Interprétabilité et explicabilité de phénomènes prédits par de l'apprentissage machine

Christophe Denis, Franck Varenne

Christophe Denis, Franck Varenne

Interprétabilité et explicabilité de phénomènes prédits par de l'apprentissage machine

Volume 3, nº 3-4 (2022), p. 287-310.

<http://roia.centre-mersenne.org/item?id=ROIA_2022__3_3-4_287_0>

Revue Ouverte d'Intelligence Artificielle
Volume 3, nº 3-4, 2022, 287-310

Interprétabilité et explicabilité de phénomènes prédits par de l'apprentissage machine

Christophe Denis[a], Franck Varenne[a]

[a] Université de Rouen, ERIAC, UFR LSH, Rue Lavoisier, 76821 Mont-Saint-Aignan, France
Adresse actuelle *:* Sorbonne Université, Laboratoire d'Informatique de Paris 6, 4, place Jussieu, Paris, France
E-mail : 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 et ceux de sortie. Après avoir proposé une clarification et une adaptation des notions d'interprétabilité et d'explicabilité telles qu'on les rencontre dans la littérature déjà abondante sur le sujet, 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. — Apprentissage machine, interprétabilité, explicabilité, epistémologie.

Les résultats, souvent spectaculaires, de l'apprentissage machine (AM) suscitent à la fois de forts espoirs, des craintes légitimes, notamment en termes d'éthique et de transformation du travail [6], et véhiculent un certain nombre de fantasmes [10]. La conception de l'algorithme de rétro-propagation du gradient a permis de mettre à jour efficacement les poids des réseaux de neurones plus profonds. Les frontières de décision des réseaux de neurones sont devenues donc plus complexes, non linéaires, mettant fin à la critique émise en 1969 par M. Minsky concernant le perceptron, qui a freiné fortement l'activité de recherche sur les réseaux de neurones au profit de l'Intelligence Artificielle symbolique. Les effets de cette avancée algorithmique ont été amplifiés par la possibilité d'accroitre la puissance de calcul en utilisant des cartes graphiques et par une production globale de données toujours croissante, permettant ainsi la mise au point de topologies de réseaux de neurones complexes. L'industrialisation de démonstrateurs développés dans des laboratoires n'est que peu souvent au rendez-vous, comme le souligne [11]. 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 [32] à une forme d'alchimie et le problème de sa transparence est considéré comme un défi scientifique majeur dans un nombre important de guides de recommandation dont [39], [15].

L'apprentissage machine bouleverse également la discipline scientifique. Depuis la conférence scientifique ECCV (European Conference on Computer Vision), en 2012, les capacités prédictives des réseaux de neurones profonds, en particulier les réseaux neuronaux convolutifs, ont bouleversé en profondeur la discipline du traitement et de la reconnaissance d'images. Ces résultats peuvent conforter en première approche la fin prophétisée de l'usage de modèle dans la discipline scientifique puisqu'il serait possible d'analyser les données sans formuler d'hypothèses les concernant [1]. Les failles de cette argumentation ont été présentées par plusieurs travaux de recherche, en particulier dans le cas des sciences sociales [36]. Pour autant, il est indéniable que de nombreuses disciplines scientifiques notamment computationnelles développent des activités de recherche orientées vers l'apprentissage machine. Au delà d'un possible effet d'aubaine, ces activités de recherche n'ont pas la même finalité. Il peut s'agir de :

(1) diminuer la puissance de calcul nécessaire pour mener des simulations numériques, en construisant un métamodèle ;
(2) augmenter les capacités prédictives d'un processus déjà basé sur une approche statistique comme par exemple pour l'étude du climat [16] ou en biologie moléculaire [12] ;
(3) combler le manque de connaissance d'un phénomène modélisé selon une approche hypothético-déductive, c'est-à-dire décrit par un modèle représentant d'une manière ou d'une autre la physique effective du phénomène (ses lois) et recourant souvent à des équations mathématiques tirées de ces mêmes lois. Il s'agit donc, alternativement à cette approche modélisant les phénomènes effectifs supposés, de chercher par exemple à prédire le comportement turbulent d'un fluide [29], la détection avec plus de précision d'un ensemble d'anomalies magnétiques en géosciences [4], et le comportement d'un système chaotique avec des temps de prédiction dépassant ceux obtenus en résolvant les équations mathématiques [30].

Nous émettons l'hypothèse que les deux premières finalités n'introduisent pas un profond changement epistémologique. En particulier, concernant la deuxième finalité, l'article [16] conclut que « *the same continuum where these various criteria of understanding come in degrees, and that therefore machine learning methods do not necessarily constitute a radical departure from standard statistical tools, as far as understanding is concerned* ». La troisième finalité est au cœur de notre projet de recherche. Il s'agit de montrer que le processus d'explicabilité de l'apprentissage machine 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'entrée 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. En effet, le fait que les éléments intervenant dans un processus computationnel ou que leurs interactions computationnelles elles-mêmes aient individuellement (ou par types) un sens n'entraîne pas pour autant le fait que la computation d'ensemble ait elle-même un sens interprétable qui conduirait par là ensuite à son explication. Concernant ce qu'il est convenu d'appeler la transparence du modèle par décomposabilité, [18] écrit ainsi à juste raison « *Note that this notion of interpretability requires that inputs themselves be individually interpretable, disqualifying some models with highly engineered or anonymous features* ». Il donne le contre-exemple des poids de certains modèles linéaires qui peuvent être à la fois intuitifs (sémantique cognitive) mais ne référer pourtant à rien de réel ou d'absolu (sémantique référentielle) dans le système cible ;

(2) la conception d'un modèle d'apprentissage machine 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. Donc, à supposer même l'existence d'une bonne interprétabilité des éléments intervenant dans le processus computationnel, le fait que le scénario causal ne soit pas explicable peut empêcher ensuite un contrôle correct du prétraitement des données (pertinence, biais dans les données). Concernant ce sujet spécifique, [18] poursuit sur le contre-exemple des poids des modèles linéaires : « *The weights of a linear model might seem intuitive, but they can be fragile with respect to feature selection and pre-processing* » ;

(3) enfin, l'explication, quand elle est possible, n'est pas pour autant assurée d'être elle-même 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 [14]. Nous montrerons ici l'intérêt de mettre en œuvre la distinction épistémologique entre fonction et usage d'un modèle [37], [38].

Concernant le caractère problématique et flottant du sens d'« interprétabilité » dans la littérature sur les modèles d'apprentissage machine, nous rejoignons donc le constat général fait par [18] lorsqu'il écrit : « *the term interpretability is ill-defined, and thus claims regarding interpretability of various models may exhibit a quasi-scientific character [...] Before we can determine which meanings might be appropriate, we must ask what the real-world objectives of interpretability research are.* ». Mais notre méthodologie se distingue de la sienne en ce qu'elle ne renonce pas d'entrée à supposer

un sens unifié à ce terme mais qu'elle réserve l'approche pluraliste et perspectiviste (*i.e.* chaque fois adaptée à la demande, au public visé et à ses compétences) pour les notions secondes d'explicabilité et de compréhensibilité. Notre méthodologie consiste dans un premier temps en particulier à s'inspirer des résultats issus de l'épistémologie des modèles en abordant les modèles à apprentissage machine au titre de modèles prédictifs. Notre contribution s'organise de la manière suivante :

- la section 1 présente les principales étapes de modélisation d'un phénomène physique avec deux approches différentes : décrire le phénomène à l'aide d'équations mathématiques, dont la forme est justifiée par des hypothèses plus ou moins précises ou approchées quant au comportement effectif (physique, biologique ou social) du système cible (approche hypothético-déductive) ou à partir de données ou de mesures le concernant (approche apprentissage machine). Nous justifierons également les termes de modèle et modélisation dans le cadre de l'apprentissage machine. L'usage de ces termes dans un tel contexte est en effet contesté : on lit parfois que les techniques d'apprentissage ne sont nullement de la modélisation. Mais nous suivrons ici plutôt la leçon de [34] qui parle précisément, dans ce même contexte, de modélisation prédictive et de « learning models ». Par ailleurs, nous verrons que, dans l'approche par AM, la modélisation prédictive du phénomène ne se résume pas à la mise au point de la méthode d'apprentissage machine sur des mesures le concernant. Car, en pratique, les mesures expérimentales ne sont pas toujours en quantité suffisante pour alimenter la phase d'apprentissage d'un réseau de neurones : on utilise alors des données simulées, fondées elles-mêmes sur un modèle. De plus, les données peuvent être pré-traitées, combinées en amont du processus d'apprentissage ou durant celui-ci. Par exemple, les travaux comme ceux présentés dans [8] ont pour objet la construction des caractéristiques interprétables par le physicien à partir de données initiales. C'est donc l'ensemble de la chaîne de modélisation à apprentissage machine qu'il convient d'expliquer comme c'est d'ailleurs le cas pour une approche hypothético-déductive [28] ;
- la section 2 présente des propositions de définitions pour l'interprétabilité et l'explicabilité de l'apprentissage machine. Il existe plusieurs définitions qui ne font pas encore consensus. Nous proposons une définition de l'interprétabilité fondée sur des aspects sémantiques et ne mobilisant pas la notion de compréhension. La notion de compréhension sera alors définie plus tard, à partir de celle d'interprétation (et non l'inverse), et sera distinguée de celle d'explication ;
- la section 3 caractérise ensuite les différentes fonctions de connaissance ou fonctions épistémiques d'un modèle. C'est dans cette section que 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 4 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 5 introduit la différence entre fonction (épistémique) et usage (pratique ou rhétorique) d'un modèle. Couramment, la demande d'explicabilité d'un modèle à AM 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 section 6 présente nos travaux en cours sur l'explicabilité de phénomènes physiques en utilisant de l'apprentissage machine.

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

1. Prédiction d'un phénomène physique en utilisant une méthode d'apprentissage machine

Les mathématiques ont pris une place prépondérante parfois même exclusive en physique notamment depuis les travaux de Galilée. En 1623, Galilée écrit que « *La philosophie [naturelle] est écrite dans cet immense livre qui se tient toujours ouvert devant nos yeux, je veux dire l'univers, mais on ne peut le comprendre si l'on ne s'applique d'abord à en comprendre la langue et à connaitre les caractères dans lesquels il est écrit. Il est écrit en langue mathématique, et ses caractères sont des triangles, des cercles et autres figures géométriques, sans le moyen desquels il est humainement impossible d'en comprendre un mot.* ». Cet extrait pose plusieurs problèmes métaphysiques au sujet du statut des mathématiques, et en particulier celui qui nous intéresse ici : s'agit-il du langage même des phénomènes naturels, qu'il est nécessaire de connaitre pour comprendre et communiquer avec la nature, ou d'un médiateur pour accéder à des connaissances sur ces phénomènes naturels ? En un sens, on peut dire que Platon avait imaginé également ce rôle de médiateur, dans le livre 6 de la République, cette fois-ci entre deux mondes ou deux lieux : un lieu accessible à nos sens (monde sensible) et un lieu accessible à l'intelligence (monde intelligible). En faisant écho en quelque sorte à cette intuition platonicienne, l'apprentissage machine peut servir à lier un « monde intelligible » censé contenir l'ensemble des lois physiques effectives, dont certaines comme la turbulence nous échappent encore, avec le monde sensible comportant quant à lui, et de notre point de vue, des processus expérimentaux entachées immanquablement d'incertitudes. Inspirée par cette distinction millénaire et

évoquée ici rapidement, notre hypothèse de recherche propose de considérer l'apprentissage machine comme une grille descriptive médiatrice entre le comportement légal d'un phénomène physique, dont la loi peut être encore mal ou non connue, comme la turbulence, et les observations le concernant.

Cependant, les mathématiques et les données peut intervenir différemment selon les approches utilisées pour prédire un phénomène. Prenons l'exemple d'un écoulement d'un fleuve dont on souhaite connaître la hauteur d'eau pour différentes valeurs de crues. Dans le cadre d'une approche hypothético-déductive : l'écoulement est modélisé à l'aide d'équations mathématiques, par exemple en utilisant les équations de Navier-Stokes. Ces équations sont déterminées le plus souvent sur des supports continus qu'il est nécessaire de discrétiser pour obtenir un schéma de calcul spatio-temporel. Ce schéma numérique est exécuté sur un ordinateur à partir de données, en utilisant par exemple une approche lagrangienne ou eulérienne qui est ensuite implémentée dans le programme informatique.

Une démarche de validation et de quantification d'incertitudes utilise d'autres données, essentiellement des mesures, pour accréditer les résultats obtenus [28]. Cette approche a fait la preuve de sa robustesse et de son efficacité dans de nombreuses disciplines scientifiques. Elle est consubstantielle liée à l'amélioration de la performance des architectures de calcul. Le schéma de calcul spatio-temporel nécessite un volume de calcul de plus en plus important pour traiter des maillages toujours mieux conçus pour d'une part respecter les conditions de Courant–Friedrich–Levy et d'autre part, prédire plus précisément les phénomènes. La première limitation de cette approche est liée à la fin de validité de la loi empirique de Moore, prédisant tous les deux ans un doublement du nombre de transistors, directement lié à leurs puissances. Cela pose aussi clairement un problème de sobriété énergétique. L'utilisation de méthodes d'apprentissage permet de diminuer le volume de calcul lors de la prédiction de phénomènes puisque seule la phase d'apprentissage a besoin d'un volume de calcul conséquent. De nombreuses disciplines scientifiques s'orientent vers l'apprentissage machine pour lever d'autres limitations de l'approche hypothético-déductive :

- l'apprentissage machine peut être utilisé pour modéliser le comportement turbulent d'un fluide [29] ;
- l'apprentissage automatique a été mis en œuvre pour mieux prédire le système chaotique Kuramoto-Sivashinsky avec des temps de prédiction dépassant ceux obtenus en résolvant les équations mathématiques [30] ;
- un autre exemple provient du domaine de la géoscience dans lequel des algorithmes d'inversion sont utilisés pour identifier des structures géologiques et reconnaître des anomalies. Ces algorithmes souffrant d'un problème de robustesse, l'utilisation de réseaux de neurones convolutifs permet de détecter avec plus de précision et plus rapidement un ensemble d'anomalies magnétiques [4].

Reprenons le cas de la prédiction de la hauteur d'eau d'un fleuve en utilisant un modèle d'apprentissage. La phase d'apprentissage, nécessitant un volume de calcul

pouvant être important, permet de fixer les paramètres du modèle. L'inférence du modèle permet d'obtenir des prédictions avec un volume de calcul beaucoup moins élevé. Ce modèle pourrait aussi prendre en compte des phénomènes de turbulence encore mal traduits sous la forme d'équations. On peut considérer, en première intention, l'apprentissage machine comme une approche inductive dans laquelle le manque de connaissance du phénomène est compensé par le traitement statistique potentiellement sophistiqué des données. Mais ce n'est généralement pas le cas pour les deux raisons suivantes :

(1) le fonctionnement des méthodes d'apprentissage machine les plus sophistiquées reste complexe à expliquer ;
(2) la mise au point d'une application basée sur de l'apprentissage machine va au delà du seul entrainement de la méthode en elle-même.

Concernant le premier point, les méthodes d'apprentissage machine les plus sophistiquées, comme les réseaux de neurones profonds ou antagonistes, généralement les plus performantes, souffrent d'un déficit de capacité à expliquer leur système cible comme aussi d'un déficit d'explicabilité en raison :

- d'absence de causalité ;
- de la difficulté à comprendre le fonctionnement des méthodes d'optimisation appliquées sur un nombre très important de paramètres ;
- de l'absence de lien direct et directement interprétable entre les prédictions et l'évolution du phénomène cible. Il est montré dans [27] que :
  - dans les cas des modèles d'apprentissage dits agnostiques, l'espace de fonctions utilisées est non seulement limité, mais surtout il n'est pas sélectionné en fonction de la nature spécifique du phénomène concerné. Il ne peut donc être justifié par aucune sorte d'induction préalable effectuée à partir du système cible spécifique étudié. Ce caractère agnostique du modèle d'apprentissage ne signifie toutefois pas qu'un tel modèle ne repose sur aucune hypothèse mathématique préalable. Comme le montre [34], les modèles d'apprentissage agnostiques de type « PAC » (probably approximately correct), par exemple, reposent sur des hypothèses mathématiques minimales et nécessaires comme l'hypothèse de plongée préalable du domaine d'objets dans une sigma algèbre (ou tribu), munie des bonnes propriétés qui nous sont nécessaires (en termes de théorie de la mesure), ainsi que l'hypothèse de définissabilité d'une distribution quelconque sur ce domaine ;
  - les données sont régularisées pour assurer la convergence des méthodes d'optimisation mais ne représentent plus toujours mathématiquement le phénomène lui-même.

Le processus de mise au point incorpore de la connaissance portant sur le phénomène en lui-même ou par analogie à travers d'autres phénomènes connus de la discipline scientifique associée. Ainsi, la prédiction d'un phénomène physique utilisant

de l'apprentissage machine ne se limite pas à l'entraînement d'une méthode d'apprentissage sur des données le concernant. Prenons à titre d'illustration une application de l'apprentissage machine en géophysique [5], dont l'objectif est de remplacer des techniques d'inversion peu robustes et peu performantes. Ces techniques d'inversion prennent comme données des cartes géomagnétiques du sous-sol obtenues par prospection, pouvant par exemple indiquer la localisation de munitions enfouies dans le sable d'une plage. Ces cartes géomagnétiques restent le plus souvent confidentielles, et en quantité limitées en raison du coût de la prospection géophysique. Il est à noter également que la topologie des anomalies dépend fortement de la nature de ce qui est recherché dans le sous-sol. Pour faire part à cette pénurie de données, l'approche retenue dans [5] est de constituer une base de données de cartes géomagnétiques par simulation, que l'on espère suffisamment générique, en utilisant des équations utilisées en géophysiques. Une approche similaire a été mise en place dans le domaine de la physique des particules pour lequel une méthode d'apprentissage machine est utilisée pour améliorer la détection de particules dans un accélérateur [8].

En résumé, le coût et la complexité, et parfois la confidentialité, des mesures expérimentales ne permettent pas toujours d'être dans une approche de données massives en accès libre. Ainsi la mise en place d'une méthode d'apprentissage pour prédire un phénomène physique inclut potentiellement :

- une phase de génération des données par simulation, puisque les mesures ou les données initiales ne sont pas toujours en quantité suffisante ou suffisamment bien réparties pour entraîner efficacement une méthode d'apprentissage. Certaines données sont ainsi générés par simulation ;
- les données sont ensuite traitées (remplacement des valeurs manquantes ou des valeurs aberrantes) puis parfois même combinées pour construire des caractéristiques plus aisément interprétables, cette phase est obligatoirement effectuée ;
- le domaine d'apprentissage de la méthode peut être différent du domaine d'utilisation. Il est dans ce cas nécessaire de mettre en place des techniques d'adaptation de domaine ou d'utiliser de l'apprentissage par transfert.

Il est donc nécessaire de définir épistémologiquement les fonctions du processus dans son ensemble et pas uniquement celles de la méthode d'apprentissage.

2. Interprétabilité et explicabilité de l'apprentissage machine

L'utilisation de l'apprentissage machine provoque de multiples interrogations allant bien au-delà de la préoccupation classique d'éthique venant des usagers des sciences et des techniques. Il existe donc une pression forte sur l'explication des résultats produits par l'apprentissage machine. L'explication peut être destinée par exemple à des ingénieurs en R&D qui utilisent le plus souvent l'apprentissage machine en mode « boîte noire ». Une autre préoccupation plus théorique et académique concerne l'évolution du savoir en raison du caractère limité et contestable du « prédire sans comprendre » [35]. Pour autant, il n'existe pas un consensus sur la définition de

l'interprétabilité et de l'explicabilité ce qui nous semble être la difficulté originelle des travaux scientifiques portant sur l'explication en apprentissage machine, comme rappelé par les deux fameux vers de Nicolas Boileau « *Ce que l'on conçoit bien s'énonce clairement, Et les mots pour le dire arrivent aisément* » [3].

L'objectif de cette section est de proposer des définitions alternatives de l'interprétabilité et de l'explicabilité de l'apprentissage machine, fondées sur une analyse préalable des définitions courantes proposées par [24] [19] [22].

2.1. Analyse de définitions de l'interprétabilité et de l'explicabilité de l'apprentissage machine

La définition suivante de l'interprétabilité a été proposée par [24] [19] [22] [25] :

Définition 2.1. — *« L'interprétabilité réfère au degré de compréhensibilité humaine d'un modèle de type boîte noire ou d'une décision »*

Cette définition de l'interprétabilité mobilise la notion problématique de compréhension, elle-même non préalablement définie. Et elle inverse le rapport habituel de détermination interprétabilité-explicabilité-compréhension.

D'après [24], les différentes significations que peut recouvrir le terme d'« explication » se scinde quant à elle ensuite en deux grands types selon que l'explication est à destination du spécialiste (modélisateur, programmeur) ou du non spécialiste (utilisateur, personne affectée par la décision). Pour le premier grand type, cette signification du terme explication rejoint en gros le domaine déjà vaste et nommé XAI (eXplainable AI) dans les publications techniques : une telle recherche d'explication favorise un travail de recherche scientifique de modélisation explicative secondaire du modèle primaire fonctionnant à base d'apprentissage. Alors que le second grand type d'explication mobilise des considérations de théories de la cognition, de psychologie cognitive et de théorie de l'argumentation ou de la persuasion en rappelant que l'explication pour le non-spécialiste ne peut être exclusivement technique ni chercher à être exhaustive mais qu'elle doit tendre à être de trois ordres : 1) contrastive, 2) sélective et 3) interactive. Les auteurs de [24] insistent en particulier sur la dimension finalement toujours interactive d'une telle explication pour le non spécialiste. Une telle explication consiste génériquement en le fait d'échanger des informations au sujet d'un phénomène. Le constat que l'on peut tirer de ces travaux de synthèse est donc, au minimum, que l'explication est chargée d'un grand nombre de fonctions différentes, comme par exemple celle de montrer une conformité du modèle d'apprentissage machine à une législation, d'améliorer la confiance du modèle dans les décisions qu'il prend ou de vérifier et améliorer les fonctionnalités du modèle. Comme l'explication ainsi proposée dans ces grandes lignes est destinée à différents acteurs (par exemple tantôt les développeurs experts, tantôt les utilisateurs). elle peut avoir parfois un rôle prioritairement pédagogique, de persuasion de bonne foi ou de mauvaise foi (manipulation, idéologie, biais accepté).

Malgré la pertinence de ces distinctions finales synthétisées par [24], les définitions conceptuelles proposées initialement ne nous semblent pas permettre de distinguer

clairement l'interprétation de l'explication. Ainsi, dans cette publication comme dans celles sur lesquelles elle s'appuie, ces deux termes sont souvent utilisés de manière inconsistante. En outre, le sens du terme interprétation lui-même reste vague : cela est dû au fait que ces publications conditionnent dès le début toute interprétation à une compréhension humaine alors que c'est l'inverse qui paraît plus vraisemblable. En fait, une interprétation n'a pas besoin d'une compréhension préalable, ni d'une explication. Le sens du terme explication, quant à lui, recouvre beaucoup d'éléments empêchant une définition claire et immédiate. Nous proposons donc ci-dessous des définitions alternatives dans le but de clarifier ces points.

2.2. Proposition de définitions alternatives

La définition de l'interprétabilité proposée par [24] [19] [22] [25] mobilise la notion de compréhension sans qu'elle soit définie, et cela alors même que la notion de compréhension semble au contraire devoir se fonder sur celle d'interprétation, plutôt que l'inverse. Signe de ce flou persistant, [25] va ensuite jusqu'à considérer que, par conséquent, les termes « interprétable » et « explicable » sont, pour lui, interchangeables.

C'est notamment pour essayer d'éviter ces flous définitionnels comme ces interchangeabilités conceptuelles qu'il est proposé dans [37] [38] la définition suivante de la compréhension d'un phénomène ou d'un calcul.

Définition 2.2. — *Compréhension d'un phénomène, ou d'un calcul (du latin cum-prehendere : saisir, appréhender ensemble) : il y a compréhension d'un phénomène quand notre esprit dispose de la possibilité d'en saisir l'ensemble et d'en unifier les manifestations successives ou diverses sous une représentation à la fois unique et aisée à concevoir et à rappeler à l'esprit.*

Quand on la définit ainsi, on voit que la notion de compréhension suppose d'abord, dans une première étape, que nous soient désignables ou mobilisables en esprit des symboles référant à certaines manifestations diverses du système cible en question. On voit ensuite que ces symboles doivent, dans une seconde étape, pouvoir être rassemblés et saisis ensemble, que ce soit statiquement ou dynamiquement (l'explication n'est pas encore en jeu ici : voir la caractérisation d'une explication plus bas), via un autre symbole qui quant à lui réfère à un geste mental d'unification opératoire, geste mental souvent mathématique et permettant une saisie intellectuelle à la fois cohérente, uniforme et facile à remobiliser intellectuellement. Cette contrainte d'unification est souvent réalisée du fait de l'existence reconnue d'une relation unifiante entre ces manifestations diverses et une norme unique, ou une valeur ou encore un principe d'optimisation. C'est seulement à l'issue de ces deux étapes qu'une compréhension – ou saisie unifiante, cohérente et synthétique du divers d'un phénomène – est acquise.

Avec cette définition et cette analyse préalable du processus de compréhension, on voit bien que l'interprétation ne peut pas supposer acquis un tel processus aussi riche, complexe et donc aussi rare. On voit en outre et en revanche que c'est la première étape qui correspond à l'interprétabilité. La compréhension suppose donc l'interprétation et l'interprétabilité, non l'inverse. L'interprétabilité, comme on va le

voir, si elle nécessite certes des symboles référentiels à la fois discrets et plus ou moins articulés, n'est pas pour autant forcément atomistique ni relevant d'une analyse ultime sinon toute compréhension fondée sur une interprétabilité devrait ensuite reposer sur une transparence cognitive intégrale, ce qui est rarement le cas et contestable dans sa faisabilité même, comme le remarque justement [18]. Mais retenons ici que, à ce stade, la simple interprétabilité n'entraîne pas nécessairement la compréhensibilité ni, *a fortiori*, la compréhension effective.

C'est précisément pour cette raison que la définition de l'interprétabilité que nous proposerons ici et que nous proposions déjà dans [9] est prioritairement conçue simplement et uniquement à partir d'une sémantique possible. Rappelons ici que la sémantique, cette capacité pour un signe d'avoir un sens, peut à son tour être cognitive ou référentielle : « une sémantique référentielle prend pour base la relation entre les formes linguistiques et ce qu'elles représentent ; une sémantique cognitive prend pour base les représentations mentales auxquelles les formes linguistiques sont associées » [33].

L'interprétabilité, au sens où nous l'entendons ici, est donc fondamentalement liée à la possibilité de relier des symboles individuels à des entités ou propriétés d'entités possibles, fictives, c'est-à-dire, simplement pensées, ou encore réelles. Une fois cela précisé, on voit aussi qu'une interprétation peut ne pas concerner uniquement un ensemble simplement non-structuré de symboles. Elle peut concerner en particulier un modèle qui, quant à lui, possède au minimum une certaine structure entre les symboles qui le constituent. Définissons donc maintenant l'interprétabilité d'un modèle.

Définition 2.3. — *Interprétabilité d'un modèle : propriété que possède un ensemble de symboles ou un modèle (une structure de symboles) de se voir composé d'éléments (signes, symboles, figures, concepts, données, etc.) qui ont chacun un sens [c'est-à-dire un référent possible] pour un sujet humain.*

On a donc obtenu une définition sémantique de l'interprétabilité d'un modèle : un modèle est interprétable quand tous ses symboles le sont. La notion de transparence souvent employée nous semble s'ensuivre. Elle ne nous paraît être qu'une modalité de la notion d'interprétabilité : un modèle, ses composants, son algorithme sont plus ou moins transparents en fonction de la plus ou moins grande capacité ou difficulté qui est la nôtre (comme concepteur, analyste ou encore utilisateur final) d'en déterminer une interprétation effective et complète.

À ce sujet, [18] objecterait qu'une telle définition de l'interprétabilité reste floue et seulement quasi-scientifique car le cardinal ou la complexité effective de l'ensemble de symboles ou de la structuration de ces symboles peuvent excéder les capacités cognitives d'un sujet humain d'embrasser d'un seul regard de l'esprit une telle complexité. Mais cette objection ne tient précisément que si on continue à mêler fautivement selon nous – la clause de compréhensibilité à celle d'interprétabilité, ce qui est prématuré. Il suffit que l'interprétation soit réalisable en principe pour qu'on évite l'objection du caractère non effectif pratiquement de l'interprétation. Remarquons ici que nous ne suivons pas non plus la caractérisation que [24] donne de la transparence puisque cette

dernière selon ces auteurs, concerne déjà, entre autres, « la compréhension [sic] mécanistique du fonctionnement du modèle (simulabilité) » alors que, selon nous, ce point concerne déjà clairement l'explicabilité du modèle dès lors que le fonctionnement du modèle est supposé en jeu.

Nous définirons précisément plus bas ce que nous entendons par explication. Mais nous pouvons nous en faire ici une première idée de façon à définir tout de suite ce qu'on peut entendre par l'explicabilité d'un modèle. Explication vient du latin explicare qui signifie déplier une structure, et donner par là à voir ses jointures extérieurement. Disons, en première approche, que nous avons affaire à l'explication d'un phénomène quand ce phénomène (ou encore calcul ou modèle en fonctionnement) est représentable en une série d'étapes ou d'états distincts qui peuvent être représentés comme immédiatement successifs ou conjointement articulés et que cette succession ou articulation est en même temps représentée comme n'étant pas contingente mais due directement aux entités du système ou à certaines de leurs propriétés intrinsèques.

À partir de là, nous pouvons proposer une définition de l'explicabilité pour un modèle : un modèle est explicable quand son fonctionnement donne lieu à des représentations des articulations entre ses états qui sont interprétables. Or, l'interprétation de relations directes (donc pas de simples corrélations) de succession ou de conjonction non contingente entre étapes ou états se fait d'ordinaire en termes de causes ou de raisons. Dans le cas d'un modèle, cette explicabilité va porter sur l'algorithme (par exemple l'algorithme d'apprentissage), ses différentes étapes et articulations, mais aussi sur ses caractéristiques d'entrée et de sortie [9]. Une explication de modèle nécessitera en premier lieu une interprétation des caractéristiques d'entrée et de sortie ainsi qu'une interprétation des relations causales plus ou moins directes entre elles.

Définition 2.4. — *L'explicabilité d'un modèle pourvu d'un algorithme est 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.*

La section suivante rappelle le nécessaire besoin de clarification épistémologique des différentes fonctions épistémiques (ou de connaissance) des modèles et elle restitue ces notions d'interprétabilité et d'explicabilité dans le cadre précis de cette clarification.

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

3.1. 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'apprentissage machine 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 [40]. 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] en faveur de l'idée selon laquelle 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 [35]. 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.

3.2. 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* » [23]. 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 [37] [38]. Les plus fréquentes sont l'analyse ou la réduction de données, la description, la prédiction et l'explication ([37], [38] : fonctions 5, 6, 7 et 8).

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 » ([37] [38] ; 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 prédites 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 existe 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 [34]. 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 contemporaines, 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 ([26] 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 : [31]), 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 utilisés 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 vérification-validation (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.

4. Explication causale, interprétabilité et explicabilité en apprentissage machine

Il existe 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 dans l'algorithme implémentant un réseau de neurones comme par exemple l'algorithme de rétropropagation du gradient servant à déterminer les pondérations du modèle. 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 modèle de 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. Les modèles à apprentissage machine se fondent sur la recherche de généralisations [34] et, assez souvent, mais pas toujours, sur l'analyse de corrélations entre les différentes données d'entraînement (entrées, feedbacks dans les cas d'apprentissage par renforcement ou supervision). Or, une corrélation

statistique pas plus qu'une généralisation d'association ou de pattern entre deux données ne signifie qu'il existe une causalité entre elles. 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 réseaux de neurones artificiels 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 [21] ne fait que confirmer ce soupçon qu'un réseau de neurones artificiel 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 [13]. 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.

5. Fonction épistémique et usage du modèle en apprentissage machine

Comme présenté en section 1, la conception d'une application à AM 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 réseaux de neurones artificiels 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.

6. 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 [22] : « 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 tout d'abord de certains travaux de psychologie cognitive [20]. 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. Nous souhaitons ensuite utiliser les concepts de pertinence et de simplicité issus de la théorie de l'information algorithmique [7], pour formaliser et quantifier la qualité d'une explication, puis de comparer celle-ci avec l'évaluation qualitative.

7. 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 [35], 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 » ([35], 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. Il y a ainsi lieu de penser qu'un choix plus riche existe que celui qui nous semble souvent proposé – mais donc probablement à tort – entre ces deux seuls extrêmes : tantôt un modèle explicatif fondé sur une théorie, tantôt un modèle purement prédictif dépourvu de toute théorie. Dans cet article, nous avons ensuite caractérisé la recherche d'interprétabilité et d'explicabilité des modèles en termes de modélisation de modèle. Pour cette étape, nous nous sommes donc prioritairement concentrés sur les questions posées par une telle recherche lorsqu'on adopte une approche de type « eXplainable AI ». Cette approche a été caractérisée par [21] comme formant un contraste avec

cette autre approche de l'explicabilité répondant quant à elle à une demande sociale forte et davantage tournée vers les utilisateurs et les non-experts. Mais nous pensons que cette seconde approche pourra justement s'enrichir, à termes, du renouvellement de perspective tel que nous le proposons ici d'abord pour l'approche XAI. Nous avons enfin montré que dans cette perspective 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 [13], sera évalué dans ce cadre, ainsi que la prise en compte de l'incertitude de modèle d'AM telle qu'effectué par exemple dans [17] en utilisant des réseaux de neurones bayésiens.

Bibliographie

[1] C. Anderson, « The End of Theory : The Data Deluge Makes the Scientific Method Obsolete », *Wired* **16** (2008), nº 7.

[2] F. Bacon, « Novum organum », 1620.

[3] N. Boileau, *L'art poétique*, Flammarion, 1998.

[4] J. Cardenas, C. Denis, H. Mousannif, C. Camerlynck & N. Florsch, « Reseaux de Neurones Convolutifs pour la Caractérisation d'Anomalies Magnétiques », in *12ème colloque GEOFCAN, Grenoble, France*, 2021.

[5] J. J. Cárdenas Chapellín, C. Denis, H. Mousannif, C. Camerlynck & N. Florsch, « Réseaux de Neurones Convolutifs pour la Caractérisation d'Anomalies Magnétiques », in *CNIA 2021 : Conférence Nationale en Intelligence Artificielle* (Bordeaux (en ligne), France), 2021, p. 84-90.

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

[7] G. J. Chaitin, *Algorithmic Information Theory*, Cambridge Tracts in Theoretical Computer Science, Cambridge University Press, Cambridge, 1987.

[8] N. Cherrier, M. Defurne, J.-P. Poli & F. Sabatié, « Embedded Constrained Feature Construction for High-Energy Physics Data Classification », in *33rd Annual Conference on Neural Information Processing Systems* (Vancouver, Canada), 2019.

[9] C. Denis & F. Varenne, « Interprétabilité et explicabilité pour l'apprentissage machine : entre modèles descriptifs, modèles prédictifs et modèles causaux. Une nécessaire clarification épistémologique », in *National (French) Conference on Artificial Intelligence (CNIA) – Artificial Intelligence Platform (PFIA)* (Toulouse, France), 2019, p. 60-68.

[10] J.-G. Ganascia, *Le mythe de la singularité. Faut-il craindre l'Intelligence Articielle ?*, Le Seuil, 2017.

[11] J.-M. Ghidaglia & N. Vayatis, « Comment faire sortir l'intelligence artificielle des labos ? », Les échos, 2019.

[12] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams & A. Aspuru-Guzik, « Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules », *ACS Central Science* **4** (2018), p. 268-276.

[13] G. L. H., D. Bau, B. Z. Yuan, A. Bajwa, M. Specter & L. Kagal, « Explaining Explanations : An Overview of Interpretability of Machine Learning », in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 2018, p. 80-89.

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

[15] High-Level Expert Group on AI, « Ethics guidelines for trustworthy AI », Tech. report, European Commission, 2019.

[16] J. Jebeile, V. Lam & T. Räz, « Understanding climate change with statistical downscaling and machine learning », *Synthese* **199** (2021), p. 1877-1897.

[17] A. Kendall & Y. Gal, « What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision ? », in *Thirsty-first Conference on Neural Information Processing Systems, NIPS 2017*, 2017, p. 5580-5590.

[18] Z. Lipton, « The Mythos of Model Interpretability », *Communications of the ACM* **61** (2018), n° 10, p. 36-43.

[19] P. J. G. Lisboa, « Interpretability in Machine Learning – Principles and Practice », in *Fuzzy Logic and Applications* (Cham) (F. Masulli, G. Pasi & R. Yager, éds.), Springer International Publishing, 2013, p. 15-21.

[20] T. Lombrozo, « Explanation and Abductive Inference », *The Oxford Handbook of Thinking and Reasoning* (2012).

[21] S. Mallat, « Understanding deep convolutional networks », *Philosophical Transactions of the Royal Society A : Mathematical, Physical and Engineering Sciences* **374** (2016), n° 2065.

[22] T. Miller, « Explanation in artificial intelligence : Insights from the social sciences », *Artificial Intelligence* **267** (2019), p. 1-38.

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

[24] B. D. Mittelstadt, C. Russell & S. Wachter, « Explaining Explanations in AI », in *FAT* '19 : Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019.

[25] C. Molnar, « Interpretable Machine Learning – A Guide for Making Black Box Models Explainable », `https://christophm.github.io/interpretable-ml-book/`, 2019.

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

[27] D. Napoletani, M. Panza & D. Struppa, « The Agnostic Structure of Data Science Methods », `https://arxiv.org/abs/2101.12150`, 2021.

[28] W. L. Oberkampf & C. J. Roy, *Verification and Validation in Scientific Computing*, Cambridge University Press, 2010.

[29] S. Pandey, J. Schumacher & K. R. Sreenivasan, « A perspective on machine learning in turbulent flows », *Journal of Turbulence* **21** (2020), n° 9-10, p. 567-584.

[30] J. Pathak, B. Hunt, M. Girvan, Z. Lu & E. Ott, « Model-Free Prediction of Large Spatiotemporally Chaotic Systems from Data : A Reservoir Computing Approach », *Phys. Rev. Lett.* **120** (2018), article no. 024102.

[31] J. Pearl, *Causality : Models, Reasoning, and Inference*, Cambridge University Press, 2000.

[32] A. Rahimi, « Machine Learning Has Become Alchemy », in *Thirsty-first Conference on Neural Information Processing Systems, NIPS 2017*, 2017.

[33] F. Récanati, *Philosophie du langage (et de l'esprit)*, Folio-Essais, Gallimard, 2008.

[34] S. Shalev-Schwartz & S. Ben-David, *Understanding Machine Learning : From Theory to Algorithms*, Cambridge University Press, 2014.

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

[36] V. Tommaso, D. Cardon & J.-P. Cointet, « Présentation », *Réseaux* **188** (2014), n° 6, p. 9-21.

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

[38] ———, *From Models to Simulations*, Routledge, 2018.

[39] C. Villani, S. Marc, Y. Bonnet, C. Berthet, A.-C. Cornut, F. Levin & B. Rondepierre, *Donner un sens à l'intelligence artificielle : Pour une stratégie nationale et européenne*, 2018.

[40] M. Zitnik, M. Agrawal & J. Leskovec, « Modeling polypharmacy side effects with graph convolutional networks », *Bioinformatics* **34** (2018), nº 13, p. i457-i466.

Abstract. — The lack of explainability of machine learning (ML) techniques poses operational, legal and ethical problems. One of the main goals of our project is to provide ethical explanations of the outputs generated by an ML-based application, considered as a black box. The first step of this project, presented in this paper, is to show that the validation of these black boxes differs epistemologically from that implemented in the framework of a mathematical and causal modeling of a physical phenomenon. The major difference is that an ML method does not claim to represent causality between input and output parameters. After having provided a clarification and an adaptation of the notions of interpretability and explainability as found in the already abundant literature on the subject, we show in this article the fruitfulness of implementing the epistemological distinctions between the different epistemic functions of a model, on the one hand, and between the epistemic function and the use of a model, on the other hand. Finally, the last part of this article presents our current work on the evaluation of an explanation, which can be more persuasive than informative, and which can therefore raise ethical problems.

Keywords. — Machine learning, interpretability, explainability, epistemology.

Manuscrit reçu le 11 mars 2021, révisé le 16 août 2021, accepté le 16 septembre 2021.