou non d'une récidive dans les deux ans la suivant. L'analyse de ces données porte, entre autres, sur les matrices de confusion (annexe 1) croisant le score de récidive : haut *vs.* bas avec la présence *vs.* absence de récidive. Ces matrices révèlent des asymétries inversées selon l'origine ethnique des personnes : celles d'origines afro-américaines présentent des taux de faux positifs (absence de récidive malgré un score élevé) plus importants que les personnes d'origine caucasienne. Cause d'un retard de libération et donc d'une plus forte désocialisation ; c'est encore un risque de prévision auto-réalisatrice.

Le diagnostic de ces problèmes en santé consiste à définir, détecter, les sources de biais dont les types sont maintenant bien identifiés. Leur résolution, ou au moins leur prise en compte, intervient à deux niveaux ; celui académique de déontologie scientifique et donc éthique, et celui réglementaire des organismes de certification pour une commercialisation et une exploitation publique à grande échelle.

Au niveau amont de la recherche académique, les études épidémiologiques de santé publique sont basées sur des cohortes dont la constitution est opérée avec une rigueur essentielle. Citons le cas de la cohorte *Constances* (Zins *et al.*, 2010) dont la mise en place sélective de volontaires sur une longue période a permis de réunir un échantillon de 200 000 personnes représentatif de la population nationale. Il en est de même pour les études basées sur un sous-ensemble de la base SNIIRAM (Schwarzinger *et al.*, 2018). La partie la plus délicate du travail n'est pas la modélisation bien balisée mais l'extraction des données pour constituer un échantillon représentatif conduisant à des résultats et des conclusions valides pour la population.

Détecter puis corriger les différentes sources de biais d'une base de données est de la responsabilité déontologique des chercheurs, essentielle à leur éthique. Cela concerne en premier chef également les relecteurs des revues scientifiques. En aval, c'est le rôle des organismes de certification. Le guide de la HAS (Haute Autorité de Santé, 2020) pour la rédaction du dossier de demande de remboursement inclut un questionnaire qui poursuit les mêmes objectifs que ceux de la liste d'évaluation du guide des experts de la CE. Le responsable du traitement est tenu de décrire toutes les dispositions prises pour s'assurer de la fiabilité, la robustesse, l'équité, la redevabilité du système d'IA concerné ; protocole analogue à celui mis en place aux USA par la *FDA* pour autoriser la commercialisation des AI/ML-SaMD (*Artificial Intelligence and Machine Learning Software as a Medical Device*) (Health, 2019).

5.2 Consentement éclairé versus Opacité des algorithmes

Le rapport Villani (Villani *et al.*, 2018) affirme que « *l'ouverture des boîtes noires de l'IA est un enjeu démocratique* » mais sans laisser entrevoir un embryon de piste pour une démarche qui peut prendre des formes multiples selon l'objectif visé et le contexte ou domaine d'application. Comme signalé en section 2.4, deux champs d'application sont déjà à considérer en fonction du type d'approche, explicative ou prédictive, mise en place.

11. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Dans de très nombreuses applications en santé, l'objectif est explicatif : trouver le gène dont la mutation est responsable d'une maladie rare ; identifier des biomarqueurs pour un diagnostic anticipé ou, en épidémiologie, les facteurs de risque d'une pathologie déterminée. Dans un cas comme de l'autre, l'objectif *explicatif* montrant l'influence d'un ou de quelques rares facteurs est prioritaire. Ce sont donc des modèles statistiques classiques qui sont utilisés pour ce type d'application : tests d'hypothèses pour les données génomiques (Lindström *et al.*, 2017), modèle de régression logistique ou modèle de Cox de durée de vie en épidémiologie. Il ne s'agit pas d'algorithmes d'apprentissage donc pas réellement d'intelligence artificielle mais comme le battage médiatique ne fait pas la différence et surtout que *l'enjeu juridique de l'accès aux données personnelles* est le même, ces modèles sont assimilés. Néanmoins, comme ils sont des modèles linéaires, pas des boîtes noires, ils permettent des interprétations détaillées pour atteindre l'objectif. La consultation de la valeur d'un coefficient, voire même simplement de son signe, suffit à la compréhension du modèle et à orienter l'explication biologique de l'influence d'un facteur.

La question de l'*explicabilité* est nettement plus critique lors du déploiement d'un apprentissage profond et même dès l'utilisation d'un réseau de neurones ou d'un algorithme d'apprentissage statistique sophistiqué. Il s'agit alors d'un modèle sous-jacent *non linéaire* d'un réseau potentiellement très complexe d'interactions entre les variables. L'effet d'une variable explicative sur la variable cible Y ne peut plus être explicité de façon détaillée. Ne nous trompons pas, certes l'algorithme est complexe mais pas plus que la réalité sous-jacente à l'analyse d'une image ou de celle du vivant incluant des effets simples, des interactions, des boucles de contre-réaction… Prenons l'exemple de l'imagerie médicale pour laquelle de très nombreuses expérimentations d'apprentissage profond ont été déployées : plus de 35000 références ont été identifiées par Liu *et al.* (2019a). Schématiquement, deux types d'explication sont à prendre en compte selon qu'elle s'adresse au chercheur ou ingénieur qui met en place l'algorithme ou à leurs usagers : le patient sur qui une image a été acquise en vue d'un diagnostic et le médecin qui le soigne. Le chercheur a besoin de comprendre finement les modes opératoires de l'algorithme, afin d'en détecter les failles et y remédier : dans quelles circonstances et à la suite de quel défaut la prévision est-elle erronée ? Quelles sont les situations insuffisamment présentes dans la base d'apprentissage pour être correctement identifiées ? Le patient, comme son médecin, ne s'intéressent pas à ce niveau de détail mais évidemment à leurs conséquences sur la qualité de la prévision. Annoncer un diagnostic avec un taux d'erreur de 1 % ou de 30 % change tout pour le patient comme pour l'équipe médicale qui doit en déduire une stratégie thérapeutique et en expliquer les conséquences. Comme l'exprime également London (2019), l'explication au patient doit être focalisée sur le risque d'erreur de cette aide au diagnostic, de la même façon que chirurgien et anesthésiste doivent expliquer à leur patient les risques afférents à une opération afin que le patient puisse exprimer son choix de façon *libre* et suffisamment *éclairée*.

Cette communication du risque d'erreur nécessite une estimation précise et sans biais mais cette tâche est, en principe, inhérente à l'entraînement d'un algorithme ; elle ne peut être évacuée et fait logiquement partie du processus de certification imposé par la FDA (Health, 2019) ou la demande de remboursement de la HAS (Haute Autorité de Santé, 2020). En revanche, au plan académique, où seule l'éthique ou la déontologie scientifique est opérante, de trop nombreuses défaillances révèlent de sérieux manques de rigueur sous la pression de publication. Liu *et al.* (2019a) témoignent que très peu de publications respectent un protocole rigoureux basé sur des estimations des erreurs utilisant des échantillons indépendants de celui d'apprentissage ; c'est une insuffisance méthodologique, donc déontologique, sans conséquence thérapeutique directe mais financière car, mêmes publiées, des études ne s'avéreront pas reproductibles lors de nouvelles analyses tout aussi coûteuses. Cela conduit Liu *et al.* (2019b) à proposer un guide sur les précautions à prendre en lisant un article d'application de l'apprentissage automatique en imagerie médicale.

5.3 Intérêt public versus risques sur la confidentialité

Le dernier point est le plus complexe à analyser. Il nécessite d'évaluer les bénéfices attendus de la recherche scientifique en santé au regard des risques encourus par l'ouverture de l'accès à des données personnelles. La réglementation européenne prévoit la protection de la confidentialité de ces données mais la garantie totale peut, soit être difficile à assurer, soit conduire à une dégradation des données et donc de la qualité des résultats escomptés.

Le *premier volet* concerne les risques de ré-identification en fonction du procédé d'anonymisation mis en œuvre. L'article L.1461-4 du code de la santé publique dispose que les données ne doivent pas contenir le nom, l'identifiant NIRPP et l'adresse des personnes. Dans le cas de pseudonymisation du HDH, les NIRPP sont cryptés pour servir de clef d'appariement. Ces précautions sont largement insuffisantes pour anonymiser des données. Une ré-identification partielle, c'est-à-dire celle d'un sous-ensemble des personnes de la base de données, peut être obtenue à partir des seules informations précisant la date de naissance, le code postal et le sexe de ces personnes. Si en plus le nombre d'enfants est connu, l'unicité du profil et donc l'identification devient très probable. De nombreux auteurs analysent ces risques (Rubinstein and Hartzog, 2016) ou en font la démonstration (Rocher *et al.*, 2019 ; Narayanan and Shmatikov, 2008). Supprimer quelques informations est donc insuffisant, il est nécessaire d'apporter suffisamment de flou dans les données les plus personnelles, par exemple en discrétisant l'âge en tranches, la localisation en grandes zones, afin de contrôler le risque d'unicité dans la population et donc de ré-identification. D'autres stratégies comme la *confidentialité différentielle* (Dwork and Roth, 2013) consistent à simuler une part de données synthétiques en respectant les principales propriétés. Dans tous les cas, il s'agit de chercher un meilleur compromis entre risque de ré-identification et dégradation des analyses statistiques. Le CHU de Nantes propose de remplacer les données réelles par des données synthétiques simulées à partir de k plus proches voisins. Les simulations sont supposées suffisamment réalistes pour que les principales propriétés statistiques – distributions et corrélations des variables – et donc les principales qualités des analyses futures soient conservées, tout en rendant en principe impossible le retour aux données initiales et la ré-identification. La procédure semble intéressante mais, mise en œuvre par une entreprise privée, le descriptif détaillé, protégé par le secret commercial, n'est pas accessible. Plutôt que de proposer du code libre d'accès évaluable par un audit indépendant, il est regrettable d'ajouter une couche d'opacité limitant la confiance envers les données générées.

Par ailleurs, il est à noter que le floutage des données ou la construction de données synthétiques n'est pas applicable aux bases de données pangénomiques. Altérer les présences / absences de mutations rendraient ces données inutilisables car la proximité sur l'arbre phylogénétique (cousinage) permet au FBI de résoudre des affaires classées[12] mais n'induit pas des proximités au sens des pathologies concernées comme permettent de l'inférer des proximités au sens de mesures biologiques quantitatives. De plus, la connaissance d'une liste de SNP relativement restreinte d'une personne peut jouer le rôle d'une *empreinte génétique* (Robinson and Glusman, 2018) et constituer une clef d'accès unique à une base génomique même anonymisée et contenant des informations sensibles sur les pathologies de cette personne. C'est en France le rôle de la CNIL de s'assurer de la confidentialité des données, qu'elles soient anonymisées ou seulement pseudonymisées. Dans ce derniers cas (HDH), l'accès aux données doit être particulièrement restreint et protégé afin d'éviter toute fuite par négligence ou malveillance, fuite des dossiers médicaux mais aussi fuite discrète d'une information synthétique ou score personnel de niveau de santé susceptible d'être commercialisé auprès d'une banque, assurance…

12. https://www.oregonlive.com/crime/2019/01/portland-police-tie-texas-serial-killer-rapist-to-40-year-old-homicide-case-using-public-genealogy-data.html

Le *deuxième volet* en balance est celui de l'intérêt public ou bien commun au Canada, conséquence des recherches. L'intérêt d'une recherche académique est généralement évalué par le nombre et l'impact des publications qui en découlent. Néanmoins, cette évaluation bibliométrique impacte l'intérêt des chercheurs – promotion ou accès à des subventions – pas directement celui du public. Il s'agit d'évaluer des intérêts publics concrets et substantiels ; telle est la mission en France du Comité d'Experts de l'Intérêt Public (CEIP) de l'Institut National des Données de Santé (INDS). Ce dernier accorde l'accès à des projets de recherche spécifiques, après avis consultatif de la CNIL sur le volet de la confidentialité.

Parmi les exemples typiques d'application de l'IA en santé, quels sont ceux conduisant à des intérêts substantiels ou non ? Il serait bien trop long de dérouler une étude exhaustive du problème et seulement cinq cas illustratifs seront considérés. Les deux premiers sont le résultat de tests et modèles statistiques avec un objectif explicatif et non prédictif. Il ne s'agit pas formellement d'IA mais le point important à considérer est bien l'ouverture de l'accès aux données et aussi donc la pertinence des résultats obtenus, application d'un algorithme vedette d'IA ou pas.

Le *premier cas* concerne les études épidémiologiques classiques, maintenant appliquées à de très grandes cohortes, en utilisant des modèles statistiques explicatifs. L'analyse de la cohorte *Constances* conduit ainsi à des résultats substantiels présentés chaque année lors d'une journée scientifique[13]. La sécurité des données est essentielle mais il s'agit d'une pratique ancienne et reconnue de la recherche médicale qui ne fait que se déployer en considérant des cohortes d'effectifs nettement plus importants afin de pouvoir détecter (puissance des tests) des facteurs ou combinaisons complexes de facteurs aux effets moins prononcés.

Dans le *deuxième cas*, des batteries de tests statistiques sont appliquées sur les bases pangénomiques pour mettre en évidence la mutation du gène, ou de son promoteur, responsable d'une maladie rare. Pujol (2019), président de la Société Française de Médecine Prédictive et Personnalisée (SFMPP), en décrit les enjeux et intérêts substantiels. Il explicite le difficile débat sur l'opportunité des tests génétiques, très encadrés en France, et sur la pertinence des informations à communiquer aux couples aux différentes étapes de la conception d'un enfant. Cette réflexion est basée sur deux concepts :

- l'estimation statistique de la pénétrance d'une mutation ou probabilité de développer la maladie qui lui est associée : elle est ainsi de 100 % pour la mucoviscidose mais de 75 % de développer un cancer du sein pour une mutation d'un des gènes BRCA ;
- l'actionnabilité ou possibilités thérapeutiques médicales ouvertes par un diagnostic de risque associé à une mutation.

L'utilisation des bases génomiques à cette fin n'est pas remise en cause et est à l'origine du plan France Génomique 2025 incluant une sécurité des données également essentielle.

Le *troisième cas* concerne l'utilisation emblématique de l'apprentissage profond en imagerie médicale en vue d'automatiser le diagnostic ou plutôt l'aide à ce diagnostic. De très nombreuses publications, largement médiatisées, témoignent de leur efficacité : Esteva *et al.* (2017), De Fauw *et al.* (2018), Haenssle *et al.* (2018), Yala *et al.* (2019)... Une synthèse de ces très nombreux travaux (Liu *et al*., 2019a) alerte sur le manque de rigueur de beaucoup de comparaisons entre diagnostic automatique et humain ; celles validées par une évaluation rigoureuse de l'erreur sur des échantillons test indépendants permettent de conclure à une capacité de diagnostic comparable entre l'algorithme et un panel de spécialiste. La FDA (Health, 2019) propose un protocole de certification élaboré qui a permis d'autoriser la pré-commercialisation (Topol, 2019) de nombreux AI/ML-SaMD (*Artificial Intelligence and Machine Learning in a Software as a Medical Device*). Attention, ces dispositifs ne sont pas infaillibles. Comme cela est expliqué plus

13. https://www.constances.fr/actualites/2019/js2019.php

haut, un algorithme d'apprentissage même profond ne peut prévoir que ce qu'il connaît et a déjà rencontré. Ainsi, Oakden-Rayner *et al.* (2019) révèlent le cas d'un cancer du poumon très rare non détecté par une analyse d'image automatique alors qu'il s'agit d'un cas mortel. C'est typiquement ce qui rend indispensable, comme le prévoit la réglementation de la FDA, la mise en place d'une surveillance constante et rétroactive des dispositifs de santé, afin d'en compléter, si nécessaire, l'apprentissage. Dans un livre blanc[14] sur « *le monde des data, des algorithmes et de l'IA* », le Conseil National de l'Ordre des Médecins appelle à juste titre à une *éthique de la vigilance*.

Le *quatrième cas* est la recherche de protéines biomarqueurs, illustré par les résultats encourageants de Williams *et al.* (2019). Ils considèrent une cohorte de 17000 patients pour lesquels 5000 protéines plasmatiques sont dosées à l'aide d'une technologie récente (aptamères) plutôt que par spectrométrie (LC MS/MS). Sur ces données, des algorithmes d'apprentissage – régression avec pénalisation Lasso et *ridge*, machine à vecteurs supports, forêts aléatoires… – conduisent à des bonnes prévisions de l'état de santé du patient et des principaux risques cardiovasculaires, diabètes, meilleures que celles des modèles cliniques usuels. Néanmoins, ces résultats nécessiteraient d'être confirmés sur un autre jeu de données car la présélection des protéines et l'algorithme d'apprentissage ont été exécutés sur le même jeu de données au risque d'un biais de sélection déjà souligné par Ambroise et McLachlan (2002) sur des études transcriptomiques.

Le *cinquième cas*, la recherche sur les maladies multifactorielles utilisant des données pangénomiques, est nettement plus controversé quant à l'intérêt public qu'elle peut apporter alors que c'est celle qui, à terme, brassera le plus grand volume de données. Elle attire de plus la convoitise des acteurs majeurs du numérique qui ont tous des projets plus ou moins avancés dans ce secteur. Ces réserves sont exprimées dès 2010 dans les conclusions d'un texte[15] issu d'une réflexion de la Société Française de Génétique Humaine (Bernheim *et al.*, 2010) et cosigné par l'ensemble des sociétés savantes et associations professionnelles de génétique et génétique humaine :

> « *Si les études pangénomiques apportent une contribution essentielle à la connaissance scientifique, l'utilisation exclusive de l'information qui en résulte est dénuée de sens en matière de prédiction de santé. Elle conduit à une perception erronée du risque encouru par les individus. Il est du devoir de la communauté scientifique de ne pas servir d'alibi en matière de prédictions individuelles de risque pour les maladies multifactorielles à partir de la seule information génomique.* »

Ceci n'a pas pour autant bloqué les programmes de recherche avec la mise en œuvre de méthodes *et al*gorithmes plus sophistiqués. L'annexe 2 en propose une rapide revue montrant des résultats peu probants ou obtenus à la suite d'une démarche manquant de rigueur. Comme déjà évoqué en 2010, ces résultats ne peuvent servir de preuve d'un intérêt public substantiel ou d'alibi pour accéder à de grandes bases de données.

À ce manque de résultats, il faut ajouter dans l'autre plateau de la balance des risques accrus de ré-identification déjà mentionnés. Les données génomiques intègrent implicitement une empreinte génétique définissable à partir d'une sélection de SNP (Robinson and Glusman, 2018). Sans garantie drastique de sécurité, ces empreintes sont autant de clefs d'identification exploitables par des sociétés telles que *23andme*[16] ou les entreprises à qui les données sont

14. https://www.conseil-national.medecin.fr/sites/default/files/cnomdata_algorithmes_ia_0.pdf
15. http://atlasgeneticsoncology.org/Associations/Predictions_risques_maladies_multifactorielles.pdf
16. https://www.23andme.com/en-int/

vendues[17]. Notons qu'aux USA, sans les contraintes européennes légales du RGPD, les principaux acteurs montent des partenariats pour constituer de gigantesques bases de données de santé : *Verily Life Science* filiale d'*Alphabet* et *GSK*[18], *Aetion* et *Sanofi*[19], projet *Nithingale* de *Google* et *Ascension*[20]... Dernière étape, *Verily* signe un partenariat[21] avec *Swiss Re Corporate Solution* (société d'assurances) avec l'opportunité de développer des contrats individualisés à l'encontre du principe, basique en assurance, de mutualisation du risque.

6. Conclusion

La réglementation européenne et les lois nationales sont claires : la *discrimination* est interdite, le *consentement libre et éclairé* des patients doit être requis, sauf pour l'accès à des données personnelles lorsque l'*intérêt public* ou bien commun de la recherche est avéré. En Europe, les questions soulevées par l'utilisation d'algorithmes d'IA en santé sont donc en premier lieu moins d'ordre éthique que juridique et réglementaire.

Les disparités socio-économiques, géographiques et maintenant numériques dans l'accès aux soins sont connues. Le risque, bien identifié, est que des algorithmes de décision d'apprentissage proposant des aides automatiques à la décision viennent renforcer ces biais, en ajoutent d'autres et donc discriminent. La FDA (Health, 2019) comme la HAS imposent des processus adéquats de détection en continu de ces biais tout au long de l'utilisation des seuls dispositifs soumis à leur certification. Ce processus intègre une évaluation de leur risque d'erreur et donc des risques de mauvais diagnostic quand celui-ci est le résultat de l'exécution d'un algorithme opaque. L'information due à l'équipe médicale et au patient pour solliciter son consentement est avant tout de faire connaître l'origine de la décision, l'évaluation du risque d'erreur associée, ainsi que l'opportunité d'un diagnostic complémentaire.

En amont de l'exploitation de ces algorithmes et dispositifs de santé, la recherche doit être *scientifique* et donc ses résultats *reproductibles*. Un effort important doit être consenti par les acteurs de la recherche pour acquérir les compétences indispensables au déploiement d'algorithmes sophistiqués, puissants mais tellement sensibles à la qualité des données, leur représentativité. Il s'agit d'en maîtriser les limites tout en résistant à la pression de publication. Les données étudiées pouvant être confidentielles, il importe de rendre accessible les codes de calcul commentés (Donoho, 2017) afin de permettre une évaluation transparente de la démarche ; les outils actuels (*jupyter notebook*, dépôts *git*) le rendent facile. Le secret commercial n'est pas opposable car les données de l'apprentissage sont protégées et son résultat, le modèle, peut rester confidentiel.

Prenant en considération toute la complexité combinatoire du vivant, associant diversité des variants génétiques et diversité environnementale des conditions de vie, les algorithmes d'apprentissage statistique, même et surtout les plus sophistiqués, partent avec un lourd handicap pour atteindre les objectifs ambitieux d'une médecine personnalisée et prédictive des maladies chroniques multifactorielles. Évaluer globalement les facteurs de risque environnementaux ou génétiques d'une pathologie multifactorielle pour une population est une chose, prévoir très tôt, pour un individu, le risque qu'il déclenche une telle maladie, ou améliorer la prévision obtenue à partir de ses seuls paramètres cliniques en utilisant ses caractéristiques génomiques en est une autre. *Il y a une forme d'antinomie entre les principes de l'apprentissage statistique, basés sur des données, et les objectifs de la médecine prédictive personnalisée* ; au regard des

17. https://www.usinenouvelle.com/article/23andme-vend-l-integralite-des-donnees-genetiques-de-ses-clients-au-laboratoire-gsk-et-cree-la-polemique.N729654
18. https://www.usine-digitale.fr/article/apres-novartis-et-sanofi-verily-life-science-google-se-rapproche-de-gsk.N421917
19. https://www.clinicaltrialsarena.com/news/sanofi-and-aetion-to-integrate-real-world-data-platforms/
20. https://www.theguardian.com/technology/2019/nov/14/google-healthcare-data-ascension
21. https://www.bloomberg.com/news/articles/2020-08-25/alphabet-s-verily-plans-to-use-big-data-as-health-insurance-tool

caractéristiques génétiques et plus encore en croisant celles génétiques et environnementales, chaque humain est unique donc difficilement prédictible ; Keyes *et al.* (2015) met en évidence ces mêmes problèmes à l'aide de simulations.

L'ouverture et l'accès aux données de santé notamment génomiques pour la recherche doivent être conditionnés à des pratiques déontologiques très strictes : exiger des mesures draconiennes de sécurité dans la gestion de bases de données très sensibles, afin d'éviter toute faille de sécurité par incompétence ou même malveillance, exiger une démarche scientifiquement rigoureuse garante de la production de *résultats reproductibles* : identification et correction des biais de tout ordre, afin de produire des prévisions représentatives d'une population de référence à définir, contrôle des prétraitements pour ne pas rajouter de biais qui conduiraient à des situations irréalistes, estimation des erreurs de prévision (AUC) sur des échantillons réellement indépendants et représentatifs de l'usage projeté, publication des codes de calcul afin de faciliter les vérifications.

Ces quelques réflexions nous amènent à suggérer des recommandations à trois niveaux : accès aux données, déontologie de la recherche, réglementation des dispositifs de santé connectés. Ces recommandations sont regroupées dans le *tableau 1* inclus en introduction.

Références

Alaa A. M., T. Bolton, E. Di Angelantonio *et al.* (2019), « Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants », *PLoS ONE*, vol. 14, p. e0213653, https://doi.org/10.1371/journal.pone.0213653.

Ambroise C., G. J. McLachlan (2002), « Selection bias in gene extraction on the basis of microarray gene-expression data », *Proceedings of the National Academy of Sciences*, vol. 99, pp. 6562–6566, https://doi.org/10.1073/pnas.102102699.

Barredo Arrieta A., N. Díaz-Rodríguez, J. Del Ser *et al.* (2020), « Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI », *Information Fusion*, vol. 58, pp. 82–115, https://doi.org/10.1016/j.inffus.2019.12.012.

Bernheim A., C. Bourgain, A. Cambon-Thomsen *et al.* (2010), « Quelle valeur accorder aux prédictions de risques pour les maladies multifactorielles ? », Texte émanant de la Société Française de Génétique Humaine.

Besse P., E. del Barrio, P. Gordaliza, and J.-M. Loubes (2018), « Confidence Intervals for Testing Disparate Impact in Fair Learning », arXiv:180706362 [cs, math, stat].

Buchanan B. G. and E. H. Shortliffe (eds.) (1984), *Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project*, Addison-Wesley, Reading, Mass.

Buolamwini J. and T. Gebru (2018), « Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification », *in Conference on Fairness, Accountability and Transparency*, pp. 77–91.

Chang D., F. Gao, A. Slavney *et al.* (2014), « Accounting for eXentricities: Analysis of the X Chromosome in GWAS Reveals X-Linked Genes Implicated in Autoimmune Diseases », *PLoS ONE*, vol. 9, p. e113684, https://doi.org/10.1371/journal.pone.0113684.

Chouldechova A. (2017), « Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments », *Big Data*, vol. 5, pp. 153–163, https://doi.org/10.1089/big.2016.0047.

Commission Européenne (2018), « Le règlement général sur la protection des données - RGPD ».

Commission Européenne (2020), commission-white-paper-artificial-intelligence-feb2020_fr.pdf, https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_fr.pdf (Accessed 12 Feb 2021).

Darlington K. W. (2011), « Designing for Explanation in Health Care Applications of Expert Systems », *SAGE Open*, vol. 1, p. 21582440114086, https://doi.org/10.1177/2158244011408618.

De Fauw J., J. R. Ledsam, B. Romera-Paredes *et al.* (2018), « Clinically applicable deep learning for diagnosis and referral in retinal disease », *Nat. Med.*, vol. 24, pp. 1342–1350, https://doi.org/10.1038/s41591-018-0107-6.

Do C. B., D. A. Hinds, U. Francke, and N. Eriksson (2012), « Comparison of Family History and SNPs for Predicting Risk of Complex Disease », *PLoS Genet*, vol. 8, p. e1002973, https://doi.org/10.1371/journal.pgen.1002973.

Donoho D. (2017), « 50 Years of Data Science », *Journal of Computational and Graphical Statistics*, vol. 26, pp. 745–766, https://doi.org/10.1080/10618600.2017.1384734.

Dwork C. and A. Roth (2013), « The Algorithmic Foundations of Differential Privacy », *FNT in Theoretical Computer Science*, vol. 9, pp. 211–407, https://doi.org/10.1561/0400000042.

Esteva A., B. Kuprel, R. A Novoa *et al.* (2017), « Dermatologist-level classification of skin cancer with deep neural networks », *Nature*, vol. 542, pp. 115–118, https://doi.org/10.1038/nature21056

Eurostat (2017), « Code de bonne pratique de la Statistique européenne ».

Fjeld J., H. Hilligoss, N. Achten N *et al.* (2019), « Principled Artificial Intelligence », https://ai-hr.cyber.harvard.edu/images/primp-viz.pdf (Accessed 11 Oct 2019).

France Comité Consultatif National d'Éthique (2018), « États Généraux de la Bioéthique : Rapport de Synthèse du Comité Consultatif National d'Éthique – Opinion du comité citoyen », La Documentation Française.

Friedler S. A., C. Scheidegger, S. Venkatasubramanian *et al.* (2019), « A comparative study of fairness-enhancing interventions in machine learning », *in Proceedings of the Conference on Fairness, Accountability, and Transparency – FAT* '19*, Atlanta, GA, USA, ACM Press, pp. 329–338.

Gim J., W. Kim, S. H. Kwak *et al.* (2017), « Improving Disease Prediction by Incorporating Family Disease History in Risk Prediction Models with Large-Scale Genetic Data », Genetics, vol. 207, pp. 1147–1155, https://doi.org/10.1534/genetics.117.300283.

Guichard C. (2018), « Affaire Cambridge Analytica : Facebook chute de près de 7 % en Bourse », *in Courrier international*, https://www.courrierinternational.com/article/affaire-cambridge-analytica-facebook-chute-de-pres-de-7-en-bourse (Accessed 9 Nov 2019).

Haenssle H.A., C. Fink, R. Schneiderbauer *et al.* (2018), « Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists », *Annals of Oncology*, vol. 29, pp. 1836–1842, https://doi.org/10.1093/annonc/mdy166.

Haute Autorité de Santé (2020), « Guide : LPPR Dépôt d'un dossier auprès de la Commission nationale d'évaluation des dispositifs médicaux et des technologies de santé ».

Health C for D and R (2019), « Artificial Intelligence and Machine Learning in Software as a Medical Device », FDA.

High Level Expert Group (2019), « Ethics guidelines for trustworthy AI », *in Digital Single Market – European Commission*, https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (Accessed 9 Nov 2019).

Ho D. S. W., W. Schierding, M. Wake *et al.* (2019), « Machine Learning SNP Based Prediction for Precision Medicine », *Frontiers in Genetics*, vol. 10, p. 267, https://doi.org/10.3389/fgene.2019.00267

Ioannidis J. P. A. (2016), « Why Most Clinical Research Is Not Useful », *PLoS Med*, vol. 13, p. e1002049, https://doi.org/10.1371/journal.pmed.1002049.

James G., D. Witten, T. Hastie, and R. Tibshirani (2013), *An Introduction to Statistical Learning*, New York, NY, Springer.

Jobin A., M. Ienca, and E. Vayena (2019), « The global landscape of AI ethics guidelines », *Nature Machine Intelligence*, vol. 1, pp. 389–399, https://doi.org/10.1038/s42256-019-0088-2.

Kahn H. S. (2009), « Two Risk-Scoring Systems for Predicting Incident Diabetes Mellitus in U.S. Adults Age 45 to 64 Years », *Annals of Internal Medicine*, vol. 150, p. 741, https://doi.org/10.7326/0003-4819-150-11-200906020-00002.

Keyes K. M., G. Davey Smith, K. C. Koenen, and S. Galea (2015), « The mathematical limits of genetic prediction for complex chronic disease », *Journal of Epidemiology and Community Health*, vol. 69, pp. 574–579, https://doi.org/10.1136/jech-2014-204983.

Kim J., C. Hu, C. Moufawad El Achkar *et al.* (2019), « Patient-Customized Oligonucleotide Therapy for a Rare Genetic Disease », *The New England Journal of Medicine*, vol. 381, pp. 1644-1652, https://doi.org/10.1056/NEJMoa1813279.

Klein R. J. (2005), « Complement Factor H Polymorphism in Age-Related Macular Degeneration », *Science*, vol. 308, pp. 385–389, https://doi.org/10.1126/science.1109557.

Kraege V., J. Fabecic, P. M. Vidal *et al.* (2020), « Validation of seven type 2 diabetes mellitus risk scores in a population-based cohort. The CoLaus Study », *The Journal of Clinical Endocrinology & Metabolism*, vol. 105, n° 3, https://doi.org/10.1210/clinem/dgz220.

Larson J. and J. Angwin (2016), « How We Analyzed the COMPAS Recidivism Algorithm », *in ProPublica*, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm (Accessed 10 Nov 2019).

Lee P., M. Le Saux, R. Siegel *et al*. (2019), « Racial and ethnic disparities in the management of acute pain in US emergency departments: Meta-analysis and systematic review », *The American Journal of Emergency Medicine*, vol. 37, pp. 1770–1777, https://doi.org/10.1016/j.ajem.2019.06.014.

Lindström S., S. Loomis, C. Turman *et al*. (2017), « A comprehensive survey of genetic variation in 20,691 subjects from four large cohorts », *PLoS ONE*, vol. 12, p. e0173997, https://doi.org/10.1371/journal.pone.0173997.

Liu X., L. Faes, A. U. Kale *et al*. (2019a), « 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*, vol. 1, pp. e271–e297, https://doi.org/10.1016/S2589-7500(19)30123-2.

Liu Y., P.-H. C. Chen, J. Krause, and L. Peng (2019b), « How to Read Articles That Use Machine Learning: Users' Guides to the Medical Literature », *JAMA*, vol. 322,p. 1806, https://doi.org/10.1001/jama.2019.16489.

London A. J. (2019), « Artificial Intelligence and Black-Box Medical Decisions: Accuracy *versus* Explainability », *Hastings Center Report*, vol. 49, pp. 15–21, https://doi.org/10.1002/hast.973.

Lopez B., F. Torrent-Fontbona, R. Viñas, and J. M. Fernández-Real (2018), « Single Nucleotide Polymorphism relevance learning with Random Forests for Type 2 diabetes risk prediction », *Artificial Intelligence in Medicine*, vol. 85, pp. 43–49, https://doi.org/10.1016/j.artmed.2017.09.005.

Mieth B., M. Kloft, J. A. Rodríguez *et al*. (2016), « Combining Multiple Hypothesis Testing with Machine Learning Increases the Statistical Power of Genome-wide Association Studies », *Scientific Reports*, vol. 6, pp. 1–14, https://doi.org/10.1038/srep36671.

Montanez C. A. C., P. Fergus, A. C. Montanez *et al*. (2018), « Deep Learning Classification of Polygenic Obesity using Genome Wide Association Study SNPs », *in* 2018 International Joint Conference on Neural Networks (IJCNN), IEEE, Rio de Janeiro, pp. 1–8.

Morel M., E. Bacry, S. Gaïffas *et al*. (2019), « ConvSCCS: convolutional self-controlled case series model for lagged adverse event detection », *Biostatistics*, vol. 21, pp. 758–774, https://doi.org/10.1093/biostatistics/kxz003.

Narayanan A. and V. Shmatikov (2008), « Robust De-anonymization of Large Sparse Datasets », *in* 2008 IEEE Symposium on Security and Privacy (sp 2008), IEEE, Oakland, CA, USA, pp. 111–125.

Oakden-Rayner L., J. Dunnmon, G. Carneiro, and C. Ré (2019), « Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging », arXiv:190912475 [cs, stat].

Obermeyer Z. and S. Mullainathan (2019), « Dissecting Racial Bias in an Algorithm that Guides Health Decisions for 70 Million People », *in Proceedings of the Conference on Fairness, Accountability, and Transparency – FAT* '19*, Atlanta, GA, USA, ACM Press, pp. 89–89.

O'Neil C. (2016), *Weapons of math destruction: how big data increases inequality and threatens democracy* (First edition), New York, Crown.

Patron J., A. Serra-Cayuela, B. Han *et al*. (2019), « Assessing the performance of genome-wide association studies for predicting disease risk », *PLoS ONE*, vol. 14, p. e0220215, https://doi.org/10.1371/journal.pone.0220215.

Popejoy A. B. and S. M. Fullerton (2016), « Genomics is failing on diversity », *Nature*, vol. 538, pp. 161–164, https://doi.org/10.1038/538161a.

Pujol P. (2019), *Voulez-vous savoir ? Ce que nos gènes disent de notre santé*, Paris, Editions humensciences.

Pulit S. L., T. Karaderi, and C. M. Lindgren (2017), « Sexual dimorphisms in genetic loci linked to body fat distribution », *Bioscience Reports*, vol. 37, n° 1, p. BSR20160184, https://doi.org/10.1042/BSR20160184.

Racine E., W. Boehlen, and M. Sample (2019), « Healthcare uses of artificial intelligence: Challenges and opportunities for growth », *Healthcare Management Forum*, vol. 32, pp. 272–275, https://doi.org/10.1177/0840470419843831.

Rappaport S. M. (2016), « Genetic Factors Are Not the Major Causes of Chronic Diseases », *PLoS ONE*, vol. 11, p. e0154387, https://doi.org/10.1371/journal.pone.0154387.

Robinson M. and G. Glusman (2018), « Genotype Fingerprints Enable Fast and Private Comparison of Genetic Testing Results for Research and Direct-to-Consumer Applications », *Genes*, vol. 9, p. 481, https://doi.org/10.3390/genes9100481.

Rocher L., J. M. Hendrickx, and Y.-A. de Montjoye (2019), « Estimating the success of re-identifications in incomplete datasets using generative models », *Nature Communications*, vol. 10, p. 3069, https://doi.org/10.1038/s41467-019-10933-3.

Ross C. and I. Swetlitz (2018), « IBM's Watson recommended "unsafe and incorrect" cancer treatments », *in STAT*, https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/ (Accessed 10 Nov 2019).

Rubinstein I. S. and W. Hartzog (2016), « Anonymization and Risk », *Washington Law Revue*, vol. 91, p. 59.

Ruby J. G., K. M. Wright, K. A. Rand *et al*. (2018), « Estimates of the Heritability of Human Longevity Are Substantially Inflated due to Assortative Mating », *Genetics*, vol. 210, pp. 1109–1124, https://doi.org/10.1534/genetics.118.301613.

Rumelhart D. E., G. E. Hinton, and R. J. Williams (1986), « Learning representations by back-propagating errors », *Nature*, vol. 323, pp. 533–536, https://doi.org/10.1038/323533a0.

Schwarzinger M., B. G. Pollock, O. S. M. Hasan *et al*. (2018), « Contribution of alcohol use disorders to the burden of dementia in France 2008-13: a nationwide retrospective cohort study », The *Lancet Public Health*, vol. 3, pp. e124–e132, https://doi.org/10.1016/S2468-2667(18)30022-7.

Silver D., T. Hubert, J. Schrittwieser *et al*. (2017), « Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm », arXiv:171201815 [cs].

Topol E. J. (2019), « High-performance medicine: the convergence of human and artificial intelligence », *Nature Medicine*, vol. 25, pp. 44–56, https://doi.org/10.1038/s41591-018-0300-7.

Udler M. S., M. I. McCarthy, J. C. Florez, and A. Mahajan (2019), « Genetic Risk Scores for Diabetes Diagnosis and Precision Medicine », *Endocrine Reviews*, vol. 40, pp. 1500–1520, https://doi.org/10.1210/er.2019-00088.

Université de Montréal (2018), « La déclaration de Montréal pour le développement responsable de l'intelligence artificielle ».

Vayena E., A. Blasimme, and I. G. Cohen (2018), « Machine learning in medicine: Addressing ethical challenges », *PLoS Med*, vol. 15, p. e1002689, https://doi.org/10.1371/journal.pmed.1002689

Villani C., M. Schoeunauer, Y. Bonnet *et al*. (2018), « Donner un sens à l'intelligence artificielle : pour une stratégie nationale et européenne », La Documentation Française.

Vyas D. A., L. G. Eisenstein, and D. S. Jones (2020), « Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms », *The New England Journal of Medicine*, vol. 383, pp. 874–882, https://doi.org/10.1056/NEJMms2004740.

Wei Z., W. Wang, J. Bradfield *et al*. (2013), « Large Sample Size, Wide Variant Spectrum, and Advanced Machine-Learning Technique Boost Risk Prediction for Inflammatory Bowel Disease », *The American Journal of Human Genetics*, vol. 92, pp. 1008–1012, https://doi.org/10.1016/j.ajhg.2013.05.002.

Wiens J., S. Saria, M. Sendak *et al*. (2019), « Do no harm: a roadmap for responsible machine learning for health care », *Nature Medicine*, vol. 25, pp. 1337–1340, https://doi.org/10.1038/s41591-019-0548-6.

Williams S. A., M. Kivimaki, C. Langenberg *et al*. (2019), « Plasma protein patterns as comprehensive indicators of health », *Nature Medicine*, vol. 25, pp. 1851–1857, https://doi.org/10.1038/s41591-019-0665-2.

Wright K. M., K. A. Rand, A. Kermany *et al*. (2019), « A Prospective Analysis of Genetic Variants Associated with Human Lifespan », *G3*, vol. 9, pp. 2863–2878, https://doi.org/10.1534/g3.119.400448.

Yala A., C. Lehman, T. Schuster *et al*. (2019), « A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction », *Radiology*, vol. 292, pp. 60–66, https://doi.org/10.1148/radiol.2019182716.

Zins M., S. Bonenfant, M. Carton *et al*. (2010), « The CONSTANCES cohort: an open epidemiological laboratory », *BMC Public Health*, vol. 10, p. 479, https://doi.org/10.1186/1471-2458-10-479.

Žliobaitė I. (2017), « Measuring discrimination in algorithmic decision making », *Data Mining and Knowledge Discovery*, vol. 31, pp. 1060–1089, https://doi.org/10.1007/s10618-017-0506-1.

Annexes

Annexe 1 – Area under the curve ou AUC

Tout modèle ou algorithme d'apprentissage produit, pour chaque individu sur l'échantillon test indépendant, une prévision, sous la forme d'une probabilité $p(x_i)$ entre 0 et 1, d'avoir affaire à un mauvais payeur, de satisfaire à un emploi, de développer une pathologie... Cette probabilité est ensuite comparée à une valeur seuil *s* (par défaut 0,5) pour une prise de décision binaire. La comparaison entre les décisions, dépendant de *s*, et les vraies valeurs observées sur l'échantillon test conduit à la construction d'une table (tableau 2) de contingence ou matrice de confusion.

Tableau 2 – *Table de contingence croisant la prévision avec la valeur observée sur l'échantillon test indépendant ; toutes les quantités dépendent de la valeur seuil s choisie a priori.*

Prévision	**Observation**		
	Oui	**Non**	**Marge**
Oui	*a(s)*	*b(s)*	*a* + *b*
Non	*c(s)*	*d(s)*	*c* + *d*
Marge	*a* + *c*	*b* + *d*	*n* = *a* + *b* + *c* + *d*

Dans cette matrice, *a(s)* désigne le nombre de bonnes décisions ou vrais positifs, *d(s)* le nombre de vrais négatifs, *c(s)* le nombre de faux négatifs et *b(s)* le nombre de faux positifs. Le taux d'erreur est simplement défini par Terr = (*b* + *c*)/*n* mais est généralement insuffisant pour apprécier la qualité d'une prévision surtout si les classes sont déséquilibrées. De très nombreux indicateurs ont été définis dont la *sensibilité* ou taux de vrais positifs : TPR = *a*/(*a* + *c*) ; la *spécificité* ou taux de vrais négatifs : TNR = *d*/(*b* + *c*) ; le taux de faux positifs : FPR = *b*/(*b* + *c*) qui est encore *un moins la spécificité*. En faisant varier le seuil *s*, il est possible de tracer la courbe ROC (*receiver operating caracteristic*) exprimant TPR en fonction de FPR, dont la figure 1 donne un exemple. Plus la courbe se rapproche du cadre supérieur avec une croissance rapide, meilleure est la prévision avec la détection nette (valeur élevée de *s*) d'une grande proportion de vrais positifs en limitant la part de faux positifs ; l'AUC (entre 0,5 et 1) est l'aire délimitée par cette courbe. Si la courbe est proche de la diagonale (AUC = 0,5), la prévision n'est qu'un tirage à pile ou face. Cet indicateur permet de comparer les qualités de prévision de différents modèles ou algorithmes, celle-ci est jugée « bonne » au-delà de 0,8, excellente au-delà de 0,9.

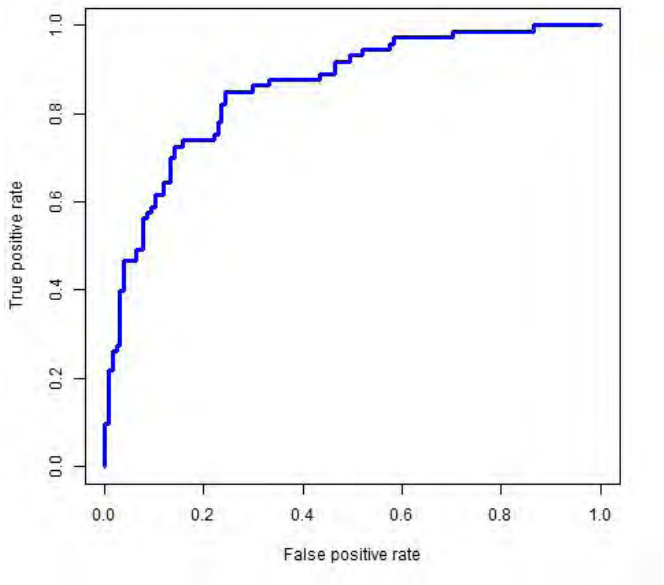


Figure 1 – *Exemple de courbe ROC pour la prévision de la variable binaire : dépassement du seuil de pic d'ozone. Les deux taux de vrais et faux positifs sont représentés en fonction de la valeur décroissante du seuil s de décision. Cette courbe aide à choisir, choix politique, la valeur de s conduisant à une détection d'un taux raisonnable de vrais positifs (pollution effective) pour la santé publique au regard d'un taux de faux positifs (absence de pollution) économiquement admissible.*

Annexe 2 – Analyse pangénomique des maladies multifactorielles

L'analyse ou médecine génomique appliquée aux maladies multigéniques, multifactorielles, généralement chroniques (cardiovasculaires, obésité, diabète, certains cancers…) et, plus idéalement encore, à l'étude de la durée de vie (Wright *et al*., 2019), soulève un ensemble de questions sur la pertinence des résultats, leur intérêt thérapeutique et leurs finalités au regard des risques encourus. Rappaport (2016) montre que les facteurs génétiques ne sont pas les facteurs majeurs des maladies chroniques. Pujol (2019) explique que, contrairement aux maladies monogéniques, la pénétrance d'une combinaison d'un nombre même important de variants génétiques est très faible, généralement quelques pourcents. Patron *et al*. (2019) proposent un outil pour estimer la qualité prédictive d'études pangénomiques publiées. Appliqué à 569 études, leur outil montre que très peu produisent une AUC (cf. définition annexe 1) plus grande que 0,75 ; la prédictibilité reste très nettement inférieure à celle obtenue avec les seules mesures cliniques classiques.

Ho *et al*. (2019) font la promotion de l'utilisation de l'apprentissage automatique, avec des maladresses sur sa présentation, par rapport à des facteurs de risques linéaires basés sur des variants jugés significativement influents ; ils citent quelques travaux plus prometteurs. Wei *et al*. (2013) obtiennent une qualité prédictive raisonnable (AUC de 0,83) de maladies inflammatoires de l'intestin en utilisant une régression (linéaire) logistique pénalisée (Lasso) pour opérer la sélection de quatre à cinq cents SNP après une pré-sélection (tests multiples) de 10800 parmi près de 179000. Lopez *et al*. (2018) utilisent l'algorithme des forêts aléatoires pour évaluer le risque d'occurrence du diabète de type 2 (T2D) pour aboutir à une AUC de 0,85. Les données initiales de ces deux analyses ont été abondamment nettoyées et sélectionnées pour produire des résultats tout à fait honorables sur des données largement prétraitées, mais, comme le rappelle Liu *et al*. (2019b), sont-ils reproductibles ? Qu'en serait-il sur de nouvelles données réelles, en général moins propres et pas nécessairement issues du même protocole

technologique ? Malgré ce qui est avancé, Mieth *et al*. (2016) s'intéressent à l'objectif de sélection de SNP influents mais pas à une prévision. Montañez *et al*. (2018) obtiennent pour la prévision de l'obésité une qualité (AUC de 0,99) qui éveille la suspicion. Cette qualité est obtenue en considérant 2465 SNP, issus d'un premier filtrage (tests) de 241000, alimentant un réseau de neurones avec deux couches cachées entraîné sur seulement 1200 individus. Ces quelques chiffres laissent présager une **situation de sur-apprentissage** mais qui ne peut être infirmée ou confirmée avec certitude sans disposer du code de l'analyse. Néanmoins, le descriptif succinct de la démarche confirme ce doute. Elle suit en détail les étapes du début du tutoriel du logiciel H2O[22] utilisé qui introduit une confusion dans les rôles des échantillons d'apprentissage, validation et test. La démarche semble donc manquer de rigueur et nous pouvons douter de la reproductibilité de la prévision sur un autre échantillon complètement indépendant.

Par ailleurs Lopez *et al*. (2018) regardent l'amélioration de la prévision du T2D par l'ajout de variables cliniques en ne respectant pas les standards cliniques comme les valeurs du glucose ou de l'insuline sanguin à jeun. Une revue plus récente et mieux documentée sur la même pathologie (Udler *et al*., 2019) conduit à des résultats plus nuancés dans l'utilisation de scores polygéniques : des valeurs d'AUC systématiquement plus faibles que celles fournies par des variables cliniques et une amélioration non significative de ces dernières lorsque les variables génomiques sont ajoutées. Lors du suivi d'une autre cohorte, Kraege *et al*. (2020) obtiennent un AUC de 0,85 très concurrentiel sur la base d'un seul score clinico-biologique (Kahn, 2009). Une des questions est donc de savoir si une analyse génétique, encore relativement coûteuse même si une seule suffit pour la vie, apporte une prévision du risque significativement plus précise qu'une analyse longitudinale des variables cliniques (Alaa *et al*., 2019). Il ne faut pas pour autant nier la part génétique de certaines maladies mais pour les principales maladies, notamment cardio-vasculaires (Do *et al*., 2012) ou T2D (Gim *et al*., 2017), la connaissance de l'historique familiale fait tout aussi bien. De plus, l'absence de prise en compte des conditions environnementales ou des styles de vie dans l'ensemble des études peut être une source de biais ou plutôt de confusion additionnelle en plus des biais ethniques déjà mentionnés. Une proximité phylogénétique (SNP ou historique familiale) peut être corrélée à une proximité géographique, sociologique donc de style de vie. Sans précision à ce sujet il est alors difficile de faire la part des choses entre les effets respectifs génétiques ou environnementaux et même leurs interactions potentielles.

Ceci n'empêche pas la société *Calico* (*California Life Company filiale d'Alphabet*) de continuer à financer l'étude de la détection de variants génétiques influençant la durée de vie par des modèles de Cox (Wright *et al*., 2019) appliqués aux bases pangénomiques, alors que Ruby *et al*. (2018) estiment à moins de 10 % la part génétique dans la durée de vie humaine.

L'étape suivante de cette démarche devrait déployer des algorithmes d'apprentissage statistique sur des cohortes très volumineuses associant données longitudinales environnementales, biologiques et cliniques, ainsi que génomiques. L'étude des possibles interactions entre variables environnementales, dont les effets sont connus, et génétiques nécessiterait, en raison de l'extrême complexité des phénomènes en jeu, des volumes de données considérables sans pour autant être sûr, à ce jour, de la pertinence des résultats attendus. Est-ce socialement acceptable compte tenu des risques encourus ?

Après cet aperçu illustratif et partiel, donc sans doute partial, d'une littérature très volumineuse, il est difficile d'établir une synthèse claire mettant en évidence un intérêt public substantiel qu'apporterait l'intégration des données génomiques dans la recherche sur les maladies multifactorielles chroniques.

22. http://docs.h2o.ai/h2o-tutorials/latest-stable/tutorials/deeplearning/index.html

Résumé succinct de la progression épistémologique récente de la recherche sur les maladies multifactorielles visant à intégrer ou associer différents types de données et méthodes d'analyse :

- des études épidémiologiques, basées sur des cohortes intégrant des données longitudinales, mettent en évidence, par des modèles statistiques, l'importance des facteurs environnementaux et des modes de vie ;
- des études pangénomiques, toujours basées sur des modèles statistiques, révèlent des listes importantes de variants génétiques, chacun de pénétrance faible, susceptibles d'influer sur le risque de maladie ;
- la prise en compte de ces facteurs génétiques n'améliore que très marginalement les qualités prédictives des scores cliniques, alors que des dosages de protéines (Williams *et al.*, 2019) conduisent eux à des résultats très encourageants ;
- l'utilisation d'algorithmes d'apprentissage statistique sur données génomiques, ou génomiques et cliniques, pour établir ces prévisions n'apporte pas de résultats plus probants ou soulève des questions quant à la rigueur de la mise en œuvre de ces outils sophistiqués, efficaces, mais excessivement sensibles à toute forme de biais : représentativité et nettoyage des données, gestion des échantillons d'entraînement, validation et test, validation sur un échantillon test réellement indépendant. Le principal risque est la construction de résultats spécifiques aux données étudiées, des artefacts non reproductibles. Ainsi, la notion de biais de sélection a déjà été identifiée par Ambroise et McLachlan (2002) dans les études transcriptomiques lorsqu'un algorithme de prévision est entraîné sur une sélection de variables sans intégrer la sélection opérée dans l'estimation de l'erreur de prévision. Mieth *et al.* (2016) induisent le même type de biais en présélectionnant une liste de SNP à l'aide d'un algorithme de SVM linéaire avant d'opérer la sélection classique par tests multiples mais avec une correction de Bonferroni inadaptée car calculée sur le seul nombre de tests.

Deep Learning : des usages contrastés
Une contextualisation de l'ouvrage de Goodfellow, Bengio et Courville


Rémi ADON

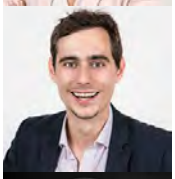

Florian ARTHUR


Guillaume BAQUIAST


Guillaume HOCHARD


Abdellah KAID GHERBI


Aurélia NÈGRE


Antoine SIMOULIN


Fouad TALAOUIT-MOCKLI

Quantmetry


Nicolas BOUSQUET[1]

EDF R&D - Laboratoire d'IA industrielle SINCLAIR & Sorbonne Université

TITLE

A critical analysis of the use of deep learning within the socio-economic world

RÉSUMÉ

Cet article propose une revue critique de la traduction française de l'ouvrage *Deep Learning*, par Ian Goodfellow, Yoshua Bengio et Aaron Courville (Goodfellow et al., 2016 ; MIT Press), publiée sous le titre *L'apprentissage profond* (Éditions Florent Massot, 2018). Celle-ci est devenue célèbre pour avoir été la première traduction scientifique d'envergure coproduite par une intelligence artificielle. Alors que l'apprentissage profond connaît une évolution rapide, cet ouvrage et les idées qu'il véhicule restent profondément d'actualité. Nourrie par de nombreux retours d'expérience portant sur l'usage réel de l'apprentissage profond au sein des entreprises, la mise en perspective de ce corpus de méthodes et d'outils vis-à-vis d'approches plus traditionnelles s'articule autour de trois thématiques-clés : le traitement d'images, l'analyse de séries temporelles et le traitement automatisé du langage naturel. Deux enjeux cruciaux pour une adoption massive mais éclairée y sont également discutés, qui nous semblent contextualiser utilement les apports techniques de l'ouvrage : l'intelligibilité de l'apprentissage profond et l'optimisation énergétique des ressources de calcul.

Mots-clés **:** *apprentissage profond, industrialisation, modèles et algorithmes, méthodes statistiques, analyse critique.*

1. nicolas.bousquet@sorbonne-universite.fr

ABSTRACT

This article provides a critical review of the French translation of the book *Deep Learning*, by Ian Goodfellow, Yoshua Bengio and Aaron Courville (Goodfellow et al., 2016; MIT Press), published under the title *L'apprentissage profond* (Éditions Florent Massot, 2018). This became famous for being the first large-scale scientific translation co-produced by an artificial intelligence. While deep learning is undergoing rapid evolution, this book and the ideas it conveys remain profoundly topical. Based on a large amount of feedback on the real use of deep learning within companies, the perspective of this corpus of methods and tools in relation to more traditional approaches revolves around three key themes: image processing, analysis of temporal data and automated processing of natural language. It also discusses two issues that are crucial for massive but mature adoption, and which seem to us to usefully contextualize the technical contributions of the book: the intelligibility of deep learning-based approaches and the energy optimization of computing resources.

***Keywords*:** *deep learning, industrialization, models and algorithms, statistical methods, critical analysis.*

1. Introduction

1.1. L'essor de l'apprentissage profond

L'ouvrage *Deep Learning*, paru fin 2016, revêt une importance particulière. À notre connaissance, il reste en effet l'un des seuls à offrir, après une introduction aux concepts mathématiques et aux bases de l'apprentissage automatique (*machine learning* ou ML), un panorama vaste et accessible de l'état de l'art scientifique en apprentissage profond (AP), domaine qui connaît aujourd'hui un engouement extraordinaire. La puissance de structuration des algorithmes d'AP, largement fondés sur l'emploi de réseaux de neurones artificiels, suggère de multiples champs applicatifs et les impose aujourd'hui comme les composants phares des intelligences artificielles *connexionnistes* (Schuman et al., 2017b), c'est-à-dire fondées sur l'exhibition de corrélations fines entre les phénomènes produisant les données. Ces outils se révèlent également si adaptés à la gestion des importants volumes de données générés par les activités numériques que des processeurs spécifiquement créés pour l'emploi optimisé de ces algorithmes commencent à être commercialisés à grande échelle (Esser et al., 2016; Schuman et al., 2017a). De ce fait, l'attractivité des métiers liés à l'exposition, la mise en forme et le traitement des données ne cesse de croître.

Quelques années après sa publication originale, cet ouvrage reste fondamental car il répond à l'exigence d'appréhender simultanément le problème de l'apprentissage automatique sous différents angles : ainsi, l'aspect informatique vise à construire des *pipeline* opérationnels de collecte, stockage et traitement des données, tandis que l'aspect statistique ambitionne de proposer un choix de *modèle* explicatif ou génératif sous-jacent à un problème de décision. La traduction française de cet ouvrage a notamment pour objectif de faciliter l'appropriation par les différents domaines professionnels concernés par l'exploitation de données massives ; comme l'écrit Cédric Villani, « *on pense toujours plus facilement [les sciences] dans sa langue natale* » (Villani, 2016).

1.2. Les risques potentiels d'un changement de paradigme

Si cette puissance des algorithmes d'AP couplés à des architectures et des méthodes de prétraitement de données spécialisées est indéniable, le réalisme de la mise en œuvre actuelle et l'utilité pratique de certains d'entre eux, en regard d'autres approches disponibles, sont ici questionnés. Le point de vue développé dans cet article est celui de praticiens expérimentés, confrontés à des problèmes de traitement de données relevant de métiers aussi variés qu'assureur, banquier d'affaire, industriel de l'énergie, professionnel de santé, logisticien, géologue ou chargé de ressources humaines. Pour certains métiers, la popularité de l'AP paraît leur imposer de se renouveler en profondeur, quitte à balayer un vaste ensemble de techniques et de méthodologies historiques. Un exemple frappant d'une telle injonction concerne ainsi le traitement de l'image, pour lequel des méthodes traditionnelles peuvent cependant rester très performantes pour la résolution de problèmes spécifiques, comme la *segmentation**[2] (Féron et Mohammad-Djafari, 2005; Pereyra et McLaughlin, 2015) – celle-ci pouvant justement favoriser l'usage de l'AP, puisqu'elle permet une labellisation[3] plus rapide (Zhang et Xu, 2018). De même, le problème du traitement des données censurées en étude de survie, généralement bien opéré par des algorithmes du type *Expectation-Maximization* (EM), est maintenant appréhendé par des outils d'AP sans que la compréhension du mécanisme probabiliste de

2. Les astériques présents dans le texte renvoie à des définitions proposées dans un lexique en fin d'article.
3. Voir § 3.1.1 pour une explication détaillée de ce terme.

reconstruction de ces données soit explicitée (voir par exemple Katzman et al., 2018). L'impératif permanent de rapidité, répondant à l'attente d'une audience fascinée par des résultats impressionnants en intelligence artificielle, y est sans doute pour quelque chose. Mais un tel dynamisme, illustré par l'augmentation exponentielle du nombre de contributions, risque – par la pression qu'elle place sur les chercheurs, notamment – d'aboutir à des pertes de qualité et de pertinence méthodologique (Bengio, 2020), potentiellement dommageables pour l'usage réel de l'AP et son enseignement (Lipton et Steinhardt, 2019).

De façon sous-jacente, nous percevons au contact des applications métiers que la très forte publicité donnée aux outils d'AP risque d'isoler des communautés scientifiques et de se priver de solutions utiles, éprouvées, moins gourmandes en données et en ressources. Notre conviction est qu'il est naturel, et souhaitable, que l'avancement de la science offre un nombre croissant de solutions à un problème donné, mais qu'il est capital de perpétuer la connaissance et l'usage d'approches traditionnelles. Celles-ci peuvent conserver un avantage comparatif par rapport aux dernières méthodologies proposées, ne serait-ce qu'en facilité d'interprétation[4]. Le développement et l'usage intensif de ces algorithmes s'accompagne de questionnements croissants sur leur intelligibilité, leur transparence, leur loyauté* et plus généralement sur la nature du produit informationnel résultant du traitement des données (Besse et al., 2017). Des modèles explicatifs, causaux, exhibant les propriétés principales d'un phénomène permettent un tel dialogue entre l'homme et la machine « boîte noire » (Bhatt, 2018).

Nous pensons donc que l'AP mérite encore de nombreux approfondissements théoriques ainsi que des retours d'expériences toujours plus variés, et qu'il est nécessaire de contextualiser son usage dans les applications métiers. L'objectif de cet article est de contribuer à ces réflexions.

1.3. Contributions de l'article

Un résumé des principaux contenus et messages de l'ouvrage suit cette introduction. Nous considérons ensuite trois champs d'application importants de l'apprentissage profond – le traitement d'image, celui de signaux temporels et de données textuelles – illustrés par des situations « de terrain ». Ces exemples permettent de percevoir les gains de performance, mais aussi les contraintes qu'amènent les méthodologies fondées sur les réseaux de neurones profonds, et les solutions traditionnelles utilisables en première intention. Les choix de solution répondent à des considérations pratiques multiples, et ne sont pas seulement liés à la disponibilité de machines puissantes ou à la taille des données disponibles. Une section de discussion est enfin consacrée à deux sujets d'ouverture peu développés dans l'ouvrage, au sujet desquels entreprises et société civile prêtent de plus en plus d'attention : l'intelligibilité des algorithmes d'AP et le coût environnemental de leur mise en œuvre au sein d'outils d'intelligence artificielle.

2. Résumé de l'ouvrage

Le monde de l'apprentissage automatique profond est abordé en trois parties. La première est consacrée aux bases d'un cursus de mathématiques appliquées et aux fondations statistiques de l'apprentissage. Des outils classiques (maximisation de vraisemblance, règle de Bayes, ...) jusqu'aux concepts les plus avancés (astuce du noyau, modèle de représentation), l'essentiel

4. Il est d'ailleurs assez révélateur de percevoir que les approches *post hoc* d'interprétation des modèles et algorithmes d'AP reposent essentiellement sur des métriques de comparaison avec des approches statistiques classiques ; voir § 4.2.

des idées de modélisation et d'estimation définissant les approches supervisées, non supervisées et par renforcement de l'apprentissage statistique est présenté selon une approche historique.

La notion de supervision renvoie à l'existence de données $(\mathbb{x}, \mathbb{y}) = (x_i, y_i)_{i\in\mathcal{I}}$, où $\mathcal{I}$ désigne un échantillon dit d'*apprentissage*. Si $\mathbb{x}$ décrit des caractéristiques observables du phénomène à l'étude, le *label*[5] $\mathbb{y}$ décrit une caractéristique décisionnelle vis-à-vis de ce phénomène, idéalement reliée à $\mathbb{x}$ par des relations causales inconnues : quelle action y doit-on préconiser, voire automatiser si on observe x, sachant qu'on connaît $(\mathbb{x}, \mathbb{y})$? Ce label $\mathbb{y}$ peut, typiquement, décrire une classe de phénomènes (approche par classification, voir Figure 1a) ou l'étalement d'une valeur ou d'un vecteur de valeurs d'intérêt (approche par régression). Une démarche *non supervisée*, privée de $\mathbb{y}$, exploite la géométrie du nuage de points $\mathbb{x}$ pour en déduire des modèles statistiques sur $\mathbb{x}$ (ex. : mélanges gaussiens) ou des partitionnements de ce nuage de points. Elle sous-tend donc qu'une décision $\mathbb{y}$ peut être formalisée à partir de cette classification (Figure 1b). Le *renforcement*, quant à lui, consiste à permettre une exploration progressive de l'environnement du couple $(\mathbb{x}, \mathbb{y})$ et s'approche de la réalité d'une prise de décision à effectuer séquentiellement dans un contexte où les incertitudes sont graduellement réduites (Figure 1c).

Ces trois approches délimitent les limites conceptuelles de ce qu'on nomme l'apprentissage automatique (on « apprend » en une fois ou progressivement les corrélations liant $\mathbb{x}$ à $\mathbb{y}$). Celui-ci comprend l'apprentissage machine (AM, ou *machine learning*) traditionnel et l'AP.

Dans la réalité des applications pratiques, les deux premières approches sont souvent mises en interaction[6]. L'apprentissage par renforcement présente cette particularité d'être encore réservé à des cas d'étude très délimités[7], en dépit de l'intérêt qu'il suscite. Sa faible maturité implique qu'il reste nettement moins étudié dans *L'apprentissage profond*, et le lecteur intéressé pourra plutôt se référer à Szepesvári (2010), l'une des bibles du domaine.

Transverses au domaine de l'apprentissage automatique, les enjeux cruciaux de régularisation, généralisation et de montée en échelle (ou *scalabilité*) – c'est-à-dire la limitation des risques de sur-apprentissage*, la pertinence de l'application des outils à de nouvelles données $\tilde{\mathbb{x}}$ et la gestion du fléau de la dimension* et de la taille croissante des données – permettent de saisir l'intérêt des approches par réseaux de neurones artificiels (RNN) dits *profonds*, qui font l'objet de la seconde partie du livre et qui constituent les outils fondamentaux de l'AP.

Les RNN constituent une classe de modèles de traitement du signal, qui s'inspirent vaguement du fonctionnement des neurones biologiques. Ils sont composés d'unités inter-connectées disposées en couches successives, décrites comme des neurones artificiels (Figure 2). Dans leurs formes les plus simples (dite de *propagation avant*), par le biais de techniques d'optimisation formelles[8] dites d'*entraînement* (un statisticien dirait d'*estimation*) , ces structures peuvent réaliser des approximations de correspondances *a priori* inconnues entre un ensemble d'entrées observées $\mathbb{x}$ et des sorties (labels) $\mathbb{y}$ connues. Étant donné une entrée $\mathbb{x} = (x_1, \dots, x_n)$, un neurone génère un signal de sortie

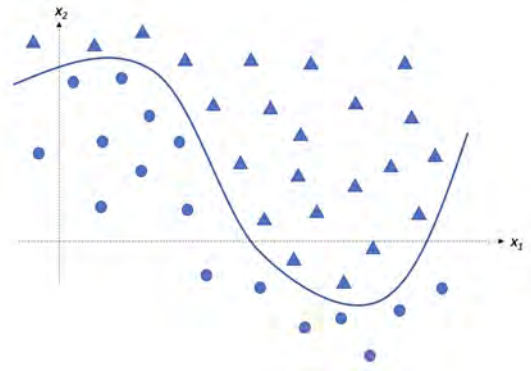
$$y = \sigma\left(\sum_{i=1}^{n} w_i x_i - b\right), \tag{1}$$

5. On parle aussi d'*étiquette* pour désigner $\mathbb{y}$.

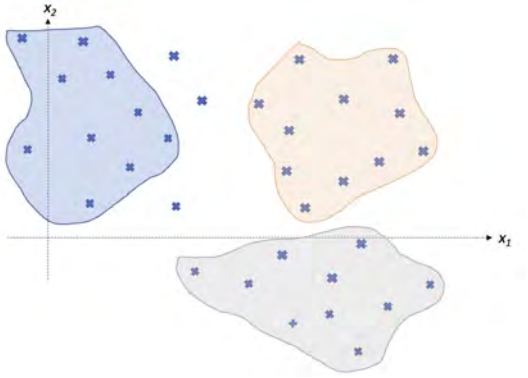
6. Ainsi, l'apprentissage *semi-supervisé*, d'usage courant, est un apprentissage qui dispose de labels pour une partie seulement des données d'entrée.

7. En particulier, on doit pouvoir opérer des simplifications fortes sur l'espace des configurations de l'environnement à explorer (ex. : son cardinal est fini et reste faible) afin d'éviter une explosion combinatoire, et/ou bénéficier d'une puissance de calcul exceptionnelle ; c'est en effet le cas des jeux de stratégie pour lequel ce type d'apprentissage a été mis en valeur.

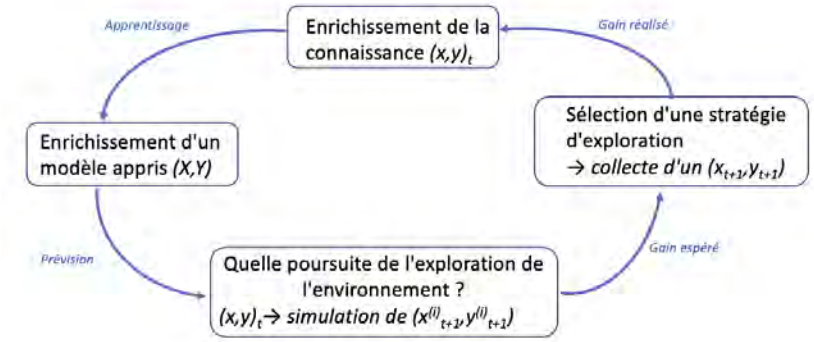
8. Fondées sur la connaissance explicite des gradients du modèle vis-à-vis de ses paramètres, et *rétropropageant* l'erreur entre prévision du modèle et observation des labels $\mathbb{y}$.



(a) *$\mathbb{y}$ est une famille binaire de formes (rond, triangle), le modèle à apprendre à partir d'un échantillon $(\mathbb{x}, \mathbb{y})$ connu est la surface de classification (en ligne pleine).*



(b) *Des regroupements de points géométriquement proches (clustering) sont opérés, qui pourront être étudiés par modélisation statistique. Certains peuvent être difficiles à classifier.*



(c) *L'apprentissage par renforcement est une approche itérative (markovienne) de la construction du lien entre une information $\mathbb{x}$ et un label $\mathbb{y}$, qui se fonde sur une fonction de décision (gain apporté par la connaissance des $\mathbb{y}$, tel que la diminution de l'erreur de modélisation $\mathbb{x} \to \mathbb{y}$). Son optimisation requiert la collecte d'un couple (x, y) le plus informatif possible, à chaque étape de l'approche, et se fonde sur le modèle courant de représentation de l'environnement $(\mathbb{x}, \mathbb{y})$. Dans ce schéma, t est une variable d'itération temporelle.*

FIGURE 1 – *Illustrations d'approches supervisée (a) et non supervisée (b) en classification, pour $\mathbb{x}$ de dimension 2. Illustration du principe d'apprentissage par renforcement (c).*

où la *fonction d'activation* σ est une transformation non linéaire, w_i est un poids associé à l'influence de la donnée x_i et b un biais. La succession de couches, chacune possédant plusieurs neurones traitant parallèlement le signal d'entrée, permet d'offrir des compositions de fonctions sur des partitions de ce signal. Cette grande flexibilité apportée par des fonctions d'activation simples permet théoriquement aux RNN de reproduire tout comportement reliant continûment $\mathbb{x}$ et $\mathbb{y}$ (Cybenko, 1989). En réalité, le choix multi-couches permet d'ôter de nombreuses hypothèses sur le choix de σ pour conserver cette propriété d'approximation universelle (Hornik, 1991; Hanin, 2019; Kidger et Lyons, 2020), cependant sans fournir d'indication forte sur l'architecture (tel un nombre optimisé de neurones par couche).

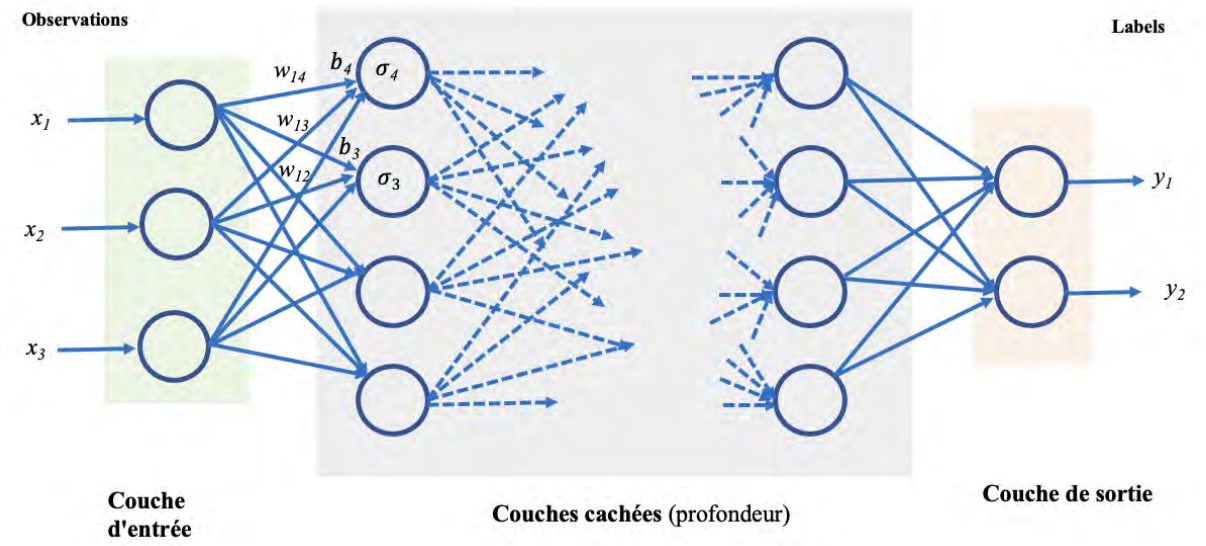


FIGURE 2 – *Illustration d'un réseau de neurones artificiels (RNN). Un RNN est défini par des couches successives, des biais et des fonctions d'activation, transformant un signal multivarié en une sortie (éventuellement) multivariée.* σ *est la fonction d'activation, apportant la non-linéarité à l'équation (1),* ω_{ij} *est le poids à l'observation* x_i *et au neurone* ij*, et* b *est un paramètre de biais.*

Les RNN dont le nombre de couches cachées (profondeur) est de 1 peuvent représenter aisément un grand nombre de transformations de $\mathbb{x}$ vers $\mathbb{y}$, et constituent un prolongement naturel de transformations non linéaires usuelles en statistique, comme la fonction logistique. Toutefois, il est nécessaire d'augmenter cette profondeur (et l'apprentissage est alors dit *profond*) pour apprendre les relations de corrélation (*motifs* ou *patterns*) les plus fins reliant signal d'entrée et signal de sortie, et améliorer la prévision y lorsqu'une nouvelle donnée x se présente. On augmente ce faisant le nombre de paramètres (ω, b) à estimer et celui des *hyperparamètres*[9] à optimiser.

Plus généralement, l'AP se différencie de l'apprentissage machine (AM) classique par le recours à des outils (modèles, algorithmes) construits spécifiquement pour rechercher des corrélations impossibles à déterminer *pratiquement* par des outils de l'AM ; ces outils mettent en jeu plus de deux couches cachées dans des réseaux de neurones, des paramètres et hyperparamètres beaucoup plus nombreux, et requièrent donc des données massives. L'AP se démarque également de l'AM par sa démarche de conception logicielle : il s'agit d'assembler des réseaux de blocs fonctionnels paramétrés en les calibrant à partir de données *via* une certaine forme d'optimisation fondée sur des gradients explicitement connus[10] : c'est une forme de *programmation différentiable* (Innes et al., 2018), à laquelle des langages de programmation spécifiques sont dédiés.

L'objectif des praticiens de l'AP est alors d'établir un compromis entre architecture de modèle et méthode d'optimisation afin de capturer un grand nombre de motifs présents dans les données, faiblement ou non directement accessibles au sens commun : texture et propriétés géométriques des images, structuration sous-jacente d'un texte, etc. De nombreuses méthodes de régularisation permettent de limiter le sur-apprentissage, telle l'augmentation du jeu de données en traitement d'image. Le lecteur trouvera donc dans cette seconde partie de

9. C'est-à-dire les paramètres de réglage des algorithmes d'entraînement.
10. D'après Yann Le Cun, Facebook, 5 janvier 2018.

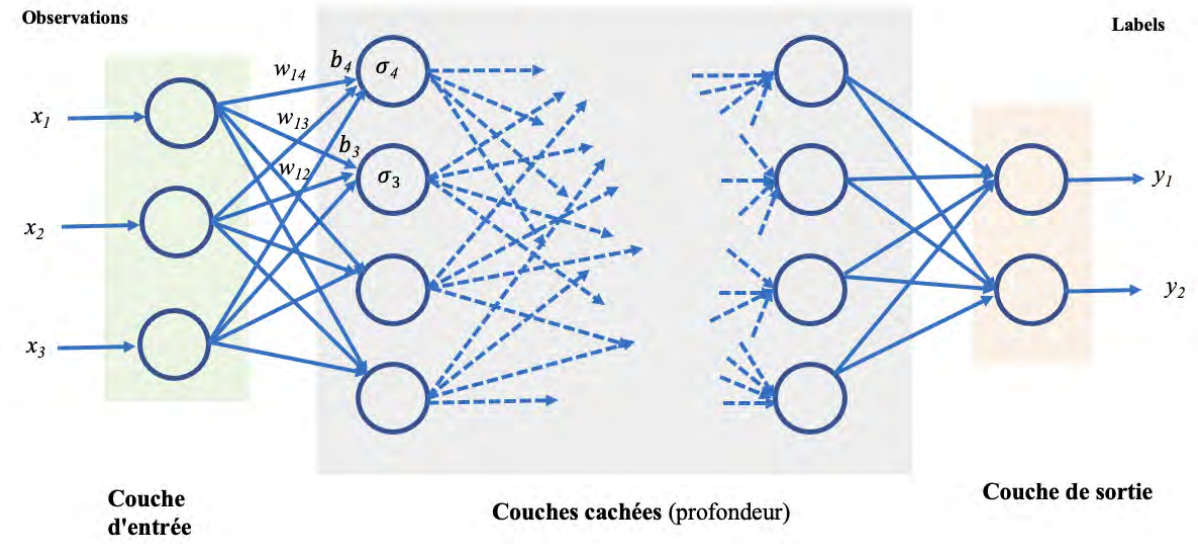


FIGURE 2 – *Illustration d'un réseau de neurones artificiels (RNN). Un RNN est défini par des couches successives, des biais et des fonctions d'activation, transformant un signal multivarié en une sortie (éventuellement) multivariée. σ est la fonction d'activation, apportant la non-linéarité à l'équation (1), ω_{ij} est le poids à l'observation x_i et au neurone ij, et b est un paramètre de biais.*

Les RNN dont le nombre de couches cachées (profondeur) est de 1 peuvent représenter aisément un grand nombre de transformations de $\mathbb{x}$ vers $\mathbb{y}$, et constituent un prolongement naturel de transformations non linéaires usuelles en statistique, comme la fonction logistique. Toutefois, il est nécessaire d'augmenter cette profondeur (et l'apprentissage est alors dit *profond*) pour apprendre les relations de corrélation (*motifs* ou *patterns*) les plus fins reliant signal d'entrée et signal de sortie, et améliorer la prévision y lorsqu'une nouvelle donnée x se présente. On augmente ce faisant le nombre de paramètres (ω, b) à estimer et celui des *hyperparamètres*[9] à optimiser.

Plus généralement, l'AP se différencie de l'apprentissage machine (AM) classique par le recours à des outils (modèles, algorithmes) construits spécifiquement pour rechercher des corrélations impossibles à déterminer *pratiquement* par des outils de l'AM ; ces outils mettent en jeu plus de deux couches cachées dans des réseaux de neurones, des paramètres et hyperparamètres beaucoup plus nombreux, et requièrent donc des données massives. L'AP se démarque également de l'AM par sa démarche de conception logicielle : il s'agit d'assembler des réseaux de blocs fonctionnels paramétrés en les calibrant à partir de données *via* une certaine forme d'optimisation fondée sur des gradients explicitement connus[10] : c'est une forme de *programmation différentiable* (Innes et al., 2018), à laquelle des langages de programmation spécifiques sont dédiés.

L'objectif des praticiens de l'AP est alors d'établir un compromis entre architecture de modèle et méthode d'optimisation afin de capturer un grand nombre de motifs présents dans les données, faiblement ou non directement accessibles au sens commun : texture et propriétés géométriques des images, structuration sous-jacente d'un texte, etc. De nombreuses méthodes de régularisation permettent de limiter le sur-apprentissage, telle l'augmentation du jeu de données en traitement d'image. Le lecteur trouvera donc dans cette seconde partie de

9. C'est-à-dire les paramètres de réglage des algorithmes d'entraînement.
10. D'après Yann Le Cun, Facebook, 5 janvier 2018.

l'ouvrage tous les concepts nécessaires à la conception, à la configuration et à l'optimisation d'un modèle de réseaux de neurones profonds.

La troisième et dernière partie de *L'apprentissage profond* présente enfin des pistes de recherches actives au moment de l'écriture de l'ouvrage – et qui le restent encore aujourd'hui. Différents modèles de réseaux, comme les auto-encodeurs ou les réseaux génératifs par antagonisme (GAN), qui agissent en binôme pour produire des données synthétiques difficilement discernables des données réelles, y sont présentés. Différentes méthodologies y sont détaillées, qui visent à étendre les domaines d'utilisation de l'apprentissage profond, notamment à de tels problèmes de génération de données et plus généralement d'apprentissage non supervisé. Les exemples évoqués sont nombreux, qui vont de la production de musique à l'augmentation de la résolution d'une image.

L'ouvrage est pensé pour un vaste public et abondamment illustré. Surtout, la présentation des principaux concepts se place en permanence dans un contexte *applicatif* : on ne peut parler des objets mathématiques *vecteur*, *matrice*, *tenseur* ou *opération de convolution* sans décrire la réalité physique de leur représentation dans l'espace mémoire ou la mémoire dynamique d'une machine. Différant formellement de l'approche théorique privilégiée par les statisticiens (telle qu'on peut la trouver dans l'ouvrage-phare de Hastie et al. (2001)), elle ne s'exonère jamais des contraintes posées par l'incarnation du calcul. L'exemple de la fonction d'activation $K-$dimensionnelle *softmax*[11]

$$\sigma(\mathbb{z})_j = \frac{\exp(z_j)}{\sum_{k=1}^{K} \exp(z_k)}$$

est révélateur : les auteurs rappellent au § 6.2.2 qu'elle possède autant de raisons pratiques que théoriques d'être utilisée pour approcher le comportement d'un vecteur de probabilité, d'où son usage intensif en classification. L'apprentissage se situe d'ailleurs au croisement de l'informatique, de la statistique, de l'optimisation, de l'analyse numérique et de la géométrie ; dans l'esprit des auteurs, il apparaît donc vain de privilégier un formalisme plutôt qu'un autre, et plus simple et naturel d'aborder le sujet sous l'angle de la mise en œuvre concrète. Ainsi, si pour un statisticien ou un expert en traitement du signal il n'existe pas de différence formelle entre un modèle d'AM et un modèle d'AP, comme indiqué précédemment, le second diffère du premier en ce sens qu'il tente non seulement d'élaborer plusieurs niveaux de représentations de l'information (produire un *vocabulaire*), mais aussi de les articuler entre elles, au moyen d'un nombre colossal de paramètres et par l'usage de briques logicielles différentiables.

Cet ouvrage permet enfin de saisir avec simplicité les difficultés actuelles de compréhension des mécanismes de l'AP – et au-delà, de l'IA dite *connexionniste*, tirant parti de motifs détectés dans les données. Il faut pour cela revenir aux fondations de l'apprentissage statistique.

Fondamentalement, celui-ci est bâti sur la définition puis la minimisation en θ d'une *fonction de coût* $L(\mathbb{x}, \mathbb{y}; \theta)$ entre des données représentant un phénomène d'intérêt Σ et un modèle $\mathcal{M}$ de ces données, paramétré par θ. Le vecteur de paramètres guide la forme, les propriétés essentielles, l'architecture, etc. de $\mathcal{M}$. Si l'on connaissait exhaustivement Σ (par exemple toutes les variétés de données-clients d'un assureur), on pourrait imaginer pouvoir prendre une décision optimale pour chaque nouvelle situation x se présentant (ex. : décider quel contrat d'assurance convient le mieux à tel client). Mais cette connaissance exhaustive étant inatteignable, le concepteur cherche à prendre la moins mauvaise décision possible conditionnellement à la connaissance des données disponibles. La fonction $L(\mathbb{x}, \mathbb{y}; \theta)$ résume donc, pour ce concepteur, les conséquences décisionnelles d'une erreur d'apprentissage liée au choix de $\mathcal{M}(\theta)$, où

11. où $\mathbb{z} = (z_1, \ldots, z_K)$ est typiquement un signal de sortie de l'avant-dernière couche d'un RNN utilisé pour une tâche de classification.

θ est estimé à partir des données disponibles.

Cette fonction de coût incorpore des métriques probabilistes, les données étant considérées comme des représentations de variables aléatoires. Plus généralement, cette fonction de coût dépend de l'objectif d'apprentissage et de la nature des données. L'apprentissage repose donc sur un choix de représentation (modèle), un choix de fonction de coût et une procédure d'optimisation. Nos choix de fonction de coût sont limités par notre capacité à comprendre l'espace des représentations possibles des données, et nous ne pouvons guère avoir de garantie sur l'exhaustivité de notre interprétation. Par ailleurs, les procédures d'optimisation sont définies de façon à combattre le manque de *(quasi-)convexité* globale en θ des fonctions de coût (Greenberg et Pierskalla, 1971). Or cette (quasi-)convexité est nécessaire pour obtenir un *optimum* global, et donc la garantie d'un apprentissage meilleur que tout autre fondé sur la même fonction de coût et les mêmes données (Figure 3).

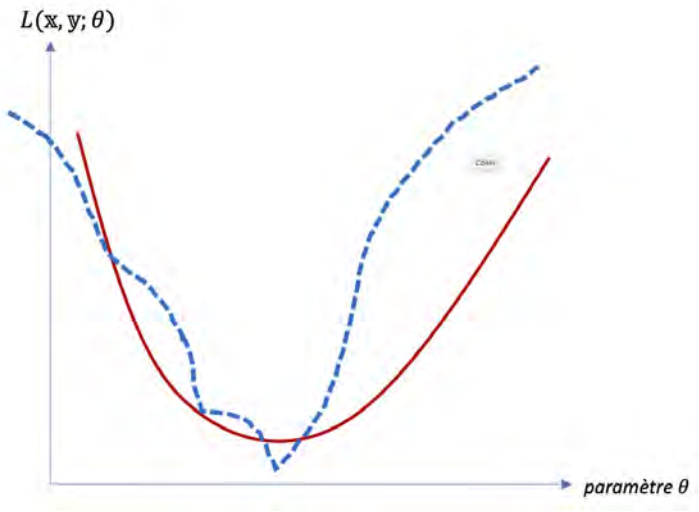


FIGURE 3 – *Pour un vecteur de paramètres θ réduit à la dimension 1, deux exemples de comportement souhaité de fonctions de coût : la recherche d'un minimum global en θ de ces fonctions est pertinent et réalisable par des algorithmes spécialisés si l'on sait ou l'on apprend qu'elles sont convexes (courbe pleine) ou quasi-convexes (courbe pointillée), cette dernière propriété étant bien plus générale que la convexité. Dans les deux cas, l'existence d'un minimum global (et donc d'une « meilleure » représentation possible étant donné le choix de modèle) est assurée.*

Dans le cas de l'AP, les algorithmes les plus avancés permettent d'obtenir des *optima* locaux, qui se révèlent proches des globaux sur de nombreux exemples bien connus, mais il est encore difficile d'en comprendre la raison ; la forme des fonctions de coût est en effet, en grande dimension, délicate à appréhender. Rajoutons également que nous n'avons encore guère, par ailleurs, de garanties théoriques sur le comportement exploratoire des algorithmes et leur atteinte réelle de ces optima, d'autant plus que nous savons que la plupart des réseaux de neurones ne peuvent être estimés d'une façon unique (ils manquent d'identifiabilité de par leur construction).

Dans son récent ouvrage de témoignage et de vulgarisation, Yann Le Cun (Le Cun, 2019, chap. 4) défend cependant l'idée que d'un point de vue pratique, le problème n'en est guère un. Un bon comportement largement prouvé de façon empirique suffirait en général à légitimer une IA fondée sur de l'AP. Pourtant, on peut présumer que pour des applications de l'IA en gestion de systèmes critiques*, l'une des ambitions majeures des programmes de re-

cherche internationaux actuels[12], les étapes de certification s'appuient aussi sur des garanties théoriques plus robustes.

Remercions donc les auteurs pour exhiber ainsi les limitations théoriques actuelles de ces méthodes *implémentées* ; ils font œuvre de salubrité publique en plaçant des mots clairs sur des incertitudes qui imprègnent encore la puissance de ces outils, et motivent ainsi les développements actuels de la recherche sur les propriétés de généralisation des IA connexionnistes. Quelques années après sa publication originale, en dépit de l'effervescence du domaine scientifique qu'il cherche à couvrir, cet ouvrage reste profondément d'actualité.

3. Analyse critique de trois types de cas d'études

Les principales applications industrielles de l'apprentissage profond portent actuellement sur l'analyse d'images, l'exploration de séries temporelles – et en particulier les séries de données produites par des capteurs – ainsi que le traitement automatisé du langage naturel. Dans cette section, ces cas d'usage sont analysés en considérant deux cas de figure distincts.

- Dans le cadre d'une **preuve de concept**, une forte contrainte de temps couplée à une contrainte sur la puissance du matériel à disposition s'exercent sur les ingénieurs, de façon récurrente, afin de produire des résultats. Le temps nécessaire à l'entraînement des réseaux de neurones profonds peut se révéler particulièrement problématique, et les arbitrages en faveur de modèles moins gourmands en ressources sont fréquents.

- Dans le cadre du **développement d'une solution logicielle** possédant une brique analytique, la réalisation est itérative ; le développement d'une application fonctionnelle et intuitive pour les utilisateurs est d'abord privilégié, sans forcément inclure des algorithmes complexes. Une seconde étape consiste en l'ajout de fontionnalités analytiques simples, permettant de fournir une première version de noyau d'intelligence. L'ajout d'une brique d'apprentissage profond est enfin étudié en fonction du cas d'usage et de la nécessité de raffiner les fonctionnalités des modèles.

En Annexe B, des typologies de cas d'usage, issus de la littérature ou que nous avons étudiés ces dernières années, sont détaillées afin de mieux illustrer les bénéfices et limitations des outils décrits dans les sous-sections suivantes. Par ailleurs, ces dernières comprennent des paragraphes techniques, comprenant de nombreuses références utiles et qui peuvent être sautés en première lecture, et des paragraphes plus généraux, incorporant conclusions et recommandations pratiques.

3.1. Détection et classification de situations dans des images

3.1.1. Cas d'étude, outils et apports de l'apprentissage profond

L'analyse d'images est certainement le domaine d'ingénierie qui a connu, grâce à l'apprentissage profond, la progression la plus spectaculaire depuis les années 1990 et les premières lectures automatiques de chèque (Jayadevan et al., 2011). Elle est très majoritairement fondée sur une classe particulière de RNN, les *réseaux de neurones convolutifs* (RNC, ou CNN en anglais), qui s'inspirent du comportement de captation de la structure d'une image par l'œil

12. Tels les projets *Explainable AI* (https://www.darpa.mil/program/explainable-artificial-intelligence), DEEL (www.deel.ai) ou *Partnership on AI* (https://www.partnershiponai.org).

en réduisant le signal d'entrée par une opération nommée convolution ; voir Le Cun et Bengio (1995) pour plus de détail, et la Figure 4.

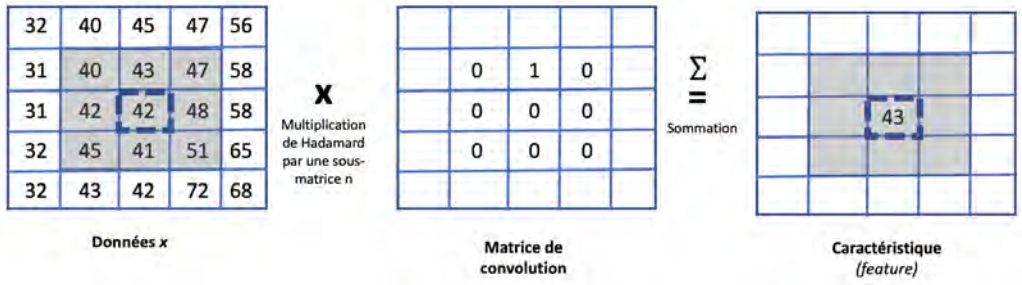


FIGURE 4 – *RNC prend en entrée une image x incarnée par une matrice de pixels auxquels sont associées des valeurs numériques entières codant des couleurs. Il opère un filtrage de sous-matrices de x (de dimension 3 x 3 dans cet exemple) en les multipliant (produit de Hadamard) par une matrice de poids ω (ici tous nuls sauf 1), puis en opérant une opération de convolution, qui revient à sommer sur ces poids. Le résultat de ce filtrage est l'activation d'un neurone portant une information résumant la variation des valeurs des pixels de la sous-matrice (dite* méga-pixel*) ; cette information résumée est appelée* caractéristique *ou* feature*.*

Les industries agroalimentaire (Liu et al., 2018), aéronautique (Kim et al., 2016), automobile (Li et al., 2018) et militaire (Chen et Wang, 2014), la santé (Litjens et al., 2017) et la sécurité (Akçay et al., 2016), le secteur bancaire, les administrations publiques (Anastasopoulos et Whitford, 2018) et la gestion environnementale en font un usage de plus en plus intensif. La *labellisation*, c'est-à-dire l'étiquetage des caractéristiques d'une image en vue de la classification d'un exemple type, puis la *classification* des images elles-mêmes en constituent les traitements les plus courants. Elles sont aussi bien utilisées pour tester la présence de défauts de production sur des produits en fin de chaîne industrielle (Wang et al., 2018) que les variations du couvert végétal, synonymes de déforestation sauvage (Zhao et al., 2017; Rakshit et al., 2018). La mise à disposition de banques d'images labellisées, telles qu'`ImageNet`, a permis de développer considérablement la classification supervisée. Toutefois, la grande variété de ces images limite la précision du pouvoir de classifieurs ainsi construits lorsqu'ils sont appliqués sur des images fortement contextualisées, et l'accroissement de la base de données s'avère généralement indispensable pour atteindre des performances proches de celles de l'humain. L'amélioration des techniques de reconnaissance optique de caractères (OCR) a ainsi fortement bénéficié des tests de sécurité CAPTCHA exécutés par les utilisateurs d'Internet depuis les années 2000, dont le résultat permet de labelliser des images de symboles (von Ahn et al., 2008).

L'AP appliqué aux images consiste surtout à localiser (segmenter) et analyser des objets au sein de ces images, généralement par le biais de RNC. L'utilisation des banques d'images permet de proposer aux utilisateurs des réseaux pré-entraînés[13], qu'il convient de réajuster (ou réentraîner) sur une banque d'images spécifiques au problème considéré. Certaines couches de ces réseaux permettent de capturer des particularités de l'image, telles un style de peinture sur la photo d'un tableau ; associé à des étapes de positionnement des objets, le réentraînement peut ainsi permettre de produire des algorithmes capables de reconnaître la proximité de style entre deux tableaux (Lesaffre, 2018). Au fur et à mesure du temps, les problèmes d'ingestion et de conservation des images et du traitement du temps réel ont conduit les chercheurs à

13. C'est-à-dire des R(C)NN qui ont déjà fait l'objet d'un premier entraînement sur des données présentant des ressemblances structurelles avec celles qui font l'objet de l'étude, et dont certains paramètres / hyperparamètres vont être conservés ; le second entraînement ne concerne donc qu'une partie des paramètres, ce qui permet d'utiliser des échantillons labellisés plus petits. Cette tâche de ré-entraînement est un exemple d'apprentissage par transfert*.

proposer des approches d'exploration rapides, telles Faster R-CNN (Ren et al., 2015) ou YOLO (Redmon et Farhadi, 2017), ainsi que la construction de banques de *masques*, permettant d'appliquer à des fractions d'image des algorithmes de reconnaissance et de séparation de motifs possiblement superposés (He et al., 2017). *Via* l'utilisation d'un dictionnaire de formes, il est par exemple devenu possible de détecter et dimensionner des toitures de bâtiments à partir de prises de vues aériennes, afin de calculer la surface potentiellement utilisable pour l'installation de panneaux solaires, ou encore de reconnaître des cellules d'un type particulier dans des amas cellulaires à partir d'images médicales (Ghouzam et Valverde, 2018).

En détectant et reproduisant les liens complexes entre les pixels d'une image à l'aide d'une hiérarchie de concepts, allant du pixel individuel aux motifs et aux formes, mais aussi en permettant de traiter des types d'image variés (2D, 3D (Maturana et Scherer, 2015), niveaux de gris, infrarouge (Gundogdu et al., 2016), etc.), les réseaux de neurones profonds ont ainsi permis une progression fulgurante de la performance des modèles d'analyse d'images depuis une vingtaine d'années. Depuis 2012, les vainqueurs des plus grandes compétitions mondiales de reconnaissance d'objet, telles que l'`ImageNet Large Scale Visual Recognition Challenge` (ILSVRC), utilisent systématiquement des réseaux de neurones profonds.

Les dernières années ont vu émerger enfin des possibilités d'applications très spécifiques, qui restent encore, pour la majeure partie d'entre elles, au stade de prototype avancé et relativement peu généralisable : compression d'image (Valenzise et al., 2018), cryptage (Dowlin et al., 2016), extraction et transfert de style (Gatys et al., 2016), colorisation (Zhang et al., 2016) et production de nouvelles images *via* des réseaux profonds dits *génératifs* faisant usage de mécanismes antagonistes (GAN) (Huang et al., 2018). Ces derniers, à l'origine des fameux *deepfakes*, sont aujourd'hui surtout considérés comme des aides prometteuses à la conception (par exemple de biens de consommation (Deverall et al., 2017), d'architectures (As et al., 2018) ou de composants industriels (Oh et al., 2019)), et des outils potentiellement intéressants pour produire des données anonymisées (Huang et al., 2018) – deux domaines en plein essor économique.

3.1.2. Limitations et approches alternatives

L'utilisation massive des RNC adaptés au traitement d'images est limitée en pratique par la disponibilité d'outils pré-entraînés, issus de grands laboratoires publics ou privés, et de banques d'images spécifiques. Ainsi, l'art des praticiens est de sélectionner des architectures de réseau par rapprochement avec des données et des contraintes opérationnelles connues, de conserver certains paramètres-clés issus des premiers entraînements (apprentissage par transfert*) et au contraire d'en réestimer d'autres.

Dès lors qu'il s'agit de travailler sur des données fortement structurées (ex. : IRM du cerveau, imagerie satellite à haute résolution) et/ou à caractère personnel (ex. : photographies d'identité, imagerie de santé) ou encore d'usage restreint par le secret industriel (ex. : images de drone survolant des installations), il devient difficile de se procurer en un temps raisonnable des jeux labellisés de grande taille. Si la recherche en labellisation rapide des éléments présents dans les images (*annotation*) – et non plus la simple classification de l'image – connaît un engouement important (He et al., 2018), leur disponibilité reste le problème principal des industriels et des institutions. Les limitations des jeux de données peuvent en effet aboutir à des biais importants, voire dangereux : des chercheurs ont récemment montré que des algorithmes de reconnaissance faciale par AP proposés par Microsoft ou IBM ont été entraînés avec des images de diversité trop faible et aboutissaient à des biais de genre et de couleur

de peau (Buolamwini et Gebru, 2018)[14]. Google a récemment rencontré d'autres difficultés en imagerie médicale, du fait d'un écart entre données d'entraînement d'excellente qualité et données de routine clinique (Heaven, 2020).

Dans un cadre semi-supervisé (voir § 2) ou non supervisé, actuellement plus réaliste pour ce type de données, on ne peut oublier des méthodes statistiques plus anciennes, visant à segmenter les images par le biais d'une modélisation par mélanges de l'information « cachée » (c'est-à-dire de la structure des objets que l'on recherche) (Aas et al., 2007). Un exemple typique d'information cachée est la finesse de séparation des objets au niveau pixel : est-elle nette, ou au contraire doit-on faire l'hypothèse d'un voisinage diffus ? Typiquement fondés sur une hypothèse de Markov et estimés à partir de techniques d'augmentation de données (Celeux et al., 2003), ces modèles statistiques ont notamment été appliqués avec succès à de nombreux problèmes en imagerie de santé (Féron et Mohammad-Djafari, 2005). Par ailleurs, des approches statistiques supervisées telles que les méthodes (ou machines) à noyaux graphiques (Harchaoui et Bach, 2007), qui souffrent certes du fléau de la dimension*, se révèlent cependant précieuses sur des images de faible dimension et requièrent moins d'images pour l'estimation ; elles mériteraient ainsi d'être employées en coopération avec des outils d'apprentissage profond pour approfondir la qualité d'une représentation particulière. Leur industrialisation au travers de langages informatiques de haut niveau (Python, Scala, etc.) reste cependant à étendre en regard des enjeux liés à l'automatisation croissante du traitement de l'information et d'accélération de diagnostic.

Préalablement au travail de segmentation, de nombreuses méthodes traditionnelles (voir Shapiro et Stockman (2003) pour une revue), aisément utilisables, permettent de produire des caractéristiques (*featuring*) à partir desquelles des modèles simples peuvent répondre à de nombreux besoins des industriels.

- L'*histogramme des couleurs* apporte en général assez d'information pour construire rapidement des modèles performants détectant des images obstruées par des nuages, ou la présence de tout autre élément ayant des signatures de couleurs spécifiques au sein des images.

- La *cascade de Haar* (Viola et Jones, 2001) permet par exemple de détecter des personnes et des visages en se fondant sur des caractéristiques construites en calculant des différences entre les sommes des valeurs des pixels de plusieurs zones. Faciles et rapides à calculer, elles peuvent être utilisées pour des cas d'usage de détection sur des flux vidéos, même sous contrainte de temps réel.

- L'*histogramme de gradient orienté* (Dalal et Triggs, 2005) génère des caractéristiques permettant aussi d'entraîner des modèles aux résultats intéressants sur la détection de formes. Dans ce cas, on construit des histogrammes d'orientation des gradients au sein de fenêtres de quelques dizaines de pixels.

Ces techniques ont été utilisées pour résoudre plusieurs cas concrets, portant notamment sur la labellisation multi-classes d'images satellites. Les performances d'un réseau pré-entraîné avec réentraînement des couches supérieures et celles de forêts aléatoires calibrées sur l'histogramme des couleurs ont été comparées. Ce dernier modèle a permis d'obtenir de bons résultats[15] très rapidement, mais de façon hétérogène selon le label, en fonction de la capacité du modèle à trouver une signature de couleurs propre au label. L'utilisation du réseau

14. Ajoutant à cela les difficultés éthiques posées par l'usage de ces outils, en l'absence de législation claire, ces entreprises ont décidé en 2020 de stopper leur activité de recherche en la matière.

15. Score F2 de 0.86 ; pour des précisions sur la définition et l'interprétation des scores de type $F\beta$ usuels en classification binaire, voir par exemple Powers (2011).

pré-entraîné permettait d'atteindre de meilleures performances[16], mais avec un coût en termes de développement et de temps d'entraînement bien supérieur : quelques minutes en regard de plusieurs jours. Il faut noter par ailleurs que l'entraînement d'un modèle non pré-entraîné durant un temps similaire ne permettait pas d'atteindre de semblables performances.

3.1.3. Recommandations pratiques et conclusions

Développer ou non une solution utilisant des réseaux de neurones profonds, dans un cadre supervisé, repose donc majoritairement sur les deux axes suivants.

- **La facilité à construire des caractéristiques explicatives**. Pour des problématiques simples de détection d'objets ou de classification d'images, des histogrammes des couleurs suffisent généralement pour entraîner un modèle simple d'apprentissage atteignant de bonnes performances.

- **La spécificité du problème**. Supposons que le problème et le format d'images soient communs, telle la classification d'images standard représentant des voitures ou des vélos. De nombreux modèles pré-entraînés très performants, de réutilisation rapide, sont déjà disponibles, et la phase très coûteuse de création d'un jeu de données n'est en général pas nécessaire. Si les images disponibles sont communes mais que le modèle est peu spécifique, et nécessite par exemple l'ajout de nouvelles classes d'objets, il est alors possible d'obtenir assez rapidement des résultats corrects à l'aide d'un apprentissage spécifique des dernières couches des modèles d'apprentissage profond pré-entraînés. Le gain de cet apprentissage par transfert résulte de la réutilisation de poids optimisés pour une tâche proche (Sharif Razavian et al., 2014). Il importe cependant de bien comprendre l'architecture du réseau, notamment en exhibant des cartes de caractéristiques* par couche. En cas d'images très spécifiques – des images microscopiques par exemple – alors l'entraînement complet d'un réseau de neurones est nécessaire et les architectures habituelles ne sont parfois pas adaptées. Des jeux de données d'apprentissage de très grande taille sont indispensables pour ce faire.

Rappelons également que la mise en œuvre de techniques d'apprentissage profond dépend fortement des ressources matérielles disponibles. Hors pré-traitement des images et développement d'un réseau profond, un entraînement complet sur des images de grande taille peut durer plusieurs semaines sur des processeurs standards.

Retenons enfin de ce parcours des cas d'usage que l'AP dédié au traitement automatique des images et des flux d'images est devenu globalement mature, à condition que la diversité des phénomènes qu'elles représentent soit bien délimitée, et que des données labellisées existent en grand nombre. L'absence de biais et la capacité de généralisation des outils d'AP en dépend, et il n'est pas évident de définir une typologie précise de ce qu'est le périmètre d'une image. Une image de chien prise sur un fond neutre n'apporte pas la même information qu'une image du même chien sur fond de verdure, ou de neige[17] Il est donc aisé d'introduire des biais préjudiciables par la sélection des données, tout en pouvant difficilement les contrôler.

Les réseaux de neurones convolutifs sont donc devenus les outils fondamentaux de ce type d'AP dans un cadre supervisé ; ils sont à présent bien compris et très largement utilisés, au-delà même du domaine du traitement d'image (nous les retrouverons notamment en traitement de série temporelle). Ils se spécialisent de plus en plus dans l'apprentissage des structures

16. Score F2 de 0.92.

17. En référence à un exemple célèbre de confusion entre un loup et un husky photographié sur fond neigeux, du fait que les images de loups disponibles dans la base d'entraînement avaient été réalisées dans un paysage enneigé ; voir par exemple Besse et al. (2019).

géométriques internes à ces images, mais deviennent d'autant plus gourmands en données d'entraînement labellisées, coûteuses par nature. Afin de proposer des architectures de réseaux pertinentes, sinon sobres et donc moins boulimiques en données, des outils de statistique classique permettent de produire des pré-traitements très efficaces. Par ailleurs, les approches statistiques markoviennes à état latent, qui permettent de mener des analyses non supervisées ou semi-supervisées, restent largement détachées des schémas algorithmiques d'AP ; il nous semble que leur capacité de labellisation et leur frugalité en termes de données sont encore insuffisamment connues des praticiens de l'AP, qui gagneraient à les étudier afin de les employer en interaction avec leurs propres outils.

3.2. Analyse de signaux temporels

3.2.1. Quelques cas d'usage fondamentaux

Les séries temporelles peuvent être générées par de nombreux processus (Forestier et al., 2017; Jones et Lorenz, 1986; Kegel et al., 2018) et nécessitent une analyse spécifique en raison de propriétés propres, notamment les possibles saisonnalités, auto-corrélations des séries et tendances (Shumway et Stoffer, 2017). La typologie des cas d'usage considérés majoritairement par les entreprises est la suivante.

- **Prévision.** Il est parfois crucial de connaître l'évolution d'une quantité à l'avance, afin de prendre des décisions ou anticiper d'éventuelles difficultés. La prévision de séries temporelles est particulièrement éprouvée dans le secteur de la finance (Sezer et al., 2020) et pour les séries économiques en général (Makridakis et al., 2009), ou bien encore dans le domaine de la vente, pour prévoir la demande ou un volume de vente à un horizon temporel donné (Fildes et al., 2019). L'incertitude autour de la valeur prévue peut être, selon les cas d'usage, aussi importante que la prévision en elle-même (Makridakis et al., 2009), et ce en particulier pour les problèmes à forte asymétrie (cas de la prévision de demande intermittente avec de forts volumes (Seeger et al., 2016)). L'état de l'art en matière de prévision converge aujourd'hui vers des modèles probabilistes, comme en témoigne la compétition de prévision M4 (Makridakis et al., 2018) et les dernières approches intégrant des réseaux de neurones visant à apprendre les paramètres d'une densité de probabilité (Salinas et al., 2019) ou d'une fonction quantile (Gasthaus et al., 2019).

- **Classification.** L'objectif est de réaliser une correspondance entre des segments de séries temporelles et un ensemble de catégories. Il peut s'agir de classifier des segments d'une même série (par exemple, identifier les périodes de sommeil au cours de la nuit d'un individu (Chambon et al., 2018)) ou de plusieurs séries (par exemple, la reconnaissance d'une personne en particulier par un assistant vocal (Längkvist et al., 2014)). La détection et l'identification de formes caractéristiques permettant l'attribution d'une série temporelle à une classe (Schäfer et Leser, 2020) peuvent se révéler, selon les cas d'application, particulièrement critiques – tel le suivi de la série temporelle des battements cardiaques d'un patient. *De facto*, la classification de cette série le plus tôt possible peut permettre un diagnostic plus précoce et une meilleure adaptation du traitement.

- **Segmentation.** On peut également souhaiter segmenter de façon non supervisée des séries temporelles afin d'identifier des groupes homogènes. Un exemple typique de cas d'usage est l'identification de profils clients, par exemple de consommateurs d'électricité pour un fournisseur d'énergie (Benítez et al., 2014). Dans le secteur de la santé, ce genre

de technique peut être utilisé pour classer des profils d'IRM fonctionnelles (Wismüller et al., 1998).

Par ailleurs, deux cas d'usage fréquents peuvent être associés à plusieurs des catégories ci-dessus :

– **Maintenance prévisionnelle.** Après la maintenance corrective, puis préventive, la tendance est aujourd'hui à la maintenance prédictive (ou prévisionnelle). Elle a pour objectif de limiter les coûts en cas de panne, les effets domino au sein d'un réseau ou encore la durée d'interruption d'un service. Ce sujet peut être abordé de deux manières : soit on s'intéresse à l'estimation du temps restant avant une panne (ce qui correspond traditionnellement à de l'analyse de survie (Talamo et al., 2019)), soit on cherche à opérer une classification à $t + x$ (panne / susceptibilité de panne / non-panne), x étant l'horizon temporel souhaité (Jahnke, 2015).

– **Détection d'anomalie.** Au-delà de la prévision d'une panne ou d'une fraude à partir d'un historique de données labellisées (Ferdousi et Maeda, 2006), on peut souhaiter identifier de nouveaux types d'anomalies au fil de l'eau : on parle alors d'apprentissage en ligne (*online*), à l'opposé de l'apprentissage par lot (*batch*), qui permet de détecter plus tôt les anomalies, en traitant les données et en mettant à jour le modèle en continu (Guo et al., 2016). C'est par exemple le cas dans les systèmes anti-fraude, où les typologies de fraude sont changeantes par nature (Seyedhossein et Hashemi, 2010). On les identifie alors comme des déviations par rapport à une norme[18] définie conjointement par le *data scientist* et les experts métiers. Ces anomalies peuvent être ponctuelles, contextuelles ou collectives (Choudhary, 2017).

3.2.2. Outils et apports de l'apprentissage profond

La nature particulière des séries temporelles implique des difficultés spécifiques. Deux points de mesure égaux à t n'engendrent pas forcément la même prédiction à $t + x$, en raison de la prise en compte des effets saisonniers et des tendances. De plus, les séries peuvent être très bruitées (en particulier les séries physiques issues de capteurs (Yao et al., 2017; Martí et al., 2015)) et peuvent être amenées à être appréhendées sous une forme multidimensionnelle (ainsi, des données issues de stations météorologiques et sismiques peuvent être utilisées conjointement pour la prévision climatique (Groves-Kirkby et al., 2006)). La stationnarité de la série temporelle est également un prérequis important pour l'utilisation de certains modèles (Dickey, 2005; Brockwell et Davis, 2016), en particulier les modèles statistiques de type ARMA, ARIMA (Woodward et al., 2017), ou le lissage exponentiel.

Enfin, une difficulté majeure est la nécessité assez récurrente de disposer de variables explicatives du phénomène étudié afin de caractériser ses différents cycles, le plus souvent construites manuellement. La connaissance métier apparaît donc primordiale afin de définir de telles variables *a priori*.

Les RNN profonds permettent de pallier certaines de ces difficultés. Ainsi, ils peuvent relativement aisément **prendre en compte des processus non-linéaires** dans la modélisation des données, ce qui permet de réduire le temps de pré-traitement (Schörgenhumer et al., 2019). Par ailleurs, ils permettent de **diminuer les hypothèses restrictives** sur les données, telle la stationnarité ou encore l'hypothèse des risques proportionnels, permettant là aussi de gagner du temps de pré-traitement (débruitage, suppression de la saisonnalité, etc. (Kusdarwati

18. Ou plus généralement une fonction de coût.

et Handoyo, 2018)). Les principaux RNN profonds qui, à présent, ont fait l'objet de multiples expérimentations réussies sont les suivants :

- Les *réseaux de neurones récurrents* (RNR), et en particulier les réseaux *Long Short Term Memory* (LSTM) (Gers et al., 1999), ont vocation à analyser des séquences et résolvent le problème de la dépendance à long terme inhérente à certaines séries temporelles, dans lesquelles deux points de mesure éloignés conservent une dépendance significative (Hochreiter et Schmidhuber, 1997). Les RNR sont construits sur un principe simple issu de l'étude des systèmes dynamiques : chaque élément de la sortie du réseau est une fonction des éléments précédents de la sortie. Le réseau permet alors d'approximer cette fonction de récurrence, soit indirectement en incorporant des connexions récurrentes entre unités cachées (voir Figure 5), soit directement via des relations de rétroaction entre la sortie et les couches cachées ; fonction que l'on traduirait, dans un cadre statistique classique, comme un opérateur markovien.

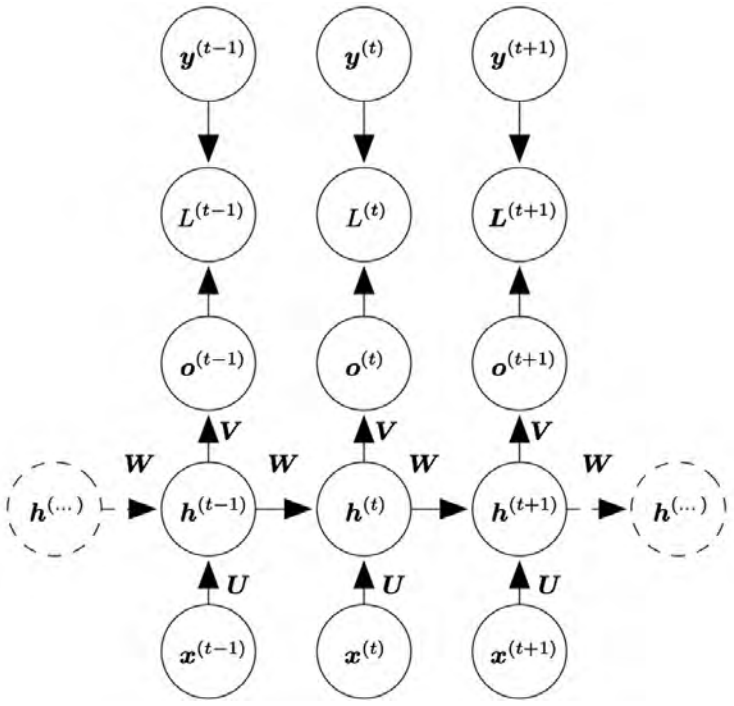


FIGURE 5 – *Illustration d'un réseau récurrent (RNR) à propagation avant présentant le calcul de la fonction de perte d'entraînement d'un réseau récurrent, qui associe une séquence d'entrée de valeurs* $\mathbb{x} = (\dots, x_{t-1}, x_t, x_{t+1}, \dots)$ *à une séquence correspondante de valeurs de sortie* $\mathbb{o} = (\dots, o_{t-1}, o_t, o_{t+1}, \dots)$*. Une perte* **L** *mesure une distance entre* $\mathbb{o}$ *et la cible (label) correspondante* $\mathbb{y} = (\dots, y_{t-1}, y_t, y_{t+1}, \dots)$*. Le RNR comprend des connexions cachées paramétrées par des poids* **U***, des connexions récurrentes entre couches cachées paramétrées par des poids* **W** *et dont l'ensemble des configurations est représenté par* $h(t)$*, et des connexions entre la sortie des couches cachées et les sorties du réseau* $\mathbb{o}$ *par des poids* **V***. Graphe inspiré de la Figure 10.3 de* Goodfellow et al. *(2016).*

Les LSTM octroient une grande liberté de paramétrage, ce qui, actuellement, peut parfois freiner leur utilisation en conditions réelles. Le temps d'entraînement reste également assez élevé par rapport aux méthodes d'apprentissage automatique plus traditionnelles. Une alternative peut être d'utiliser des *Gated Recurrent Units* (GRU) qui possèdent moins de paramètres que les LSTM et sont plus rapides à entraîner. Dans certains cas, tel que la prévision de flux routiers, les performances obtenues via des GRU sont mêmes

meilleures que via des LSTM (Fu et al., 2016).

- Les réseaux convolutifs (RNC) issus de l'analyse d'image se révèlent également intéressants pour la classification (Qian et al., 2020) et la prévision (Borovykh et al., 2017) de séries temporelles. Cette adaptation a notamment pour origine l'idée qu'une série temporelle peut être décrite comme une suite de valeurs qui peut être structurée, possiblement après transformation, sous la forme d'une image (Hatami et al., 2018)[19]. Généralement plus rapides à entraîner que les réseaux récurrents, les RNC ont néanmoins quelques inconvénients qui peuvent rendre difficile leur application aux séries temporelles (ex. : discordance de taille entre l'entrée et la sortie du réseau, entrée du réseau de taille fixe). Ils s'adressent donc à des applications très spécifiques.

- L'*auto-encodeur** (Sun et al., 2016) est également un RNN intéressant que nous avons utilisé dans plusieurs contextes, tel celui de la détection de défauts sur des séries temporelles (Lv et al., 2017). Sa capacité à extraire des variables latentes par reconstruction du signal original a fait ses preuves, ce qui en fait un candidat idéal pour la segmentation et la réduction de dimension, notamment pour les séries temporelles. Il peut être vu dans ce dernier cas comme une généralisation de l'analyse en composantes principales (ACP). L'auto-encodeur possède aussi des applications moins usuelles : il est par exemple possible de l'utiliser en amont d'un modèle, afin de contrôler la cohérence des données à différents moments.

3.2.3. Limitations et approches alternatives

D'une manière générale, l'expérience nous a appris que le déploiement de techniques d'AP, dans le cas spécifique du traitement des séries temporelles, doit être très précautionneux, et qu'en première intention, des approches plus classiques se révèlent robustes et compétitives.

Dans un contexte de **prévision**, les modèles paramétriques (ARMA et ses variantes de la famille Box-Jenkins (Brockwell et Davis, 2016), Holt-Winters (Chatfield, 1978) ou les modèles de Markov cachés (MacDonald et Zucchini, 1997)) ont fait leurs preuves depuis longtemps. Des approches non paramétriques, comme le krigeage par processus gaussiens (Espinasse et al., 2011) sont également indiquées. L'objectif est alors de trouver une fonction d'approximation des données par inférence bayésienne. Il y a donc, potentiellement, une infinité de paramètres en jeu.

Les outils de l'AM, par exemple les forêts aléatoires (Dudek, 2015) ou le *gradient boosting* (Taieb et Hyndman, 2014), restent également très compétitifs en pratique. Dans les cas de classification supervisée, la difficulté principale est de construire des variables explicatives (Deb et al., 2017). La transformée de Fourier (Elbir et al., 2018) et la transformée en ondelettes (Emanet, 2009) peuvent être d'une aide précieuse dans ce cas de figure. Lorsque la classification devient un problème de segmentation non supervisée (Zakaria et al., 2012), après avoir sélectionné une mesure de similarité adéquate (par exemple le *Dynamic Time Warping* (Cuturi et Blondel, 2017)), on procède de façon usuelle à une réduction de dimension (Keogh et Pazzani, 2000).

En **détection d'anomalie**, enfin, une technique traditionnellement utilisée est la modélisation par machines à vecteurs de support (SVM, ou *Séparateurs à Vastes Marges*) à une classe (Ma et Perkins, 2003). Dans les cas les plus simples, un modèle fondé sur des règles métier et des critères statistiques simples peut être utilisé (Shipmon et al., 2017) avec des résultats très

19. Ce parallèle prend plus de sens encore lorsqu'on considère une vidéo, qui est une série temporelle d'images.

honorables, et aisément intelligibles : par exemple, les points de mesure dont la distance à la médiane mobile dépassent un certain seuil peuvent être considérés comme des anomalies.

3.2.4. Recommandations pratiques et conclusions

Plusieurs considérations sont à prendre en compte pour s'orienter ou non vers l'apprentissage profond lorsqu'on souhaite résoudre un problème impliquant le traitement de séries temporelles.

Quantité et qualité des données disponibles. Il est conseillé d'avoir à disposition un volume de données conséquent pour tirer profit des réseaux de neurones (Remus et O'Connor, 2001). En cas de quantité faible (< 700 points), krigeage et méthodes à noyau apparaissent encore préférables pour la prédiction. La complexité (en $O(n^3)$) de ces types de méthodes les rendent difficilement utilisables en pratique avec plus de données (Hensman et al., 2013). Des méthodes statistiques simples, telles que ARIMA ou Holt-Winters, sont également indiquées dans un contexte où peu de données sont disponibles (Burger et al., 2001).

En présence de valeurs manquantes et d'un manque d'historique, certaines techniques ont fait leur apparition, telles que la factorisation de matrices et le filtrage collaboratif (Xie et al., 2016). Dans le secteur de la vente de détail, la prévision de vente de nouveaux produits ayant peu d'historique n'est pas une tâche aisée avec des modèles statistiques ou par apprentissage machine classique. De nouvelles approches, permettant l'intégration de ces produits, dits *cold starts*, intègrent des réseaux de neurones (Alexandrov et al., 2020; Salinas et al., 2019), mais elles restent à déterminer.

Caractérisation des données ou du phénomène. Il faut parfois transformer les séries temporelles *via* une méthode de Box-Jenkins (Helfenstein, 1986), certaines hypothèses (de stationnarité par exemple) devant être vérifiées pour permettre l'emploi de modèles comme ARMA. Si la qualité ou la complexité des données rend cette transformation complexe, il peut être plus intéressant d'utiliser un réseau de neurones ; leur intérêt est double : comme indiqué précédemment, ils peuvent permettre d'alléger la tâche souvent fastidieuse de la création de variables explicatives (Le Cun et Bengio, 1995). Cependant, certaines études ont montré que les performances d'un modèle SARIMA peuvent égaler celles d'un réseau de neurones (Camara et al., 2016). Dans ce cas de figure, le modèle le moins complexe (SARIMA) doit être préféré pour respecter le principe de parcimonie.

Horizon de prévision. Pour un horizon de prévision court terme, les modèles paramétriques tels que ARMA donnent des résultats satisfaisants et sont simples à mettre en œuvre. Dans le cas contraire, on favorisera plutôt le LSTM pour sa capacité à utiliser les dépendances de long terme. C'est aussi le cas du krigeage dans une certaine mesure (Haji Ghassemi et M., 2014). En fonction de la multiplicité des horizons temporels à prévoir, le choix d'une méthode intégrant des réseaux de neurones peut être pertinent. En effet, dans une approche directe de prévision multi-horizons, les modèles sont démultipliés avec le nombre d'horizons à prédire (Bontempi et al., 2013). Cela entraîne des coûts d'entraînement et de maintenance trop élevés. Une alternative est alors d'utiliser une approche récursive, la prévision à l'instant précédent *t+1* étant utilisée en entrée pour prédire l'instant *t+2*. Cette approche a pour défaut principal la propagation d'erreur, augmentant avec le nombre d'horizons temporels à prédire. Une hybridation directe-récursive des deux méthodes est possible (Taieb et al., 2012) pour tenter de contourner la limitation de l'approche précédente, en utilisant plusieurs modèles pour chaque horizon temporel et en incorporant la prévision à l'instant *t+1* en entrée du modèle de prévision de l'instant *t+2*. L'approche « séquence à séquence », permise par les réseaux de neurones, permet de conserver un seul modèle ayant pour objectif de prédire plusieurs ho-

rizons temporels en une seule phase d'inférence (Mariet et Kuznetsov, 2019). Ces modèles sont plus lents à entraîner et requièrent plus de données que les modèles précédents, mais proposent une solution élégante pour la prévision multi-horizons.

Vers des méthodes hybrides. Une récente avancée de l'état de l'art, permise en partie par la compétition de prévision M4 (Makridakis et al., 2018), montre que la communauté dirige actuellement ses efforts de recherche vers des méthodes hybrides, tirant parti du meilleur des deux mondes, en combinant approches statistiques et modèles d'apprentissage classique et profond. L'approche gagnante de la compétition M4 (Smyl et al., 2018) combine des réseaux de neurones récurrents et un modèle de Holt-Winters, les paramètres globaux (poids du réseau de neurones) et les paramètres des séries temporelles (composantes initiales de la saisonnalité et coefficients de lissage) étant appris à l'entraînement par descente de gradient. Cette approche a été développée sur la base du constat que les réseaux de neurones, dans un contexte de prévision incluant des séries temporelles aux saisonnalités complexes et hétérogènes, ne capturent pas de manière performante la saisonnalité. Cette méthode, gourmande en données et ressources computationelles, a récemment été implémentée sur GPU (Redd et al., 2019).

D'autres méthodes, fondées sur la décomposition, visent à simplifier les séries temporelles avant de les présenter à une réseau de neurones. Bien que les algorithmes de désaisonnalisation ont été conçus pour atteindre d'autres objectifs que celui d'être un bon pré-traitement pour les réseaux de neurones, cette étape de décomposition apparaît comme bénéfique pour les ensembles de données provenant de sources de données disparates (Bandara et al., 2020). Cette méthodologie combine une série de techniques de décomposition multisaisonnière pour compléter la procédure d'apprentissage des réseaux LSTM.

D'autres méthodes d'hybridation existent, combinant modèles statistiques (ARMA) et modèle d'apprentissage machine (*Gradient Boosting Machine*) (Hochard et Blanche, 2019). Elles montrent une meilleure performance sur les premiers horizons temporels de prévision fournie par le modèle ARMA, et une meilleure performance sur des horizons temporels plus lointains permis par l'algorithme d'apprentissage machine classique. Ce modèle peut s'écrire :

$$y_{pred} = \alpha(t) \cdot y_{pred,ARMA} + (1 - \alpha(t)) \cdot y_{pred,GBM},$$

où $\alpha(t) = \exp(-\frac{t}{\lambda})$, λ étant appris par optimisation. D'autres approches originales sont en plein développement, qui s'appuient notamment sur l'apport d'un réseau de neurones pour estimer (apprendre) les paramètres de modèles ARMA (Callot, 2019).

Le dynamisme observé dans la littérature des quelques dernières années montre que le domaine de la prévision de séries temporelles évolue vite et voit aujourd'hui deux mondes historiquement opposés dans leurs approches converger vers la complémentarité.

3.3. Compréhension du langage naturel

À l'instar du traitement automatique des images, le traitement automatique du langage (TAL) a connu un important changement de paradigme avec l'avènement de l'apprentissage profond (Manning, 2015). Il est considéré comme un vecteur-clé de l'automatisation de bon nombre de processus, les données textuelles constituant de loin la majorité des données produites chaque jour dans le monde (Delen, 2014, chap. 6). En particulier, éventuellement associé à des outils de traduction automatisée de plus en plus puissants (Fan et al., 2020), le TAL se révèle précieux pour établir rapidement des bases de connaissances sur de nouveaux marchés, dialoguer avec des partenaires ou des clients (Zhang, 2017) ; son intérêt économique en fait logiquement aujourd'hui l'une des compétences les plus recherchées par les entreprises de l'économie digitale.

3.3.1. Quelques cas d'usage fondamentaux

Outre des cas d'usage bien connus comme l'élaboration de *chatbots*, qui nécessite deux briques de compréhension du langage (compréhension des messages d'un utilisateur et génération d'une réponse) (Mnasri, 2019), les principales applications du TAL rencontrées ces dernières années ont essentiellement trait à la connaissance client, la santé et l'optimisation de processus opérationnels. Ainsi,

- **Extraction de sujets**. On souhaite faire ressortir les thèmes principaux de courriels de réclamations de clients afin de comprendre les principales sources d'insatisfaction[20].
- **Classification de documents**. On peut classifier l'objet d'un mail en « privé » ou « professionnel », pour respecter la vie privée des employés. On peut également faire de l'analyse de sentiments, comme dans le cadre de campagnes marketing.
- **Extraction d'entités nommées**. Dans de nombreux contextes, il est intéressant d'extraire des informations de manière automatisée. Ainsi, les compétences majeures du CV d'un candidat, ou bien les chiffres clés d'un document financier, sont des entités couramment cherchées.
- **Moteur de recherche**. Il est important de pouvoir identifier rapidement les informations pertinentes dans une base de documents. Par conséquent, les cas d'usage de moteur de recherche dans des documentations textuelles sont assez fréquents, notamment dans l'industrie pharmaceutique, les industries complexes ou encore le droit.

3.3.2. Outils et apports de l'apprentissage profond

Le TAL place le problème de la compréhension d'un énoncé dans un cadre temporel : une phrase est une séquence de mots, lue progressivement. Mais à la différence d'une série temporelle, la lecture d'un énoncé peut être réalisée dans les deux sens, et plusieurs parcours de l'énoncé sont en général nécessaires pour arriver à exprimer des éléments de contexte qui vont permettre d'aider à l'interprétation du message en vue de réaliser les tâches décrites au § 3.3.1. Les progrès en TAL (ou NLP en anglais) s'appuient sur trois piliers principaux, listés ci-dessous.

- Le développement de **modèles spécifiques au langage**. Les modèles récurrents à mémoire de type LSTM (Hochreiter et Schmidhuber, 1997; Cheng et al., 2016) et les mécanismes d'attention* (Vaswani et al., 2017) ont permis de modéliser les relations latentes entre les constituants d'un énoncé et, ces dernières années, de mieux capturer la structure séquentielle d'une phrase que des approches statistiques par chaîne de Markov (Goodfellow et al., 2016). Les états cachés de ces modèles sont utilisés comme des représentations globales d'un énoncé pour résoudre de nombreux cas d'usage, tels la traduction (Figure 6), l'analyse de sentiment ou la classification.
- **L'apprentissage auto-supervisé.** De nombreuses méthodes d'apprentissage de représentations s'appuient uniquement sur des corpus non labellisés. Ces méthodes tirent parti de la structure du texte pour construire des représentations qui sont utilisées *a posteriori* pour des cas d'usages. Ces formes d'apprentissage s'auto-enrichissant, de par la structure temporelle des données, diffèrent ainsi des traditionnelles approches supervisées d'AP qui supposent l'accès à des données labellisées (Glasmachers, 2017).

20. Voir par exemple https://github.com/MAIF/melusine et De Javel (2019) pour un exemple de package de TAL dédié à cette tâche.

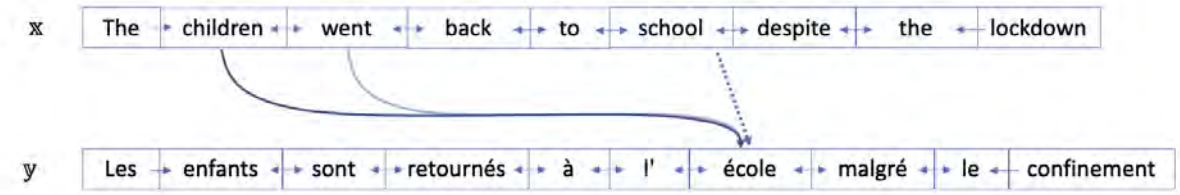


FIGURE 6 – *Illustration de l'utilisation d'un mécanisme d'attention dans une tâche de traduction. Celle-ci peut être réalisée par un auto-encodeur qui comprend deux parties : (a) l'encodage – une réduction de dimension – de l'information initiale* $\mathbb{x} \rightarrow \mathbb{z}$*, puis le décodage de cette information* $\mathbb{z} \rightarrow \mathbb{y}$*. Le vecteur* $\mathbb{z}$*, produit par une première série de transformations de type (1), est nommé représentation latente (ou vecteur de variables latentes). Un décodage terme-à-terme risque d'aboutir à une phrase maladroite voire incompréhensible. Pour éviter cela, pour chaque mot ou séquence de mots de* $\mathbb{x}$ *est construit un mécanisme d'attention permettant de contextualiser la traduction de cette entrée dans la sortie* $\mathbb{y}$*. Ces mécanismes d'attention alimentent donc le décodeur* $\mathbb{z} \rightarrow \mathbb{y}$*. Dans le cas présent, celui-ci utilise l'importance relative des mots « children » et « went » pour aider à générer le mot « école » dans une position adéquate dans la phrase* $\mathbb{y}$*.*

– Les **plongements** ou *embeddings* (Mikolov et al., 2013; Bojanowski et al., 2017), qui sont des représentations sémantiques des mots. Elles sont obtenues en entraînant des algorithmes à prévoir un mot en fonction des mots dans son contexte. Elles offrent de riches propriétés linguistiques et sémantiques, notamment le fait que deux mots de sens proches seront représentés par des vecteurs proches dans l'espace de représentation (Figure 7). Ces méthodes s'étendent à des structures plus complexes et cherchent à composer les représentations des mots pour obtenir des représentations sémantiques des phrases ou des documents. Les modèles de langues (Merity et al., 2018) proposent de prédire le mot suivant en fonction des précédents. Ils permettent de construire des représentations étonnamment robustes et des modèles génératifs extrêmement performants comme par exemple Open GPT ou ELMo (Radford et al., 2019; Peters et al., 2018). Des extensions des méthodes *d'embeddings* cherchent quant à elles à prévoir les phrases suivantes ou précédentes en fonction de la phrase courante pour apprendre des *embeddings* de phrases (Logeswaran et Lee, 2018).

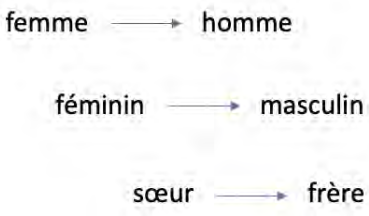


FIGURE 7 – *Les techniques de plongement, comme* word2vec *(Mikolov et al., 2013), proposent des représentations des mots dans un espace de dimension inférieure (à celle de l'espace de tous les mots possibles) dans lequel le sens des mots les rapproche, selon une distance statistique. Ici, la distance est similaire entre des mots différant par le genre.*

– **Le transfert de modèles pré-entraînés**. Finalement, des modèles plus versatiles, les *transformers** comme BERT ou XLNet (Devlin et al., 2019; Yang et al., 2019), sont

également entraînés par le biais de tâches auxiliaires qui s'apparentent à de l'auto-supervision : prévoir des mots masqués ou l'ordre d'apparition de deux phrases. Les *transformers* ont l'avantage d'utiliser simultanément plusieurs mécanismes d'attention en évitant les boucles de réccurence des LSTM, ce qui permet dès lors de paralléliser le traitement de suites de données (Vaswani et al., 2017). Ces modèles entraînés sur de gigantesques corpus peuvent ensuite être utilisés pour d'autres tâches auxiliaires *via* un mécanisme de *transfer learning* (Howard et Ruder, 2018). Leurs performances restent évaluées sur des benchmarks académiques, notamment la tâche GLUE, qui regroupe des ressources très variées, issues de textes de droit, de critiques de films, etc. (Wang et al., 2019).

Dans la majorité des situations, le besoin client implique de **combiner plusieurs algorithmes de compréhension du langage**. Par exemple, pour l'analyse de mails de retours de clients, on pourra commencer par une extraction de sujets non-supervisée. L'analyse de ces sujets pourra nous amener à définir des catégories, et à classifier un nouveau mail dans une de ces catégories (problème de classification supervisée). Les variables explicatives du problème de classification pourront être les plongements des mots (voir paragraphe suivant). On pourra enfin y ajouter la présence ou non d'entités nommées, extraites de manière automatique.

3.3.3. Limitations pratiques, défis et approches alternatives

Si les scores obtenus sur les *benchmarks* académiques dépassent aujourd'hui les performances humaines, des limites importantes subsistent encore, et freinent nettement l'industrialisation de méthodes performantes dans de nombreux contextes métiers.

Bien qu'il soit ainsi possible de construire des représentations fines du texte sans données labellisées, les cas d'usages finaux nécessitent donc d'importantes quantités de données labellisées, coûteuses à obtenir. Par ailleurs, des corpus existent, tel SNLI pour l'inférence (Bowman et al., 2015) ou pour la détection de paraphrases (Marelli et al., 2014), mais ces ressources sont anglophones et il n'existe généralement pas d'alternatives aussi riches pour d'autres langues. L'usage de modèles spécifiques au français, tel le modèle CamemBERT (Martin et al., 2019), débute tout juste.

Par ailleurs, les représentations textuelles sont généralement construites sur de gigantesques corpus dont les biais sont capturés dans les représentations (Bolukbasi et al., 2016) ou reproduits par les algorithmes (Barocas et Selbst, 2016). On observe notamment des biais pour le genre féminin/masculin. En outre, les prévisions peuvent se révéler inconsistantes entre elles (Li et al., 2019) et dans un problème de classification, deux exemples pourtant contradictoires peuvent être classifiés par erreur dans la même catégorie. Les algorithmes s'appuient sur un comportement statistique et ne possèdent pas de sens commun. Dans le cas de la traduction ou du résumé automatique, l'algorithme génère des mots statistiquement probables mais le sens global de la phrase peut être en contradiction avec l'énoncé original (Cao et al., 2018). Ces biais sont d'autant plus difficiles à corriger que les algorithmes sont complexes et peu intelligibles ; il reste encore aujourd'hui difficile d'expliquer une prévision et l'influence du jeu d'apprentissage sur cette prévision.

Enfin, la puissance de calcul colossale nécessaire pour entraîner ou réentraîner des modèles de TAL limite la capacité des acteurs socio-économiques à développer « leur » modèle pour leur domaine d'affaires, et présente un impact environnemental important (Strubell et al., 2020) (voir § 4.1). Certaines techniques moins évoluées, mais plus contrôlables, restent encore largement utilisées en pratique. Citons par exemple :

– l'usage d'*expressions régulières* ou *RegEx*, simples règles logiques sur des chaînes de caractères, qui restent très utiles pour nettoyer des textes et/ou implémenter des règles métier ;

– l'utilisation de la technique TF-IDF (*Term Frequency - Inverse Document Frequency*), utilisée pour la représentation vectorielle de documents (Robertson, 2004) ;

– l'utilisation de la technique LDA (*Latent Dirichlet Allocation*), qui permet l'extraction non-supervisée de sujets (L Griffiths et Steyvers, 2004) ;

– l'emploi de modèles de type HMM (*Hidden Markov Model*) et CRF (*Conditional Random Field*), qui permettent de prendre en compte la séquence des mots et de reconnaître des entités nommées (Sutton et McCallum, 2007).

3.3.4. Recommandations pratiques et conclusions

Le choix de développer ou non une solution utilisant des outils d'AP va être aiguillé selon différents axes, assez similaires à ceux de l'analyse d'images.

Il faut tout d'abord considérer l'*adéquation au besoin métier.* Si l'on considère par exemple le problème de l'extraction non-supervisée de sujets d'un corpus de textes, un LDA répond simplement au besoin. Par ailleurs, il peut être intéressant d'utiliser des règles métier, pour des raisons de simplicité ou de compréhension du modèle. Dans cette situation, de simples Regex peuvent déjà résoudre le problème de façon tout à fait satisfaisante.

Par ailleurs, la *spécificité du problème* joue également un grand rôle dans la décision de s'orienter ou non vers une solution d'AP. Le texte et le besoin métier correspondent-ils à des formats assez usuels ? Lorsque c'est le cas, on peut alors utiliser des modèles existants et déjà entraînés pour de la reconnaissance d'entités nommées.

Dans le cas contraire, les données labellisées sont-elles suffisamment nombreuses pour entraîner un réseau ? Ainsi, un RNR produisant une classification de sentiments aura typiquement un nombre de paramètres à entraîner en

$$O\,(\text{nombre de mots} \times \text{nombre de sentiments})\,.$$

Enfin, les *ressources matérielles disponibles* constituent le dernier élément critique déterminant cette orientation, et la faisabilité de l'AP sur un cas d'étude particulier. Elles s'articulent autour du temps alloué au développement et de la volumétrie des données disponibles.

Dans le cadre de la création d'un chatbot, par exemple, le temps de développement peut être long, surtout si on ne souhaite pas utiliser d'API[21] déjà existantes. La création d'une base d'apprentissage peut également s'avérer longue et fastidieuse dans le cas typique de l'annotation d'un corpus de textes.

Dans l'optique de développer une première solution logicielle en temps contraint, il sera alors préférable d'utiliser des méthodes classiques qui sont moins demandeuses en temps de développement, de collecte de données et d'entraînement.

21. *Application Programming Interface*, ou interface de programmation applicative, qui permet à des applications de communiquer entre elles. Un exemple courant d'API utilisant de l'AP est un traducteur automatique en ligne, comme *www.deepl.com*.