Interprétabilité et explicabilité pour l'apprentissage machine : entre modèles descriptifs, modèles prédictifs et modèles causaux. Une nécessaire clarification épistémologique

Christophe Denis, Franck Varenne

Article paru dans les *Actes de la CNIA PFIA 2019*, Toulouse, AFIA, J. Lang (dir.), pp. 60-68

Interprétabilité et explicabilité pour l'apprentissage machine : entre modèles descriptifs, modèles prédictifs et modèles causaux. Une nécessaire clarification épistémologique.

C. Denis[1] F. Varenne[2]

[1] Sorbonne Université, 4 Place Jussieu, 75005, Paris
[2] Université de Rouen, UFR LSH, Rue Lavoisier, 76821 Mont-Saint-Aignan, ERIAC & IHPST

christophe.denis@lip6.fr franck.varenne@univ-rouen.fr

Résumé

Le déficit d'explicabilité des techniques d'apprentissage machine (AM) pose des problèmes opérationnels, juridiques et éthiques. Un des principaux objectifs de notre projet est de fournir des explications éthiques des sorties générées par une application fondée sur de l'AM, considérée comme une boîte noire. La première étape de ce projet, présentée dans cet article, consiste à montrer que la validation de ces boîtes noires diffère épistémologiquement de celle mise en place dans le cadre d'une modélisation mathématique et causale d'un phénomène physique. La différence majeure est qu'une méthode d'AM ne prétend pas représenter une causalité entre les paramètres d'entrées, qui peuvent être de plus de haute dimensionnalité, et ceux de sortie. Nous montrons dans cet article l'intérêt de mettre en œuvre les distinctions épistémologiques entre les différentes fonctions épistémiques d'un modèle, d'une part, et entre la fonction épistémique et l'usage d'un modèle, d'autre part. Enfin, la dernière partie de cet article présente nos travaux en cours sur l'évaluation d'une explication, qui peut être plus persuasive qu'informative, ce qui peut ainsi causer des problèmes d'ordre éthique.

Mots Clés

IA, apprentissage machine, interprétabilité, explicabilité, causalité, modèles descriptifs, modèles prédictifs, modèles causaux, épistémologie

Abstract

The lack of validation and of explainability of some Machine Learning (ML) models involves operational, legal and ethical issues. One of the main objectives of our research project is to provide ethical explanations of the outputs produced by a ML based application, considered as a black box. The first step of this project, presented in this article, is to underline the epistemic differences between the validation of an ML model and of a causal mathematical model. ML is based on statistical correlations between input - which could have a high dimension - and output parameters without building causality links between them, unlike most mathematical models in science and engineering. This absence of causality is the major drawback of ML, making difficult the validation and the explanation of some ML methods, generally the most efficient ones. Our scientific contribution is to highlight, in this context, the epistemic distinctions between the different functions of a model, on the one hand and between the function and the use of a model, required to build explanation. Our current work in the evaluation of the quality of an explanation, which could be more persuasive than informative and consequently generates ethical problems, is reported in the last part of the article.

Keywords

AI, machine learning, interpretability, explainability, causality, descriptive modeling, predictive modeling, causal modeling, epistemology

1 Contexte et motivations

Depuis 2010, l'Intelligence Artificielle (IA) numérique ou connexionniste fondée sur de l'apprentissage machine (AM) produit des résultats impressionnants, principalement dans les domaines de la reconnaissance de forme, du traitement naturel du langage et de la perception, succédant à la domination de l'IA symbolique centrée sur le raisonnement logique. Ces succès ont entraîné un engouement médiatique : on parle souvent d'un *renouveau de l'IA*, voire de l'apparition d'une nouvelle forme d'IA. Il s'agit plutôt de la fin de l'hibernation de l'IA connexionniste, dont l'origine remonte à la critique du perceptron par Minsky en 1969. Elle s'explique par plusieurs phénomènes concomitants : augmentation de la puissance de calcul, production et traitement performant d'un volume de plus en plus important de données et algorithmes efficaces de pondération des réseaux de neurones profonds. Une analyse pertinente du rapport entre IA symbolique et IA connexionniste a été proposée dans [3] : *« alors que les concepteurs des machines symboliques cherchaient à insérer dans le calculateur et le monde et l'horizon, la réussite actuelle des machines connexionnistes tient au fait que de façon presque opposée, ceux qui les fabriquent vident le calculateur pour que le monde se donne à lui-même son propre horizon »*. L'IA bouleverse des pans entiers de la société humaine comme l'économie, le salariat, la justice et la médecine.

Les résultats, souvent spectaculaires, de l'AM suscitent à la fois de forts espoirs, des craintes légitimes, notamment en termes d'éthique et de transformation du travail [4], et véhiculent un certain nombre de fantasmes [6]. Le déficit de

transparence de ces méthodes d'apprentissage, notamment signalé dans le rapport [25], permet de le considérer, pour certains domaines, comme un colosse au pied d'argile. L'industrialisation de démonstrateurs développés dans des laboratoires n'est que peu souvent au rendez-vous, comme le souligne [8]. L'acceptabilité opérationnelle de telles applications est largement conditionnée par la capacité des ingénieurs et décideurs à comprendre le sens et les propriétés des résultats produits par ces outils. On se heurte au manque de compréhension de leurs mécanismes de décision ou d'aide à la décision. De plus, la délégation décisionnelle croissante proposée par les outils d'IA rivalise avec des règles métier éprouvées, en nombre limité, constituant parfois des systèmes experts certifiés. L'apprentissage machine est comparé dans [20] à une forme d'alchimie et le problème de sa transparence est considéré comme un défi scientifique majeur dans [25].

La première difficulté rencontrée est la polysémie du mot transparence. La transparence d'un algorithme peut faire référence à deux types de propriétés selon [19] : extrinsèques, comme par exemple la loyauté et l'équité, ou intrinsèques, comme l'interprétabilité et l'explicabilité. En nous installant en amont du choix conceptuel de [15] selon qui *"interpretability is the degree to which a human can understand the cause of a decision"* ([15], p.10), nous entendrons ici par interprétabilité pour un sujet humain la capacité d'une représentation à se voir composée d'éléments (signes, figures concepts, données, etc.) qui ont un sens pour le sujet en question. L'explicabilité dénotera ici soit l'explicabilité de l'algorithme de l'AM, soit celle des sorties de l'AM, c'est-à-dire la capacité de déploiement et d'explicitation de cet algorithme ou de ses sorties en série d'étapes reliées entre elles par ce qu'un être humain peut interpréter sensément comme des causes ou des raisons. Nous nous intéressons plus précisément dans cet article au problème de l'explicabilité des sorties produites par l'AM. De telles sorties peuvent être en particulier des suggestions de décisions, de prédictions ou d'actions [19]. Il est à noter que le déficit d'explicabilité se retrouve au delà de l'AM : par exemple dans le cadre d'une modélisation mathématique d'un phénomène physique mal connu (incertitudes épistémiques) ou d'un système expert possédant un nombre important de règles.

Notre projet de recherche consiste à construire des explications de méthodes d'AM que nous considérons ici comme une boîte noire. Dans ce contexte plus précis, nous définissons l'explicabilité comme la capacité à fournir pour un ensemble d'utilisateurs une explication des résultats obtenus par l'AM adaptée à leurs connaissances scientifiques et métier. Les évaluations humaines d'une explication introduisent un biais cognitif envers les raisonnements les plus simples [9]. Il est donc nécessaire sur le plan éthique d'évaluer le meilleur compromis entre explication intelligible et explication persuasive. La première étape de ce projet, celle qui est principalement présentée dans cet article, consiste à montrer que le processus d'explicabilité de cette boîte noire diffère épistémologiquement de celle mise en place dans le cadre de la modélisation mathématique et causale d'un phénomène physique. La différence majeure est qu'une méthode d'AM ne prétend pas représenter une causalité entre les paramètres d'entrées et ceux de sortie, cela, malgré le recours aux termes trompeurs, issus de la théorie statistique, de variables dites « explicatives ». Nous soutenons que c'est en grande partie cette absence de représentation d'une causalité qui est à l'origine des trois points de fragilité de l'apprentissage machine déjà signalés et étudiés dans la littérature :

1. l'interprétabilité du processus computationnel - ou de ses éléments - n'assure pas à elle seule son explicabilité;
2. la conception d'un modèle d'AM nécessite un prétraitement et une sélection des données. Quand ces données ne sont pas reliées à un scénario causal explicite, cela a pour effet de masquer les choix ontologiques qui accompagnent inévitablement les choix de formats, de données ou de leurs prétraitements ;
3. enfin, l'explication, quand elle est possible, n'est pas pour autant assurée d'être dépourvue de biais et peut se révéler être un dispositif servant davantage un usage rhétorique qu'une fonction épistémique objective, à savoir un usage pour la persuasion et la mise en confiance des utilisateurs [9]. Nous montrerons ici l'intérêt de mettre en œuvre la distinction épistémologique entre fonction et usage d'un modèle [23], [24] ;

2 Contribution et organisation de l'article

Notre contribution s'organise de la manière suivante :

— la section 3 présente tout d'abord la différence épistémique, en termes d'usage général des données, entre une prédiction d'un phénomène fondée sur sa modélisation mathématique (modèle hypothético-déductif) et une prédiction fondée sur un apprentissage machine (modèle inductif). Nous rappelons qu'une application fondée sur de l'AM n'utilise pas les données de la même manière que l'approche hypothético-déductive. Les deux ne se nourrissent donc pas des données de manière équivalente ou indifférente : des contraintes ontologiques spécifiques en termes de format et de traitement de données interviennent déjà lors des choix de conception.
— la section 4 caractérise ensuite les différentes fonctions de connaissance ou fonctions épistémiques d'un modèle. Nous nous interrogeons sur les critères qui permettent de décider si et quand un processus de modélisation relève d'une explication causale. Nous rappellerons succinctement les liens que la philosophie contemporaine des sciences voit entre explication causale et mécanisme. Nous y montrons comment cette classification peut être adaptée aux applications fondées sur l'AM.
— la section 5 marque les similitudes et les différences entre l'explication causale *par* un modèle d'AM et l'explication causale *d'un* modèle d'AM. Cela nous permettra de distinguer plus clairement trois

grands types d'explicabilité dans ce contexte : tournée vers le système cible, tournée vers le concepteur, tournée vers l'utilisateur. Comme suite à ces distinctions, les rapports entre différents types d'interprétabilité et d'explicabilité pourront être élucidés.

— la section 6 introduit la différence entre fonction (épistémique) et usage (pratique ou rhétorique) d'un modèle. Couramment, la demande d'explicabilité mélange ces deux notions. Nous montrerons qu'il n'est certes pas possible de séparer complètement dans les faits fonction et usage (comme il n'est pas possible de séparer complètement l'aspect *ethos* - confiance - et l'aspect *logos* - rationnel - d'un discours persuasif), mais qu'il peut être nécessaire de savoir les séparer conceptuellement, c'est-à-dire de savoir exprimer la différence entre les types de savoirs qui les fondent pour se rendre capable de distinguer et de clarifier les sources de fragilité de l'AM.

La synthèse de cet article et la suite de notre projet de recherche sont présentées en dernière partie.

3 Conception d'une application fondée sur de l'AM

Nous présentons tout d'abord le rôle différent joué par les données dans une prédiction fondée tantôt sur de l'AM, tantôt sur une formulation mathématique explicite du phénomène à prédire. Bien que les données jouent un rôle majeur dans la conception d'une application procédant par AM, des contraintes ontologiques en termes de format et de traitement, engendrées elles-mêmes par des choix de conception, influencent le comportement et l'explication de cette application.

3.1 Rôle des données dans le cadre d'une machine hypothético-déductive et d'une machine inductive

Prenons l'exemple d'un fluide dont on souhaite estimer la vitesse d'écoulement dans un canal. Deux options sont possibles pour obtenir la prédiction de la vitesse :

— utiliser *une machine hypothético-déductive* : l'écoulement est modélisé à l'aide d'équations mathématiques, par exemple en utilisant les équations de Navier-Stokes. La discrétisation des ces équations sur ordinateur conduit à un programme. Les données servent seulement à instancier un problème déjà résolu par le programme sur la machine hypothético-déductive. Dans une démarche de validation et de quantification d'incertitudes, d'autres données, essentiellement des mesures, sont utilisées pour caler et pour valider le programme [17].

— utiliser une *machine inductive* : le programme utilisé pour prédire la vitesse de l'écoulement peut résulter d'un AM prenant en entrée une base de données. Ici, les données servent à la fois à construire et à valider le programme. Cependant, la finalité d'un modèle inductif n'est pas uniquement la prédiction mais aussi l'explication du phénomène. Actuellement, cela n'est pas toujours le cas puisque certaines techniques d'AM, généralement les plus performantes au niveau statistique, souffrent d'un déficit de capacité à expliquer leur système cible comme aussi d'un déficit d'explicabilité.

3.2 Etapes de conception d'une application d'AM

Nous nous plaçons sans perte de généralité dans le cadre de l'AM supervisé. La conception d'un modèle d'apprentissage n'est pas un processus automatique. Elle ne se résume pas au choix de la méthode d'apprentissage (régression linéaire, arbre de décision, réseau de neurones, etc.). La figure 1 présente les principales étapes intervenant lors de la conception d'une application d'AM. En particulier, le prétraitement, le choix et la combinaison des variables dites explicatives sont la source de choix ontologiques résultant de ces manipulations. Il est à juste titre indiqué dans [6] « *qu'il y a une connaissance implicite cachée derrière la formulation des données que l'on utilise, autrement dit, des dogmes que l'on entre, par-devers soi, dans les machines* ».

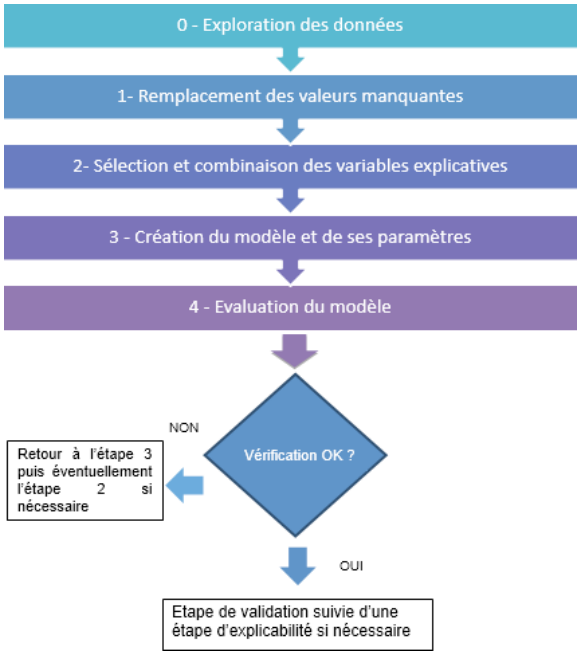

FIGURE 1 – Principales étapes intervenant lors de la conception d'une application d'AM

La conception comprend une phase de vérification et de validation [5] :

— *étape de vérification* : s'assurer que le modèle d'AM permet d'obtenir la précision statistique souhaitée et possède de bonnes capacités de généralisation ;

— *étape de validation* : justifier le choix du modèle d'AM en particulier lors de la sélection ou de la combinaison de variables explicatives.

4 Fonctions principales d'un modèle : réduction de données, description, prédiction, explication

4.1 Sur la fin programmée des modèles face au déluge des données

L'argumentation portant sur la fin des modèles et des hypothèses théoriques exprimée dans [1] a entraîné des débats sur le rôle de la démarche scientifique dans un monde marqué par une production effrénée de données, d'une part, et par une augmentation continue de la puissance de calcul, d'autre part. Le traitement petaflopique, bientôt exaflopique voire quantique d'un déluge de données rendrait obsolète la modélisation ou la théorie scientifique : *"Petabytes allow us to say: 'Correlation is enough.' We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.* [1]*"*. Indéniablement, le traitement massif de données permet d'obtenir des avancées remarquables dans de nombreux domaines, notamment en santé. Un enjeu par exemple est la prévention des interactions médicamenteuses dangereuses pour des patients atteints de maladies complexes ou concomitantes. Les essais cliniques, en nombre relativement restreint, ne permettent pas de prédire toutes les interactions en raison d'une combinatoire élevée. L'utilisation d'une technique d'AM supervisé recourant à un nombre important de données pharmacogénomiques et de populations de patients a permis d'améliorer considérablement la prévention des effets indésirables en polypharmacie [26].

L'argumentation de [1], également reprise par d'autres auteurs, s'inscrit dans le courant empiriste de la pensée scientifique moderne. En 1620, Francis Bacon argumente dans [2] que la démarche scientifique ne doit pas être fondée prioritairement sur des hypothèses (modèle déductif dominant la science depuis Aristote) mais sur des données expérimentales (modèle inductif). Or, l'explicabilité de la prédiction reste malgré tout une fonctionnalité importante du modèle inductif. Sinon, la performance statistique de la prédiction sans explicabilité conduit à une forme de sophisme pragmatique : est jugée abusivement réaliste ou conforme au réel une représentation qui permet seulement - pragmatiquement - de le prédire [22]. À bien y regarder, le traitement performant statistique d'un important volume de données n'est pas l'annonce de la fin des modèles mais l'annonce d'une utilisation accrue d'autres types de modèles et pour lesquels un travail épistémologique affiné s'impose plus que jamais.

4.2 Caractérisation d'un modèle et des fonctions d'un modèle

Dans son caractère le plus général, un modèle peut être défini comme un objet médiateur auquel on adresse une question au sujet d'un objet cible qu'on ne peut interroger directement : « *pour un observateur B, un objet A* est un modèle d'un objet A dans la mesure où B peut utiliser A* pour répondre à des questions qui l'intéressent au sujet de A* » [14].

La prédiction et l'explication d'un phénomène sont des fonctions de connaissance parmi d'autres. Parmi, les nombreuses fonctions de connaissance que peut faciliter la médiation d'un modèle, il est possible d'en identifier et d'en classer une vingtaine [23], [24]. Les plus fréquentes sont l'analyse ou la réduction de données, la description, la prédiction et l'explication ([23], [24] : fonctions 5, 6, 7 et 8). La figure 2 évoque l'utilisation d'un modèle mathématique causal simulé numériquement pour expliquer et prédire un phénomène physique.

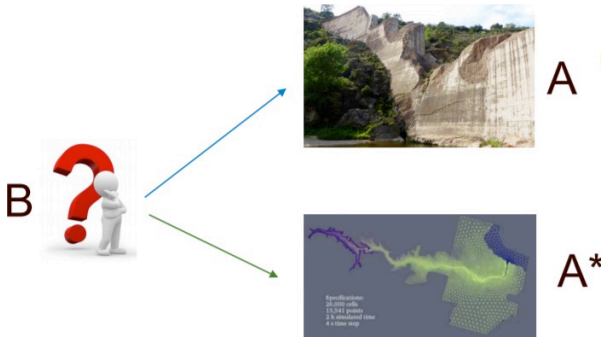

FIGURE 2 – Cas d'un modèle à la fois explicatif et prédictif

Un modèle d'analyse de données ne décrit pas encore des structures propres au système cible mais seulement des structures entre des signaux auxquels ils donnent lieu. Les modèles de réduction de données structurent, élaguent ou classent les données sans permettre encore de description ni d'interprétation de ces classifications en termes d'ontologies interprétables et valant pour le système cible de manière signifiante. Ils servent de représentation intermédiaire de la seule structure informationnelle des données du système cible, mais pas directement de la structure des propriétés intrinsèques du système cible ni des relations mutuelles entre ces propriétés : ils traitent les données comme des *signaux* non comme des *signes*. Un signal indique, qualifie ou quantifie une interaction. Un signe désigne, qualifie ou quantifie une propriété. Un signal est le résultat de la détection ou de la mesure par capteur physique d'un phénomène d'interaction entre l'objet cible et son environnement physique (qui est au minimum son cadre spatial, temporel ou spatio-temporel). Dans l'approche « signal », les propriétés physiques intrinsèques de l'objet cible sont certes supposées mais leur nature peut demeurer largement inconnue, alors que l'approche « signe » entend rendre compte d'une propriété du système cible et de sa valeur, en recourant à cet autre type de médiateurs que sont les instruments de mesure. Les modèles de réduction de données sont donc faiblement prescriptifs ontologiquement. Ils préparent l'utilisation d'autres modèles : les modèles à fonction de description ou d'explication du système cible.

Tout en dépassant déjà la considération superficielle de la seule structurelle informationnelle des données, les modèles

descriptifs et prédictifs sont toutefois nommés encore « phénoménologiques » ([23], [24] ; fonction 7) : ils facilitent la reproduction ou production de structures de données dont l'apparence (la phénoménologie) est jugée fidèle à certaines structures des propriétés observables du système cible. Ils réalisent cette production par des moyens intelligibles, déductifs ou calculatoires. Un modèle descriptif structure des données qui, séparément, ont déjà un sens minimal, c'est-à-dire qui sont interprétables en termes de propriétés au regard de la connaissance minimale que l'on a, par ailleurs, du système cible. En revanche, la structure que le modèle descriptif propose pour ces propriétés (leur relation mutuelle représentée dans le modèle) peut être complètement phénoménologique, c'est-à-dire ne rien signifier de robuste, ne renvoyer à rien de réel, *i.e.* ne pas se fonder elle-même sur une propriété profonde de structure du système cible.

Un modèle prédictif est un cas particulier de modèle descriptif. Il décrit le système à travers deux types minimaux de données qui le représentent (le décrivent) partiellement sans pour autant encore l'expliquer : les données prédictives - qui servent dans l'algorithme ou le modèle - et les donnés comportementales ou prédictives qui servent à évaluer la qualité de la prédiction, donc la qualité du modèle. Un modèle descriptif peut en effet être statique ou dynamique. Une dynamique au sens large (*i.e.* permettant de distinguer des conditions initiales et un état final, ou des variables d'entrée et des variables de sortie, ou encore des variables prédictives et des variables prédites) peut être reproduite de manière elle aussi purement descriptive, *i.e.* sans qu'une séquence temporelle soit représentée de manière ontologiquement significative (explicative par exemple) pour les séquences d'états du système cible. Quand cette dynamique permet non seulement de décrire correctement le comportement observable du système dans les cas connus d'entrées/sorties mais aussi d'interpoler ou d'extrapoler correctement une description de son comportement observable à partir de données qui n'ont pas été utilisées pour calibrer le modèle (données nouvelles, période de temps non encore testée), le modèle descriptif (à AM, de régression, de classification, etc.) se trouve être également ce qu'on appelle un modèle prédictif.

Il y a deux grands types de modèles prédictifs : à régression au sens large (dès lors qu'ils servent à prédire des variables quantitatives), et de classification, ces derniers servant à prédire une variable qualitative ou, plus largement, à estimer la probabilité d'un événement. Dans les modèles prédictifs de classification, il est utile de distinguer les modèles discriminatifs qui ne font pas d'hypothèse *a priori* sur la distribution (régression logistique, perceptron, SVM), et les modèles génératifs se fondant sur l'hypothèse de l'existence d'une forme paramétrique précise pour une distribution sous-jacente [21]. Ces derniers permettent l'adoption d'une approche bayésienne qui, si les *priors* sont acceptés et se révèlent féconds à l'usage, peuvent nous mener à l'idée que le modèle est explicable. Il n'en reste pas moins que le fondement de l'explicabilité de tels modèles reste épistémique car n'entendant pas reposer sur une hypothèse ontologique de lois de la nature ou de causalité.

En philosophie des sciences contemporaine, il n'existe pas de consensus sur la différence précise entre expliquer et comprendre. Une grande partie des auteurs s'accorde cependant sur le fait d'associer l'explication à la causalité et la compréhension à l'unification d'une diversité de phénomènes sous un principe unique ([16] p. 18). En se fondant sur cette idée toujours discutable mais également fréquemment acceptée qu'une explication réfère à une causalité (nous n'entendrons pas ici causalité au sens où l'entend la pratique de l'inférence causale, même si la sophistication récente de sa stratégie de formalisation des propositions contrefactuelles au moyen d'une approche structurelle se révèle pragmatiquement efficace : [18]), on peut dire qu'un modèle mathématique ou algorithmique est explicatif d'un objet cible lorsque :

1. il est au moins partiellement prédictif pour ce système ;
2. il offre une représentation interprétable, c'est-à-dire signifiante et accessible à un esprit humain non aidé, des éléments dont il est composé et des processus élémentaires d'interaction qu'il met en œuvre ;
3. ces éléments et processus élémentaires sont supposés eux-mêmes représenter plus ou moins iconiquement des éléments et des processus d'interaction causale (ou mécanismes) intervenant réellement et majoritairement dans le système cible lui- même.

Ainsi, la modélisation mathématique causale d'un phénomène physique repose sur un ensemble de causes X reconnues par les physiciens pour être réellement et majoritairement à l'origine du phénomène physique Y. Plus précisément, les équations mathématiques dérivent d'un modèle mécaniste théorique obtenu à partir des lois théoriques hypothétiques de la physique. Le modèle mécaniste théorique est alors un ensemble de relations structurelles, causales, entre des variables X décrivant le lien entre Y et X. C'est le cas de la grande majorité des modèles théoriques utilisées pour représenter les phénomènes. Le phénomène Y n'est certes pas entièrement explicable par l'ensemble des causes X puisque, d'une part, il existe des incertitudes (aléatoires ou épistémiques) sur l'évaluation des causes X et que, d'autre part, certaines causes peuvent être inconnues dans l'état actuel de la connaissance scientifique. La procédure de V&V d'un code de calcul fondé sur une modélisation physique d'un phénomène a pour vocation de répondre successivement aux deux questions suivantes : 1) étape de vérification : est-ce que le programme informatique résout correctement les équations mathématiques choisies ? 2) étape de validation : est-ce que l'on a choisi les bonnes équations et les bons paramètres d'entrée? On voit ici que la validation dépend étroitement de la fonction épistémique attendue du modèle. Un modèle explicatif ne devra pas être seulement validé dans sa capacité à reproduire certains comportements du système cible. Il faudra aussi évaluer sa capacité à représenter pas à pas, de manière correcte, *i.e.* approximativement réaliste, non seulement les états successifs du système cible mais aussi chaque étape de calcul, chaque opération du processus lui-même.

5 Explication causale, interprétabilité et explicabilité en apprentissage machine

Il y a plusieurs types d'explication qui peuvent intervenir dans l'évaluation d'un modèle. On doit distinguer d'abord l'explication du système cible par le modèle de l'explication du modèle lui-même et de son fonctionnement. C'est cette seconde explication qui est l'enjeu de cet article. Mais il y a des liens possibles entre les deux. D'abord, il peut exister un mécanisme causal réellement existant et affectant le système cible. Ce mécanisme peut reposer sur des lois connues, en être la déduction, le résultat calculatoire : par exemple une interaction locale entre deux astres repose sur les lois de Newton, une interaction entre deux atomes en chimie reposent sur l'équation de Schrödinger, etc. On peut alors modéliser le système cible en modélisant de manière fidèle la causalité même affectant ce système cible. On le fait en représentant de manière iconique (*i.e.* au moins termes à termes) les principaux éléments en interaction et leurs principales interactions : par là, le modèle est fidèle au moins à l'individuation des éléments naturels réels comme une planète, un atome, même s'il y a une part d'idéalisation dans leur représentation individuelle comme celle qui consiste à supposer que la masse de la planète est entièrement située sur le point géométrique centre de gravité de la planète. Dans ce type de modèle, la première forme d'explication intervient au sens d'abord où c'est un modèle expliquant le système cible : il est explicatif (voir plus haut). C'est-à-dire qu'en même temps qu'il effectue ses computations, il explicite, il rend visible pas à pas le processus qui affecte le système cible de manière assez fidèle et suffisamment réaliste au vu des connaissances que nous avons par ailleurs de ses éléments, de leurs propriétés, des lois de la nature qui les affectent et des mécanismes d'interaction que ces lois déterminent (loi de la physique, de la chimie, voire de la biologie). Mais, de manière similaire, dans le cas d'un système expert modélisant une décision médicale, par exemple, les bases de données et les règles de raisonnement sensées et appliquées pas à pas dans le modèle expliquent le processus même de la décision. Notons que, dans le cas de la décision humaine motivée, on peut considérer qu'une raison joue le même rôle qu'une cause dans un système physique soumis à des lois physiques.

Dans tous ces cas favorables, une des conséquences est que le modèle est non seulement explicatif mais aussi explicable. Cela veut dire que le processus de computation suivi par le modèle implémenté dans le programme est également interprétable et explicable en lui-même. Il est interprétable car l'ontologie du modèle (ses représentations) renvoie à des ensembles d'entités et de propriétés reconnues comme existant réellement dans le système cible auquel on a accès par ailleurs sous une forme interprétable. Dans ce cas de modèle explicable, on utilise donc la connaissance préalable que l'on a 1) de la structuration réelle du système cible (l'ontologie qu'on lui reconnaît), 2) du fait que le modèle utilise cette structuration et n'utilise qu'elle dans ses processus, 3) du fait que les processus du modèle sont également supposés réalistes, 4) du fait que ce dépliement processuel pas à pas converge mathématiquement (théorème de convergence) vers les résultats, pour, au final, décider que le modèle non seulement explique son système cible mais qu'il est également interprétable et explicable en lui-même.

Dans ce cas favorable d'un modèle expliquant son système cible (explication *par* le modèle), c'est par l'effet d'une transitivité de la représentation de l'ontologie, des structures et des processus, que l'on peut également conclure à une explicabilité du modèle lui-même (explication *du* modèle). Cette explicabilité du modèle est ici assurée et légitimée par notre connaissance des lois qui affectent réellement le système cible. On peut prendre l'exemple d'une simulation numérique en mécanique des structures ou en mécanique des fluides. Quand un tel modèle est validé, même s'il est traité numériquement, il reste à la fois explicatif, interprétable et explicable. Pour un modèle de décision experte à base de règles motivées et une à une significatives au regard de règles métier, l'explicabilité du modèle est également assurée du fait de la représentation de cette causalité généralisée (raisonnement symbolique significatif) dans le modèle lui-même.

À première vue, on pourrait se dire qu'on peut adapter ce processus de validation aux modèles d'AM puisqu'on peut considérer que leur nature est également mathématique et numérique. En effet, il existe bien des équations mathématiques pour concevoir un réseau de neurones comme par exemple l'algorithme de rétropropagation du gradient servant à déterminer les pondérations. De ce point de vue, une validation de l'algorithme, c'est-à-dire la preuve qu'il assure bien la fonction de prédiction désirée, entraînerait une confiance sur le calcul des poids du réseau et *de facto* sur le réseau de neurones. Mais, dans le cas de l'AM, à la différence des cas précédents, l'explicabilité du modèle n'est pas aussi facile à assurer car elle ne peut pas être directement héritée du fait que le modèle serait explicatif. L'explicabilité du modèle que l'on recherche doit être fondée autrement pour deux raisons. Premièrement, comme pour un modèle d'analyse de données standard, en AM, le modèle contrôlant les relations entrées/sorties n'entend pas représenter, même de manière seulement stylisée, un scénario causal d'interaction pas à pas opérant sous l'effet de lois ou de règles motivées. Le modèle se fonde sur l'analyse de corrélations entre les paramètres d'entrée. Or, une corrélation statistique entre deux paramètres ne signifie pas qu'il existe une causalité entre eux. Deuxièmement, la situation de l'AM est pire encore que celle des modèles classiques d'analyse de données : car le modélisateur ne cherche même pas à ce que les conditions minimales d'exercice d'un hypothétique modèle explicatif soient réunies. L'ontologie sous-jacente aux données et à leur structure peut en effet être complètement inconnue ou fictionnelle. On ne connaît pas d'ontologie robuste et objective du domaine qui soit explicitement prescrite et sur laquelle pourrait éventuellement s'exercer un ensemble de mécanismes déterminés par des lois. Ainsi, on ne peut pas s'appuyer d'emblée sur une reconnaissance préalable, ne serait-ce que descriptive, de la structure interne des données (car les données sont dites mal structurées ou bien leur structure significative - s'il en est une - nous est inaccessible). Enfin, quand bien même une structure serait

perceptible dans les données, une technique comme les RN par exemple met en œuvre des modèles non linéaires reliant les valeurs prédictives et les valeurs prédites. Les valeurs prédictives interagissent fortement : donc on ne peut plus parler de simples corrélations. Dans le cas d'un modèle non linéaire à arbres de décision, les étapes élémentaires restent certes interprétables une à une, mais le processus d'ensemble n'est pas pour autant aisément sensément résumable : il n'est pas compréhensible.

En analyse des données classique, toutefois, le modèle d'analyse repose sur des hypothèses globales et minimales - qu'on peut dire métaphysiques - de symétries temporelles ou spatiales liées à l'environnement de captation des données. C'est ce genre d'hypothèse qui autorise l'approche par traitement de signal, très fréquente en ingénierie : analyse linéaire, analyse de Fourier, transformée en Z, analyses non paramétriques, etc. Mais ces hypothèses métaphysiques minimalistes ne sont même pas toujours possibles en AM. Le rapprochement récent entre l'analyse par ondelettes et les réseaux de neurones convolutionnels [12] ne fait que confirmer ce soupçon qu'un RN quelconque (non convolutionnel) est en général plus neutre encore et moins-disant d'un point de vue métaphysique et causal que les approches par analyse de données paramétriques ou non paramétriques. Ainsi, les modèles à AM ne peuvent pas hériter directement leur interprétabilité et leur explicabilité du caractère réaliste et causal des interactions qu'ils modélisent dans leur calcul. Car, ils sont a priori dépourvus d'un tel ancrage réaliste et causaliste. Ce défaut fragilise les pratiques de vérification, de validation, mais aussi de diffusion et d'appropriation par les utilisateurs, d'où la demande d'interprétabilité et d'explicabilité de ces modèles. Au vu des distinctions faites précédemment, remarquons que la demande d'interprétabilité d'un modèle d'AM revient finalement à demander la construction d'un modèle descriptif de ce modèle d'AM. La demande d'explicabilité d'un modèle d'AM, quant à elle, revient à demander d'en construire un modèle explicatif.

Remarquons enfin que dans la demande d'explicabilité, il y entre souvent en même temps la demande de compréhension du modèle. On recherche alors des grands principes unificateurs permettant de penser et représenter de manière unitaire le fonctionnement global, la logique globale, du fonctionnement du modèle. La légitimation et l'acceptabilité du modèle va souvent de pair avec sa compréhensibilité. On cherche alors à construire un modèle de compréhension du modèle d'AM (fonction 9 des modèles selon [23], [24]). Plus encore que l'interprétabilité, la compréhensibilité est très sensible aux compétences de la personne à laquelle elle s'adresse. Aristote avait souligné que la rhétorique existe pour mettre en confiance et persuader les personnes qui ne peuvent suivre de manière attentive de longues chaînes de raisonnement : une explication pas à pas, même si elle est disponible et même si elle est causale, ne leur suffit pas ; il faut alors que les experts pratiquent la rhétorique et leur fournissent une représentation prenant la forme d'un grand mouvement simple et uniforme de pensée supportable par un esprit humain non aidé. Pour satisfaire les demandes d'interprétabilité et d'explicabilité d'un modèle d'AM, on cherche ainsi souvent à en construire un modèle second qui recourt à une remathématisation, une factorisation ou tout autre simplification de ses multiples couches humainement inextricables de représentation, d'une part, de computation, d'autre part [7]. Par là, on cherche à simplifier et remodéliser de manière humainement maniable (voir fonction 9 dans [23]) un comportement de modèle sinon inextricable. Cette simplification est donc une modélisation de second degré, un modèle de modèle d'AM. Cette modélisation peut ensuite elle-même s'accompagner de différents usages. Elle peut en effet être recherchée pour une vérification, une validation ou servir encore à un dispositif explicite de persuasion - plus ou moins biaisé - à destination des utilisateurs.

6 Fonction épistémique et usage du modèle en apprentissage machine

Comme présenté en section 3, la conception comprend des choix techniques le plus souvent décidés empiriquement et qui déterminent eux-mêmes des choix ontologiques. L'objectif de l'étape de validation consiste à justifier les choix effectués lors de la conception du modèle d'apprentissage, y compris au regard de la fonction (pour l'AM, le plus souvent, une fonction de prédiction). C'est là ce que nous appelions le troisième point de fragilité de l'AM : la conception d'un modèle d'apprentissage nécessite toujours un formatage, une sélection puis un prétraitement des données. Comme ces données ne sont pas reliées d'entrée de jeu à une ontologie explicite ni à un scénario causal explicite (voir sections précédentes) mais que l'on peut continuer à dire qu'on en propose une sorte d'interprétabilité puis une sorte d'explicabilité pour le processus qui les traite ensuite, cette interprétabilité et cette explicabilité proposées peuvent avoir pour effet de masquer davantage encore les choix de format et de représentation qui structurent implicitement les données initiales et leurs prétraitements. Le problème peut venir de la confusion suivante : ce n'est pas parce qu'on a réussi à modéliser de manière explicative ou compréhensive un modèle par ailleurs purement prédictif que l'on a rendu ce modèle explicatif. On a pu expliquer le fonctionnement interne de ce modèle prédictif mais pas le fonctionnement du système cible initial. On n'a pas non plus davantage confirmé le caractère réaliste de l'ontologie de ce modèle prédictif. Ainsi, les structures implicites de données peuvent être à l'œuvre sans que l'on sache dans quelle mesure ni à quel niveau c'est le cas, même quand le modèle remplit son office. Le succès d'un modèle de prédiction peut éventuellement être une indication que les éléments qu'ils postulent explicitement pour effectuer ses calculs (par exemple : des neurones très simplifiés) reflètent finalement quelque chose qui serait réellement à l'œuvre dans le système cible : c'est ainsi ce qui fonde l'opinion biomimétiste de Yann Le Cun. Mais le théorème d'universalité concernant les RN peut aussi nous engager plutôt à penser qu'il s'agit simplement d'une autre forme, parmi d'autres, d'automate universel de calcul, simplement plus commode à utiliser en pratique pour certaines formes de données et de questions qu'on leur adresse. Un usage normatif et prescriptif des modèles de prédiction en AM peut ainsi s'exercer non seulement du fait du caractère prescriptif du choix de modélisation lui-même, pour peu

qu'il s'accompagne d'une interprétabilité jugée acceptable parce que suffisamment performante et compréhensible par les utilisateurs, mais aussi du fait que cette interprétabilité peut masquer la non explicitation des choix ontologiques demeurés latents dans les données d'apprentissage.

7 Évaluation de la qualité d'une explication

Il existe une taxonomie de méthodes pour produire des explications sur les résultats produits par de l'AM [7]. La forme de l'explication doit être évaluée et choisie pour minimiser les biais cognitifs de l'utilisateur, comme indiqué dans [13] : *"The motivations and benefits of different types of transparency can vary significantly depending on context, and objective criteria are difficult to identify."*

L'explication doit en particulier permettre :

- pour un développeur, de comprendre le fonctionnement de l'application afin de la déboguer ou de l'améliorer ;
- pour un utilisateur, de comprendre le périmètre d'utilisation et les hypothèses sous-jacentes donnant des clés de lecture des résultats obtenus ;
- pour un expert, de statuer sur un audit lors d'un incident.

Il existe en outre un problème éthique du fait de la possible dérive consistant à produire des explications davantage persuasives que transparentes. Nous avons commencé un travail de recherche visant à rendre possible la mesure de la qualité d'une explication. Au vu de nos analyses préalables, il nous apparaît désormais clairement que cette mesure devra se faire en fonction de l'usage et du destinataire de l'explication. Pour cela, nous nous inspirerons notamment de certains travaux de psychologie cognitive [11]. Un protocole d'évaluation qualitative d'explications d'une application basée sur de l'AM sera alors défini et testé sur un panel d'utilisateurs.

8 Conclusion et perspectives

Le déficit d'explicabilité des techniques d'apprentissage machine profond pose des problèmes d'ordre opérationnel, juridique et éthique. Notre projet de recherche a pour objectif de fournir et évaluer des explications de méthodes d'apprentissage machine considérées comme une boîte noire. La première étape de ce projet, celle qui est présentée dans cet article, consiste à montrer que la validation de cette boîte noire diffère épistémologiquement de celle mise en place dans le cadre de la modélisation mathématique et causale d'un phénomène physique. La différence majeure est qu'une méthode d'apprentissage machine ne prétend pas représenter une causalité entre les paramètres d'entrées et ceux de sortie, cela, malgré le recours aux termes trompeurs, issus de la théorie statistique, de variables « explicatives ». Nous soutenons que c'est en grande partie cette absence de représentation d'une causalité qui est à l'origine des trois points de fragilité de l'apprentissage machine déjà signalés et étudiés dans la littérature. Cette première analyse nous a conduits à distinguer plusieurs fonctions pour les modèles : analyse de données, description, prédiction, explication. Nous avons distingué également deux approches des données : en termes de signaux ou en termes de signes. Dans une approche purement « signal », le modèle prend pour objet d'étude et de traitement le seul niveau de la structure informationnelle du système cible. Dans les approches « signe », les modèles s'engagent sur le lien entre les données et une certaine ontologie plus ou moins réaliste du système cible, réalisme souvent fondé sur des théories scientifiques et sur des hypothèses métaphysiques associées de symétrie et d'invariance. Selon que cet engagement réaliste en reste aux entités et aux propriétés ou qu'il en passe ensuite aux structures voire aux liens causaux, on a affaire à des modèles descriptifs, prédictifs ou explicatifs.

Remarquons en passant que [22], souvent cité dans ces débats tournant autour de prédire et expliquer, n'est justement pas si clair à ce sujet. Lorsqu'il soutient une conception métaphysique continuiste et causaliste, qu'il s'oppose ensuite aux approches discrétisées et statistiques supposées par principe ne pouvoir que décrire, il écrit en effet : *« il n'y a de science [explicative] que dans la mesure où l'on plonge le réel dans un virtuel contrôlé. Et c'est par l'extension du réel dans un virtuel plus grand que l'on étudie ensuite les contraintes qui définissent la propagation du réel au sens de ce virtuel »* ([22], p. 122). Cependant, il n'est justement pas certain que l'approche topologique générale du réel qu'il propose, avec cette métaphore de la plongée dans un espace topologique différentiel à la fois général et supposé par là causalement contraignant, soit en réalité elle-même autre chose qu'une « approche signal » qui veut se faire passer pour une « approche signe » : dans quelle mesure, en effet, une telle plongée est-elle assurée d'être davantage qu'une simple insertion dans un cadre spatio-temporel, cadre lui aussi surimposé, avec ses choix ontologiques et ses biais donc, pour servir à une approche signal ? Mais cela reste ici un débat seulement connexe à notre contribution. En tous les cas, la prise de conscience de la réelle diversité des modèles alternatifs aux modèles exclusivement théorico-explicatifs (diversité également et significativement sous-estimée par [1]) permet justement de poser à nouveaux frais ce genre de questions et d'y répondre plus précisément au regard du contexte technique et de la fonction épistémique du modèle chaque fois recherchée.

Dans cet article, nous avons ensuite caractérisé la recherche d'interprétabilité et d'explicabilité des modèles en termes de modélisation de modèle. Nous avons enfin montré que les modèles explicatifs sont d'emblée explicables, mais que les modèles prédictifs à AM ne le sont pas directement le plus souvent, bien qu'ils puissent être expliqués secondairement par d'autres modèles : ces derniers rendent les modèles d'AM explicables, sans pour autant - ni toujours ni directement - légitimer les ontologies mobilisées par ces modèles ni pouvoir montrer qu'ils expliquent leur système cible. Les usages que l'on fait des modèles interprétant ou expliquant les modèles à AM, aussi impressionnants soient-ils, ne doivent donc pas faire oublier les fragilités persistantes des modèles qu'ils modélisent.

Cet article a présenté ainsi la première étape de notre projet de recherche dont l'objectif final est de définir un cadre épistémique pour l'explicabilité et la validation d'applications d'AM. Pour cela, un ensemble de techniques d'explicabilité, notamment répertoriées dans [7], sera

évalué dans ce cadre, ainsi que la prise en compte de l'incertitude de modèle d'AM telle qu'effectué par exemple dans [10] en utilisant des réseaux de neurones bayésiens.

Références

[1] C. Anderson, The End of Theory: The Data Deluge Makes the Scientific Method Obsolete, *Wired: Science*, 2008.

[2] F. Bacon, *Novum Organum*, 1620.

[3] D. Cardon, J-P. Cointet, et A. Mazières, La revanche des neurones. L'invention des machines inductives et la controverse de l'intelligence artificielle, *Réseaux*, 2018.

[4] A. Cassili, *En attendant les robots. Enquête sur le travail du clic*, Le Seuil, 2019.

[5] C. Denis, Interprétabilité et validation d'applications métiers basées sur de l'IA statistique, *Journée Ethique et IA, PFIA*, 2018.

[6] J-G. Ganascia, *Le Mythe de la Singularité. Faut-il craindre l'intelligence artificielle ?*, Le Seuil, 2017.

[7] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter and L. Kagal, Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning, *CoRR, http ://arxiv.org/abs/1806.00069*, 2018.

[8] J-M. Ghidaglia, N. Vayatis, Comment faire sortir l'intelligence artificielle des labos ?, *Les Echos*, 2019.

[9] B. Herman, The Promise and Peril of Human Evaluation for Model Interpretability, *Thirsty-first Conference on Neural Information Processing Systems*, 2017.

[10] A. Kendall, Y. Gal, What Uncertainties Do We Need in Bayesian Deep Learning for Computer, *Thirsty-first Conference on Neural Information Processing Systems*, 2017.

[11] T. Lombrozo, *Explanation and Abductive Inference*, The Oxford Handbook of Thinking and Reasoning, 2012.

[12] S. Mallat, Understanding deep convolutional networks. *Phil. Trans. R. Soc*, 2016.

[13] T. Miller, Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 2019.

[14] M. Minsky, Matter, Mind and Models, *Proc. of the International Federation of Information Processing Congress*, 1965

[15] C. Molnar, *Interpretable Machine Learning*, 2018.

[16] M. Morrison M., *Reconstructing Reality: Models, Mathematics, and Simulations*, Oxford University Press, 2015.

[17] W. L. Oberkampf, Christopher J. Roy, *Verification and Validation in Scientific Computing*, Cambridge, 2015.

[18] J. Pearl, *Models, Reasoning, and Inference*, 2009.

[19] M. Pegny, M. I. Ibnouhsein, Quelle transparence pour les algorithmes d'apprentissage machine ?, *Revue d'Intelligence Artificielle*, 2019.

[20] A. Rahimi, Machine Learning has become alchemy, *Thirsty-first Conference on Neural Information Processing Systems*, 2017.

[21] S. Shalev-Schwartz, S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms,* 2014.

[22] R. Thom, *Prédire n'est pas expliquer*, Flammarion 1991.

[23] F. Varenne, Modèles et simulations dans l'enquête scientifique : variétés traditionnelles et mutations contemporaines, *Modéliser & Simuler. Épistémologies et pratiques de la modélisation et de la simulation*, Tome I, F. Varenne, M. Silberstein (dir.), Matériologiques, 2013.

[24] F. Varenne, *From Models to Simulations*, Routledge, 2018.

[25] C. Villani, M. Schoenauer, Y. Bonnet, C. Berthet, A. C. Cornut, F. Levin, B. Rondepierre, Donner un sens à l'intelligence artificielle *Mission Villani sur l'intelligence artificielle*, 2018.

[26] M. Zitnik, M. Agrawal, J. Leskovec, Modeling polypharmacy side effects with graph convolutional net- works, *Bioinformatic*, 2018.