Explicabilité en Intelligence Artificielle ; vers une IA Responsable

Daniel Racoceanu, Mehdi Ounissi, Yannick L. Kergosien

Explicabilité en Intelligence Artificielle ; vers une IA Responsable

Instanciation dans le domaine de la santé

Explainability in Artificial Intelligence; towards Responsible AI

par **Daniel RACOCEANU ***

Professeur des Universités, HDR, PhD, M.Sc., Dipl.Ing.

Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, F-75013, Paris, France

par **Mehdi OUNISSI**

Chercheur, M.Sc.

Sorbonne Université, Sorbonne Center for Artificial Intelligence (SCAI), Institut du Cerveau - Paris Brain Institute - ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, F-75013, Paris, France

par **Yannick L. KERGOSIEN**

Professeur Honoraire des Universités, HDR, MD

Université de Cergy-Pontoise, Cergy, France

Résumé

Essentielle pour une adoption efficace comme pour une utilisation avisée et objective de l'Intelligence Artificielle (IA), l'explicabilité est un véritable verrou de l'évolution de ces technologies, en particulier concernant l'apprentissage automatique et profond. Sans une réelle explicabilité des algorithmes proposés, ces technologies resteront une boîte noire pour les professionnels de santé (et pas seulement), chercheurs, ingénieurs, techniciens - qui assument (et vont continuer à assumer) la pleine responsabilité de leurs actes. De plus en plus, les ingénieurs exploitants et concepteurs d'outils d'IA devront donc faire preuve de responsabilité, en fournissant des algorithmes permettant de garantir l'explicabilité des modèles proposés. Cet article présente les motivations d'une IA explicable, les principales caractéristiques du paysage conceptuel de l'explicabilité en IA, les grandes familles de méthodes pour l'explicabilité - avec un focus sur quelques méthodes parmi les plus courantes, pour finir sur un aperçu des opportunités, challenges et perspectives de ce domaine passionnant de l'interaction homme-machine. En effet, c'est uniquement par une bonne compréhension des challenges associés à cette révolution technologique que nous pourrons la transformer en atout pour nos entreprises ainsi que pour l'ensemble de nos acteurs, partenaires et clients humains.

Abstract

Essential for a good adoption, as well as for a wise and unbiased use, explicability is a real technology lock to the evolution of Artificial Intelligence (AI), in particular concerning Machine and Deep Learning. Without an effective explicability of the proposed algorithms, these techniques will remain a black box for health (and not only) professionals, researchers, engineers and technicians - who assume (and will continue to assume) the full responsibility of their actions. Increasingly, engineers and designers of AI tools will have to demonstrate their responsibility by providing algorithms that guarantee the explicability of the proposed models. This article presents the motivations of an explainable AI, the main characteristics of the conceptual landscape of explainability in AI, the major families of explainability methods - with a focus on some of the most common methods, to finally present some of the opportunities, challenges and perspectives of this exciting field of human-machine interaction. Indeed, only through a good understanding of the challenges associated with this technological revolution that we will be able to transform AI into assets for our companies as well as for our human actors, partners and customers.

Mots-clés

Intelligence Artificielle Explicable, Apprentissage Automatique, Apprentissage Profond, Intelligence Artificielle Responsable

Keywords

Explainable Artificial Intelligence, Machine Learning, Deep Learning, Responsible Artificial Intelligence

***Auteur correspondant :** Daniel Racoceanu, email : daniel.racoceanu@sorbonne-universite.fr

Table des matières

1 Introduction

L'Intelligence Artificielle (IA) moderne connaît un essor sans précédent durant ces dernières décennies. De nombreux domaines applicatifs trouvent ainsi une dynamique nouvelle, grâce à ces technologies révolutionnaires. Cependant, l'adoption de ces techniques se trouve très souvent limitée par le manque d'éléments de traçabilité et de retour d'expérience vis-à-vis des experts. Ceux-ci se sentent donc frustrés de part ce manque de retour, alors que la mise en place-même de l'outil leur demande de fournir un effort considérable de formalisation et de mise à disposition d'une expertise colossale. Certains auteurs parlent donc d'une tendance "boite noire" (black-box évolution), peu souhaitable pour une utilisation traçable, interprétable, explicable, et ultimement, responsable de ces outils.
Le besoin d'explications quant à la manière dont un système intelligent opère est d'autant plus important que les performances du système dépassent – au moins dans un domaine spécialisé – les capacités humaines, et cette question a été abordée dès l'époque des systèmes experts. Les récents systèmes d'apprentissage profond peuvent atteindre des performances étonnantes et leur grand nombre de paramètres rend d'autant plus difficile la compréhension des solutions auxquelles il parviennent, quand bien même ces paramètres sont tous accessibles. Cependant, l'actualité du sujet de l'explicabilité pour les systèmes intelligents vient moins de véritables percées – encore attendues – dans la résolution de ce problème que de la nouveauté juridique – qui s'impose en particulier aux acteurs de l'IA – que constitue l'inclusion dans le Règlement Général sur la Protection des Données (RGPD, règlement européen) d'obligations d'explications pour le traitement automatique de données personnelles. Nous adoptons donc une démarche de technologique traditionnelle à un domaine très mobile et proposons, à côté d'exemples, un cadre conceptuel guidant l'approche du praticien dans la recherche de solutions.

2 Motivations d'une intelligence artificielle explicable

Durant la dernière décennie, du point de vue conceptuel, une tendance évidente se fait remarquer dans la littérature liée à l'intelligence artificielle. Il est ainsi intéressant de noter le besoin latent de modèles d'IA interprétables au fil du temps (ce qui est conforme à l'intuition, car l'interprétabilité est

une exigence dans de nombreux domaines). Ce n'est par contre qu'à partir de 2017-2018 que l'intérêt pour les techniques d'explication des modèles d'IA s'est répandu dans la communauté scientifique et R&D (Fig. 1)

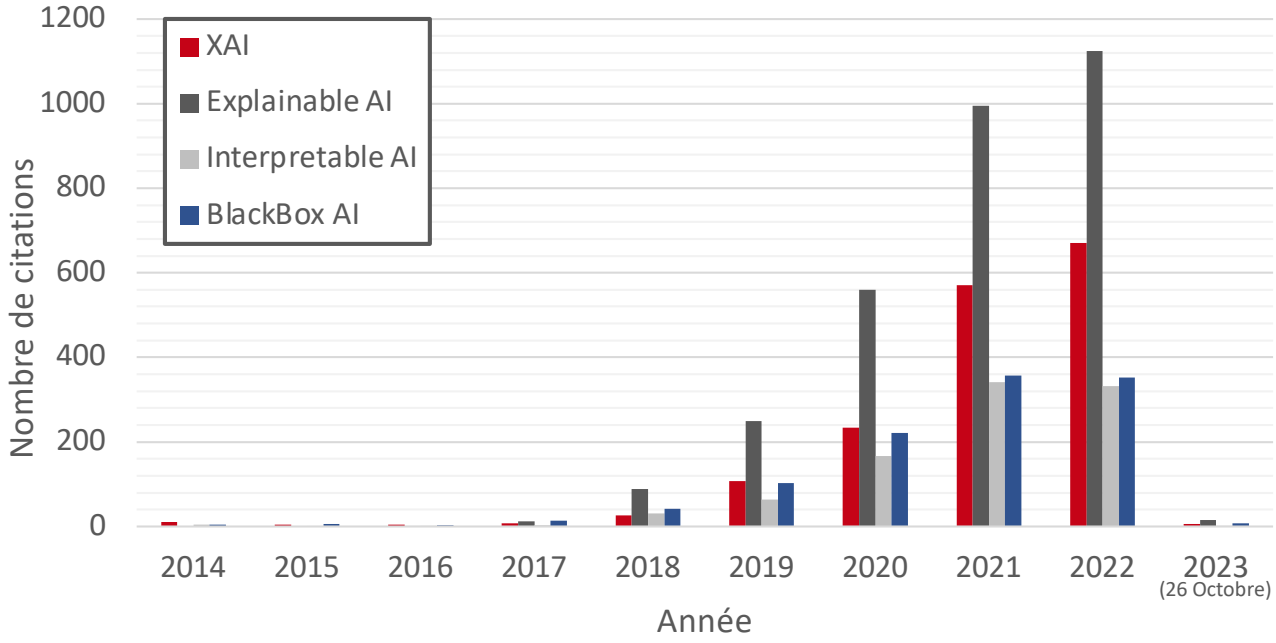

FIGURE 1 – Évolution du nombre de publications dont le titre, le résumé et/ou les mots-clés font référence au domaine de l'intelligence artificielle explicable (XAI) au cours des dernières années. Données extraites de Scopus® (26 octobre 2022) en utilisant les termes de recherche indiqués dans la légende lors de l'interrogation de cette base de données.

2.1 Définition de l'Intelligence Artificielle eXplicable (XAI)

L'intelligence artificielle explicable (en anglais : "eXplainable Artificial Intelligence" - XAI) est un ensemble de méthodes et de processus permettant aux utilisateurs de comprendre les hypothèses, le principe de fonctionnement, les résultats et les conclusions générés par les algorithmes d'apprentissage automatique. L'objectif est de faciliter ainsi la prise de conscience et la cristallisation de la confiance opérationnelle et à long terme dans ces technologies. L'IA explicable est capable de décrire le modèle d'IA utilisé, l'impact attendu, ainsi que les biais potentiels. Elle aide à caractériser l'exactitude, la transparence, l'équité, l'éthique et les résultats des modèles dans la prise de décision assistée par l'IA. Pour une entreprise ou une organisation, l'explicabilité de l'IA est essentielle pour être en mesure d'instaurer la confiance lors de la mise en production ainsi que la maintenance des modèles d'IA. Enfin, l'explicabilité de l'IA aide également l'entreprise / l'organisation à adopter une approche d'IA responsable.

2.2 Intérêt des entrepreneurs pour l'intelligence artificielle explicable

2.2.1 Bâtir la confiance en IA à travers l'explicabilité

Il est essentiel qu'une organisation comprenne parfaitement les processus décisionnels de l'IA, avec une surveillance des modèles et une responsabilisation de l'IA, afin de ne pas lui accorder une confiance aveugle, tout en tirant un maximum d'avantages (de plus en plus stratégiques, de nos jours) de ces remarquables technologies. L'IA explicable peut aider les humains à comprendre, assimiler, valider, exploiter et partager les algorithmes d'apprentissage automatique.

Comme nous l'avons précisé précédemment, les modèles d'apprentissage automatique sont souvent considérés – par le commun des utilisateurs – comme étant difficiles (voire impossible) à interpréter. A titre d'exemple, les réseaux neuronaux utilisés dans l'apprentissage en profondeur sont, en effet, parmi les plus difficiles à comprendre pour un humain. Les biais potentiels constituent un risque pour l'entraînement des modèles d'IA. En outre, les performances des modèles d'IA peuvent dériver ou se dégrader dans le temps, dû aux données de production, souvent (très) différentes des données d'entraînement. Il est donc crucial pour une entreprise de surveiller et de gérer en permanence les modèles, afin de maintenir et faire évoluer l'explicabilité de l'IA, tout en mesurant l'impact de l'utilisation de ces algorithmes, sur l'entreprise. L'IA explicable contribue également à stimuler la confiance des utilisateurs finaux, l'auditabilité des modèles et l'utilisation efficace et effective de l'IA. Elle atténue également les risques de non-conformité, ainsi que liés à la légalité, la sécurité et la réputation induites par l'IA de production.

2.2.2 Ouvrir la voie vers l'IA Responsable (RAI)

L'IA explicable est l'une des conditions essentielles à la mise en œuvre d'une IA responsable (en anglais "Responsible Artificial Intelligence" - RAI). Elle permet le déploiement efficace, à grande échelle, des méthodes d'IA dans des organisations réelles en garantissant l'équité, l'explicabilité des modèles et la responsabilité des suggestions guidées par l'IA. Pour aider à adopter l'IA de manière responsable, les entreprises doivent donc intégrer (au sein de l'entreprise ainsi que dans son écosystème) des principes éthiques dans les applications et les processus d'IA, en créant des systèmes d'IA basés sur la confiance et la transparence. Ceci est donc à même de faire évoluer les mentalités non seulement parmi ses propres employés mais aussi - par induction - parmi ses partenaires, collaborateurs et aussi parmi ses clients.

2.2.3 Objectifs de l'IA Explicable (XAI)

Les objectif de l'IA explicable peuvent donc être regroupés autour de quelques centres d'intérêt clefs :

- Exploitation de l'IA en toute confiance : L'IA sera de plus en plus utilisée par des non-experts du domaine informatique. Pour ce faire, les modèles doivent donner la possibilité de générer une confiance lors de l'utilisation et l'interprétation des résultats proposés, afin de conforter la décision finale engagée par l'utilisateur, vis-à-vis de son environnement professionnel.
- Accélération des résultats de l'IA : Les utilisateurs et les concepteurs des systèmes d'IA doivent avoir la possibilité d'agir sur des leviers algorithmiques et conceptuels (paramétrisation) permettant une accélération de l'obtention des résultats élaborés par l'IA, en fonction du caractère temps-réel dur, ferme ou doux de la problématique abordée [1].
- Évaluation continue des modèles : Avec l'IA explicable, une entreprise peut maintenir et améliorer les performances des modèles, tout en aidant les parties prenantes à comprendre les comportements des modèles d'IA. L'étude des comportements des modèles par le suivi de leur état de déploiement, de leur équité, de leur qualité et de leurs dérives est essentielle à la mise à l'échelle de l'IA. L'évaluation continue des modèles permet à une entreprise de comparer les prévisions des modèles, de quantifier les risques des modèles et d'en optimiser les performances.

  Un exemple intéressant est celui de la société BlackRock : leurs modèles (i.e. Aladdin [2]) analysent continuellement les données des réseaux sociaux pour aiguiller les investissements. La portée de ces modèles peut donc s'avérer très importantes, avec des conséquences financières, industrielles, sociétales et sociales considérables, sur l'ensemble d'un écosystème.

1. Temps réel, temps souple et temps réel ferme
2. Aladdin

2.2.4 Chantiers à venir en XAI : concepts, métriques, feuille de route

Vu le rôle croissant de l'IA dans notre vie professionnelle et personnelle, il devient urgent et indispensable de bénéficier d'une explication / interprétation des résultats et opinions générés par celle-ci. La littérature du domaine fait apparaître, clairement, la nécessité d'un concept unifié de l'explicabilité des modèles d'IA. Même si générique, ce type de conceptualisation doit, en même temps, traduire les besoins exprimés par les utilisateurs, sur le terrain. Par ailleurs, cet effort conceptuel doit proposer une feuille de route réaliste, consensuelle, commune à tous les système d'IA XAI.
Afin de permettre la synthèse d'une opinion éclairée concernant l'IA explicable, nous fixons les concepts qui nous semblent essentiels pour une bonne compréhension de ce domaine émergeant [1].
Pour l'évolution du domaine, il est impératif de mettre en place une feuille de route à partir de laquelle la communauté serait en mesure de contribuer par de nouvelles techniques et méthodes.
Une autre caractéristique clé nécessaire pour relier un certain modèle à ce concept concret est l'existence d'une métrique (ou groupe de métriques) permettant une comparaison significative de l'adéquation d'un modèle à la définition d'explicable. Sans un tel outil, toute affirmation au sujet de l'explicabilité en IA ne peut fournir une base solide sur laquelle appuyer une stratégie solide.
Ces métriques d'explicabilité, comme les métriques classiques de performances (sensibilité, spécificité, précision, rappel, score F1, caractéristique de fonctionnement du récepteur - ROC, aire sous la courbe - AUC, ...), doivent exprimer les performances du modèle dans un certain aspect de l'explicabilité. Certaines tentatives ont été publiés autour des mesures de XAI, comme présenté dans [2, 3].
En général, les mesures XAI devraient évaluer la qualité, l'utilité et la satisfaction des explications, l'amélioration du modèle mental du public induite par les explications du modèle, et l'impact des explications sur les performances du modèle et sur la confiance et la confiance du public.

2.3 L'explicabilité, nouvelle obligation réglementaire pour l'IA

2.3.1 Craintes inspirées par l'IA, éthique de l'IA, promesses de l'explicabilité

Avec ses résultats étonnants et l'arrivée de ses applications concrètes l'I.A. suscite dans la société des craintes fondées qui justifient que son emploi soit encadré. La plupart des entrepreneurs devraient profiter de ces progrès juridiques qui facilitent l'acceptation de l'IA par le public, et c'est dans cette direction qu'ont été prises des initiatives gouvernementales américaines [4] et française[5] où les aspects éthiques sont mis en avant. Inspirant les nouvelles normes juridiques, la recherche en éthique de l'IA, ainsi qu'en I.A. Ethique (I.A. sous contraintes éthiques), s'est récemment constituée en domaine reconnu avec ses journaux, ses sociétés savantes, ses filières de formation. En plus des mémorandums fondateurs déjà cités on peut mentionner l'initiative IEEE (cf. infra), et les productions des centres de recherche d'Oxford ou Munich, et en France du Comité de Réflexion sur l'Ethique de l'ALLISTENE (Alliance des Sciences et Technologies du Numérique). Ses thèmes concernent principalement les moyens de rendre l'IA compatible en pratique avec les droits humains, spécialement en ce qui concerne l'équité, la sécurité, la confidentialité, la dignité.
Par exemple, comme le développe un rapport de la Banque d'Angleterre[6], l' attribution automatisée de prêts bancaires faisant appel au profilage risque de discriminer certaines populations. Elle doit s'assurer qu'elle évite le recours à des variables fondées sur la couleur de peau, l'appartenance ethnique ou la religion. En France, il est interdit d'enregistrer de tels attributs personnels dans un fichier informatisé (cf. infra), mais – outre que d'autres pays n'ont pas nécessairement de telles restrictions – il est dans une grande mesure possible de faire de tels attributs l'objet d'une reconstruction à partir de données autorisées et donc d'effectuer un profilage interdit. Or le ciblage de tels critères discriminatoires peut être utilisé d'une manière inapparente (et *a fortiori* involontaire) par les algorithmes résultant de procédures d'apprentissage qui n'incluraient pas ces contraintes légales, par exemple pour optimiser un profit sous des seules contraintes économiques. Cet exemple pose déjà un problème de responsabilité des décisions et montre quels sont les acteurs parties possibles d'un éventuel conflit : (1) le concepteur de l'algorithme de décision, qu'il l'ait conçu rationnellement

ou – surtout – qu'il l'ait obtenu par apprentissage (paramétré par lui en utilisant éventuellement une bibliothèque qu'il maîtrise incomplètement), (2) l'utilisateur de l'algorithme, (3) l'individu auquel s'applique la décision et qui peut être lésé, (4) une éventuelle autorité de certification ou d'agrément de l'algorithme. L'analyse de ce cas n'est pas complète sans ajouter (5) les individus à l'origine des données utilisées pour l'apprentissage, si les données n'ont pas suffisamment agrégées pour ne permettre qu'un traitement statistique, et surtout (6) les annotateurs des données, experts "métier" en général distincts des concepteurs, fournisseurs d'une partie essentielle, sémantique, de la valeur de la base de données destinée à l'apprentissage. Ces annotateurs dépendent souvent eux-mêmes (7) d'outils d'annotation et (7bis) du choix d'un lexique d'annotation (élaboré ou importé) qui peut être à l'origine d'observations discriminatoires. A tout ceci peut s'ajouter (8) l'effet d'un apprentissage par transfert (transfer learning, cf. infra) important par le biais d'un pré-apprentisage par le modèle de concepts possiblement discriminatoires. A tout ceci pourrait s'ajouter la couche supplémentaire de complexité de l'apprentissage fédératif (cf. infra). Comme nous le verrons plus loin, la loi veille à ménager à l'individu lésé par une décision automatisée la possibilité d'un recours individuel par un humain qui doit pouvoir lui fournir des *explications* quant à la décision qui le concerne. Mais la loi française inclut aussi dans certains cas une obligation, très nouvelle, d'"*explicabilité*" des algorithmes. En contribuant à l'analyse d'algorithmes obtenus par apprentissage automatisé et à l'analyse d'éventuels résultats inattendus de ces algorithmes l'explicabilité devrait permettre de vérifier certaines des contraintes légales ou de sécurité sur le fonctionnement des algorithmes, d'informer les utilisateurs responsables et de gagner leur confiance. Les applications de l'explicabilité dépassent cependant le pur cadre des obligations réglementaires et intéressent aussi les entrepreneurs de l'I.A. spécialement pour des produits interagissant avec des humains, notamment en langage naturel, comme dans le cas des robots humanoïdes pour l'assistance à des personnes dépendantes, un domaine ou l'éthique et la prise en compte de la dignité des humains constituent un des sujets majeurs de recherche, tant du point de vue juridique qu'informatique.
A ce jour et en France, les obligations légales concernant l'I.A. trouvent principalement leur origine dans les trois textes qui suivent.

2.3.2 Loi du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés, et sa modification du 6 août 2004

Cette loi était motivée par les progrès au sein de la société de l'Informatique appliquée au traitement de données. Elle concerne principalement les fichiers de données à caractère personnel, c'est-à-dire de données concernant des individus identifiables directement ou indirectement, quand ces données font l'objet de fichiers informatisés. Il est à noter que depuis le vote de cette loi les données collectées auprès d'individus se sont considérablement enrichies avec la généralisation des terminaux de paiement, du web et des magasins en ligne, du téléphone mobile et des réseaux sociaux, et que la ré-identification des individus à partir de leurs données dépasse souvent la plupart des mesures d'anonymisation. La loi définit l'autorité de contrôle des données personnelles, la CNIL, les obligations des acteurs constituant des fichiers informatisés (déclaration des fichiers, interdiction d'y faire figurer certaines informations relative à l'appartenance raciale, ethnique, religieuse, politique, syndicale), et les droits des individus dont les données font l'objet d'un fichier informatisé (droit d'être informé, droit d'accès, droit de rectification, droit d'effacement).
La Loi n 2004-801 du 6 août 2004 relative à la protection des personnes physiques à l'égard des traitements de données à caractère personnel modifie le texte précédent en apportant notamment une contrainte sur le traitement automatisé des données individuelles qui s'applique spécialement à l'I.A.. Son article 10 stipule en effet que : "Aucune décision de justice impliquant une appréciation sur le comportement d'une personne ne peut avoir pour fondement un traitement automatisé de données à caractère personnel destiné à évaluer certains aspects de sa personnalité." et que "Aucune autre décision produisant des effets juridiques à l'égard d'une personne ne peut être prise sur le seul fondement d'un traitement automatisé de données destiné à définir le profil de l'intéressé ou à

évaluer certains aspects de sa personnalité."

2.3.3 Règlement général sur la protection des données : RGPD

Le RGPD est un règlement européen. Ceci veut dire qu'au contraire d'une directive, qui doit, pour s'appliquer, faire l'objet d'une transposition dans le droit de chacun des états membres, ce règlement s'applique directement, tel que publié au Journal officiel de l'Union Européenne, et simultanément dans la totalité des états membres dès sa date d'entrée en vigueur. Le RGPD est en partie suscité par les progrès de l'I.A. et dispose que "En tout état de cause, un traitement de ce type devrait être assorti de garanties appropriées, qui devraient comprendre une information spécifique de la personne concernée ainsi que le droit d'obtenir une intervention humaine, d'exprimer son point de vue, d'obtenir une explication quant à la décision prise à l'issue de ce type d'évaluation et de contester la décision." De plus "Afin d'assurer un traitement équitable et transparent à l'égard de la personne concernée, compte tenu des circonstances particulières et du contexte dans lesquels les données à caractère personnel sont traitées, le responsable du traitement devrait utiliser des procédures mathématiques ou statistiques adéquates aux fins du profilage, appliquer les mesures techniques et organisationnelles appropriées pour faire en sorte, en particulier, que les facteurs qui entraînent des erreurs dans les données à caractère personnel soient corrigés et que le risques d'erreur soit réduit au minimum, et sécuriser les données à caractère personnel d'une manière qui tienne compte des risques susceptibles de peser sur les intérêts et les droits de la personne concernée et qui prévienne, entre autres, les effets discriminatoires à l'égard des personnes physiques fondées sur la l'origine raciale ou ethnique, les opinions politiques, la religion ou les convictions, l'appartenance syndicale, le statut génétique ou l'état de santé, ou l'orientation sexuelle, ou qui se traduisent par des mesures produisant un tel effet. La prise de décision et le profilage automatisés fondés sur des catégories particulières de données à caractère personnel ne devraient être autorisés que dans des conditions spécifiques."

2.3.4 Loi n 2021-1017 du 2 août 2021 relative à la bioéthique. Explicabilité

Cette loi, incluse dans le Code de la santé, comporte des dispositions relatives à d'importantes applications médicales de l'I.A. Elle introduit une obligation d'*explicabilité* pour les algorithmes ou systèmes intelligents prenant des décisions qui concernent des individus. Il est à noter que le néologisme "explicabilité" y est employé sans que sa signification ne soit explicitée. Les moyens de parvenir à cette propriété restent abstraits. Précisément, son article 17 introduit les innovations suivantes :

- I. – Le professionnel de santé qui décide d'utiliser, pour un acte de prévention, de diagnostic ou de soin, un dispositif médical comportant un traitement de données algorithmique dont l'apprentissage a été réalisé à partir de données massives s'assure que la personne concernée en a été informée et qu'elle est, le cas échéant, avertie de l'interprétation qui en résulte.
- II. – Les professionnels de santé concernés sont informés du recours à ce traitement de données. Les données du patient utilisées dans ce traitement et les résultats qui en sont issus leur sont accessibles.
- III. – Les concepteurs d'un traitement algorithmique mentionné au I s'assurent de l'explicabilité de son fonctionnement pour les utilisateurs.

Les utilisateurs mentionnés en III sont ici des professionnels de santé ; c'est à eux que s'adressent les explications sur le fonctionnement du traitement algorithmique et ce sont eux qui géreront l'information au patient ainsi que le recueil du consentement.

2.3.5 Autres textes : exemple du problème de la responsabilité des décisions

Un exemple montrera que d'autres textes, même anciens, peuvent s'appliquer à des situations où intervient l'IA. Dans la section précédente on pourrait imaginer que le dispositif intelligent puisse directement gérer l'interaction avec le patient, un peu comme des systèmes intelligents interagissent directement avec les clients de sites de vente en ligne. Ceci cependant constituerait une infraction au Code de la santé au titre d'un exercice illégal de la médecine. Il est pour l'instant exclu que des actes médicaux soient délégués à des machines, et le problème de la responsabilité des concepteurs et fournisseurs des dispositifs dans un tel cas se poserait avec une acuité au moins aussi grande que dans le cas des véhicules autonomes.

2.3.6 Projet de Règlement européen : règles harmonisées concernant l'IA

En plus de textes pouvant s'ajouter au RGPD dans des pays membre de l'U.E., des législations relatives à l'I.A. sont en cours d'élaboration dans l'U.E., aux Etats-Unis et au Royaume Uni, avec d'importants enjeux économiques et stratégiques. A ce propos la Commission européenne a déclaré "Il est donc particulièrement crucial que l'UE prenne les devants pour élaborer de nouvelles normes mondiales ambitieuses".Le projet de règlement européen relatif à l'IA est accessible. Il fait suite au livre blanc de la Commission «Intelligence artificielle – Une approche européenne axée sur l'excellence et la confiance» qui proposait déjà certaines orientations réglementaires. Ce projet a fait l'objet d'intéressants commentaires par le Conseil économique et social européen [7]. Mentionnons aussi la "Résolution du Parlement européen du 20 octobre 2020 contenant des recommandations à la Commission concernant un cadre pour les aspects éthiques de l'intelligence artificielle, de la robotique et des technologies connexes" où il est affirmé que le futur cadre réglementaire "devrait s'appuyer sur le droit et les valeurs de l'Union et être guidé par les principes de transparence et d'explicabilité, d'équité, et de responsabilité", mais sans autre occurrence du mot "explicabilité" dans le texte de 68 pages.

2.3.7 Le mouvement normatif hors UE : la feuille de route des Etats-Unis

Le mouvement mondial de prise de conscience par le politique de l'importance – tant sociétale qu'économique – de l'IA, et des problèmes éthiques et réglementaires que pose son développement, doit certainement beaucoup à l'initiative de l'éxécutif des Etats-Unis en 2016, accompagnée par la publication par la Maison Blanche du livre blanc "Se préparer au futur de l'Intelligence Artificielle" [8]. Il s'agit d'une feuille de route pour le développement de l'IA aux Etats-Unis qui comprend aussi des recommandations aux ministères et agences gouvernementales, qui assureront les financements, parmi lesquelles la mise en place de bases de données ouvertes ou de plateformes de calcul pour l'accès de nouveaux acteurs à l'IA, la généralisation de l'enseignement de l'IA, mais aussi la mise en place de formations à l'éthique en IA et leur inclusion dans l'enseignement de l'IA. Au lieu de légiférer d'emblée en utilisant un critère abstrait comme l'explicabilité, ce pays lance simultanément (en 2016) par son agence la DARPA un appel à projets de recherche sur le thème "IA explicable" [9, 10], et encourage diverses sociétés professionnelles à développer de nouvelles recommandations et standards de bonnes pratiques. Par exemple, toujours en 2016, l'IEEE crée l'"Initiative mondiale pour les considérations éthiques en IA et systèmes autonomes" [11] et publie un appel à commentaires sous la forme du rapport "Conception éthiquement alignée" [12]. Comparé à ses répliques française et européenne, ce mouvement adopte donc une approche ascendante très efficace par la grande implication d'emblée de l'ensemble des professionnels et chercheurs. L'effort produit immédiatement des normes applicables après que les professionnels aient participé à leur élaboration, et c'est sur cette base solide et cette expérience que pourront venir des textes plus généraux au niveau législatif.

2.3.8 Travaux et recommandation de l'UNESCO, directive de l'OCDE

Plusieurs organisations internationales travaillent aux problèmes éthiques posés par l'IA et essayent d'anticiper le mouvement réglementaire mondial. Après avoir produit l'intéressant rapport de sa COMEST (Commission Mondiale d'Ethique des Connaissances Scientifiques et Techniques) [13], l'UNESCO a émis une Recommandation sur l'Ethique de l'intelligence Artificielle [14]. Reprenant en partie l'approche française, ce texte utilise la notion d'explicabiilté pour la recommander, en l'explicitant un peu plus, mais la lie à la transparence, aussi prônée : "L'explicabilité des systèmes d'IA renvoie également à l'intelligibilité des intrants, des extrants, du fonctionnement des différents modules algorithmiques et de leur contribution aux résultats des systèmes. L'explicabilité est donc étroitement liée à la transparence, puisqu'il convient de rendre les résultats et les sous-processus qui y conduisent intelligibles et traçables, en fonction du contexte." De son côté, l'OCDE avait publié en 2019 une directive relative l'IA[15] qui recommande elle aussi transparence et explicabilité.

3 Le paysage conceptuel de l'explicabilité

3.1 Les paradoxes de la transparence

Les systèmes intelligents sont souvent traités de boîtes noires, ce qui est péjoratif tout comme l'opacité qui relève de la même métaphore. La transparence semble être l'antidote à ce défaut, mais elle peut aussi bien signifier l'invisibilité, comme dans "ce changement de solution est transparent pour l'utilisateur" que la visibilité totale, comme celle de l'accès au code source d'un logiciel. Or même les réseaux de neurones profonds dont le fonctionnement semble le plus opaque donnent souvent un accès complet à toutes les caractéristiques (structure et fichier des poids du modèle) qui permettent d'en reproduire les résultats. L'explication de leur fonctionnement nécessite plutôt une véritable construction comme pour la compréhension d'un code source non commenté. Nous n'utiliserons plus le terme de transparence qui nous semble ici à la fois ambigu et irrelevant, mais remarquerons cependant que l'argument du découplage entre transparence et explicabilité a offert un argument au législateur en défaveur d'une obligation de rendre publics les algorithmes d'IA, ménageant ainsi une possibilité de confidentialité dans le cadre d'une politique de propriété industrielle (rapport assemblée nationale).

3.2 Reproductibilité, auditabilité

La reproductibilité d'un algorithme, bien que désirable, n'assure pas son explicabilité, tout comme en Génie Logiciel la reproductibilité d'un bogue ne suffit pas à en déterminer ni comprendre la cause. Si un algorithme déterministe produit toujours le même résultat en partant du même état initial, les causes de non déterminisme en apprentissage automatique sont nombreuses : utilisation d'algorithmes stochastiques d'optimisation (mais encore peut-on fixer les graines des générateurs pseudo-aléatoires pour atteindre la reproductibilité), parallélisme massif des processeurs dont le séquençage n'est pas toujours maîtrisé, base de données d'apprentissage et ordre de présentation des exemples, éventuel apprentissage par transfert dépendant d'autres acteurs. Il faut bien entendu distinguer la reproductibilité de l'algorithme de décision final, souvent déterministe, de la reproductibilité de l'apprentissage, plus problématique. C'est heureusement l'algorithme de décision final qui fait l'objet des règlements, à charge pour le couple fournisseur-utilisateur d'avoir vérifié son fonctionnement et sa maintenance. Cependant, on doit prévoir la généralisation d'algorithmes continuellement mis à jour par un apprentissage permanent utilisant un flux de données nouvelles.

3.3 L'interprétabilité

De nombreux travaux portent sur l'interprétabilité des systèmes intelligents et ce mot reste utile pour la recherche bibliographique en XAI. Cependant, ce terme est alors utilisé soit dans un contexte statistique, où, sans être formalisé, il a mené à une heuristique importante, soit, comme dans l'article de Doshi-Velez et al. [16], au sens de "la capacité à expliquer ou à présenter en termes compréhensibles à un humain", ce qui revient à déplacer le problème à celui de l'explication et de la compréhension, que nous abordons plus loin. L'intitulé "IA explicable" semble heureusement prendre l'avantage sur celui d'"IA interprétable". Nous éviterons donc, sauf dans la sous-section "apprentissage statistique" de la section "méthodes", d'utiliser le terme dans ce sens général imprécis, d'autant plus que le mot "interprétation" admet depuis Tarski une formalisation fondamentale en logique mathématique, utilisée aussi avec les logiques de description en modélisation de la connaissance (ontologies, web sémantique, bio-informatique). C'est dans cette acception que nous appliquerons le concept d'interprétation à une première analyse des phénomènes de compréhension, en gardant à l'esprit – les juristes le savent aussi – qu'une interprétation n'est pas nécessairement unique.

3.4 La causalité

Expliquer une décision dans laquelle de nombreux facteurs ou critères ont potentiellement été impliqués peut être vu comme un exposé des "causes" de la décision. La notion de causalité a fait l'objet d'études depuis l'antiquité (par Aristote notamment, qui distinguait les causes formelle, matérielle, efficiente et finale) et à notre époque dans des disciplines relevant tant des sciences humaines (philosophie, histoire, sociologie, sciences politiques, économie, ...) qu'exactes (physique, biologie, statistique, logique,...) [17]. En sciences juridiques, cette notion est importante pour les questions de responsabilité (au sens légal, à distinguer du terme plus récent d'"IA responsable"), ce qui nous renvoie aux motivations – déjà exposées – de l'IA explicable. Chacune de ces disciplines a ses propres règles et il est bon de préciser pour chaque auteur dans quel cadre il se place, notamment afin d'éviter les abus d'autorité (le triste précédent de la condamnation de la recherche sur les perceptrons, prédécesseurs des réseaux de neurones, par l'extrapolation abusive d'un résultat de Minsky et Papert, tenants d'une IA symbolique, doit être gardé à l'esprit) [18]. Ainsi, au sein des théories statistiques de la causalité, qui se donnent pour but de retrouver des relations causales entre variables – et non pas de simple associations – à partir de données observationnelles et particulièrement de considérations d'indépendance conditionnelle, on pourra tempérer l'assurance de J. Pearl quant aux Réseaux Bayésiens par les critiques de statisticiens comme C. Dawid [19] et apprécier dans des sommes récentes [20] les difficultés d'utilisation de ces méthodes (appelées Modèles Graphiques par l'école de Lauritzen qui explique qu'ils n'ont rien de bayésien). Plus simples à comprendre et à mettre en oeuvre, les théories contrefactuelles de la causalité ont grandement diffusé dans les sciences politiques et juridiques. Ces domaines sont très pertinents pour notre sujet dont l'humain est la cible. Ainsi [21] non seulement indique une méthodologie contrefactuelle pour donner des explications à l'individu sujet d'une décision automatisée (cf. la sous-section causalité 4.2.6 dans la section méthodes) mais propose en outre trois objectifs à une telle explication : (1) aider l'individu à comprendre, (2) le guider dans d'éventuelles contestations juridiques de la décision, (3) le guider dans une éventuelle stratégie d'adaptation à l'algorithme (dans l'exemple d'un refus de prêt bancaire, on peut lui faire savoir à quel point il devrait réduire ses différents postes de dépense, ou baisser sa demande, ou diminuer la fréquence de ses situations de découvert bancaire, etc). Il est vrai que l'article est un plaidoyer pour la conservation de la confidentialité des algorithmes et essaye donc de dresser un inventaire des demandes légitimes pour montrer qu'on peut y satisfaire sans ouvrir la boîte noire.

3.5 Explication et compréhension

3.5.1 Proposition d'un formalisme conceptuel

Dans la voie d'un formalisme pour la spécification des fonctions d'explication d'un système intelligent, nous proposons quelques définitions :
Nous appellerons *explication* le message envoyé par le système qui doit s'expliquer (la *source*), et reçu par le demandeur (la *cible*), éventuellement en réponse à une requête (demande d'explication plus ou moins précisée). Il est bon de ne pas restreindre les cibles à des humains ; il peut s'agir d'autres systèmes intelligents.
Nous appellerons *compréhension* certains changement d'état chez la cible, celle-ci étant vue comme un automate fini. En fonction de ses effets sur la cible, l'explication peut être plus ou moins satisfaisante : une métrique pour la qualifier pourra s'appuyer sur la structure et l'état initial de la cible. L'effet d'une explication semble s'apparenter à celui d'une projection par sa propriété d'idempotence : si une compréhension totale est atteinte après une première explication, une deuxième explication ne modifie plus l'état.
La recherche de métriques sur la qualité d'une explication peut s'appuyer sur des interrogatoires et méthodes de psychologie estimant un degré de satisfaction de la cible, ou quantifier quelques autres aspects de l'explication : adéquation de la réponse à la requête, niveau d'expertise de la cible, existence de vocabulaires de référence, complétude de la réponse (dans un sens à préciser).

3.5.2 Un exemple en radiologie

Analysons ce que peuvent être l'explication et la compréhension entre humains pour les généraliser ensuite à d'autre système intelligents. Soit un radiologue (la source) devant expliquer son diagnostic de tumeur maligne du sein à un correspondant (la cible) à la suite d'un examen mammographique. Le compte rendu radiologique pourrait comprendre le passage : "sur le cliché de face, opacité arrondie mal limitée à contour spiculé de 15mm de diamètre dans le quadrant supéro-externe, sans déformations architecturales", et plus loin la conclusion "probable adénocarcinome, classification BIRADS 5".
La partie descriptive permet à un expert de retrouver la lésion sur les images et de confirmer le diagnostic, éventuellement de la discuter ou le contester. Un correspondant moins expérimenté peut demander des explications complémentaires. A la requête "où se trouve la lésion ?" la source peut répondre par l'annotation du centre d'un disque approchant l'image ronde. Des flèches peuvent indiquer le contour, d'autres flèches indiquer les spicules. Le concept abstrait d'"image arrondie" est donc instancié en un disque parfaitement défini au sein de l'image. La question de l'existence de spicules est remplacée par l'appréciation du caractère spiculé d'un segment de contour parfaitement identifié. On peut dire que l'explication a complètement instancié les concepts abstraits utilisés dans la description verbale, ce qui constitue une interprétation au sens logique (qui fait correspondre à des symboles de variables des éléments d'un ensemble appelé domaine : ici à la variable "masse arrondie" on a fait correspondre un unique disque dans le plan). A la question "pourquoi ce type de tumeur ?" la source peut répondre par des références bibliographiques justifiant une telle déduction à partir des éléments de description invoqués et maintenant compris, ce qui, cette fois-ci, explique un raisonnement déductif au sein d'un formalisme logique.
L'aide informatisée au diagnostic (CAD : Computer Assisted Diagnosis) a d'abord utilisé des méthodologies où l'on automatisait par traitement d'images les tâches de reconnaissance des primitives déjà connues des radiologues, avant d'utiliser un système logique (à base de règles, par exemple) pour déduire des résultats de cette première étape une gamme de diagnostics avec éventuellement des poids apparentés à des probabilités. La "radiomique" (radiomics) plus récente reprend ce schéma en remplaçant l'étape logique par des méthodes d'apprentissage statistique ou un réseau de neurones à couches combinatoires. Dans ces systèmes calqués sur la démarche de l'humain (reconnaissance visuelle puis raisonnement), les explications peuvent reprendre la démarche exposée ci-dessus : anno-

tation d'image pour la justification des détections de primitives, puis aide à la compréhension d'une déduction. Au contraire, dans les systèmes à base d'apprentissage profond (Deep Learning), l'optimisation qui correspond à l'apprentissage s'effectue de bout en bout (des images aux diagnostics) sans la contrainte de s'appuyer sur des classifications pré-existantes, et les traitements associés aux différentes couches sont paramétrés par les poids et seuils des noeuds du réseau. Tous ces paramètres sont parfaitement accessibles, mais – nous l'avons remarqué ci-dessus – cette transparence totale n'est que de peu d'aide pour comprendre comment le système opère. On peut cependant tenter de reprendre l'approche hiérarchique précédente. Quand les couches basses utilisent des opérations de convolution – ce qui traduit une invariance spatiale – il est naturel de rechercher une détection de primitives dans le traitement qu'elles opèrent. L'exploration du traitement effectué par ces couches peut s'appuyer sur la technique des cartes d'attention (cf. ci-après). L'interprétation de telles cartes pour les couches combinatoires qui leur font suite demande plus de travail car le domaine d'interprétation n'est plus conceptuellement aussi proche de l'espace imagé ou de ses avatars multi-résolution.

3.6 Intelligence artificielle explicable

L'Intelligence Artificielle Explicable (XAI) est un ensemble de processus et de méthodes qui permettent aux utilisateurs humains de comprendre les résultats et les conclusions créés par les algorithmes d'apprentissage automatique et de leur faire confiance. L'IA explicable est utilisée pour décrire un modèle d'IA, son impact attendu et ses biais potentiels. Elle aide à caractériser la précision, l'équité, la transparence et les résultats des modèles dans la prise de décision assistée par l'IA. L'IA explicable est cruciale pour une organisation qui souhaite instaurer la confiance lors de la mise en production de modèles d'IA. L'explicabilité de l'IA aide également une entreprise / organisation / unité à adopter une approche responsable du développement de l'IA [3]
XAI correspond à une approche qui propose une décision accompagnée par une justification adaptée à l'application (sous forme d'un texte argumenté, visualisation, simplification, par exemple). L'explicabilité comprend un mode interactif qui prend on compte le retour d'expérience de l'utilisateur dans un schéma d'amélioration continue (apprentissage par renforcement, après le déploiement du modèle) ;
Nous pouvons distinguer, deux catégories d'explications :

- explication générée / guidée par l'expert
- explication automatisée - générée par le système d'IA (en plus de sa décision)

4 Grandes familles de méthodes pour l'explicabilité

4.1 Propriétés des méthodes d'explicabilité

Les méthodes décrites pour l'explicabilité en I.A. sont en général classées *a priori* en fonction de quelques propriétés qui indiquent dans une certaine mesure leur cas d'usage. Nous rendons compte de la terminologie la plus courante.

4.1.1 Méthodes locales, méthodes globales

Les méthodes *locales* s'attachent à expliquer la décision d'un système intelligent pour un cas individuel, c'est-à-dire pour une entrée particulière. On pourrait les appeler ponctuelles, mais elles concernent souvent aussi tout un voisinage de l'entrée considérée au sein d'un espace d'entrées, ce qui coïncide alors avec l'usage topologique standard du mot "local". Les méthodes dites *globales* s'adressent à l'explication du fonctionnement du système sur l'ensemble des entrées possibles, et

3. https://royalsociety.org/topics-policy/projects/, https://www.ibm.com/fr-fr/watson/explainable-ai

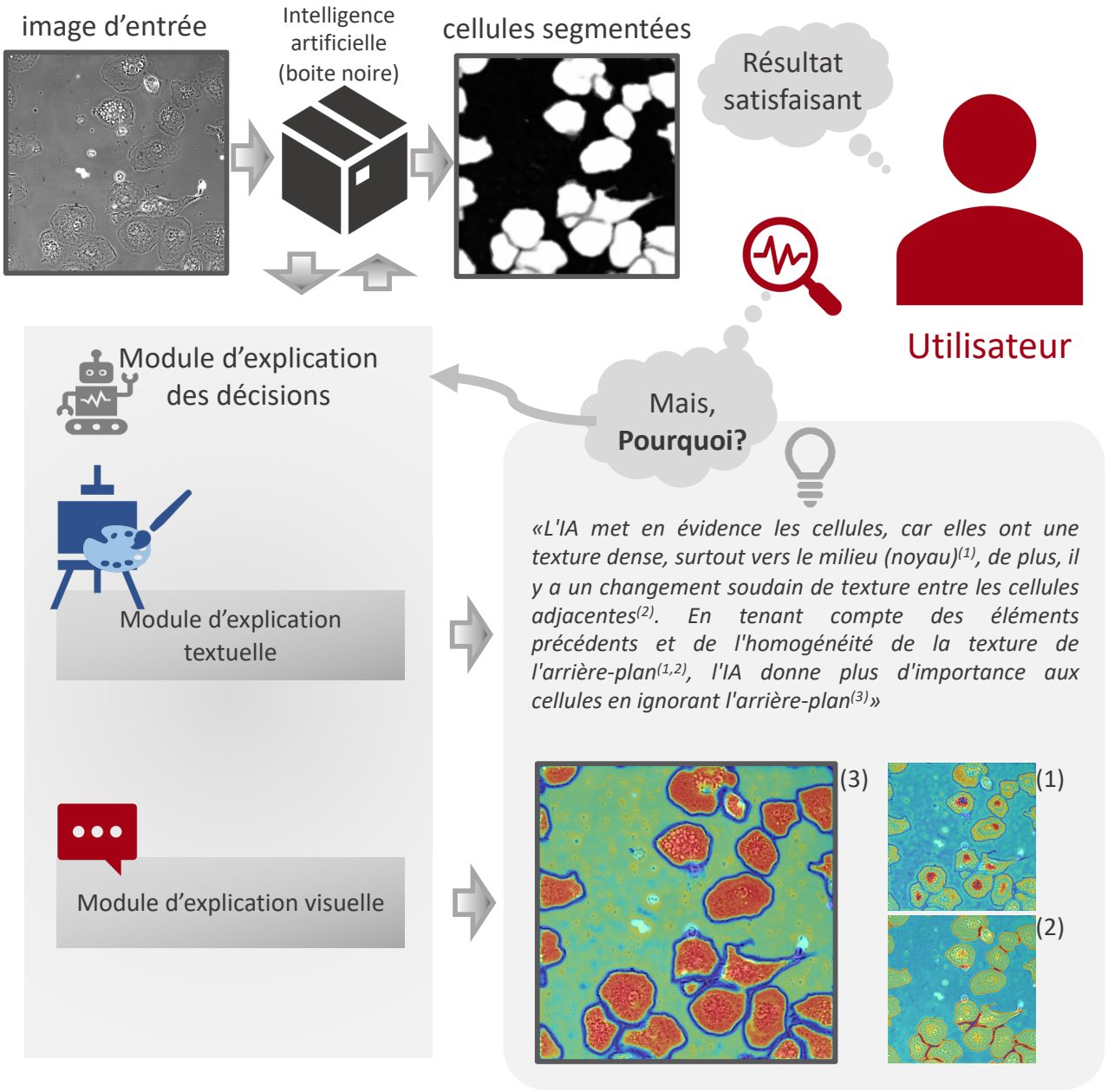

FIGURE 2 – Double boucle d'apprentissage : différents niveaux d'intégration du retour de l'expert

FIGURE 3 – Principes et composantes de l'IA responsable : équité, éthique, confidentialité, traçabilité, redevabilité, sécurité, empreinte carbone.

éventuellement à l'algorithme permettant de les obtenir par apprentissage. On voit que l'obligation légale d'explication pour une décision individuelle met plutôt en jeu des méthodes locales fournies par l'acteur qui met en oeuvre le système intelligent, alors que les méthodes globales concernent surtout les rapports entre fournisseurs de systèmes intelligents et leurs clients, ou encore entre concepteurs et leurs interlocuteurs.

4.1.2 Méthodes agnostiques, méthodes spécifiques

Les méthodes *agnostiques* considèrent le système intelligent comme une "boîte noire" et ne nécessitent pas d'information particulière sur lui. Elles sont utilisables sans transfert de propriété intellectuelle et peuvent donc constituer un minimum légal, notamment en fin de chaîne. Les méthodes *spécifiques* sont au contraire adaptées à un système particulier et sont plutôt développées en synergie avec lui, concernant plutôt l'amont de la conception.

4.1.3 Méthodes *post hoc*, méthodes *ex ante*

Moins fréquemment utilisés, ces termes, repris du vocabulaire de la statistique appliquée, distinguent l'explicabilité d'une décision déjà rendue par un système déjà conçu, et l'explicabilité prise en compte dès la conception du système intelligent. Cette dernière approche est moins répandue, plutôt restreinte jusqu'ici aux besoins des concepteurs de systèmes intelligents, mais constitue certainement une importante direction de recherche. Dans le cas où l'explicabilité *ex ante* est recherchée en contraignant le système à n'emprunter que des composants utilsables en vue d'une explication, on conçoit que la solution obtenue par optimisation soit possiblement de valeur inférieure à la solution obtenue sans ces contraintes, ce qui rend compte d'un arbitrage souvent évoqué sur des arguments empiriques entre explicabilité et performance [22].

4.2 Composants à expliquer, composants utilisables

L'explicabilité en I.A. est un sujet de recherche actif et encore récent ; en quelque sorte une cible mouvante. Au lieu d'en dresser un inventaire exhaustif nous allons tenter d'en tracer les lignes directrices telles qu'on peut les percevoir à ce jour. Avant de présenter quelques études de cas dans une section suivante, nous parcourons d'abord les méthodes d'apprentissage (sans renforcement, pour simplifier) les plus courantes en I.A., et ceci depuis deux points de vue : d'une part les algorithmes issus de différents types d'apprentissage peuvent poser des problèmes d"explicabilité différents, d'autre part, puisqu'en pratique l'explicabilité doit souvent reposer sur une génération automatisée d'explications, différents types d'apprentissage peuvent être mobilisés pour la synthèse des systèmes fournissant les explications.

4.2.1 Apprentissage statistique, heuristique de l'interprétabilité

On regroupe sous le terme d'apprentissage statistique différentes méthodes prolongeant les outils de l'analyse statistique des données. La distinction entre statistique descriptive et statistique interprétative est classique, et la notion d'interprétation est courante en statistique appliquée. La statistique interprétative relie en général certaines hypothèses relatives au phénomène étudié à des modèles probabilistes explicites qui permettent d'engendrer à volonté des données aléatoires de même type que les données observées. En comparant quantitativement la distribution des données observées aux distributions générées par les différent modèles (que ces distributions soient connues par des moyens analytiques ou des simulations numériques) on décide du choix d'un modèle et donc de la validité des hypothèses correspondantes. Cette méthodologie est pratiquée par une large communauté. Il existe ainsi un cadre consensuel d'interprétation statistique et la pertinence des explications fournies par différents modèles peut être mesurée à l'aide d'outils compris par les praticiens de différentes sciences expérimentales comme la Médecine, la Biologie ou la Psychologie.

Les méthodologies numériques (simulation, méthodes de Monte Carlo, rééchantillonage, méthodes bayésiennes) on permis l'utilisation de modèles complexes et de s'affranchir des approximations (gaussiennes par exemple) de la statistique classique. Aux modèles linéaires, comme la régression linéaire, et linéaires généralisés (régression logistique par exemple) sont venues s'ajouter les méthodes d'arbres pour la régression et la classification (CART de Breiman), puis les méthodes de "bagging" (aggrégation d'arbres rééchantillonllés : Bootstrap AGGregation) et de forêts aléatoires (random forests) combinant (par exemple par un vote) les décisions de populations d'arbres. Les méthodes de boosting comme XGBoost permettent elles aussi d'améliorer une méthode de base. La théorie de l'apprentissage statistique de Vapnik a permis, de manière décisive, de décrire quantitativement la possibilité pour un système d'apprentissage de généraliser correctement à partir des données, notamment en évitant le phénomène de sur-ajustement (overfitting). Issues de cette recherche, les machines à vecteurs d'appui (SVM : support vector machines) s'apparentent cependant au méthodes linéaires généralisées dans la mesure où elles ramènent un problème de classification non linéaire à la séparation de deux nuages de points par un hyperplan.
Ces méthodes d'apprentissage statistique sont partagées par une communauté d'utilisateurs ayant un langage commun (statistique) et font l'objet de plateformes logicielles comme scikit-learn ou les bibliothèques du langage R, qui incluent aussi des outils d'explicabilité comme la valeur de Shapley. On peut organiser l'ensemble de ces approches en une progression depuis des méthodes élémentaires bien comprises des utilisateurs vers des méthodes dont les décisions, prises à partir de plusieurs méthodes élémentaires, demandent un effort de compréhension. Les approches d'explicabilité empruntées par les chercheurs de l'apprentissage statistique sont intéressantes même pour l'abord des problème *a priori* plus difficiles posés par l'apprentissage profond. Le cas des forêts aléatoires est éclairant. Si la décision finale d'un arbre de décision est compréhensible, c'est parce qu'elle résulte simplement d'une suite de décisions élémentaires ayant chacune une signification pour l'humain, ces décisions élémentaires étant sélectionnées au sein d'un répertoire que l'utilisateur apporte avec leurs significations directement compréhensibles (dans d'autre types d'apprentissage cet apport de sémantique vient des d'annotations). Pour expliquer la décision prise par une forêt à partir des arbres qui la composent, certains auteurs ont proposé de construire un arbre de décision – ce type d'algorithme étant plus facilement analysable en vue d'une explication – reproduisant avec une bonne approximation les décisions de la forêt.
On peut regrouper avec ces méthodes l'approche connue en Imagerie Médicale sous le terme de radiomique (radiomics) dans laquelle on fait commencer l'analyse d'images radiologiques par une batterie de filtres ou traitement d'mage classiques dont les résultats, sortes de traits élémentaires (features) constituent les entrées d'un système d'apprentissage statistique. Ce qui nous importe ici est que les différents filtres élémentaires sont dotés d'une signification connue et comprise par les différents utilisateurs : détecteur de bord ou de point isolé au sein de l'image, analyseur de texture par exemple. Seule la dernière étape, combinatoire, est apprise et pose un problème d'explication. De la même manière, les SVM nécessitent qu'un "noyau" d'analyse élémentaire leur soit fourni sur lequel elle feront leur optimisation. Au contraire, l'apprentissage profond dont nous parlons dans la prochaine section, soumettant tous ses composants à l'apprentissage réclame l'explication de tous ses composants en plus de leur décision globale.

4.2.2 Apprentissage automatique profond (deep learning)

L'apprentissage profond (deep learning) est une évolution de la méthodologie des réseaux de neurones artificiels (ANN : artificiel Neural Networks), ou neuromimes, explorée depuis les perceptrons multi-couches (Rosenblatt), elle-même issue de travaux de modélisation neuro-biologique (Rashevski, McCulloch-Pitts, Hebb). Après avoir été mise sous le boisseau à la suite de rivalités avec l'IA symbolique (cf. remarques ci-dessus à propos de J. Pearl) elle a refait surface dans les années 80 puis s'est imposée par ses résultats spectaculaires dans les années 2000. Pour aller au delà de la courte présentation qui suit le lecteur est renvoyé à l'un des nombreux ouvrages de référence [23, 24].

Chaque neurone formel constituant du réseau est défini par une fonction faisant correspondre à un ensemble d'entrées numériques (comme arrivant par les dendrites) un seule sortie numérique (comme propagée par l'axone) : à une combinaison linéaire des entrées (dont les coefficients sont appelés *poids*) on applique une fonction non linéaire, appelée *fonction d'activation*, qui a pu consister en un simple seuillage ou une fonction tangente hyperbolique, mais en pratique est une fonction linéaire par morceaux comme la fonction RELU (REctified Linear Unit) : $x \mapsto max(0, x)$ ou une de ses variantes. La fonction d'entrées-sortie résultant de ces deux étapes est donc totalement spécifiée par un nombre fini de paramètres numériques. De tels neurones formels sont formellement assemblés en un réseau qui en général se décompose en couches successives (par analogie avec l'organisation anatomique du système visuel qui a beaucoup inspiré la recherche dans le contexte du traitement d'images) depuis la couche d'entrée jusqu'à la couche de sortie du réseau. Une couche intermédiaire prend donc ses entrées parmi les sorties de la couche précédente et envoie ses sorties vers les entrées de la couche suivante. Parmi les multiples arrangements possibles de connections entre deux couches successives, on distingue en particulier les *couches combinatoires* (ou *complètement connectées*) : toutes les connections possibles vers les sorties de la couche précédente existent, et les poids sont indépendants. Au contraire dans les couches dites *de convolution* chaque neurone de la couche n'est relié qu'aux sorties des neurones constituant un voisinage, ou masque, de son homologue dans la couche précédente, les masques ayant tous même taille, même forme, et même ensemble de poids pour le calcul de la fonction d'entrées-sortie pour le neurone cible (noyau de convolution). Il y a donc dans ce type de liaison entre deux couches beaucoup moins de paramètres : les poids associés au masque de l'unique noyau de convolution servant à tous les neurones et les paramètres de l'unique fonction d'activation utilisée par tous les neurones. Très souvent on associe à chaque couche n canaux de sortie (par analogie avec les $n = 3$ composantes d'une image en couleur), c'est-à-dire que chaque neurone est remplacé par n neurones qui calculent chacun le résultat d'un canal, et on fait calculer simultanément par une couche une opération de convolution par canal : les canaux correspondent alors aux résultats d'opérations de convolution par différents noyaux et on peut leur faire jouer le rôle de différentes caractéristiques locales (local features) de l'image, par exemple la présence de bords dans différentes directions. La couche suivante prendra tous ces canaux en entrée de ses neurones. Dans de nombreux modèles, à mesure que l'on s'éloigne de l'entrée les couches ont moins de neurones et plus de canaux (mais cette tendance peut s'inverser dans la deuxième partie du réseau comme dans l'encodeur-décodeur, cf. *infra*). La diminution de résolution d'une couche à la suivante est obtenue par une couche de "max pooling" qui opère une sorte de décimation : la couche de départ étant scindée en carreaux jointifs qui s'envoient chacun sur un unique neurone de la couche d'arrivée, on affecte à chaque neurone d'arrivée le maximum des valeurs des neurones du carreau de départ. Au total, en s'éloignant de la couche d'entrée les neurones deviennent moins nombreux mais porteurs d'une information sémantique plus grande et dépendant d'une région plus vaste de la couche d'entrée.

On peut simuler numériquement la réponse d'un réseau de neurones formels et calculer la sortie (nombre, vecteur, tableau, suivant la forme de la dernière couche) associée à une entrée (nombre, vecteur, ou tableau, suivant la forme de la couche d'entrée), c'est-à-dire calculer la fonction d'entrée-sortie du réseau. Un des principes de l'apprentissage de tels neuromimes est de modifier progressivement les paramètres des neurones (entre autres leurs poids) pour approcher une fonction d'entrée-sortie de réseau désirée. Le plus souvent on utilise une variante de méthode de gradient (à l'origine la rétro-propagation de gradient ou "back-prop") qui à partir d'une fonction coût évaluant la différence entre le tableau des sorties actuelles et le tableau de sorties désirées, et connaissant les caractéristiques des paramètres d'entrées-sortie actuels de chacun des neurones, calcule une modification souhaitable de tous ces paramètres. Dans le cadre d'un apprentissage supervisé, on donne successivement au système en entrée les éléments d'un corpus d'exemples pour lesquels on dispose à chaque fois d'une supervision que l'on prend comme sortie désirée. On souhaite du système que non seulement il apprenne à associer aux entrées du corpus les sorties désirées, mais qu'il généralise à des entrées nouvelles des sorties convenables, faute de quoi on dit qu'il a simplement "appris par coeur"

les réponses associées au corpus sans avoir su généraliser, ce qui se produit souvent quand l'ensemble des paramètres accessibles pour construire la fonction d'entrée-ortie est trop grand, menant à un sur-ajustement (overfitting) bien abordé par la théorie de Vapnik ou les méthodes de régularisation. On peut ainsi comprendre que les couches de convolution soient plus facilement entraînables que les couches combinatoires.
L'apprentissage est dit profond quand le nombre de couches est élevé. La simulation de tels réseaux profonds a nécessité la mise au point de certaines techniques, notamment l'utilisation de fonctions d'activation adaptées et la résolution du problème des gradients évanescents (vanishing gradients) par exemple à l'aide de "connections multi-couches" (skip connections).
Expliquer le fonctionnement de ces réseaux est parfois relativement aisé pour certaines de leurs composantes. Un réseau destiné à l'analyse d'images consistant en quelques couches de convolution suivies de quelques couches combinatoires approche, pour les premières couches, les calculs mis en oeuvre par les solutions radiomiques dans la mesure ou ces couches, pour lesquelles la convolution a été imposée (ce qui revient à faire respecter par le système une contrainte d'invariance par translation), retrouvent souvent des traitements proches des filtres d'image, eux aussi à base de convolution, que propose l'expert concevant un système radiomique. Le fonctionnement des couches suivantes peut s'aider de visualisation comparant les sorties d'une couche à l'image d'entrée ou à la couche précédente car toutes ces couches ont des formats comparables bien que de résolutions en général différentes. L'explication devient plus difficile à mesure que l'on s'éloigne de l'entrée et encore plus pour les couches combinatoires, mais la recherche d'explications fait partie du travail de recherche sur l'apprentissage profond. On peut par exemple, comme en neurophysiologie du système visuel, reconstruire le champ récepteur d'un neurone profond en cherchant les sous-images qui, placées en entrée, maximisent la réponse du neurone profond étudié. Certaines des méthodes récemment développées pour les réseaux de neurones, comme les méthodes d'attention, sont en même temps des moyens d'atteindre certaines performances et des moyens de visualisation pouvant aider à la compréhension du fonctionnement.
Il faut remarquer que les outils mathématiques qui ont trouvé leurs utilité pour expliquer le fonctionnement de neuromimes ne se réduisent pas aux concepts de la statistique, mais relèvent souvent de la théorie de l'approximation, du non linéaire, de la géométrie différentielle dans des espaces de grande dimension avec l'appel, au moins heuristique, à des notions comme les variétés différentiables et leurs plongements. L'interprétabilité au sens des statisticiens peut donc trouver ses limites dans ces nouveaux domaines et le risque de "démonstration" abusive du caractère intrinsèquement non explicable de l'apprentissage profond par les spécialistes d'autres domaines de l'IA est réel. Par exemple, le fait que la réponse d'un neuromime puisse être très sensible à d'infimes variations de l'entrée a pu être invoqué comme rendant préférables les "systèmes déductifs logiques, système probabilistes qui n"ont pas les mêmes problèmes car ils sont plus explicables car on ne perd pas la causalité entre entrée et sortie" (émission La Méthode scientifique, France Culture du 30/03/2022 https ://www.franceculture.fr/emissions/la-methode-scientifique/intelligence-artificielle-par-dela-le-bien-et-le-mal). Une discussion plus approfondie pourrait par exemple reprendre les outils de la démonstration de l'importance de la stabilité topologique pour les signes utilisés en radiologie [25].

4.2.3 Apprentissage supervisé, apprentissage non supervisé

Nous avons jusqu'ici mis en avant les méthodes d'apprentissage supervisé. C'est de l'annotation des données d'apprentissage que provient la supervision. Cette annotation peut comporter des erreurs et une partie de l'effort d'explication peut être dirigé vers la détection de ces erreurs dans une optique d'assurance qualité avec retour de l'annotation aux experts pour confirmation ou correction. Quand elle est disponible, la sélection des exemples annotés les plus "responsables" d'une décision individuelle est aussi un des moyens d'explication.
Paradoxalement, certaines méthodes d'apprentissage non supervisé peuvent aussi être utilisées dans le cadre qui nous occupe, par exemple pour la détection de transactions bancaires frauduleuses ou

dans le domaine de la cybersécurité, notamment dans ce dernier cas parce que les caractéristiques des attaques sont souvent inconnues ("zero day attacks"). On ne cherche alors qu'à détecter les transactions inhabituelles et l'un des réseaux de neurones souvent utilisé avec succès est l'auto-encodeur. Ce réseau comporte un nombre de couches qui peut être grand (et alors entrer dans le domaine de l'apprentissage profond) avec une entrée et une sortie de même format. On entraîne le réseau à reconstituer son entrée (qui pourrait consister en les paramètres de la transaction) le plus exactement possible à sa sortie, avec la particularité, qui enlève au problème toute trivialité, que certaines couches intermédiaires sont de taille volontairement très réduite, ce qui force le réseau à élaborer une compression de l'entrée. On parle de forme en sablier du réseau car depuis l'entrée, les couches diminuent progressivement en taille puis ré-augmentent jusqu'à la sortie. Dans ce type d'apprentissage, c'est la distribution de probabilité des entrées qui est modélisée, et la responsabilité d'une décision individuelle peut être argumentée en comparant le corpus des exemples, ainsi que leur paramètres, qui ont servi à l'apprentissage et le contexte dans lequel a été prise la décision. On conçoit que, mal utilisée, une méthode ce type pourrait facilement mener à des discriminations abusives, surtout quand on sait que la richesse des données individuelles disponibles permet souvent de reconstituer assez précisément des variables dont l'enregistrement est interdit.

4.2.4 Modularité, apprentissage par transfert, apprentissage multi-tâche.

Comme dans le reste de l'industrie, l'IA tend à modulariser et standardiser ses produits tout en permettant de les assembler en assez de combinaisons pour répondre aux attentes des utilisateurs. Dans le cas des neuromimes, nous appellerons *composant cognitif* un modèle pré-entraîné sur des données qui ne sont pas celles de l'utilisateur (il peut s'agir ou non de données publiques). En vue d'un usage particulier on peut alors utiliser, à côté de la classique combinatoire des composants, une adaptation finale du composant par un deuxième apprentissage, supplémentaire, cette fois sur les données de l'utilisateur et partant des poids obtenus à l'issue du premier apprentissage. Cette technique, connue sous le nom d'*apprentissage par transfert* (transfer learning), est efficace même quand le premier corpus ne semble pas très proche des données qui serviront à l'inférence, ce qui le rattache à l'apprentissage multi-tâche [26]. L'effet du premier apprentissage peut donc faire partie des éléments à prendre en compte pour une explication du système final ; c'est pourquoi l'accès au corpus du premier apprentissage est désirable.
L'apprentissage par transfert ou par apprentissage multi-tâche peut aussi être une technique d'élaboration de systèmes explicatifs, après avoir été au départ un moyen d'étude des phénomènes de facilitation de l'apprentissage par l'adjonction d'apprentissages simultanés [27, 28, 29]. Nous abordons ici une technique qui est propre aux neuromimes et s'éloigne des paradigmes de l'apprentissage statistique. Elle a été utilisée dans le cadre de la synthèse de commentaires de programmes aux fins d'ingénierie inverse [30], et ceci s'applique notamment à la détection de maliciels, ce qui entre dans notre cadre d'étude. On fait donc apprendre simultanément (ou suivant une séquence appropriée, pour rappeler l'apprentissage par transfert) la tâche de commentaire des programmes désassemblés, et l'apprentissage d'une API (interface de programmation d'applications) dont on possède des fragments parfaitement documentés [31]. Cette puissante technique est donc une forme de transfert de connaissances au réseau par des opérations de très haut niveau d'assemblage de tâches qui permet d'introduire de la sémantique (ici les commentaires, ailleurs des annotations) depuis différents corpus.

4.2.5 Apprentissage fédératif

L'apprentissage fédératif est une solution qui a été proposée pour permettre à plusieurs acteurs possédant chacun des données de collaborer pour l'apprentissage d'un modèle en préservant la confidentialité de leur données. Il est utilisé notamment en imagerie médicale pour faire collaborer à l'apprentissage d'un modèle (par exemple à but diagnostic) différentes entités de soin au sein d'une

étude multi-centrique tout en préservant la confidentialité des données au niveau de chaque entité de soin, la mise en commun des données des patients présentant un risque jugé inacceptable et de trop grands problèmes de responsabilité [32]. On ne regroupe donc pas l'ensemble des images et dossiers cliniques des patients dans un même site, ne transmettant depuis chaque site participant que les données différentielles strictement nécessaires à l'apprentissage collaboratif, avec des mesures de sécurité appropriées. Un tel outil ajoute une couche de complexité aux problèmes d'explicabilité et appelle des recherches particulières sur des méthodes qui seront sans doute spécifiques et *ex ante*. A l'actif ce ces entreprises, notons que l'effort de spécification, passant souvent par une modélisation des acteurs, préalable à la mise en place d'un système fédératif contribue positivement à l'explicitation de nombreux facteurs à prendre en compte dans les études d'explicabilité.

4.2.6 Causalité

Le processus d'explication est lié à un des mécanismes suivants [33] :

- d'abduction : l'inférence hypothétique telle que définie par Peirce [34]. Ce serait le plus souvent le "début", l'hypothèse à vérifier, bien souvent par induction ;
- de retrospection : voir si des explications du passé peuvent fonctionner – vérification contrefactuelle ;
- de prospection : s'interroger sur ce qui pourrait arriver, en relation avec la prédiction.

Malgré une attraction croissante des ingénieurs vers l'utilisation de l'IA, la plupart des projets basés sur l'apprentissage automatique se concentrent sur la prédiction des résultats plutôt que sur la compréhension de la causalité des évènements intrinsèques ou associés (induits, adjacents ...). En effet, l'apprentissage automatique est excellent pour trouver des corrélations dans les données, mais pas pour identifier une causalité. Il est donc essentiel de ne pas tomber dans le piège consistant à assimiler corrélation et causalité.
Ce problème limite considérablement la capacité à se fier à l'apprentissage automatique pour la prise de décision. D'un point de vue commercial, il est nécessaire de disposer d'outils capables de comprendre les relations causales entre les données et de créer des solutions d'apprentissage automatique, en mesure d'être facilement (sinon directement) généralisables.

Problèmes avec l'apprentissage automatique :
Dans leur état actuel, les algorithmes de ML (en anglais "Machine Learning") peuvent être biaisés – souffrant d'un manque relatif d'explicabilité – et sont limités dans leur capacité à généraliser les modèles qu'ils trouvent dans un ensemble de données d'apprentissage pour plusieurs applications. Il est donc très important d'améliorer la capacité de généralisation de ces modèles.
La capacité de généralisation représente la capacité du modèle à s'adapter correctement à de nouvelles données inédites, issues de la même distribution que celle utilisée pour créer le modèle. De plus, les approches actuelles de ML ont tendance à sur-adapter les données (surapprentissage, surajustement, ou encore surinterprétation - en anglais "overfitting"). En effet, soit par manque ou déséquilibre des données, soit à force de trop les pousser, les modèles essaient d'apprendre parfaitement le passé, au lieu de découvrir les relations réelles/causales qui continueront à se maintenir au fil du temps, en donnant un caractère durable et efficient, à la modélisation).
Dans le domaine de la santé, les modèles soutiennent simplement que les symptômes surviennent en présence d'une maladie et que la maladie génère la présence de symptômes.
À l'heure actuelle, les systèmes d'IA les plus performants sont des modèles d'apprentissage profond (en anglais "deep learning" - DL), qui exploitent des ensembles plus volumineux de données, avec plus d'exemples de différentes situations possibles. Il pourrait être tentant de simplement s'appuyer sur plus de données (big data), mais ce serait une erreur.
Même si nous pouvons observer une corrélation, cela ne prouve, toutefois, pas une causalité.

La nouvelle science des causes et des effets met en évidence les principales limites des solutions actuelles d'apprentissage automatique et le défi de l'inférence causale [35]. Les auteurs notent que le battage médiatique selon lequel les mégadonnées résoudront bon nombre des grands défis auxquels nous sommes confrontés, est déplacé.
Parce que le DL s'est trop concentré sur la corrélation sans causalité, les données ne répondront pas à la question lorsque le problème s'éloigne de situations très étroites. En fait, beaucoup de données du monde réel ne sont pas générées de la même manière que les données que nous utilisons pour former des modèles d'IA. En d'autres termes, le DL est bon pour trouver des modèles en termes de données, mais ne peut pas expliquer comment ils sont connectés.
La plupart des solutions sont incapables de généraliser au-delà du domaine des exemples présents dans un ensemble de données.
Pour un nombre croissant d'applications métiers, la capacité du ML à trouver des corrélations est plus que suffisante (ex : prédiction de prix, classification d'objets, meilleur ciblage, etc.). En effet, les systèmes de ML sont excellents dans l'apprentissage des connexions entre les données d'entrée et les prédictions de sortie, mais manquent de raisonnement sur les relations de cause à effet ou les changements d'environnement. Les modèles ML qui pourraient saisir les relations causales seront plus généralisables.
La causalité représente l'influence par laquelle un événement, un processus ou un état, une cause, contribue à la production d'un autre événement, processus ou état, un effet, où la cause est en partie responsable de l'effet, et l'effet est en partie dépendant de la cause.
La capacité à découvrir les causes et les effets de différents phénomènes dans des systèmes complexes nous aiderait à élaborer de meilleures solutions dans des domaines aussi divers que les soins de santé, la justice et l'agriculture. En effet, les acteurs de ces domaines ne devraient pas prendre de risques lorsque les corrélations sont confondues avec la causalité.

Inférence causale et cas d'utilisation :
En tant qu'être humains, nous pensons souvent en termes de cause à effet - si nous comprenons pourquoi quelque chose s'est passé, nous pouvons changer notre comportement pour améliorer les résultats futurs.
En d'autres termes, notre objectif est d'essayer d'apprendre la causalité à partir des données (quelle était la cause et quel était l'effet). Comme mentionné précédemment, dans de nombreux cas d'utilisation, la corrélation suffit jusqu'à présent. Cependant, l'inférence causale nous permettrait d'aller plus loin et de comprendre ce qui se passerait si nous décidions de modifier certaines des hypothèses sous-jacentes de notre modèle.
Comprendre la cause et l'effet rendrait les systèmes d'IA existants plus intelligents et plus efficaces. Par exemple, "pensez à un robot qui comprend que laisser tomber des objets les fait casser n'aurait pas besoin de jeter des dizaines de vases sur le sol pour voir ce qui leur arrive".
De plus, la capacité à comprendre la causalité nous aiderait à créer des modèles économiques ainsi que de nouvelles startups spécialisées dans l'aide aux entreprises pour mieux comprendre leurs données. La causalité nous aiderait ainsi à identifier de nouvelles pistes basées sur des éléments auxquels nous n'avons jamais pensé.

Raisonnement contrefactuel
Le raisonnement contrefactuel est un type de raisonnement qui consiste à changer – de façon mentale – l'issue d'un évènement, en modifiant l'une de ses causes.
Dans l'apprentissage automatique interprétable, les explications contrefactuelles peuvent être utilisées pour expliquer les prédictions d'instances individuelles. L'"événement" est le résultat prédit d'une instance, les "causes" sont les valeurs de caractéristiques particulières de cette instance qui ont été entrées dans le modèle et ont "causé" une certaine prédiction [36].
Prenons deux cas illustratifs :

1. Robert fait une demande de prêt et est rejeté par le logiciel bancaire (basé sur l'apprentissage automatique). Il se demande pourquoi sa demande a été rejetée (la banque n'a généralement aucun intérêt à la transparence ...) et comment il pourrait améliorer ses chances d'obtenir un prêt. La question du "pourquoi" peut être formulée comme un contrefactuel : quelle est la plus petite modification des caractéristiques (évolution professionnelle, revenu, prêts en cours, âge, ...) qui changerait la prédiction de rejet à approbation ?
2. Lucie veut louer son appartement, mais elle ne sait pas à quel prix le louer. Elle décide donc d'entraîner un modèle d'apprentissage automatique pour prédire le loyer. Après avoir saisi tous les détails concernant la taille, l'emplacement, les facilités, les accès (transports, vies rapides, commerces de proximité), etc., le modèle lui indique qu'elle peut facturer 800 EUR. Elle s'attendait à 900 EUR ou plus, mais elle fait confiance à son modèle et décide de jouer avec les valeurs des caractéristiques de l'appartement pour voir comment elle peut améliorer la valeur de celui-ci. Elle découvre que l'appartement pourrait être loué pour plus de 900 EUR s'il était plus grand de 10 m2. Une connaissance intéressante, mais non exploitable, car elle ne peut pas agrandir son appartement. Enfin, en modifiant uniquement les valeurs des caractéristiques qu'elle contrôle (cuisine intégrée oui/non, animaux domestiques autorisés oui/non, type de sol, etc.), elle découvre que si elle autorise les animaux domestiques et installe des fenêtres mieux isolées, elle peut facturer 900 EUR. Lucie utilise intuitivement des contrefactuels pour modifier le montant du loyer plausible.

Les contrefactuels sont des explications conviviales pour l'homme, étant sélectifs, c'est-à-dire qu'ils se concentrent généralement sur un petit nombre de changements de caractéristiques. L'explication contrefactuelle d'une prédiction décrit la plus petite modification des valeurs des caractéristiques qui fait évoluer la prédiction vers un résultat prédéfini.
Il existe des méthodes d'explication contrefactuelle à la fois agnostiques et spécifiques à un modèle. L'exemple suivant est basé sur un jeu de données allemand sur le risque de crédit qui peut être trouvé sur la plate-forme de défis d'apprentissage automatique [4]. Les auteurs [37] suggèrent de minimiser simultanément une fonction de perte (en anglais, "loss function") à quatre critères (C_1 à C_4), en entraînant un SVM (avec noyau à base radiale) pour prédire la probabilité qu'un client présente un bon risque de crédit :

C_1 Une première exigence évidente est qu'une instance contrefactuelle produise la prédiction prédéfinie aussi fidèlement que possible. Il n'est pas toujours possible de trouver une instance contrefactuelle avec la prédiction prédéfinie. Par exemple, dans un contexte de classification avec deux classes, une classe rare et une classe fréquente, le modèle peut toujours classer une instance dans la classe fréquente. Il pourrait être impossible de modifier les valeurs des caractéristiques pour que l'étiquette prédite passe de la classe fréquente à la classe rare. Nous voulons donc assouplir l'exigence selon laquelle la prédiction du contrefactuel doit correspondre exactement au résultat prédéfini. En classification, nous pourrions rechercher un contrefactuel dans lequel la probabilité prédite de la classe rare est augmentée à 10% au lieu des 2% actuels (par exemple). La question est alors de savoir quels sont les changements minimaux à apporter aux caractéristiques pour que la probabilité prédite passe de 2% à 10 % (ou proche de 10 %).

C_2 Un contrefactuel doit être aussi similaire que possible à l'instance en ce qui concerne les valeurs des caractéristiques. La distance entre deux instances peut être mesurée, par exemple, avec la distance de Manhattan ou la distance de Gower si nous avons des caractéristiques discrètes et continues. Le contrefactuel doit non seulement être proche de l'instance originale, mais il doit également modifier le moins de caractéristiques possible. Pour mesurer la qualité d'une explication contrefactuelle selon cette métrique, nous pouvons simplement compter le nombre de caractéristiques modifiées ou, en termes mathématiques sophistiqués, mesurer la distance entre l'instance contrefactuelle et l'instance réelle.

4. Kaggle : https://www.kaggle.com/

C_3 Il est souvent souhaitable de générer des explications contrefactuelles diverses afin que le sujet de la décision ait accès à de multiples façons viables de générer un résultat différent. Dans l'exemple du prêt, une explication contrefactuelle pourrait suggérer de doubler le revenu pour obtenir un prêt, tandis qu'une autre explication contrefactuelle pourrait suggérer de déménager dans une ville voisine et d'augmenter le revenu d'un petit montant pour obtenir un prêt. On peut noter que si le premier contrefactuel est possible pour certains, le second est plus facilement réalisable pour d'autres. Ainsi, en plus de fournir à un sujet de décision différentes façons d'obtenir le résultat souhaité, la diversité permet également aux individus "divers" de modifier les caractéristiques qui leur conviennent.

C_4 Une instance contrefactuelle doit avoir des valeurs de caractéristiques qui sont probables. Cela n'aurait pas de sens de générer une explication contrefactuelle pour l'exemple du loyer où la taille d'un appartement est négative ou le nombre de pièces est fixé à 100. C'est encore mieux lorsque le contrefactuel est probable selon la distribution conjointe des données, par exemple, un appartement de 5 pièces et de 20 m2 ne devrait pas être considéré comme une explication contrefactuelle. Idéalement, si le nombre de mètres carrés est augmenté, une augmentation du nombre de pièces devrait également être proposée.

L'ensemble de données compte 522 observations complètes et neuf caractéristiques contenant des informations sur le crédit et le client. L'objectif est de trouver des explications contrefactuelles pour un client avec les valeurs de caractéristiques suivantes :

âge	sexe	emploi	logement	épargne	montant	durée	achat
58	F	non-qualifié	à titre gratuit	petite	6143	48	voiture

Le SVM prédit que la femme présente un bon risque de crédit avec une probabilité de 24,2 %. Les contrefactuels doivent répondre à la question de savoir comment les caractéristiques d'entrée doivent être modifiées pour obtenir une probabilité prédite supérieure à 50 %.
Le tableau 1 présente les dix meilleurs contrefactuels (avec $\hat{f}(x')$, la prédiction du modèle pour le contrefactuel x'). Les cinq premières colonnes contiennent les changements de caractéristiques proposés (seules les caractéristiques modifiées sont affichées), les trois colonnes suivantes montrent les valeurs des critères (C_1 est égal à 0 dans tous les cas) et la dernière colonne affiche la probabilité prédite. Tous les contrefactuels ont des probabilités prédites supérieures à 50 % et ne sont pas dominés les uns par les autres. Non dominé signifie qu'aucun des contrefactuels n'a de plus petites valeurs dans tous les critères que les autres contrefactuels. Nous pouvons considérer nos contrefactuels comme un ensemble de solutions de compromis.

âge	sexe	emploi	montant	durée	C_2	C_3	C_4	$\hat{f}(x')$
		qualifié		20	0,108	2	0,036	0,501
		qualifié		24	0,114	2	0,029	0,525
		qualifié		22	0,111	2	0,033	0,513
6		qualifié		24	0,126	3	0,018	0,505
3		qualifié		24	0,120	3	0,024	0,515
1		qualifié		24	0.116	3	0.027	0.522
3	M			24	0.195	3	0.012	0.501
6	M			25	0.202	3	0.011	0.501
30	M	qualifié		24	0,285	4	0,005	0,590
4	M		1254	24	0.204	4	0.002	0.506

TABLEAU 1 – Dix meilleurs contre-factuels.

Ils suggèrent tous une réduction de la durée de 48 mois à 23 mois minimum, certains d'entre eux proposent que la femme devienne qualifiée au lieu d'être non qualifiée. Certains contrefactuels suggèrent même de changer le sexe de la femme en homme, ce qui montre un biais sexiste du modèle. Ce changement est toujours accompagné d'une réduction de l'âge entre 1 et 30 ans. Nous pouvons également constater que, bien que certains contrefactuels suggèrent de modifier quatre caractéristiques, ces contrefactuels sont ceux qui sont les plus proches des données d'apprentissage.

Niveaux de causalité / Basé sur les travaux de Judea Pearl [35] :
D'un point de vue commercial, il convient de réfléchir aux scénarios suivants :

1. Dans un contexte de commerce électronique, nous pourrions déterminer quel facteur spécifique impacte le plus la décision d'acheter un produit. Avec ces informations, nous pourrions mieux allouer des ressources pour améliorer un KPI (Key Performance Indicator) spécifique. Nous pourrions également classer l'impact de différents facteurs sur la décision d'achat. Nous pourrions déterminer si un client donné aurait acheté un produit spécifique s'il n'avait pas acheté d'autres produits au cours des deux dernières années.
2. Dans un sens plus large, nous pourrions découvrir comment et quels impacts négatifs auraient pu être évités par une stratégie commerciale donnée ? Nous pourrions également déterminer de combien devrions-nous nous attendre à ce que nos ventes augmentent en mettant en œuvre un programme de formation spécifique pour nos développeurs d'affaires. l'impact d'un programme de formation spécifique.
3. Dans le domaine agricole, on essaie souvent de prédire si le rendement des cultures d'un agriculteur sera inférieur cette année. Cependant, en utilisant des déductions occasionnelles, il deviendra de mieux comprendre quelles mesures devons-nous prendre pour augmenter la récolte.

Au-delà de ces cas d'utilisation potentiels, le développement de plus de causalité dans l'apprentissage automatique est une étape nécessaire dans la construction d'une intelligence machine plus humaine (éventuellement de l'intelligence générale artificielle).

Solutions actuelles et futures :
A ce jour, des solutions existent. Cependant, les solutions actuelles (ex : simulations Monte Carlo, analyse de chaîne de Markov, Naïve Bayes, modélisation stochastique et quelques packages open source comme DAGitty [5]) ne sont pas à la hauteur de nos attentes en matière d'applications métiers.

Structures causales de méta-apprentissage :
En 2019, Yoshua Bengio et son équipe [38] proposent une version d'apprentissage profond capable de reconnaître de simples relations de cause à effet. Ils ont utilisé un ensemble de données qui cartographie les relations causales entre des phénomènes du monde réel, tels que le tabagisme et le cancer du poumon, en termes de probabilités. Ils ont également généré des ensembles de données synthétiques sur les relations causales.
En d'autres termes, "l'algorithme résultant forme essentiellement une hypothèse sur les variables qui sont causalement liées, puis teste comment les changements apportés aux différentes variables correspondent à la théorie".

Modélisation par équation structurelle (Structural Equation Modeling - SEM) :
L'autre approche qui mérite d'être mentionnée s'appelle la modélisation par équation structurelle. Sans trop entrer dans les détails, "les mathématiques fondamentales établies par Judea Pearl et l'évolution rapide des modèles de graphes contribuent à rendre disponibles des outils de causalité".

5. http://dagitty.net/

Réseau causal bayésien
Cette méthode estime les relations entre toutes les variables d'un ensemble de données et peut être considérée comme une véritable méthode de découverte. Il permet la découverte de plusieurs relations causales en même temps.
Fondamentalement, il en résulte une carte visuelle intuitive montrant quelles variables s'influencent mutuellement, ainsi que l'étendue de leur influence. En effet, les modèles graphiques causaux permettent de simuler simultanément de nombreuses interventions possibles. Les réseaux causaux bayésiens nécessitent beaucoup de données pour capturer toutes les variables possibles.
L'autre élément important à garder à l'esprit est que l'IA causale ne fonctionne pas dans une boîte noire. Les chercheurs peuvent vérifier le raisonnement du modèle et réduire le risque de biais.
D'un point de vue commercial, cette approche permet l'incorporation de connaissances d'experts pour contrer les limites possibles d'une approche purement data-driven. Les experts métiers peuvent vous aider : Placer des conditions sur le modèle pour améliorer sa précision, Déterminer quelles variables doivent entrer dans le modèle Aidez à comprendre les résultats contre-intuitifs.

Apprentissage par renforcement vs causalité :
L'apprentissage par renforcement (ang. reinforcement learning - RL) est une méthode d'apprentissage incrémental à l'aide d'interactions avec un environnement.
Certaines figures de proue de la communauté de l'IA pensent que la RL est intrinsèquement causale, en ce sens que l'agent expérimente différentes actions et apprend comment elles affectent les performances par le biais d'essais et d'erreurs. Ce type d'apprentissage est appelé "sans modèle" car il peut apprendre des comportements efficaces sans avoir à apprendre un modèle explicite de la façon dont le monde fonctionne.
En apprentissage par renforcement, un algorithme sans modèle est un algorithme qui n'utilise pas la distribution de probabilité de transition (et la fonction de récompense) associée au processus de décision de Markov, qui, en RL, représente le problème à résoudre.
Cependant, il s'agit seulement d'apprendre la relation causale entre les actions et la performance, plutôt que la façon dont les actions affectent directement le monde. Par exemple, cela peut impliquer d'apprendre que retourner un seau d'eau plein au-dessus d'un feu l'éteint, sans comprendre la relation entre l'eau et le feu.
Comme l'a mentionné George Lawton, "si l'agent recevait un tuyau au lieu d'un seau d'eau, il ne saurait pas quoi en faire sans apprendre à partir de zéro, car il n'a pas appris la relation causale entre l'eau et le feu". RL consiste davantage à tester une croyance pour trouver un point optimal dans l'espace de recherche.

4.2.7 Métriques associées aux concepts

Étant donné que l'explication, comme l'argumentation, peut impliquer la pondération, la comparaison ou la conviction d'un public avec des formalisations basées sur la logique des (contre) arguments [39], l'explicabilité pourrait nous amener dans le domaine de la psychologie cognitive et de la psychologie des explications [40], car mesurer si quelque chose a été compris ou exprimé clairement est une tâche difficile à évaluer objectivement. En revanche, il est possible d'évaluer objectivement dans quelle mesure les éléments fondamentaux d'un modèle peuvent être expliqués. Tout moyen de réduire la complexité du modèle ou de simplifier ses résultats devrait être considéré comme une approche XAI. L'importance de ce saut en termes de complexité ou de simplicité correspondra au degré d'explicabilité du modèle qui en résulte. Un problème sous-jacent qui n'est toujours pas résolu est que le gain d'interprétabilité apporté par ces approches XAI n'est pas toujours facile à quantifier : par exemple, une simplification du modèle peut être évaluée sur la base de la réduction du nombre d'éléments architecturaux ou du nombre de paramètres du modèle lui-même (comme c'est souvent le cas, par exemple, pour les réseaux de neurones profonds). Au contraire, l'utilisation de méthodes de visualisation ou du langage naturel dans le même but ne favorise pas une quantification claire

des améliorations obtenues en termes d'interprétabilité. La dérivation de métriques générales pour évaluer la qualité des approches XAI reste un défi ouvert qui devrait être sous les projecteurs du domaine dans les années à venir.
La littérature du domaine demande clairement un concept unifié d'explicabilité de l'IA. Pour que le domaine XAI/RAI se développe de manière cohérente, il est impératif de mettre en place un terrain d'entente sur lequel la communauté soit en mesure de contribuer avec de nouvelles techniques et méthodes. Un concept commun doit traduire les besoins exprimés sur le terrain. Il doit aussi proposer une structure commune pour chaque système de XAI. Dans la proposition d'un concept d'explicabilité de Gunning [40], l'explicabilité est définie comme la capacité d'un modèle à rendre son fonctionnement plus clair pour un public. Pour y remédier, des méthodes de type post-hoc existent. Le concept décrit dans cette synthèse permet un premier terrain d'entente et un point de référence pour soutenir une discussion fructueuse sur cette question. Il est primordial que le domaine de XAI parvienne à un accord à cet égard, combinant les différents efforts.
Une autre caractéristique clé nécessaire pour relier un certain modèle à ce concept unifié est l'existence d'une métrique. Une métrique, ou un groupe de métriques, devrait permettre une comparaison significative de l'adéquation d'un modèle à la définition d'explicable. Sans un tel outil, toute affirmation à cet égard se dilue, ne fournissant pas une base d'appui solide. Ces métriques, comme les métriques classiques (précision, sensibilité, F1, Dice, Jaccquard, Tversky ...), doivent exprimer les performances du modèle concernant un certain aspect de l'explicabilité. Certaines tentatives récentes sont à noter, concernant les métriques de XAI [2, 41]. En général, les mesures XAI devraient évaluer la qualité, l'utilité et la satisfaction des explications, l'amélioration du modèle mental du public, induites par les explications du modèle, et l'impact des explications sur les performances du modèle et sur la confiance et la confiance du public. Les techniques de mesure étudiées (liste contrôle-qualité, échelle de satisfaction des explications, méthodes d'élicitation pour les modèles mentaux, mesures de calcul pour la fidélité de l'explicateur, fiabilité des explications et fiabilité du modèle) représentent un bon début dans la direction d'évaluation des techniques XAI. Malheureusement, les conclusions tirées de ces aperçus sont alignées avec les perspectives du terrain : des métriques XAI plus quantifiables et générales sont vraiment nécessaires pour soutenir les procédures et outils de mesure existants, proposés par la communauté.
La conception d'une telle suite de métriques devrait être abordée par la communauté, dans son ensemble, avant l'acceptation du concept plus large d'explicabilité. Des efforts supplémentaires se montrent ainsi nécessaires, vers de nouvelles propositions d'évaluation des performances des techniques XAI et des méthodologies de comparaison entre les approches XAI qui permettent de les instancier à de différents contextes d'application, modèles et objectifs.

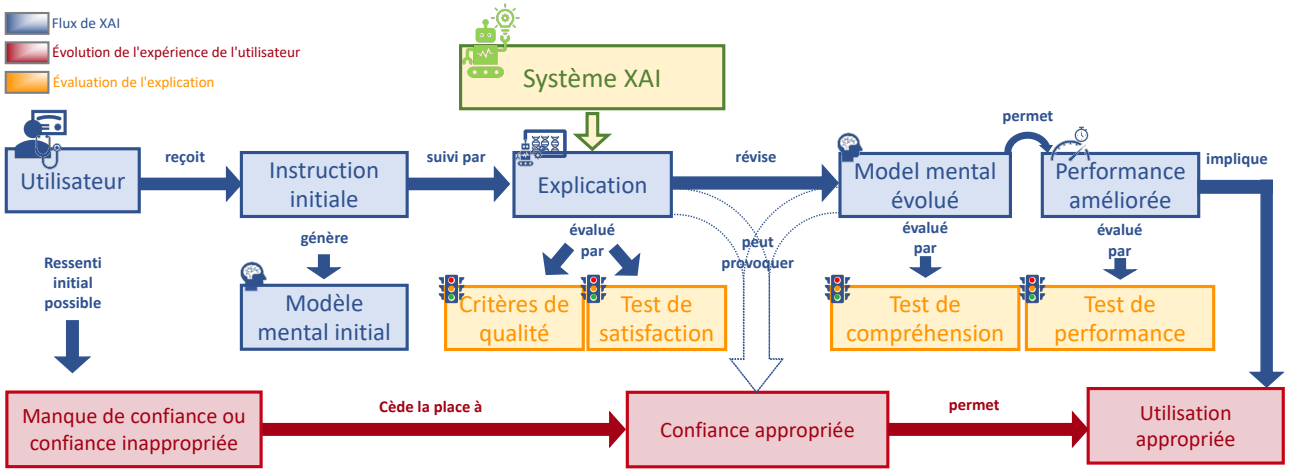

FIGURE 4 – Modèle conceptuel du processus d'explication, dans le contexte XAI [2].

La génération / création de métriques générales pour évaluer la qualité des approches XAI reste un

défi ouvert pour les années à venir.
Un des défis dans XAI consiste à établir des métriques objectives sur ce qui constitue une bonne explication. Une possibilité de réduire cette subjectivité est de s'inspirer d'expériences en psychologie humaine, en sociologie ou en sciences cognitives pour créer des explications objectivement convaincantes [42].
Des recherche existent autour des concepts et des métriques pour évaluer l'explicabilité des modèles de ML et définir les directions de recherche pour rendre les modèles d'apprentissage profond plus compréhensibles. A titre d'exemple, [2, 3] suggèrent des indicateurs comme la liste de contrôle de qualité (goodness checklist), l'échelle de satisfaction des explications (explanation satisfaction scale), les méthodes d'élicitation pour les modèles mentaux (elicitation methods for mental models), les mesures de calcul pour la fidélité de l'explicateur (computational measures for explainer fidelity) et la fiabilité de l'explication et la fiabilité du modèle (explanation trustworthiness and model reliability). Malheureusement, les conclusions tirées de ces aperçus sont alignées avec nos perspectives sur le terrain : des métriques XAI plus quantifiables et générales sont vraiment nécessaires pour soutenir les procédures et outils de mesure existants proposés par la communauté.

5 Aide à l'explication : méthodes, exemples

5.1 Méthodes générales : valeurs de Shapley

5.1.1 Définition

La valeur de Shapley a été décrite en 1953 par Shapley [43] dans le contexte de la théorie des jeux coopératifs. Elle permet, un jeu étant défini, d'évaluer la contribution moyenne d'un participant aux gains permis par les diverses coalitions, et peut ainsi servir de base à une répartition des gains entre participants fondée sur une sorte de mérite. Avant de la définir plus formellement, rappelons que le problème qui nous occupe est, en vue d'en donner une explication, celui d'attribuer une décision à certains traits élémentaires (features). Par exemple un diagnostic radiologique peut être attribué à une combinaison particulière de signes élémentaires observés au sein de l'image ; la décision de refus d'un prêt bancaire prend en compte certaines caractéristiques du demandeur. On assimile donc ces combinaisons de traits élémentaires à des coalitions, et on veut attribuer un certain mérite quantitatif quant au succès diagnostique à chacun des traits participants. Une des difficultés à laquelle se heurte l'approche statistique de ce problème est la présence possible, parmi les facteurs explicatifs disponibles, de plusieurs traits fortement corrélés, semblables à différents partenaires apportant le même avantage (par exemple la même information) à une coalition.
Un jeu coopératif à n joueurs est formalisé par l'ensemble $E = \{1, \dots, n\}$ ou ensemble des joueurs, par l'ensemble C des coalitions possibles, définies comme les parties non vides de E, donc au nombre de $2^n - 1$, et par une fonction v qui attribue une valeur, nombre réel strictement positif, à chaque coalition. La valeur de Shapley est elle aussi une fonction, mais attribue une valeur à chaque joueur i pour le jeu défini par v : on la note alors $\phi_i(v)$. Shapley a montré qu'elle est l'unique fonction satisfaisant certains axiomes (afin de traduire la sémantique recherchée) et a donné son expression :

$$\phi_i(v) = \sum_{S \in C} \frac{(s-1)!(n-s)!}{n!}[v(S) - v(S \setminus \{i\})]$$

où, pour chaque coalition S, s est le nombre d'éléments de S et $S \setminus \{i\}$ est la coalition obtenue en excluant i de S. Cette expression est donc une somme pondérée des avantages apportés aux coalitions par la présence de i. On peut l'interpréter comme une espérance en vérifiant que les poids sont bien égaux aux probabilités que de tels avantages surviennent du fait de i dans la situation suivante (rothman). Supposons que les n joueurs fassent la queue, dans un ordre aléatoire équidistribué, à l'entrée d'une pièce où ils entrent un par un. Chaque joueur qui entre étend la coalition des joueurs

déjà entrés à une coalition S et incrémente le gain des joueurs de la pièce de $v(S) - v(S \setminus \{i\})$. Soit S une coalition comprenant le joueur i, les files qui permettent à i de trouver précisément les membres de $S \setminus \{i\}$ à son entrée dans la pièce sont constituées des $s-1$ membres de $S \setminus \{i\}$, suivis de i, lui-même suivi des $n-s$ joueurs n'appartenant pas à S. Le nombre de permutations de la file qui respectent ces conditions est donc $(s-1)!(n-s)!$ et la probabilité pour i de former S avec le gain marginal $[v(S) - v(S \setminus \{i\})]$ est bien $(s-1)!(n-s)!/n!$.
La valeur de Shapley et son estimation fournissent une méthode agnostique d'explication très utilisée en pratique et disponible dans plusieurs cadres de développement d'apprentissage automatique, par exemple avec la bibliothèque Python SHAP [43]. La section suivante développe un exemple de son application au sein de scikit-learn.

5.1.2 Exemple de mise en oeuvre : classification des patients sur la base de donnée OASIS longitudinal – maladie d'Alzheimer

Dans cet exemple, nous utiliserons le jeu de données public (Open Access Series of Imaging Studies – OASIS), dont les caractéristiques ont été extraites à l'aide de données IRM longitudinales. Pour chaque patient, nous avons des informations comme le sexe, l'âge, années d'études (EDUC) et le statut socio-économique (SES), ainsi que des caractéristiques médicales : i) MMSE : mini examen de l'état mental, ii) CDR : évaluation clinique de la démence, iii) eTIV : volume intracrânien total estimé, iv) nWBV : volume cérébral total normalisé, v) ASF : facteur d'échelle de l'atlas.
Tout d'abord, nous divisons le jeu de données en deux parties, la première pour l'entraînement d'un classificateur (forêt aléatoire –bibliothèque python scikit-learn) qui vise à prédire si un patient donné est atteint de la maladie d'Alzheimer. La seconde partie est gardée cachée et sera utilisée pour tester les performances du classificateur en situation réelle.
Après l'entraînement, le classificateur donne un score F1 de 87% sur les données de test, lorsqu'il est comparé au diagnostic de l'expert humain (CDR).
D'après les résultats du test, ce modèle est capable d'identifier la maladie d'Alzheimer. Cependant, aucune indication n'est donnée sur la manière dont la décision de classification est prise (boîte noire). Pour rendre l'approche plus traçable et explicable, une des façons est de mesurer quantitativement la contribution de chaque caractéristique dans la décision finale. Nous allons utiliser les valeurs de Shapley (bibliothèque python [44]) pour évaluer l'importance des caractéristiques pour la classification des patients (voir Figure 5).
Nous constatons, sur la base de l'analyse des caractéristiques (Figure 5), que le MMSE est la caractéristique la plus importante pour le modèle de forêt aléatoire. Le fait qu'un patient ait un score MMSE faible est fortement corrélé à la démence. Il en va de même pour la deuxième caractéristique importante (nWBV) : avoir un volume du cerveau entier plus faible est également corrélé à la démence. Comme nous pouvons le voir, cette approche peut non seulement nous aider à comprendre les décisions d'un modèle, mais peut aussi donner aux médecins des indications pour les aider à mieux comprendre la démence. Cette approche peut être appliquée aux données individuelles de chaque patient pour une aide au diagnostic pour les médecins.

5.2 Méthodes de visualisation

Afin de comprendre les modèles complexes d'apprentissage profond, les scientifiques cherchent à cerner la manière dont le modèle produit une décision donnée. Ceci est essentiel, en particulier, pour les technologies d'analyse d'images (surveillance vidéo, pilotage de véhicules autonomes, biométrie, analyse d'images biomédicales, etc.).
Une approche de base consiste à utiliser une fenêtre d'occlusion coulissante qui parcourt une image d'entrée. L'occlusion [45] qui influence le score de confiance du modèle se verra ainsi accorder une grande importance lors de la production d'une carte de sensibilité [6], correspondant à la mise en valeur

6. D'autres noms utilisés dans la littérature pour le carte de sensibilité sont : carte de chaleur (heatmap), carte de

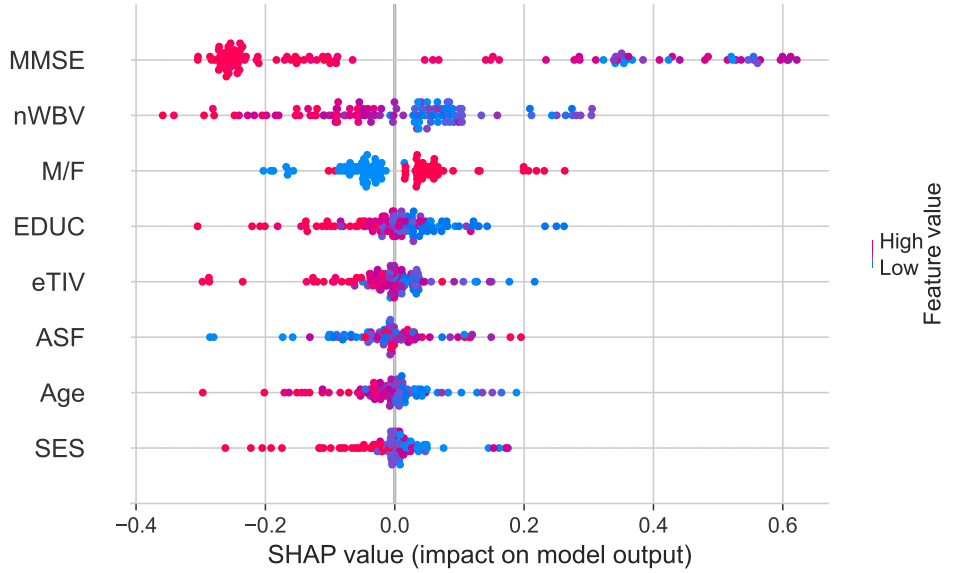

FIGURE 5 – Évaluation de l'importance des caractéristiques (jeu de données OASIS) à l'aide de la bibliothèque SHAP appliquée au modèle de forêt aléatoire. Les caractéristiques sont classées par importance (MMSE est la plus importante, SES est la moins importante). Les points rouges sont les valeurs utilisées pour classer un patient comme atteint de démence, les points bleus correspondant à des patients sains.

des parties les plus saillantes de l'image. Malgré cette approche simpliste, cette méthode donne des résultats intéressants. Cependant, elle consomme beaucoup de temps et de ressources, en raison de son processus itératif exhaustif.
Les scientifiques cherchent à rendre accessible la boîte noire que constitue le modèle d'apprentissage profond, afin de mieux comprendre ses prédictions. Cela a donné naissance à toute une famille de méthodes appelées "Class-activation mapping" (CAM) [46], famille que nous introduisons dans ce qui suit.

5.2.1 Méthodes d'activation

La cartographie d'activation de classe (CAM : class activation mapping) est une méthode de visualisation des caractéristiques pour n'importe quel réseau de neurones convolutif (CNN). Pour l'utiliser, l'architecture doit appliquer le global average pooling (GAP) aux cartes de caractéristiques convolutionnelles finales. Le GAP sera suivi d'une couche, entièrement connectée, destinée à produire les prédictions.
CAM [46] a été la première méthode à être largement utilisée par la communauté scientifique, en raison de sa capacité à donner à l'utilisateur la possibilité de voir la ou les régions (dans l'image d'entrée) utilisée pour la prédiction d'une classe. L'inconvénient de cette approche est que pour tous les CNN qui n'ont pas la couche GAP ne peuvent pas en profiter. L'introduction de cette couche GAP induit des modifications dans l'architecture du réseau, modification nécessitant un réapprentissage. En outre, CAM utilise la couche entièrement connectée pour classifier l'image, ce qui expose l'approche à un risque non négligeable d'erreurs. En effet, en passant de la représentation des caractéristiques (données par la couche la plus profonde de la convolution) aux couches entièrement connectées (aplatissement des caractéristiques), nous perdons l'information spatiale, ce qui augmente le risque de défaillance du classifieur.
Pour résoudre ce problème, EigenCam [47] propose de se baser uniquement sur les activations de

salliance (salliency map), carte d'activation (activation map). Afin d'éviter toute confusion, seules l'intitulé "cartes d'attention" sera utilisé dans la suite de l'article.

toutes les cartes de caractéristiques sans utiliser le classifieur. Comme il existe de nombreuses cartes et que toutes ne seront pas nécessairement activées, EigenCam utilise l'analyse en composantes principales (ACP) sur l'espace des caractéristiques, non seulement comme mécanisme de filtrage, mais aussi pour séparer les objets saillants dans une image donnée (ç.à.d. la première image propre contient les activations de la classe d'objets la plus imposante de l'image).
De plus, cette approche peut fonctionner sur n'importe quel modèle d'apprentissage (approche agnostique), ce qui constitue un avantage majeur par rapport au CAM classique. Cependant, cette approche perd la discrimination de classe. En effet, nous ne pouvons pas obtenir une carte d'activation sur une classe de notre choix. Par exemple, dans le cas d'un petit chat en présence d'un gros chien dans la même image, la première image propre d'EigenCam se concentre uniquement sur le gros chien et nous n'avons aucune assurance que la deuxième image propre contiendra le petit chat.
ScoreCam [48] est une autre méthode de visualisation basée sur les activations du modèle. Cette méthode évalue chaque carte d'activation sur le score de classification, en l'utilisant comme un masque d'occlusion. En résumé, les cartes importantes pour un score de classification d'objet donné auront plus d'importance, donc ces activations seront prépondérantes dans la carte d'activation finale. L'AblationCam [49] est également une approche similaire, elle supprime une carte d'activation à la fois et mesure la baisse du score de classification. Plus la baisse est importante, plus la pondération de la carte d'activation est élevée. Ces approches sont inspirées des approches basées sur l'occlusion, héritant ainsi de leur processus itératif exhaustif. Cela rend cette approche gourmande en ressources et ses performances sont directement influencées par le nombre de cartes d'activation et la taille de l'image d'entrée.

5.2.2 Méthodes du gradient

Comme la méthode CAM exige la présence de la couche GAP à l'intérieur du réseau, son potentiel d'utilisation se trouve limité. Afin de résoudre ce problème, Grad-CAM a été conçu comme une généralisation de CAM. Cette approche ne nécessite pas d'apprentissage ni de modification de l'architecture.
Grad-CAM est une méthode qui nous permet de voir l'endroit sur lequel le modèle focalise sont attention, afin de prendre ses décisions de classification. Ceci correspond à une forme d'attention post-hoc, signifiant qu'il s'agit d'une méthode produisant des cartes d'attention à partir d'un réseau de neurones déjà entraîné (une fois l'apprentissage automatique terminé et les paramètres du modèle, définis). Les couches de convolution conservent naturellement les informations spatiales qui sont perdues dans les couches entièrement connectées. Selon l'article fondateur [50], on peut s'attendre à ce que les dernières couches de convolution présentent un meilleur compromis entre la sémantique de haut niveau et les informations spatiales détaillées. Donc, l'idée est d'exploiter l'information spatiale qui est préservée par les couches de convolution, afin de comprendre quelles parties d'une image d'entrée étaient importantes pour une décision de classification.
Cependant, il s'avère que Grad-CAM ne parvient pas à localiser correctement les objets dans une image si celle-ci contient plusieurs occurrences de la même classe. Une autre conséquence de la moyenne non pondérée des dérivées partielles (intervenant dans le calcul des pondérations des cartes d'attention) que, souvent, la localisation ne correspond pas à l'objet entier, mais à des morceaux et des parties de celui-ci. Pour surmonter ce problème, GradCAM++ [51] a été conçu en utilisant les gradients de second ordre afin d'obtenir des activations indépendantes de la taille de l'objet.
Une autre variante de GradCAM appelée XGradCAM [52]. Cette approche consiste à mettre à l'échelle les gradients par les cartes d'activations normalisées. Ce qui s'est avéré donner de meilleurs résultats en comparaison avec les autres méthodes.

5.2.3 Synthèse des méthodes visuelles

Méthodes	**Description**	**Limitations**	**Vitesse d'exécution**
Occlusion [45] (2013)	L'approche est basée sur l'occlusion d'une partie de l'image d'entrée avec un patch (noir/blanc/gris), dans le but de voir comment cela affecte le score de classification du réseau : plus la classification est dégradée, plus la partie occultée est importante pour classifier l'objet. Ceci est fait de manière itérative et exhaustive (sur l'ensemble de l'image), afin de générer une carte globale de sensibilité.	Un seul patch glissant n'est pas suffisant pour déterminer la réaction réelle du modèle (exemple : sensibilité du modèle à une paire de régions). L'utilisation de plus d'un patch augmentera le nombre d'itérations de manière exponentielle. De plus, dans chaque cas d'utilisation, des paramètres doivent être réglés (par exemple, la taille, le rembourrage, le chevauchement des patches, etc.).	GPU : 83.7062s ± 0.2522s. CPU :40-50 min (taille : patch=32, stride=14)
CAM [46] (2016)	Cette approche génère les poids d'activation en utilisant la couche GAP. Ces poids sont utilisés pour pondérer les cartes de caractéristiques finales (par combinaison linéaire pour une classe donnée).	Une modification doit être apportée aux réseaux qui n'ont pas de couche GAP suivie d'une couche entièrement connectée. Par conséquent, dans ce cas, le réseau doit être ré-entraîné.	GPU : 0.0827s ± 0.0043s. CPU : 1.7095s ± 0.0681s (bibliothèque :torchcam [53])
GradCam [50] (2017)	Cette approche génère les poids d'activation pour une classe donnée, en calculant les gradients moyens pour chaque carte de caractéristiques correspondante. La carte résultante est passée par la fonction d'activation ReLU, puis elle est mise à l'échelle afin que sa taille corresponde à celle de l'image d'entrée.	La surface d'un objet (ou partie d'un objet) dans les cartes de caractéristiques, influe directement sur les résultats, car avoir une petite surface (par exemple, avoir plusieurs objets identiques dans une image) signifie un score faible. Par conséquent, cela affecte la pertinence de la carte d'attention du réseau.	GPU : 0.5821s ± 0.0194s. CPU : 3.6152s ± 0.0210s
GradCam++ [51] (2018)	Identique à GradCam mais elle utilise des gradients de second ordre. Cette approche est considérée comme une généralisation de GradCam.	GradCam++ obtient de meilleures performances par rapport à GradCam, mais, malheureusement, sa capacité à distinguer les classes est réduite (plus de détails dans l'annexe de [52]).	GPU : 0.5819s ± 0.0244s. CPU : 3.6429s ± 0.0191s
EigenCam [47] (2020)	Au lieu de faire appel au gradient, cette approche utilise la décomposition en valeurs singulières sur les cartes de caractéristiques, afin de calculer les composantes principales. Le premier vecteur propre est utilisé pour générer une carte d'activation (en principe, cela coïncide avec la classe dominante présente dans l'image d'entrée).	La principale limite de cette approche est qu'elle ne prend pas en compte les couches du réseau après la dernière couche convolutionnelle. Vu que l'approche prend la première "image propre" (en anglais "eigen image") des cartes d'activation après l'ACP, cela entraîne une perte de la discrimination de classe.	GPU : 0.3697s ± 0.0034s. CPU : 0.5816s ± 0.0072s

ScoreCAM [48] (2020)	Cette approche n'utilise pas de gradient. Elle est basée sur la perturbation de l'image d'entrée en utilisant les parties activées des cartes de caractéristiques (64 cartes de caractéristiques signifient 64 variantes d'images d'entrée), ceci tout en mesurant comment la perturbation influence le score de probabilité de classification du réseau (de manière itérative).	Le temps de calcul est proportionnel au nombre de cartes de caractéristiques finales que nous avons (2048 cartes pour Resnet50 [54] correspondent à 2048 boucles). Cela signifie que l'approche a une faible efficacité d'inférence, malgré l'utilisation d'une implémentation "inférence par lots" de la méthode sur un GPU. Cette approche consomme beaucoup de mémoire de GPU pendant l'inférence.	GPU : 29.7567s ± 0.2631s. CPU : environ 14-15min (GPU/CPU calcul en utilisant une taille de lot de 32)
Ablation-CAM [49] (2020)	Cette approche n'utilise pas de gradient : elle utilise l'ablation. Elle met à zéro les cartes de caractéristiques et mesure leur influence sur la sortie du réseau. Cette influence quantifie l'importance de chaque carte de caractéristiques, correspondante. La carte d'attention est calculée en additionnant les cartes de caractéristiques multipliées par les poids d'influence correspondants. Ensuite, le combo " la fonction ReLU + upsampling " est appliqué, afin que la taille de la carte coïncide avec la taille de l'image.	Étant donné que cette approche utilise un processus itératif (plusieurs passages successifs : 2048 boucles pour Resnet50 [54]), elle prend plus de temps. Cela réduit drastiquement son efficacité d'inférence. Malgré l'utilisation de la parallélisation (avec plus de ressources informatiques), l'approche présente toujours une faible efficacité d'inférence (temps de calcul élevé).	GPU : 26.9703s ± 0.1641s. CPU : environ 14-15min (GPU/CPU calcul en utilisant une taille de lot de 64)
XGradCam [52] (2020)	Cette approche normalise les activations des cartes de caractéristiques, puis les pondère en utilisant les gradients. À la fin, la fonction d'activation ReLU est appliquée, suivie d'un sur-échantillonnage pour générer une carte ayant la même taille que l'image d'entrée. XGradCam est identique à CAM pour les réseaux avec couche GAP (GAP-CNNs), mais il peut être appliqué à n'importe quel modèle CNN. Ainsi, XGrad-CAM est considéré comme une généralisation de CAM.	Les gradients sont bruyants par nature, et la plupart du temps, seuls les gradients positifs sont utilisés, ce qui entraîne une approximation dont il faut tenir compte. Mais plus important encore, les approches basées sur le gradient reposent exclusivement sur l'espace des caractéristiques et le classifieur (couches entièrement connectées). En d'autres termes, elles supposent que le modèle est parfait, ce qui peut s'avérer fatal lorsque la classification n'est pas assez précise, et par conséquent, entraîner l'échec complet de l'approche.	GPU : 0.5698s ± 0.0257s. CPU : 3.6353s ± 0.0136

TABLEAU 2 – Brève description des méthodes de visualisation avec leurs limites et leurs principaux inconvénients.L'évaluation de la vitesse d'exécution a été réalisée en utilisant la configuration suivante (valeurs moyennes sur 10 tests consécutifs) : **(GPU)** Nvidia Tesla V100S-PCIE-32GB **(CPU)** 10 coeurs Intel(R) Xeon(R) Gold 6126 CPU @ 2.60GHz **(RAM)** 20GB de RAM. Le modèle utilisé est le Resnet50 [54] (pré-entraînée sur ImageNet [55]), avec l'image d'entrée (1024x683x3) et sans redimensionnement (bibliothèques python [56, 53]).

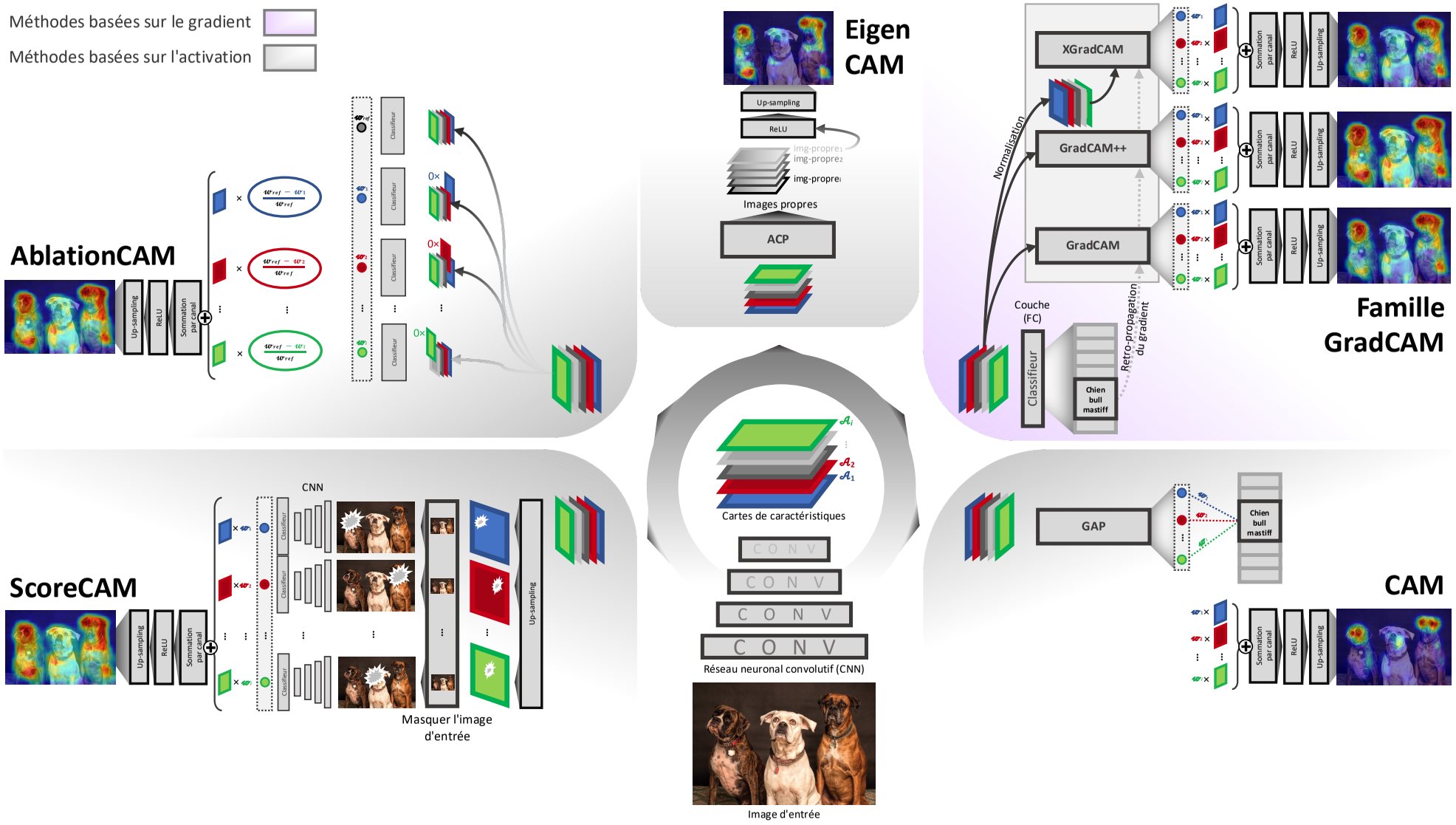

FIGURE 6 – Les méthodes de visualisation des réseaux de neurones profonds.

5.2.4 Exemple de recherche d'explication par visualisation

Grâce aux progrès majeurs de la physique et de l'informatique, les scientifiques ont désormais à leur disposition de puissants microscopes, ouvrant de nouvelles possibilités qui leur permettent de voir les maladies et leurs causes, plus près que jamais.
Les biologistes et les cliniciens ont désormais accès à davantage de détails cellulaires, ce qui fait de la quantification de l'imagerie cellulaire, un domaine à part entière.
En raison de l'importance de la compréhension des cellules, les tâches de base – telles que le comptage des cellules – ou les tâches plus complexes – telles que la séparation et le suivi des cellules – , sont toutes considérées comme des tâches fondamentales dans les études biomédicales. En effet, l'utilisation des méthodes manuelles traditionnelles, induit une perte de temps importante pour les chercheurs, ingénieurs et utilisateur.
Afin de résoudre ce problème et d'automatiser ces tâches répétitives, la vision par ordinateur et le traitement d'images jouent un rôle important. De plus, grâce aux récentes avancées dans le domaine de l'apprentissage automatique et de l'intelligence artificielle en général, les chercheurs construisent des outils qui permettent d'atteindre des performances similaires à celles des humains, avec une rapidité remarquable.
En effet, ceci est essentiel pour aider les biologistes et les cliniciens à obtenir une quantification de haute qualité, afin de réaliser des études rapides, non biaisées, efficaces et reproductibles. Par ailleurs, cela permet de répondre à des questions fondamentales sur les maladies, afin de mieux les comprendre, mais aussi pour obtenir des rapports automatisés sur les patients, consolidant ainsi l'action des médecins.
Pour de telles tâches, U-Net et ses variants sont largement utilisés et considérés comme l'architecture/modèle de référence en matière de segmentation sémantique pour l'imagerie biomédicale 2D/3D.
Même si cette approche offre une efficacité impressionnante, elle est considérée comme opaque pour les décisions à fort enjeu, notamment par le manque de détails sur la façon dont les prédictions sont faites. Ceci est considéré comme la principale limitation de la plupart des modèles d'apprentissage profond et de l'apprentissage automatique, dans son ensemble.

Instanciation à la segmentation cellulaire en contraste de phase

Afin d'illustrer ces notions, nous présentons un exemple de segmentation cellulaire, l'une des tâches les plus complexes de l'analyse d'images biomédicales. L'objectif est d'expliquer visuellement, étape par étape, les prédictions du modèle U-Net et de l'une de ses variantes.
Le contexte de l'étude est la segmentation des cellules microgliales (cellules immunitaires du cerveau et de la moelle épinière) dans leur dynamique, afin de fournir une aide aux biologistes dans leur compréhension des maladies neurodégénératives.
S'agissant de cellules immunitaires, toute utilisation d'objets étrangers fluorescents modifierait le comportement des cellules, n'étant donc pas souhaitable. Dans ce contexte in vivo, la microscopie à contraste de phase sans agent fluorescent est alors choisie pour capturer les mouvements des cellules microgliales.
La figure 7 illustre la difficulté de l'étude, vue la forme irrégulière des cellules microgliales, ainsi que leur faible contraste par rapport à l'arrière-plan.

Apprentissage supervisé du U-Net et de sa variante basée sur l'attention

Nous avons préparé une soixantaine d'images 7 et leurs vérités de terrain, minutieusement annotées (manuellement) et validées par un expert biologiste. Ceci sera utilisé comme notre jeu de données d'entraînement. Afin d'amplifier notre ensemble de données, nous avons utilisé des techniques de base d'augmentation des données (telles que la rotation et les retournements).

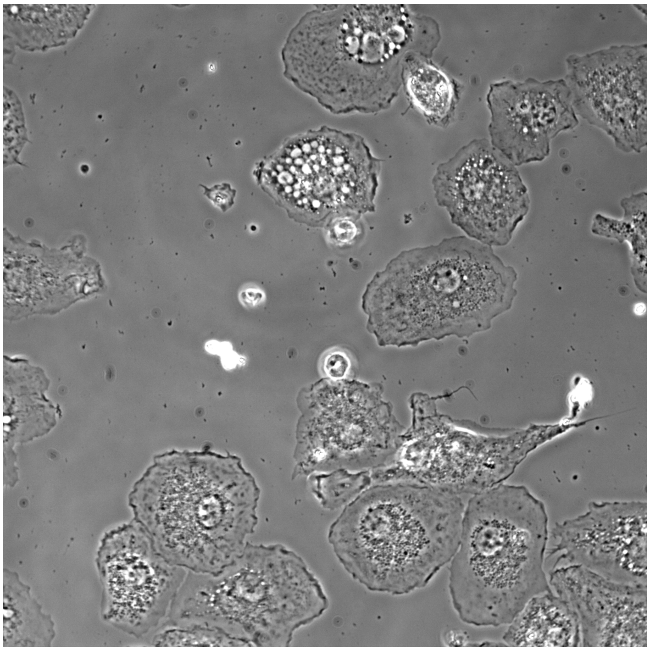

FIGURE 7 – Cellules microgliales enregistrées par microscopie à contraste de phase.

Comme l'entraînement des modèles d'apprentissage profond est stochastique, le U-Net classique et l'Attention-U-Net ont été entraînés dans les mêmes conditions (Tableau 3).

Modèle	DICE [%]
Attention-UNET	94,33% $\pm$ 0,5318
UNET	94,17% $\pm$ 0,7152

TABLEAU 3 – Résultats sur trois entraînements consécutifs en utilisant une graine aléatoire différente à chaque fois pour le modèle U-Net et sa variante Attention-U-Net.

Afin de visualiser l'effet du mécanisme d'attention, les cartes d'activation ont été enregistrées pendant la phase d'entraînement des deux modèles (U-Net et Attention-UNet). Les mêmes hyper-paramètres ont été utilisés pour les deux modèles, ainsi qu'une graine aléatoire commune. Après 1400 itérations, les cartes d'activation moyennes ont été extraites (voir figure 8). Qualitativement, on constate que le mécanisme d'attention joue un rôle majeur en filtrant le bruit de fond et donne plus d'importance aux cellules et à leurs contours (connexion avec le fond et/ou d'autres cellules). Ces cartes d'attention ont un grand potentiel et peuvent être utilisées pour expliquer le comportement des modèles d'apprentissage profond. En les utilisant, nous pouvons déterminer si un modèle d'apprentissage profond donné apprend des caractéristiques génériques/utiles pour un cas d'utilisation, ce qui permet de déterminer si le modèle est digne de confiance après évaluation par les experts. En outre, cela peut aider les scientifiques sur le plan méthodologique afin d'améliorer les modèles en analysant/comprenant les cas où le modèle ne fournit pas de bons résultats, de sorte que le modèle puisse être expliqué ainsi qu'améliorer ses performances, en même temps.

Confirmation : quantification de l'influence du mécanisme d'attention sur le modèle

Afin de quantifier le rôle du mécanisme d'attention et de mieux le comprendre, la variance du Laplacien de l'image est utilisée pour quantifier le niveau de détail, c.-à-d., une valeur élevée signifie qu'il y a beaucoup de détails dans l'image et vice versa. Il s'agit d'une évaluation objective et elle sera appliquée à l'ensemble du second bloc du décodeur (bloc de convolution pour UNet et le block d'attention pour l'Attention-UNet). La carte d'attention résultante est masquée en ne conservant que les informations de l'arrière-plan. Ainsi, on quantifie le bruit de fond.

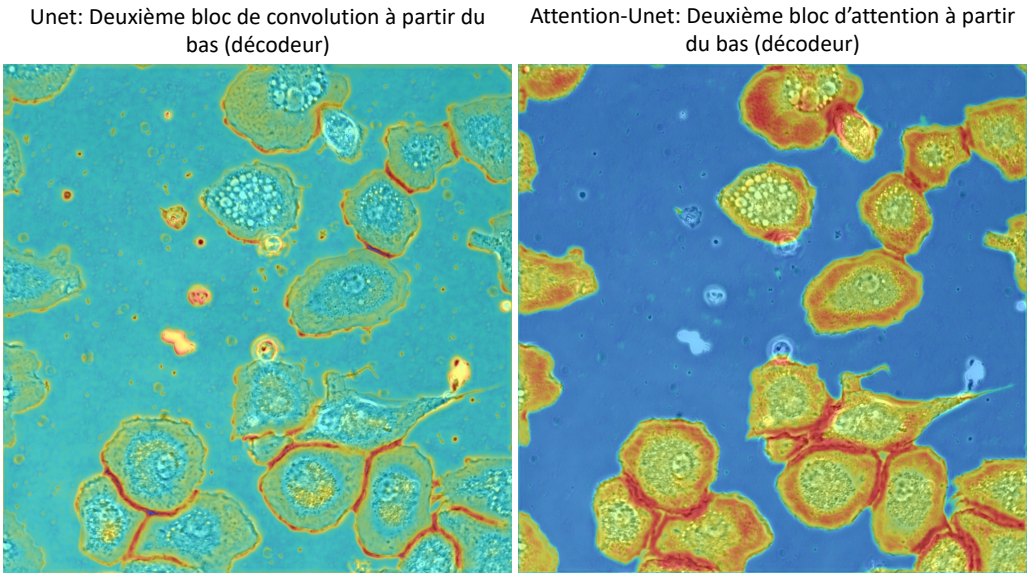

FIGURE 8 – Cellules microgliales enregistrées par microscopie à contraste de phase.

Dans la figure 9, on observe l'évolution de l'entraînement tout en quantifiant le niveau de bruit de fond (variance de laplacien appliqué ou fond de l'image), on remarque que l'Attention-UNet prend plus de temps au début pour adapter ses blocs d'attention (bruit de fond élevé au début) par rapport à l'UNet. Pourtant, après 25000 itérations, l'Attention-UNet parvient à filtrer le bruit de fond mieux que l'UNet, ce qui confirme les résultats qualitatifs visuels présentés dans la figure 8.

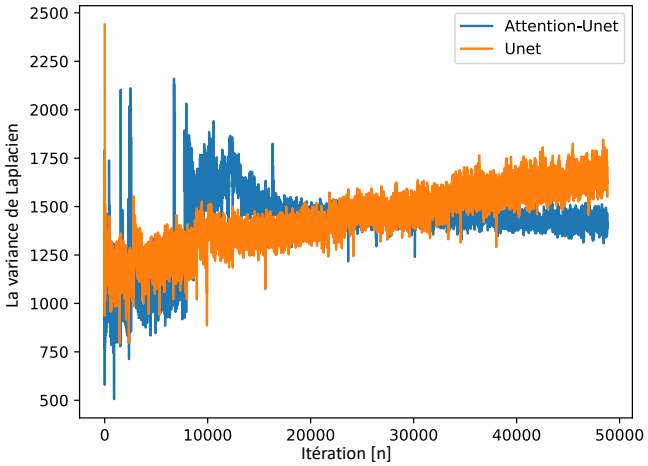

FIGURE 9 – L'évolution de la variance du Laplacien appliqué au fond de l'image pour les deux modèles U-Net et Attention-Unet (pendant l'entraînement). Ceci a été effectué sur le deuxième bloc de convolution du décodeur (une variance élevée signifie un bruit de fond élevé).

6 Perspectives

Nous proposons notre vision concernant la feuille route XAI/RAI dans un avenir proche. Deux phases sont à prendre en considération : une première phase 10 prenant en compte le retour d'expérience REX consistant, nécessitant une supervision des experts pour une intégration de nouvelles explications dans l'outil d'apprentissage automatique (en occurrence, profond) et la deuxième phase 11, dédiée à un fonctionnement quasi-autonome du système XAI (avec comme seule contrainte, la vérification régulière de la consistance, de la cohérence et de l'actualité des connaissances utilisées par

le système.
Le flux existant, le flux étendu, les contraintes de lexique / référentiel et l'explication à la demande sont mises en évidence dans les deux phases.
La première phase correspond à une (re)configuration du modèle sous le contrôle de l'expert, via une consolidation de la base d'explications.
La deuxième phase et la phase d'exploitation autonome du système XAI, en intervenant via l'interface utilisateur, pour mettre à jour directement le modèle et générer une meilleure explication.
La configuration avec intégration continue du REX constitue un verrou technologique majeur pour la communauté de l'IA. En effet, l'intégration continue, robuste du REX dans les outils d'IA/DL existants sera certainement à l'origine d'une nouvelle génération de modèles d'IA/ML/DL.

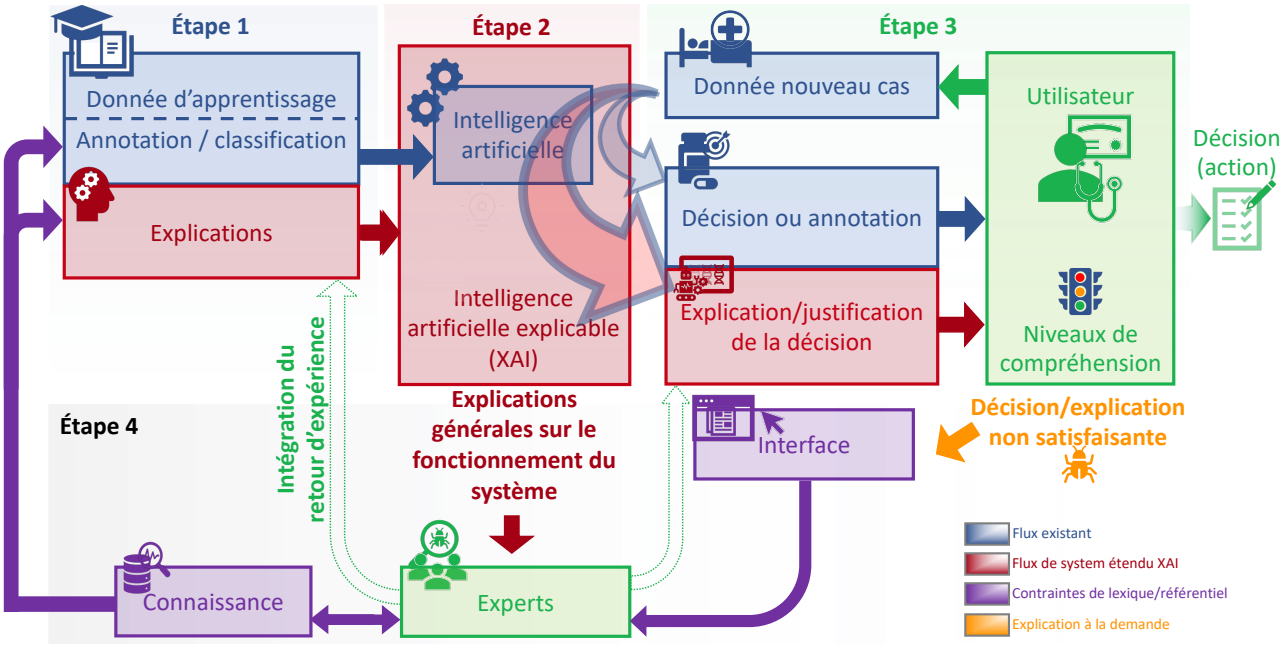

FIGURE 10 – Pipeline d'intégration et d'utilisation de l'intelligence artificielle explicable : i) phase d'apprentissage (et de ré-apprentissage) nécessitant l'appui d'experts ; intégration continue du retour d'expérience (REX) avec intégration et consolidation de nouvelles connaissances.

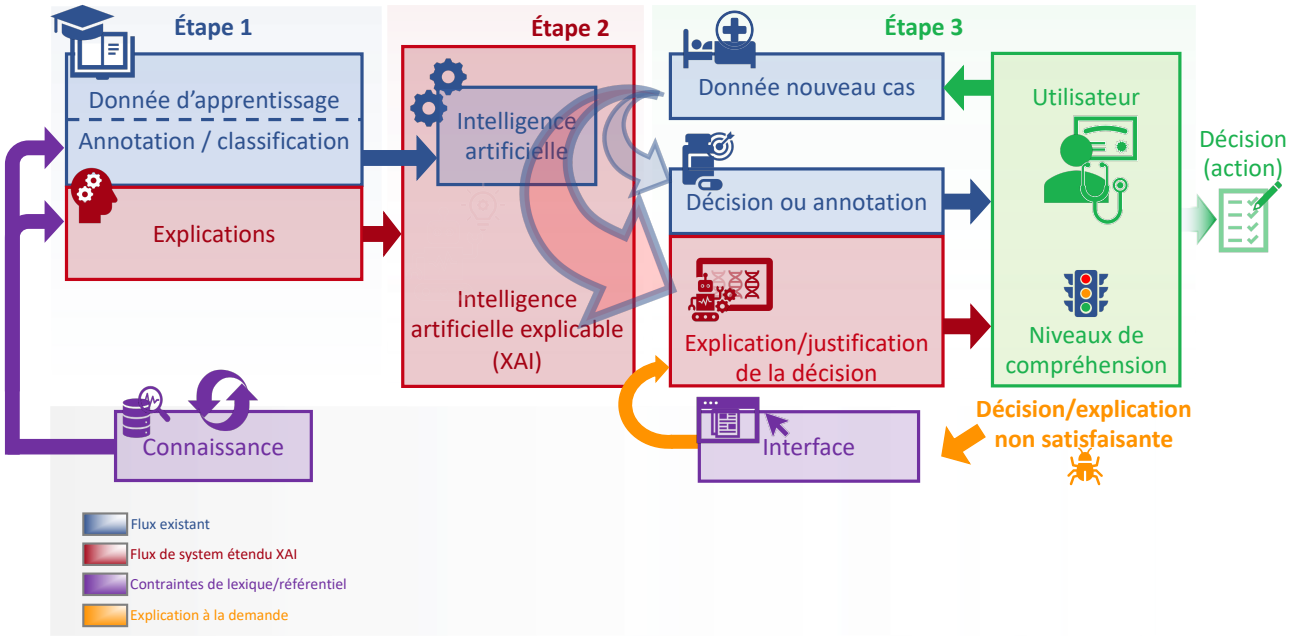

FIGURE 11 – Pipeline d'intégration et d'utilisation de l'intelligence artificielle explicable (instanciation au domaine biomédical et de la santé) : ii) phase d'exploitation/génération autonome d'explications avec apprentissage continu, par renforcement (intégration automatique de REX).

Remerciements

"Les données ont été fournies [en partie] par OASIS : Longitudinal : Principal Investigators : D. Marcus, R, Buckner, J. Csernansky, J. Morris ; P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, U24 RR021382", https ://doi.org/10.1162/jocn.2009.21407.

Références

[1] Jiaxuan WANG et al. "Learning Credible Models". In : *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (juill. 2018). arXiv : 1711.03190, p. 2417-2426. DOI : 10.1145/3219819.3220070. URL : http://arxiv.org/abs/1711.03190 (visité le 16/09/2021).

[2] Robert R. HOFFMAN et al. "Metrics for Explainable AI : Challenges and Prospects". In : *CoRR* abs/1812.04608 (2018). arXiv : 1812.04608. URL : http://arxiv.org/abs/1812.04608.

[3] Sina MOHSENI, Niloofar ZAREI et Eric D. RAGAN. "A Survey of Evaluation Methods and Measures for Interpretable Machine Learning". In : *CoRR* abs/1811.11839 (2018). arXiv : 1811.11839. URL : http://arxiv.org/abs/1811.11839.

[4] Executive Office of the PRESIDENT, J.P. M. HOLDEN et SMITH. "Preparing for the future of artificial intelligence". In : (2016).

[5] C. VILLANI et al. "Donner un sens à l'intelligence artificielle : pour une stratégie nationale et européenne". In : (2018). URL : https://hal.inria.fr/hal-01967551/document.

[6] Carsten Jung PHILIPPE BRACKE Anupam Datta et Shayak SEN. "Machine learning explainability in finance : an application to default risk analysis". In : (août 2019). URL : https://www.bankofengland.co.uk/working-paper/2019/machine-learning-explainability-in-finance-an-application-to-default-risk-analysis.

[7] Content Directorate-General for Communications Networks et European Commission Technology (European Commission). "WHITE PAPER On Artificial Intelligence - A European approach to excellence and trust". In : (fév. 2020).

[8] United States (2016) Executive Office of the President, J.P. Holdren et M. Smith. "Preparing for the future of artificial intelligence". In : (oct. 2016).

[9] Defense Advanced Research Projects Agency. "Explainable Artificial Intelligence (XAI)". In : (août 2016). URL : https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf.

[10] David Gunning et David W. Aha. "DARPA's Explainable Artificial Intelligence (XAI) Program". In : *AI Magazine* (2019).

[11] Jonathan P. How. "Ethically Aligned Design [From the Editor]". In : *IEEE Control Systems Magazine* 38.3 (2018), p. 3-4. DOI : 10.1109/MCS.2018.2810458.

[12] Kyarash Shahriari et Mana Shahriari. "IEEE standard review Ethically aligned design : A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems". In : *2017 IEEE Canada International Humanitarian Technology Conference (IHTC)*. 2017, p. 197-201. DOI : 10.1109/IHTC.2017.8058187.

[13] World Commission on the Ethics of Scientific Knowledge et Technology. "Preliminary study on the Ethics of Artificial Intelligence". In : (fév. 2019). URL : https://unesdoc.unesco.org/ark:/48223/pf0000367823.

[14] UNESCO. "Recommendation on the Ethics of Artificial Intelligence". In : (2021). URL : https://unesdoc.unesco.org/ark:/48223/pf0000380455.locale=en.

[15] Legalinstruments OECD. "Recommendation on the Ethics of Artificial Intelligence". In : (mai 2019). URL : https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

[16] Finale Doshi-Velez et Been Kim. *Towards A Rigorous Science of Interpretable Machine Learning*. 2017. DOI : 10.48550/ARXIV.1702.08608. URL : https://arxiv.org/abs/1702.08608.

[17] Helen Beebee, C. Hitchcock et P. Menzies. *The Oxford Handbook of Causation*. English. United Kingdom : Oxford University Press, déc. 2009. ISBN : 9780199279739.

[18] Mikel Olazaran. "A Sociological Study of the Official History of the Perceptrons Controversy". In : *Social Studies of Science* 26 (1996), p. 611-659.

[19] A. Philip Dawid. "Beware of the DAG!" In : *NIPS Causality : Objectives and Assessment*. 2010.

[20] Genevera I. Allen. "Handbook of Graphical Models". In : *Journal of the American Statistical Association* 115.531 (2020), p. 1555-1557. DOI : 10.1080/01621459.2020.1801279.

[21] Sandra Wachter, Brent Daniel Mittelstadt et Chris Russell. "Counterfactual Explanations Without Opening the Black Box : Automated Decisions and the GDPR". In : *Cybersecurity* (2017).

[22] Alejandro Barredo Arrieta et al. "Explainable Artificial Intelligence (XAI) : Concepts, taxonomies, opportunities and challenges toward responsible AI". In : *Information Fusion* 58 (2020), p. 82-115. ISSN : 1566-2535. DOI : https://doi.org/10.1016/j.inffus.2019.12.012. URL : https://www.sciencedirect.com/science/article/pii/S1566253519308103.

[23] Ian Goodfellow, Yoshua Bengio et Aaron Courville. *Deep Learning*. MIT Press, 2016. URL : http://www.deeplearningbook.org.

[24] M. Ekman. *Learning Deep Learning : Theory and Practice of Neural Networks, Computer Vision, Natural Language Processing, and Transformers Using TensorFlow*. Addison-Wesley Professional, 2021. ISBN : 9780137470358. URL : https://books.google.fr/books?id=W8dFzgEACAAJ.

[25] Y.L. Kergosien. “Generic sign systems in medical imaging”. In : *IEEE Computer Graphics and Applications* 11.5 (1991), p. 46-65. doi : 10.1109/38.90567.

[26] Sinno Jialin Pan et Qiang Yang. “A Survey on Transfer Learning”. In : *IEEE Transactions on Knowledge and Data Engineering* 22 (2010), p. 1345-1359.

[27] S. C. Suddarth et Y. L. Kergosien. “Rule-injection hints as a means of improving network performance and learning time”. In : *Neural Networks*. Sous la dir. de Luis B. Almeida et Christian J. Wellekens. Berlin, Heidelberg : Springer Berlin Heidelberg, 1990, p. 120-129. isbn : 978-3-540-46939-1.

[28] Rich Caruana. “Multitask Learning”. In : *Machine Learning* 28 (2004), p. 41-75.

[29] Simon Graham et al. “One Model is All You Need : Multi-Task Learning Enables Simultaneous Histology Image Segmentation and Classification”. In : *ArXiv* abs/2203.00077 (2022).

[30] Xiaotao Song et al. “A Survey of Automatic Generation of Source Code Comments : Algorithms and Techniques”. In : *IEEE Access* 7 (2019), p. 111411-111428. doi : 10.1109/ACCESS.2019.2931579.

[31] Xing Hu et al. “Summarizing source code with transferred API knowledge”. English. In : *Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI 2018*. Sous la dir. de Jerome Lang. International Joint Conference on Artificial Intelligence 2018, IJCAI 2018; Conference date : 13-07-2018 Through 19-07-2018. United States of America : Association for the Advancement of Artificial Intelligence (AAAI), 2018, p. 2269-2275. doi : 10.24963/ijcai.2018/314. url : https://www.ijcai.org/proceedings/2018/,%20https://www.ijcai.org/proceedings/2018/.

[32] Micah J Sheller et al. “Federated learning in medicine : facilitating multi-institutional collaborations without sharing patient data”. In : *Scientific reports* 10.1 (juill. 2020), p. 12598. issn : 2045-2322. doi : 10.1038/s41598-020-69250-1. url : https://europepmc.org/articles/PMC7387485.

[33] Alain Mille, Rémy Chaput et Amélie Cordier. *Une perspective historique sur l'IA explicable Document préparatoire à un tutorial AFIA juillet 2020*. Research Report. LIRIS UMR 5205 CNRS/INSA de Lyon/Université Claude Bernard Lyon 1/Université Lumière Lyon 2/école Centrale de Lyon, juill. 2020. url : https://hal.archives-ouvertes.fr/hal-03352469.

[34] C. -S. Peirce. “LA LOGIQUE DE LA SCIENCE : PREMIÉRE PARTIE : Comment Se Fixe la Croyance”. In : *Revue Philosophique de la France Et de l'Etranger* 6 (1878), p. 553-569.

[35] *The Book of Why de Judea Pearl, Dana Mackenzie - Livre audio | Scribd*. fr. url : https://www.scribd.com/audiobook/379263778/The-Book-of-Why-The-New-Science-of-Cause-and-Effect (visité le 21/12/2021).

[36] Christoph Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. https://christophm.github.io/interpretable-ml-book/. 2019.

[37] Susanne Dandl et al. “Multi-Objective Counterfactual Explanations”. In : *Lecture Notes in Computer Science* (2020). issn : 1611-3349. doi : 10.1007/978-3-030-58112-1_31.

[38] Yoshua Bengio et al. “A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms”. In : *CoRR* abs/1901.10912 (2019). arXiv : 1901.10912. url : http://arxiv.org/abs/1901.10912.

[39] Philippe Besnard et Anthony Hunter. *Elements of argumentation*. T. 47. MIT press Cambridge, 2008.

[40] David Gunning et al. “XAI - Explainable artificial intelligence”. In : *Science Robotics* 4.37 (déc. 2019). Publisher : American Association for the Advancement of Science, eaay7120. doi : 10.1126/scirobotics.aay7120. url : https://www.science.org/doi/10.1126/scirobotics.aay7120 (visité le 15/11/2021).

[41] Sina Mohseni, Niloofar Zarei et Eric D. Ragan. “A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems”. In : *ACM Transactions on Interactive Intelligent Systems* (2021).

[42] Tim Miller. “Explanation in artificial intelligence : Insights from the social sciences”. In : *Artificial Intelligence* 267 (2019), p. 1-38. issn : 0004-3702. doi : https://doi.org/10.1016/j.artint.2018.07.007. url : https://www.sciencedirect.com/science/article/pii/S0004370218305988.

[43] Scott M Lundberg et Su-In Lee. “A Unified Approach to Interpreting Model Predictions”. In : *Advances in Neural Information Processing Systems 30*. Sous la dir. d’I. Guyon et al. Curran Associates, Inc., 2017, p. 4765-4774.

[44] Scott M. Lundberg et al. “From local explanations to global understanding with explainable AI for trees”. In : *Nature Machine Intelligence* 2.1 (2020), p. 2522-5839.

[45] Matthew D. Zeiler et Rob Fergus. “Visualizing and Understanding Convolutional Networks”. In : *arXiv :1311.2901* [*cs*] (nov. 2013). arXiv : 1311.2901. url : http://arxiv.org/abs/1311.2901 (visité le 03/11/2021).

[46] Bolei Zhou et al. “Learning Deep Features for Discriminative Localization”. In : *2016 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). Las Vegas, NV, USA : IEEE, juin 2016, p. 2921-2929. isbn : 978-1-4673-8851-1. doi : 10.1109/CVPR.2016.319. url : http://ieeexplore.ieee.org/document/7780688/ (visité le 20/12/2021).

[47] Mohammed Bany Muhammad et Mohammed Yeasin. “Eigen-CAM : Class Activation Map using Principal Components”. In : *2020 International Joint Conference on Neural Networks* (*IJCNN*) (juill. 2020). arXiv : 2008.00299, p. 1-7. doi : 10.1109/IJCNN48605.2020.9206626. url : http://arxiv.org/abs/2008.00299 (visité le 28/12/2021).

[48] Haofan Wang et al. “Score-CAM : Score-Weighted Visual Explanations for Convolutional Neural Networks”. In : *arXiv :1910.01279* [*cs*] (avr. 2020). arXiv : 1910.01279. url : http://arxiv.org/abs/1910.01279 (visité le 28/12/2021).

[49] Saurabh Desai et Harish G. Ramaswamy. “Ablation-CAM : Visual Explanations for Deep Convolutional Network via Gradient-free Localization”. en. In : *2020 IEEE Winter Conference on Applications of Computer Vision* (*WACV*). Snowmass Village, CO, USA : IEEE, mars 2020, p. 972-980. isbn : 978-1-72816-553-0. doi : 10.1109/WACV45572.2020.9093360. url : https://ieeexplore.ieee.org/document/9093360/ (visité le 28/12/2021).

[50] Ramprasaath R. Selvaraju et al. “Grad-CAM : Why did you say that ?” In : *arXiv :1611.07450* [*cs, stat*] (jan. 2017). arXiv : 1611.07450. url : http://arxiv.org/abs/1611.07450 (visité le 03/11/2021).

[51] Aditya Chattopadhyay et al. “Grad-CAM++ : Improved Visual Explanations for Deep Convolutional Networks”. In : *2018 IEEE Winter Conference on Applications of Computer Vision* (*WACV*) (mars 2018). arXiv : 1710.11063, p. 839-847. doi : 10.1109/WACV.2018.00097. url : http://arxiv.org/abs/1710.11063 (visité le 28/12/2021).

[52] Ruigang Fu et al. “Axiom-based Grad-CAM : Towards Accurate Visualization and Explanation of CNNs”. In : *arXiv :2008.02312* [*cs, eess*] (août 2020). arXiv : 2008.02312. url : http://arxiv.org/abs/2008.02312 (visité le 28/12/2021).

[53] François-Guillaume Fernandez. *TorchCAM : class activation explorer*. https://github.com/frgfm/torch-cam. Mars 2020.

[54] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In : *CoRR* abs/1512.03385 (2015). arXiv : 1512.03385. URL : http://arxiv.org/abs/1512.03385.

[55] Jia Deng et al. “Imagenet : A large-scale hierarchical image database”. In : *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, p. 248-255.

[56] Jacob Gildenblat et contributors. *PyTorch library for CAM methods*. 2021. URL : https://github.com/jacobgil/pytorch-grad-cam.

Abréviations

DL Deep Learning = Apprentissage profond. 20, 21, 37

IA Artificial Intelligence = Intelligence Artificielle. 3, 21, 37

ML Machine Learning = Apprentissage automatique. 20, 21, 37

RAI Responsible Artificial Intelligence = Intelligence Artificielle Responsable. 5, 36

REX Retour d'EXpérience. 36–38

XAI eXplainable Artificial Intelligence = Intelligence Artificielle Explicable. 4, 6, 36, 37