Quelle transparence pour les algorithmes d'apprentissage machine ?

Maël Pégny, Issam Ibnouhsein

Quelle transparence pour les algorithmes d'apprentissage machine ?

Maël Mégny* Issam Ibnouhsein†

Résumé

La notion de « transparence des algorithmes » a récemment pris une grande importance à la fois dans le débat public et dans le débat scientifique. Partant de la prolifération des emplois du terme « transparence », nous distinguons deux familles d'usages fondamentaux du concept : une famille descriptive portant sur des propriétés épistémiques intrinsèques des programmes, au premier rang desquels l'intelligibilité et l'explicabilité, et une famille prescriptive portant sur des propriétés normatives de leurs usages, au premier rang desquels la loyauté et l'équité. Parce qu'il faut comprendre un algorithme pour l'expliquer et en réaliser l'audit, l'intelligibilité est logiquement première dans l'étude philosophique de la transparence. Afin de mieux cerner les enjeux de l'intelligibilité dans l'emploi public des algorithmes, nous introduisons dans un deuxième temps une distinction entre intelligibilité de la procédure et intelligibilité des sorties. Dans un dernier temps, nous appliquons cette distinction au cas particulier de l'apprentissage machine.

Abstract

Recently, the concept of "algorithmic transparency" has become of primary importance in the public and scientific debates. In the light of the proliferation of uses of the term "transparency", we distinguish two families of fundamental uses of the concept: a descriptive family relating to intrinsic epistemic properties of programs, the first of which are intelligibility and explicability, and a prescriptive family that concerns the normative properties of their uses, the first of which are loyalty and fairness. Because one needs to understand an algorithm in order to explain it and carry out its audit, intelligibility is logically first in the philosophical study of transparency. In order to better determine the challenges of intelligibility in the public use of algorithms, we introduce a distinction between the intelligibility of the procedure and the intelligibility of outputs. Finally, we apply this distinction to the case of machine learning.

*Université de Paris 1 Panthéon-Sorbonne, IHPST, 13 rue du Four, 75006 Paris
†Quantmetry, 52 rue d'Anjou, 75008 Paris, France

1 Introduction

La politique des algorithmes est couramment fort en vogue. Les algorithmes nous gouvernent, nous surveillent, nous profilent, et bientôt peut-être prendront notre emploi. L'enjeu de l'usage public des algorithmes est d'autant plus brûlant que, loin d'être des parangons d'objectivité scientifique, les algorithmes seraient racistes, homophobes, antisémites, sexistes ou classistes.

Bien des réflexions formulées à l'égard des problèmes politiques posés par l'usage des algorithmes pourraient se voir reprochées d'être un cas d'école de renommage de problèmes anciens. La philosophie, les sciences sociales et nombre d'œuvres artistiques ont déjà décrit et critiqué les absurdités et injustices provoquées par la complexité et l'opacité des processus de décision et règles qui structurent le fonctionnement des sociétés modernes. Bien des critiques aujourd'hui adressées à l'usage social des algorithmes étaient déjà formulées dans la critique de la bureaucratie, de l'usage des statistiques et des benchmarks dans le débat public et la prise de décision.

Nombre des questions et des positions développées dans la discussion de la bureaucratie ou de l'emploi des statistiques se retrouveront assurément dans la discussion politique de l'usage des algorithmes. En effet, nombre d'algorithmes ne sont qu'une automatisation des procédures auparavant exécutées à la main, sans nécessairement en modifier la nature ni même la complexité. Sans nier que l'automatisation puisse avoir un impact nécessitant un niveau de réflexion propre, il semble nécessaire de veiller à ne pas sombrer dans un technocentrisme qui masquerait l'héritage intellectuel de la critique de la bureaucratie, attribuant ainsi aux algorithmes tous les problèmes dus à la complexité, à l'opacité ou à l'absurdité des décisions prises. Le technocentrisme qui mène si rapidement à parler de gouvernance par les algorithmes, de décision prise par la technologie, ou d'intentions du programme servent alors d'écran dissimulant des enjeux politiques, éthiques et juridiques à la fois plus anciens et plus profonds.

Ces précautions prises, on peut partir à la recherche de ce qui fait la véritable nouveauté de l'usage des algorithmes dans nos sociétés. Sans prétendre épuiser le sujet, l'étude de l'apprentissage machine (AM) (*machine learning* en anglais) nous semble une voie d'étude privilégiée. L'AM, qui représente une variante particulière de l'intelligence artificielle (IA) (voir Fig. 1) [1], s'est largement développé pour remédier aux situations où « la théorie de la décision aisément comprise n'est pas suffisante » [1]. Ils ne peuvent donc être décrits comme une simple automatisation de procédures préexistantes. Ces algorithmes sont en outre réputés difficiles à comprendre, pour leur concepteur expert tout comme leur utilisateur profane, parce que leurs procédures d'optimisation sont précisément en rupture avec les modalités de raisonnement humain, et ne peuvent être aisément traduites en ses termes. Cette différence de nature technique, qu'il convient d'expliciter, pose des défis radicalement nouveaux lorsque ces algorithmes d'AM sont utilisés pour étendre le domaine de la prise de décision automatisée. Sans prétendre exclure les autres catégories d'algorithmes d'IA comme les systèmes

1. Nous ne détaillerons pas plus avant les différents types d'IA, et laisserons le soin au lecteur d'approfondir les définitions de ces classes de modèles si besoin.

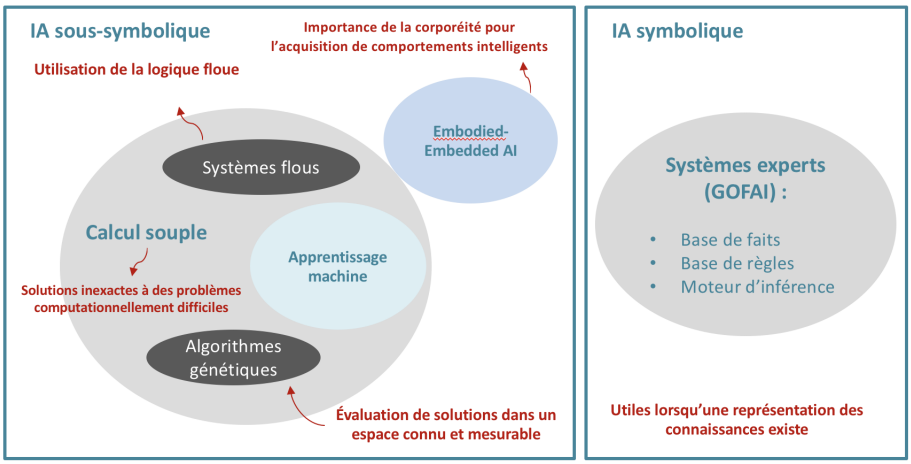

FIGURE 1 – Positionnement de l'AM parmi les techniques d'IA

experts ou les algorithmes génétiques, il nous semble pertinent, pour bien cibler les enjeux originaux posés par l'IA, de commencer par une étude centrée sur ces algorithmes d'AM. Ces derniers sont au cœur de nombre de développements scientifiques et industriels récents et ils constituent, du point de vue de l'ensemble de la communauté scientifique, la rupture épistémologique la plus nette par rapport aux procédures de décision classiques.

Les procédures d'AM posent un problème particulier pour l'un des grands enjeux du débat public sur les algorithmes, à savoir la « transparence des algorithmes » [2]. Cette notion complexe, dont la définition constitue précisément l'un des principaux enjeux de cet article, renvoie en première analyse à notre capacité à comprendre les décisions prises par un algorithme, à nous assurer qu'il fasse bien ce qu'il est censé faire, et à éviter tout effet inéquitable. Ces enjeux deviennent plus pressants pour l'industrie avec l'entrée en vigueur en mai 2018 du Règlement Général de Protection des Données (RGPD), où les questions liées à la transparence jouent un rôle crucial. Notre problème sera donc de comprendre quelles exigences de transparence peuvent être définies pour des procédures d'AM difficiles à comprendre pour leurs propres concepteurs, en donnant priorité aux enjeux épistémologiques.

Dans un premier temps, nous nous livrerons à une analyse conceptuelle générale de la transparence des algorithmes, pour en distinguer quatre sens fondamentaux (section 2). Ayant montré que l'intelligibilité des algorithmes est l'as-

2. Les récents développements en AM posent nombre de problèmes particuliers, qui ne se réduisent pas à l'application de l'exigence de transparence. Nous en évoquerons certains au cours de cet article, mais notre réflexion restera centrée sur la problématique de la transparence.

pect épistémologique fondamental de la transparence, nous explorerons certains des défis particuliers que cette notion rencontre dans l'usage public des algorithmes. Nous introduirons une distinction entre intelligibilité de la procédure et intelligibilité des sorties, et montrerons son importance cruciale pour l'explication des procédures au grand public (section 3). Enfin, nous appliquerons cette distinction au cas particulier de l'AM afin de mieux cerner les enjeux épistémologiques particuliers posés par son intelligibilité et son explicabilité (section 4).

2 La transparence des algorithmes

2.1 Transparence : exploration d'une polysémie

La transparence des algorithmes est tantôt présentée comme une vertu des algorithmes eux-mêmes, tantôt de leurs usages. Elle fait occurrence dans des contextes variés, et reçoit des définitions explicites diverses. Un effort de clarification conceptuel général semble nécessaire avant de préciser le rôle tout particulier que cette notion peut jouer pour l'AM.

Le terme de « transparence » est évidemment polysémique, et ce d'autant plus que la transparence a acquis ces dernières années le statut de vertu politique privilégiée. Sans même rentrer dans les emplois foisonnants de ce terme dans le discours public actuel, nous allons voir que même dans le contexte de l'informatique ce terme est employé de manière très diverse.

Avant de commencer notre travail de clarification, une précaution oratoire s'impose. Avec le développement de l'emploi des algorithmes dans tout un ensemble de contextes sensibles, le discours médiatique, mais aussi parfois le discours expert, utilise abondamment des catégories intentionnelles, politiques et morales pour décrire l'action des algorithmes : les algorithmes sont loyaux, équitables, ou biaisés, racistes, sexistes, homophobes, etc. Soyons clairs : l'IA forte constitue toujours un fantasme lointain, qui bien souvent pollue le débat plus qu'il n'y contribue. Un algorithme doit fondamentalement être conçu comme une machine, et il ne peut pas plus être raciste qu'un moteur d'avion. Mais l'usage des algorithmes peut indiscutablement produire des effets de cet ordre, et un algorithme peut être explicitement conçu pour avoir un tel effet. Si l'usage d'un algorithme déborde bien souvent les intentions de son concepteur, il est cependant déterminé par un ensemble de caractéristiques techniques intrinsèques de cet algorithme : impossible d'obtenir l'effet voulu avec un programme boggé ou trop complexe pour être exécuté sur l'architecture choisie, par exemple. Lorsque nous parlerons par la suite de « propriété intrinsèque » d'un algorithme, nous parlerons des propriétés qui peuvent être légitimement considérées comme des propriétés de l'algorithme lui-même sur la base de ses caractéristiques techniques, indépendamment de son contexte d'usage, et sans aucune attribution d'intentionalité à l'algorithme lui-même.

Il faut ensuite remarquer que le discours public a tendance à ne pas distinguer « algorithme »et « programme ». Les termes sont parfois interchangeables

même dans le discours savant, mais il existe une distinction intuitive entre ces deux termes. Un algorithme est une entité mathématique décrivant une procédure ; le programme est une entité technique, qui implémente un algorithme dans un langage de programmation donné. Il n'est pas à l'heure possible de donner une compréhension rigoureuse de cette distinction, puisque cela supposerait d'être capable d'identifier un algorithme à travers différents langages de programmation, et de répondre, au moins partiellement, à la question de la nature des algorithmes [2]. Mais cette distinction est cependant opérée à un niveau intuitif, et elle joue un rôle heuristique important dans la pratique. La distinction entre programme et algorithme a-t-elle un rôle à jouer dans les débats sur les effets politiques des algorithmes ? Ou est-elle sans pertinence ? Il s'agit là d'une question difficile, que nous allons ignorer dans le cadre de ce travail.

Pour mieux cadrer le débat, il est nécessaire de distinguer sous le terme de « transparence » quatre notions indépendantes.

2.2 La loyauté des algorithmes

L'algorithme est *loyal* si la fonctionnalité affichée par son fournisseur [3] auprès des utilisateurs correspond à la fonctionnalité connue du fournisseur. Si le fournisseur dissimule une fonctionnalité dont il a une claire conscience qu'elle est remplie par l'algorithme, alors l'algorithme sera déloyal. Nombre d'usages actuels des algorithmes font malheureusement preuve d'une telle déloyauté : tarification volatile, personnalisation de l'offre commerciale réalisée selon des critères différents des critères affichés, recommandation de trajets basés sur la présence de points d'intérêts commerciaux, réponse à des requêtes sur un moteur de recherche donnant la priorité à des liens sponsorisés plutôt qu'à la pertinence, etc. On peut également parler de « déloyauté » lorsque l'algorithme, même s'il remplit les fonctionnalités affichées, remplit en outre une autre fonctionnalité dont l'utilisateur n'a pas été clairement informé, comme par exemple la collecte de géolocalisation de l'utilisateur d'un smartphone à des fins commerciales [4]. Cette forme de déloyauté est aggravée si elle est implémentée par des applications qui n'ont aucunement besoin de collecter cette information pour fonctionner. On remarquera que dans le cas de la loyauté, l'attribution de cette propriété à

3. Nous emploierons les termes génériques de « fournisseur »et d'« utilisateur »pour désigner les parties prenantes dans une prestation algorithmique. L'utilisateur n'est pas forcément un client (notamment dans le cas d'un logiciel libre) et le fournisseur n'est pas nécessairement une entreprise, mais peut aussi être une association, un particulier, une fondation, une administration etc. Le fournisseur n'est pas non plus nécessairement le concepteur, mais l'on supposera dans la suite qu'il assumera les mêmes responsabilités que le concepteur auprès de l'utilisateur. Le rapport du CERNA ([3], 22-23) distingue à juste titre entre le concepteur, qui conçoit le modèle, et l'entraîneur d'un modèle d'IA, qui choisit la base de données, la structure et entraîne le modèle : il peut s'agir dans la pratique de personnes différentes. Mais nous considérons par défaut un modèle instancié, dont l'apprentissage est achevé : nous n'utiliserons donc pas cette distinction. Nous emploierons à nouveau le terme de "concepteur" lorsque nous traiterons uniquement du développement d'un programme et non de la relation de prestation (sections 3 et 4).

4. Pour une discussion plus approfondie des exemples d'algorithmes déloyaux, voir le site de TransAlgo [4].

l'algorithme est quelque peu abusive : l'algorithme n'est pas loyal ou déloyal, puisque cette notion suppose une intention de tromper qu'on ne peut guère attribuer à un algorithme. C'est la relation entre le fournisseur et l'utilisateur de l'algorithme, ou le fournisseur lui-même, qu'on peut en toute rigueur qualifier de loyal ou déloyal.

2.3 L'équité des algorithmes

Un algorithme est *équitable* si les résultats qu'il produit n'induisent pas un effet discriminant ou biais à l'égard d'une catégorie particulière de la population [5]. La discrimination peut être introduite, de manière consciente ou inconsciente, prévisible ou imprévisible, au niveau de la conception de l'algorithme. Elle peut aussi être introduite, dans le cas de l'AM, au niveau de l'apprentissage sur un jeu de données, et peut donc dépendre des discriminations inscrites dans les données elles-mêmes. Là aussi, les exemples abondent. Ainsi, les développeurs d'une application de la ville de Boston pour signaler les dégâts sur la voie routière ont dû corriger un biais favorisant les quartiers les plus fortunés : la collecte des données était faussée par le fait que la probabilité de posséder un smartphone et de télécharger l'application était bien plus grande dans les populations les plus aisées ([6], 51-52). Dans des cas bien plus graves, les suggestions de requête ou les résultats des moteurs de recherche peuvent présenter des biais sexistes, racistes, ou antisémites (voir par exemple [7]). La présence d'effets discriminatoires peut être devenir extrêmement difficile à prévoir, et il faut faire preuve de prudence avant d'attribuer des intentions malignes au fournisseur de l'algorithme.

Il est délicat de décider si l'équité est ou non une propriété intrinsèque de l'algorithme. D'une part, il n'est pas question d'attribuer à l'algorithme une intentionnalité discriminatoire. D'autre part, la propriété d'inéquité ne peut être systématiquement attribuée au fournisseur, ni à la relation unissant fournisseur et utilisateur, dans la mesure où l'inéquité peut être produite de manière invo-

5. Cette définition, tout comme les trois autres, n'a pas la prétention d'être une définition juridique. On remarquera qu'elle recouvre à la fois les notions du droit américain de *disparate treatment* (utilisation explicite ou indirecte d'une variable protégée dans la décision) et de *disparate impact* (effet discriminant des résultats de l'algorithme sans traitement d'une variable protégée). Notre définition ne prend pas en compte le caractère intentionnel de l'effet discriminant, ou la possibilité pour l'accusé de plaider la nécessité de la procédure pour ses activités. Pour une introduction à ces questions de droit américain en AM, voir [5]. Nous ne discuterons pas non plus les problèmes philosophiques redoutables posés par la notion d'équité, et ses liens avec la notion de discrimination. Pour une introduction à ces questions, voir Reuben Binns, "Fairness in Machine Learning : Lessons from Political Philosophy", Proceedings of Machine Learning Research 81 :1–11, 2018 Conference on Fairness, Accountability, and Transparency. Pour une discussion du cas majeur des outils prédictifs dans le système judiciaire, voir Ben Green, "'Fair' Risk Assessments : A Precarious Approach for Criminal Justice Reform", FATM/L 2018. Pour les difficultés liées aux diverses formalisations de la notion d'équité (*fairness*), leurs compatibilités et leurs relations aux notions intuitives, voir Pratik Gajane, Mykola Pechenizkiy, "On Formalizing Fairness in Prediction with Machine Learning", FATM/L 2018 et Christina Wadsworth, Francesca Vera, Chris Piech, "Achieving Fairness through Adversarial Learning : an Application to Recidivism Prediction", FATM/L 2018. et Roel Dobbe,Sarah Dean, Thomas Gilbert, Nitin Kohli "A Broader View on Bias in Automated Decision-Making : Reflecting on Epistemology and Dynamics", FATM/L2018

lontaire et même imprévisible. L'inéquité doit être attribuée aux effets sociaux de l'algorithme. Ces effets sociaux dépendent de caractéristiques techniques de l'algorithme, mais ne s'y réduisent pas *a priori* : ils dépendent de l'interaction entre l'algorithme et son contexte social d'usage. Il suffit d'imaginer un spambot sélectionnant des noms de personnes en fonction de leurs consonances pour leur envoyer un courriel. Selon qu'il s'agisse d'insultes racistes ou d'un extrait des *Illuminations*, on parlera ou non d'effets discriminatoires de l'algorithme. Pourtant, d'un point de vue technique, les deux algorithmes peuvent parfaitement être considérés identiques à substitution du corps de message près. Il est donc délicat de parler de « propriété intrinsèque de l'algorithme » pour l'équité des algorithmes. L'équité est une propriété de l'usage social de l'algorithme, qui ne se réduit pas à ses caractéristiques techniques. Dans l'impossibilité d'une attribution simple de la propriété d'équité, nous continuerons donc à parler d'équité des algorithmes *cum grano salis*.

2.4 L'explicabilité des algorithmes

Un algorithme est *explicable* s'il est possible de donner à l'ensemble des utilisateurs, quelque soit leur bagage éducatif, une vision claire des procédures employées et des fonctionnalités remplies par l'algorithme, afin de permettre un usage informé[6]. Cette notion pose des problèmes de pédagogie de systèmes technologiques complexes, et de la définition de la connaissance nécessaire à un usage informé. Il faut souligner dans ce cadre que les technologies d'information et de communication sont aussi utilisées par des enfants, ce que le RGPD prend explicitement en compte ([9], chapitre II, article 8). L'explicabilité est donc un enjeu non seulement pour s'adresser à un public sans aucun bagage technique en informatique, mais aussi à des personnes n'ayant pas les capacités psychologiques, les compétences cognitives ni même l'expérience de la vie d'un adulte. Pourtant, il sera également possible d'obtenir le consentement éclairé d'un enfant au titre du RGPD. Cette propriété est également cruciale pour les algorithmes participant à la prise de décision administrative, où l'opacité de fonctionnement induite par la complexité des procédures peut être légitimement attaquée par les citoyens. La loi pour une République Numérique crée ainsi l'obligation d'une mention explicite de l'existence d'un traitement algorithmique dans une prise de décision administrative individuelle, et la communication sur demande des règles de ces traitements et des caractéristiques principales de sa mise en œuvre ([10], article 4). De manière plus générale, dans tout secteur où les algorithmes jouent ou vont jouer un rôle important dans des prises de décision affectant de manière significative la vie de chaque citoyen (*credit score* aux USA, assurance, offre d'emploi, accès aux formations, etc.), l'explicabilité des algorithmes va devenir un enjeu démocratique majeur, notamment pour permettre la possibilité de recours ou de demande d'informations.

6. Le terme de transparence (*transparency*) est utilisé en ce sens par Abdohalli et al. ([8], 31) quand ils la définissent dans le cas particulier des recommandations automatiques comme "revealing the reasoning behind the system's recommandation".

L'explicabilité représente un enjeu tout particulier pour les modèles d'AM récents comme l'apprentissage profond. Pour ces modèles, non seulement le fournisseur lui-même peut être incapable d'expliquer dans le détail comment l'algorithme a effectué son apprentissage et comment il prend ses décisions, mais il n'existe à l'heure aucune technique systématique permettant de demander aux systèmes de fournir une explication claire pour une décision donnée[7]. Si ces algorithmes étaient utilisés dans des contextes où ils devraient prendre des décisions socialement importantes, comme conduire une voiture, accorder un prêt ou un entretien d'embauche, ou même refuser une mise en liberté, il serait parfois impossible de fournir la justification de la décision prise aux parties concernées. La prise de décision deviendrait une pure boîte noire impossible à contester en dehors du rejet global de l'emploi du modèle, piétinant le droit du citoyen à la demande d'information et à la formulation d'un recours. Si l'on reconnaît qu'une telle situation est juridiquement inacceptable, comme c'est clairement le cas dans le RGPD, l'explicabilité devient donc un enjeu majeur de l'industrialisation de ces algorithmes. Cet enjeu est d'autant plus important qu'on ne peut exclure la possibilité de limites fondamentales à la compréhension des décisions prises par ce type de modèles. Certains chercheurs vont jusqu'à envisager que, tout comme le comportement d'un individu humain ne peut parfois qu'être partiellement rationalisé tout en étant doué de sens, il est possible qu'une partie des décisions prises par l'IA ne puisse être capturée par une explication verbalisable, et soit seulement descriptible comme une nouvelle forme de décision instinctive [11]. Nous reviendrons sur ces enjeux dans les sections suivantes.

L'explicabilité est-elle une propriété intrinsèque de l'algorithme ? D'une part, l'explicabilité semble bien dépendre de propriétés techniques intrinsèques de l'algorithme, comme le montrent nos dernières remarques sur l'AM. D'autre part, il n'est pas évident qu'on puisse donner une seule explication d'un algorithme donné : la pédagogie pourrait avoir à s'adapter au public visé, produisant différentes explications pour différents groupes d'utilisateurs. Dans cette optique l'explicabilité dépend elle aussi du contexte social d'usage. Mais il faut garder à l'esprit que la prolifération de représentations diverses d'un même algorithme pourrait poser des problèmes graves de communication, voire de responsabilité légale. Seule une réflexion approfondie sur la pédagogie des algorithmes, qui va bien au-delà de la portée de cet article, pourra décider s'il est préférable de promouvoir une explication fixée pour un algorithme donné, ou si l'explicabilité doit être comprise comme une propriété de la relation entre les propriétés intrinsèques de l'algorithme et le bagage éducatif des utilisateurs. Nous nous contenterons ici de poser le problème.

2.5 L'intelligibilité de l'algorithme

Un algorithme est *intelligible* s'il est possible au concepteur de comprendre son fonctionnement et de vérifier s'il satisfait bien les propriétés désirées. L'intelligibilité est une forme d'explicabilité fondamentale, qui porte sur la capacité

7. Pour une bonne vulgarisation de ces questions, voir [11].

des concepteurs à s'expliquer l'algorithme qu'ils conçoivent. Ses limitations sont dues aux limites de l'état de l'art scientifique, voire à des limites scientifiques fondamentales.

L'explicabilité comme l'intelligibilité peuvent renvoyer à l'anglais *(human) interpretability*, concept qui a pris son essor dans la littérature des toutes dernières années et se voit maintenant muni d'une existence institutionnelle, avec un Workshop dédié de l'*International Conference on Machine Learning*. L'intelligibilité est parfois aussi nommée *intelligibility*. La DARPA a également lancé une initiative appellée *Explainable Artificial intelligence* (XAI), qui comprend entre autres un aspect *explainable artificial machine* : la notion d'explication employée recouvre à la fois l'intelligibilité et l'explicabilité [12].

L'intelligibilité constitue un enjeu majeur de la discipline informatique en général. Il n'est pas exagéré de dire que l'une des questions les plus fondamentales de l'informatique est « comment savoir si le programme fait bien ce qu'il est censé faire ? » Il se pose de manière particulièrement aïgue pour les IA, et en particulier les IA basées sur des mécanismes d'apprentissage profond dont le comportement précis échappe encore à la communauté scientifique. Mais il se pose aussi déjà pour les programmes les plus conventionnels. Nombre d'algorithmes commerciaux séquentiels et déterministes peuvent poser de graves problèmes de spécification et de vérification. On sait que la majorité des programmes n'est pas prouvée, et que la fiabilité des méthodes empiriques de certification de programmes est une question complexe.

Examinons à présent les relations entre nos quatre concepts. La distinction entre l'explicabilité et l'intelligibilité ne fait pas problème, puisqu'elle renvoie à des différences manifestes de connaissances scientifiques. La distinction entre l'équité et la loyauté peut être plus complexe à établir selon les cas, mais la distinction conceptuelle n'est pas problématique : un fournisseur peut parfaitement livrer un algorithme déloyal mais équitable, ou produire un algorithme inéquitable tout en étant parfaitement loyal sur sa spécification, les effets discriminants pouvant être extrêmement durs à anticiper.

Ces concepts doivent être regroupés en deux familles différentes. Les concepts d'explicabilité et d'intelligibilité renvoient à des propriétés épistémiques des algorithmes. Les concepts de loyauté et d'équité renvoient à des propriétés normatives ou prescriptives, de la relation entre fournisseur et utilisateur, et des effets sociaux des usages. Pour grossir le trait [8], on a donc affaire à deux familles de significations bien distinctes du concept de transparence : une famille descriptive renvoyant à des propriétés des algorithmes, une famille prescriptive renvoyant à des propriétés de leurs usages.

La publicité du code source n'est une condition ni nécessaire ni suffisante de la transparence des algorithmes. Elle n'est pas suffisante puisqu'elle ne satisfait pas immédiatement aux conditions d'intelligibilité et d'explicabilité. Elle n'est sans doute pas nécessaire parce que l'intelligibilité des fonctionnalités du code peut être masquée au profane, et parfois même à l'expert, par la multiplicité

8. Le trait est quelque peu grossi, dans la mesure où nous avons vu que l'explicabilité pourrait être dépendante du contexte social d'usage. Mais dans l'attente d'une étude approfondie de ce problème, la classification est éclairante en première approche.

des détails de l'implémentation. L'exigence de transparence des algorithmes ne remet pas en cause le logiciel propriétaire, et ne doit pas être confondue avec les revendications libristes.

Le glissement d'une famille de significations à une autre est rendu naturel par une dépendance conceptuelle essentielle. L'intelligibilité est évidemment la condition de possibilité de l'explicabilité : il faut comprendre quelque chose pour pouvoir l'expliquer. De manière plus générale, les propriétés épistémiques sont la condition de possibilité des propriétés normatives. Pour communiquer de manière loyale sur les fonctionnalités et pour garantir l'absence d'effets inéquitables, le fournisseur de l'algorithme doit avoir la capacité scientifique de comprendre son programme, d'anticiper ses effets, et de pouvoir l'expliquer à l'utilisateur. La notion d'intelligibilité se trouve donc au fondement de la notion de transparence.

Or cette capacité à comprendre de manière fondamentale les algorithmes est difficile à garantir en pratique pour les algorithmes conventionnels, et elle est à l'heure actuelle impossible à garantir pour nombre de procédures d'AM. Le fonctionnement interne complexe et en large partie mystérieux de ces dernières dans certains cas d'application entraîne une opacité essentielle au plus haut niveau de la connaissance scientifique, qui rend pour le moins problématique la vérification des trois autres propriétés. L'opacité de l'AM devient donc un problème fondamental pour la conception des normes régissant l'usage des algorithmes.

Il convient de noter que la prévisibilité du comportement de l'algorithme, qui est une composante de l'intelligibilité, est déjà un trait absent dans de nombreuses créations techniques, en particulier pour certains algorithmes classiques considérés comme bien compris, parce que bien spécifiés et certifiés. L'imprévisibilité peut être due par exemple à l'application de l'algorithme à un type de données dont la particularité n'a pas été bien comprise par le concepteur (gestion d'exceptions), ou simplement au grand nombre d'opérations nécessaires à l'obtention du résultat. Dans le cas de l'AM, une autre forme d'imprévisibilité est cependant introduite par le fait que le concepteur même de l'algorithme ne dispose pas d'une représentation claire du mécanisme de prise de décision utilisé par la machine. On créé ainsi des machines à la fois imprévisibles et insondables (*imprevisible and inscrutable*, [11]).

2.6 Positionnement par rapport à la littérature

Notre analyse de la notion de transparence algorithmique vise à prolonger et à raffiner les distinctions conceptuelles opérées par la littérature existante, notamment le rapport du CERNA [3]. Ce rapport définit ainsi la loyauté :

> La loyauté signifie que les systèmes se comportent comme leurs concepteurs le déclarent.

Cette définition, si elle est en soi parfaitement saine, ne permet pas immédiatement de distinguer les problématiques liées à la tromperie délibérée de celles liées à un manque de compréhension de son système par le fournisseur.

La responsabilité du fournisseur peut certes être engagée s'il fournit, sans aucune intention maligne, un système qui ne correspond pas à ses fonctionnalités affichées. Mais dans le contexte de discussion de l'AM, où le manque de compréhension du système par son développeur même est un enjeu majeur, il nous semble important de distinguer les enjeux d'offuscation des enjeux de maîtrise intellectuelle.

L'explication est définie en les termes suivants par le rapport du CERNA (*ibid*, 17) :

> Expliquer un algorithme est faire comprendre à ses utilisateurs ce qu'il fait, avec assez de détails et d'arguments pour emporter leur confiance. Cette tâche est difficile même dans le cas d'un algorithme dépourvu de capacité d'apprentissage, comme l'illustre le débat autour de l'algorithme d'admission post-bac APB. En outre il convient de distinguer preuve et explication : ainsi Gilles Dowek donne l'exemple simple de la multiplication de 12345679 par 36, dont le seul calcul du résultat (444444444) n'explique pas aux yeux d'un esprit mathématique pourquoi ce résultat ne comporte que des 4.

Cette définition, si elle n'est pas problématique en elle-même, ne distingue pas explicitement les enjeux d'explicabilité (explication à l'utilisateur) des enjeux d'intelligibilité (compréhension experte d'une preuve et d'un résultat mathématique) : une telle distinction est absolument nécessaire pour les problématiques de notre travail. En outre, elle ajoute une couche de complexité supplémentaire en introduisant la dimension de la confiance de l'utilisateur. Le rapport du CERNA a raison de souligner que le consentement n'est pas uniquement fondé sur la compréhension rationnelle (*ibid*, 22), mais aussi sur la confiance. Sans prétendre qu'il soit aisé de traçer une limite claire entre saine pédagogie et rhétorique persuasive, nous ignorerons dans ce travail la dimension irrationnelle de l'obtention du consentement, et considérerons l'explication comme une entreprise visant exclusivement l'obtention d'une compréhension rationnelle de la part de l'utilisateur. Il s'agit assurément d'une abstraction grossière face aux complexités de la pratique, qui n'est justifiée que par la volonté de restreindre et de clarifier la portée de notre travail.

3 Problèmes de l'intelligibilité

Nous allons à présent interrompre notre réflexion générale sur la transparence des algorithmes, pour nous concentrer sur le problème particulier de l'intelligibilité. Ce dernier problème peut devenir particulièrement épineux pour les fournisseurs de solutions et services basés sur de l'AM depuis l'entrée en vigueur du RGPD dans le droit européen en mai 2018. Ce texte impose entre autres un « droit à l'intervention humaine » via des instances de contrôle, et un « droit à l'explication » pour l'utilisateur impacté par un algorithme d'aide à la décision ([9], Considérants, 71). Ceci constitue une contrainte majeure pour les fournisseurs, et fait de l'explicabilité une propriété contraignant les réalisations

industrielles. Qui plus est, indépendamment de toute contrainte légale, il est bon de rappeler qu'un utilisateur est moins susceptible d'utiliser un dispositif qu'il ne comprend pas, et dont il ne peut prévoir le comportement.

L'exigence d'explicabilité dans notre terminologie ne doit pas être confondue avec la production d'explications plausibles mais fausses. Certains fournisseurs pourraient ainsi produire de telles explications afin d'augmenter l'acceptabilité de leurs produits, éventuellement en flattant les préjugés existants. Le RGPD tente de prévenir toute tentative d'enfumage en imposant que les explications soient données en des termes clairs, précis, utilisant une langue naturelle ([9], Considérants, 39). L'enjeu scientifique de l'intelligibilité et l'enjeu pédagogique de l'explicabilité deviennent ainsi des enjeux juridiques, économiques et politiques majeurs, non pas uniquement pour l'AM mais en particulier pour l'AM.

Il faut souligner la force du cadre juridique proposé par l'Union Européenne, qui interdit de fait nombre de pratiques de l'industrie, comme la monétisation des données sans consentement éclairé des utilisateurs actuellement très répandue parmi les géants américains. Un tel cadre légal n'aura de sens que si, d'une part, il n'impose pas de contraintes irréalistes aux fournisseurs, et que s'il est d'autre part suffisamment bien conçu pour permettre au législateur de peser sur les industriels et de poursuivre et de sanctionner les infractions. Contrairement à ce que l'on pourrait penser, les GAFAM sont en avance de phase de plusieurs années pour ce qui relève de l'accès aux données par l'utilisateur, la portabilité des données, ou encore le droit à l'oubli, en comparaison avec des acteurs classiques de l'économie européenne comme les banques ou les assurances. Mais ils se montrent également beaucoup plus intrusifs dans la vie privée des utilisateurs de leurs plateformes, et une bonne part des revenus de Google et Facebook sont tirés de la monétisation de ces données privées.

Le nouveau cadre juridique proposé par le RGPD est ainsi voué à redistribuer les cartes dans la chaîne de valeur de la donnée, et il faut souligner l'importance pour le milieu industriel de ne pas adopter une attitude purement négative face à l'apparition de ces nouvelles réglementations. Contrairement à une idéologie à la mode, les lois et règlements ne sont pas qu'un frein à l'innovation. Pour reprendre la formule de Bryce Goodman et al. [13], les problèmes posés par la régulation européenne "are good problems to have". Si la loi est bien conçue et n'impose pas de contraintes irréalistes ou trop précoces, elle peut être un stimulant pour des travaux de grande valeur, et source d'une réflexion profonde sur les propriétés des algorithmes. Du point de vue des applications, les défis scientifiques posés par le droit à l'explication peuvent mener à une interaction enrichie avec l'utilisateur et à une plus grande faculté de communication au sein des chaînes hiérarchiques [14].

Nous allons commencer par quelques remarques transverses sur l'intelligibilité en AM, avant de poser une distinction fondamentale : celle entre intelligibilité de la procédure et intelligibilité des sorties.

3.1 Aspects transverses de l'intelligibilité en AM

Nombre de discussions de l'intelligibilité de l'AM peuvent enfermer le lecteur dans une opposition quelque peu rigide entre l'AM, doté d'un grand pouvoir prédictif mais inintelligible, et les algorithmes conventionnels, limpides mais moins puissants. C'est cette opposition qu'il faut commencer par nuancer.

La tension entre l'exigence d'intelligibilité et l'ambition des algorithmes de dépasser les performances sensorielles et cognitives humaines sur certaines tâches est certes indiscutable. Dans Hara et al. [15], on voit que les *additive tree models* doivent clairement leur performance à la division de l'ensemble des entrées en de nombreuses sous-régions (> 1000), ce qui rend le résultat difficilement compréhensible. La substitution d'un arbre de décision rend le modèle plus simple, plus compréhensible mais aussi moins puissant au niveau prédictif [9]. L'exigence d'intelligibilité pourrait ainsi mener à renoncer au pouvoir prédictif qui fait toute la richesse des modèles d'AM.

Mais comme le remarque Lipton ([1], 5), un modèle d'AM réputé complexe comme un réseau de neurones n'est pas forcément moins intelligible qu'un modèle appartenant à d'autres classes, comme les modèles linéaires, les modèles basés sur des règles, ou les arbres de décision. Une régression avec des centaines de paramètres covariables pourra être très difficile à interpréter. La profondeur des arbres de décision, des règles lourdes et une haute dimensionnalité du modèle peuvent les rendre bien moins compréhensibles qu'un modèle de réseau de neurones compact, implémentant une porte logique par exemple. Un modèle d'AM n'est donc pas systématiquement moins intelligible qu'un modèle d'une autre classe : il faut encore préciser de quels modèles on parle, quelles données sont utilisées, et ce que l'on cherche à rendre intelligible.

En outre, Krause et al. [16] soulignent à juste titre que l'intelligibilité n'est pas forcément une qualité nécessaire à l'usage de l'AM. Dans certains cas d'usage on sera parfaitement heureux avec un modèle prédictif puissant, comme par exemple le jeu d'échecs. L'explicabilité demeurera cependant une règle juridique prudentielle, qui s'imposera par défaut à tout programme, mais force est de constater que certains contextes d'usage poseront bien moins d'enjeux d'intelligibilité et d'explicabilité que d'autres. Le rapport du CERNA sur l'éthique de l'AM [3] suggère à juste titre d'adopter une attitude de compromis entre intelligibilité et pouvoir prédictif, dont l'équilibre exact sera déterminé par le contexte d'usage.

De ce point de vue, le développement de l'AM pourrait mener à renoncer à certaines formes d'intelligibilité plutôt qu'à les imposer à tout prix, comme on a pu renoncer à trouver un modèle classique derrière les phénomènes quantiques. Si une telle attitude peut être admissible au niveau scientifique, reste à voir quelles exigences d'intelligibilité sont indispensables dans certains contextes d'interaction avec un utilisateur, et donc dans quels cas l'intelligibilité sera une qualité préférable au pouvoir prédictif.

Les modèles d'AM ne doivent cependant pas être réduits à leur puissance

9. On a dans cet exemple un facteur de taille clair dans l'intelligibilité, mais ce ne sera pas toujours le cas, et ce n'est pas propre à l'AM.

prédictive, et l'intelligibilité ne doit pas toujours être conçue comme une exigence antithétique au pouvoir prédictif qui serait le seul attrait de ces modèles. Au contraire, nombre de modèles d'AM sont censés nous aider à comprendre, en particulier à comprendre les grands ensembles de données [16]. Mais pour rendre les données intelligibles, ces modèles doivent eux-mêmes être intelligibles : l'intelligibilité est alors la propriété cible du développement de l'AM, et non une exigence surajoutée.

Enfin, l'AM ne mène pas toujours à des formes de raisonnement exotiques incompréhensibles au commun des mortels : il peut au contraire mener à la découverte d'un trait simple et intuitif des données à côté duquel nous étions passé. Ainsi le *malware detection neural network* (voir [17], cité dans [18]) a obtenu d'excellentes performances de détection de logiciels malveillants en utilisant la qualité grammaticale des commentaires de code comme critère. Rendre l'AM intelligible peut être ainsi un moyen de réaliser la force d'idées simples et intuitives.

3.2 Intelligibilité de la procédure et intelligibilité des sorties

Comme le souligne Lipton [1], la notion d'interprétabilité ou intelligibilité dans notre terminologie, mérite d'être élaborée. Mais face à un concept si complexe, et si riche de connotations, il existe un véritable risque de prolifération terminologique, qu'on distingue l'intelligibilité selon ses objets (intelligibilité globale du modèle, intelligibilité des étapes de calcul, des composantes du modèle), ses produits (production de certification, d'artefacts visuels ou textuels explicatifs, déboggage), ou ses modalités (intelligibilité comme compréhension qualitative du lien entrées-sorties, ou dans un formalisme rigoureux). Une telle prolifération terminologique, outre les graves problèmes de communication et de systématisation qu'elle pose, est aussi symptomatique de la difficulté à hiérarchiser les différents problèmes soulevés par la notion, et à en donner une compréhension ordonnée.

Sans prétendre épuiser cette question épineuse, il nous semble crucial de distinguer entre l'intelligibilité de la procédure et l'intelligibilité des sorties de cette procédure[10]. La première prétend comprendre la procédure dans son ensemble, la seconde porte sur une exécution donnée de cette procédure. Bien qu'elles ne soient pas logiquement indépendantes, les deux questions doivent cependant être distinguées, dans la mesure où l'on peut prétendre avoir une bonne compréhension théorique d'un modèle sans comprendre la sortie particulière d'une exécution, et ce aussi bien en IA que dans la programmation la plus conventionnelle. À ces deux formes d'intelligibilité correspondent deux formes d'explicabilité, répondant à deux questions différentes : est-il possible d'expliquer l'algorithme à un public profane, et est-il possible d'expliquer une sortie particulière au public profane ?

10. Cette distinction est déjà partiellement anticipée dans [19] et [20]. Nous adoptons le terme générique de « sortie » pour désigner aussi bien les décisions, prédictions, résultats numériques, graphiques et actions sur leur environnements produits par les algorithmes.

La distinction entre ces deux niveaux d'analyse doit être constamment maintenue à l'esprit lorsqu'on examine d'autres aspects de l'intelligibilité. Ainsi, lorsqu'on parle comme Lipton [1] de l'intelligibilité des étapes du calcul, parle-t-on des étapes de la procédure en général ou des étapes d'une exécution particulière ? Lorsqu'on parle des défis d'explicabilité au grand public d'un modèle comme l'apprentissage profond, parle-t-on de la difficulté de faire la vulgarisation des réseaux de neurones, ou parle-t-on de la difficulté à donner des raisons simples et claires pour une décision particulière ?

Cette relative autonomie des deux intelligibilités est essentielle à la fois au niveau théorique et au niveau pratique. Lorsqu'un consommateur ou un administré demande une explication sur une décision le concernant, sa demande est dans la grande majorité des cas une demande d'explication locale, arrimée à une intelligibilité de la sortie de la procédure de décision. Par exemple, si un administré demande pourquoi une aide lui a été refusée, la réponse attendue ne consistera pas en un cours de droit administratif sur les fondements juridiques de la sécurité sociale, ou d'une vue d'ensemble des procédures de décision de la Caisse nationale des Allocations Familiales. Elle prendra le plus souvent une forme du type : « cette aide est réservée aux couples avec plus de deux enfants ». Ce type d'explication a l'avantage d'être à la fois simple et de permettre la formulation d'un recours. S'il était au contraire nécessaire d'exposer l'intégralité du fonctionnement d'un processus bureaucratique pour fournir une explication, l'interaction avec les usagers et leur droit à l'explication seraient fortement compromis.

Ceci nous permet de voir une propriété essentielle des procédures bureaucratiques [11] ordinaires. Dans ces procédures, conçues grossièrement comme un enchaînement de « si...alors...sinon... », et donc comme un arbre de décision, la décision finale peut être conçue comme une combinaison de décisions élémentaires, simples et compréhensibles. Ces procédures jouissent donc de deux propriétés essentielles. La première est leur compositionalité : la décision peut être analysée en une composition de plusieurs sous-décisions. La seconde est leur élémentarité : l'analyse de la décision s'arrête sur des sous-décisions simples et compréhensibles par tous. Ce sont ces propriétés qui facilitent l'explicabilité d'une décision, malgré l'immense complexité des systèmes bureaucratiques modernes. Si la complexité de la procédure bureaucratique est fonction de sa taille, l'explication d'une sortie consiste le plus souvent à en extraire un court passage pertinent. Le caractère compositionnel des procédures de décisions bureaucratiques permet d'autonomiser l'exigence d'intelligibilité de la procédure, de plus en plus difficile quand la taille croît, de l'exigence d'intelligibilité d'une décision donnée, qui dépend d'un critère de décision contextuel, souvent trivial.

Par conséquent, même si l'explicabilité de la décision ne garantit pas l'explicabilité de la procédure, ce trait n'est pas forcément problématique dans l'usage parce que celle-ci n'est pas systématiquement nécessaire à celle-là. La compo-

11. Nous entendrons ici le qualificatif « bureaucratique » dans son sens sociologique le plus large, qui décrit toute activité fondée sur l'exécution de procédures encadrées par des lois et règlements écrits. Ce sens large concerne aussi bien des activités administratives que des décisions juridiques, comptables ou commerciales.

sitionalité et l'élémentarité permettent d'extraire une sous-partie pertinente de la procédure, extraction fondée à la fois sur l'exclusion des branches inactivées de cette procédure —inutile de rappeler les clauses applicables aux étrangers de l'UE à un citoyen français— mais aussi à l'exclusion d'éléments activés mais non-pertinents —inutile de rappeler sa nationalité à l'administré qui se fait refuser une aide parce qu'il est célibataire [12].

Toutes les procédures bureaucratiques ne sont pas modélisables comme des arbres de décision. Certaines procédures prennent la forme d'attribution de score à un utilisateur, la décision finale étant fondée sur l'atteinte d'un seuil par le dit score. Le *credit score* américain est sans doute l'exemple le plus connu, mais la Mairie de Paris emploie aussi un système de scorage pour l'attribution de logements sociaux. De telles procédures semblent imposer une plus grande dépendance conceptuelle de l'intelligibilité de la sortie à l'intelligibilité de la procédure, puisque le score final est la somme de différents sous-scores. Mais elles ont bien les propriétés de compositionnalité et d'élémentarité des décisions que nous avons trouvées dans les arbres de décision. Elles permettent donc elles aussi la formulation d'explications simples permettant d'agir sur leurs bases, comme « votre score deviendra plus élevé si vous recandidatez l'année prochaine » ou « vous pourriez atteindre le seuil si vous justifiez votre absence de ressources ».

Certains modèles statistiques employés par exemple pour l'octroi de crédit ou d'assurances peuvent poser des problèmes d'intelligibilité et d'explicabilité encore plus profonds, dans la mesure où leur modèle ne se réduit pas à une somme pondérée de diverses variables. Il est plus compliqué d'expliquer à un client pourquoi le modèle considère des variables liées, par exemple le produit de deux variables au lieu de leur somme. Les propriétés de compositionalité et d'élémentarité, même si l'on se restreint aux modèles utilisés avant l'AM contemporain, ne sont donc pas toujours évidentes à attribuer à une procédure donnée, et ce sujet mériterait une étude approfondie. Nous nous contenterons de souligner que les arbres de décision constituent à coup sûr une part majeure des procédures bureaucratiques, et que le problème d'intelligibilité et d'explicabilité des sorties que nous posons est donc un enjeu pratique de grande envergure. En outre, il n'est pas impossible d'étendre nos remarques à des procédures autres que bureaucratiques, comme nous le verrons dans les exemples étudiés dans la section 4.

Pour résumer, cette relative indépendance de l'intelligibilité de la procédure et de l'intelligibilité des sorties est fondée sur deux propriétés remarquables des procédures bureaucratiques ordinaires, à savoir leur compositionnalité et leur élémentarité, qui permettent d'extraire de la procédure une explication brève, pertinente et compréhensible d'une décision particulière. Ce sont ces deux propriétés qui permettent aux processus bureaucratiques de croître en taille sans compromettre totalement l'intelligibilité et l'explicabilité des décisions particu-

12. La sous-partie extraite de la procédure n'est pas toujours une branche de l'arbre de décision : les informations considérées comme pertinentes peuvent être extraites en divers noeuds de l'arbre, qui ne sont pas forcément voisins. Ainsi une explication pourrait être constituée par la racine de l'arbre (« êtes-vous citoyen de l'UE ») et le dernier noeud (« quel est votre revenu mensuel ? »), en ignorant tous les noeuds intermédiaires.

lières. Nous parlerons d'*explicabilité par extraits* pour désigner cette capacité à expliquer au profane une sortie particulière d'une procédure sans faire référence à l'intégralité de cette dernière, en sélectionnant un ensemble restreint d'éléments pertinents [13].

Ces dernières remarques permettent de souligner le caractère stratégique de l'intelligibilité des sorties pour l'industrialisation de l'AM. Il est certes difficile d'expliquer à un public profane [14] le fonctionnement de l'apprentissage profond, mais cette difficulté de l'intelligibilité de la procédure n'est pas un problème propre à l'AM : nombre d'algorithmes conventionnels sont d'une grande sophistication mathématique, et leur explication à un public profane poserait de graves défis. Mais c'est l'intelligibilité des sorties qui représente la majeure partie des enjeux d'explicabilité. L'intelligibilité des procédures d'AM sera d'abord un enjeu théorique avant d'être un enjeu pratique. D'un point de vue pratique, la question centrale est : puis-je justifier une décision particulière à l'aide de critères simples et intelligibles, sans faire référence à l'intégralité de la procédure [15] ?

C'est ici que l'AM peut poser des problèmes particuliers, en ce que les critères employés par la machine, notamment dans des tâches de classification, peuvent demeurer obscurs même pour l'expert (voir section 4). L'AM pose donc bien un problème spécifique d'intelligibilité des sorties, et donc d'explicabilité de ces mêmes sorties.

3.3 Droit et progrès scientifique : pour un comité dédié à l'intelligibilité et l'explicabilité de l'AM

Avant de passer à l'examen de ce problème spécifique, nous souhaiterions ouvrir une courte parenthèse sur ce que cet état de la science implique d'un point de vue institutionnel. Tout état de l'art en AM n'est qu'une photographie d'un champ en mouvement rapide, auquel manquent des résultats fondamentaux structurants. Est-il possible d'avoir des algorithmes conventionnels aussi performants que l'AM en reconnaissance d'images, de sons et en traitement de la langue naturelle ? Est-il possible de traduire un algorithme d'apprentissage profond en une procédure de décision classique, ou une telle représentation est-elle

13. Il nous faut bien souligner que nous parlons ici d'explicabilité et non d'intelligibilité. Du point de vue scientifique, qui est celui de l'intelligibilité, il n'y a pas d'autonomie complète de l'intelligibilité de la procédure et de l'intelligibilité des sorties : difficile de comprendre une sortie particulière sans avoir une idée de la procédure d'ensemble. Mais d'un point de vue pédagogique, et éventuellement juridique, il est possible d'expliquer une sortie particulière en extrayant quelques éléments de décision pertinents.

14. La notion de « profane » désigne ici simplement un utilisateur qui n'a pas le bagage éducatif pour comprendre l'intégralité de la procédure avec laquelle il interagit. La notion est donc relative au type de procédures : on peut être profane sur un type de procédures et expert sur un autre. Elle ne désigne aucun groupe sociologique particulier : le profane peut tout aussi bien être le plus humble des administrés qu'un dirigeant d'entreprise ou un décideur politique.

15. Il ne s'agit bien sûr pas de nier que certaines demandes d'explication exigent et exigeront une intelligibilité de la procédure. Mais il s'agit de souligner que le fonctionnement de nos systèmes bureaucratiques est fondé sur la possibilité de fournir, dans la majorité des cas, une explication des sorties sans fournir une explication des procédures, et que cette propriété devrait être conservée pour permettre un usage massif de l'AM dans la prise de décision.

fondamentalement impossible ? Peut-on obtenir des modèles d'AM qu'ils produisent des logs compréhensibles et des messages expliquant leurs décisions ? Toutes ces questions étant ouvertes, on ne sait pas à quel point l'irréductibilité de l'AM à des procédures plus classiques est fondamentale et définitive.

Un problème scientifique d'une telle profondeur est susceptible de demeurer ouvert pendant une longue période. Dans l'état actuel de nos connaissances scientifiques, avoir à l'égard des modèles d'AM des attentes d'explication strictement équivalentes à celles formulées pour des algorithmes conventionnels n'est pas réaliste. Une interprétation stricte du droit à l'explication reviendrait à interdire l'emploi d'une bonne partie de ces modèles. Si une telle interdiction peut bien sûr être nécessaire dans certains contextes, une interprétation plus nuancée est désirable, non seulement pour le bien de l'activité économique mais aussi pour les autres bénéfices que certains modèles peuvent apporter à leurs utilisateurs.

D'un point de vue réglementaire, en particulier pour la mise au point des standards en termes d'explicabilité des programmes, il n'est donc pas possible de mettre en place un cadre juridique à la fois définitif et fort. Il faut alors tâcher d'avoir à la fois des recommandations opérationnelles dès à présent, et de maintenir une perspective prospective. Le RGPD encourage la mise en place, par les « associations et autres organismes représentant des catégories de responsables du traitement ou de sous-traitants », de codes de conduite visant à préciser les modalités d'application du règlement, notamment les « informations communiquées au public et aux personnes concernées [16] » .

Nous appelons à la mise en place d'un groupe de travail dédié à l'intelligibilité et l'explicabilité de l'AM, servant à la fois à la mise au point de standards d'explicabilité, à la veille scientifique sur le sujet, et à la mise à jour progressive des standards au fur et à mesure des avancées scientifiques. Ce comité devra bien sûr veiller à développer la plus forte coordination possible avec les autres initiatives pertinentes, comme le projet TransAlgo en France ou le groupe de travail de la Mairie de New York sur l'emploi des algorithmes par la puissance publique.

Nous venons de montrer que, d'un point de vue pratique, ce groupe de travail devra concentrer ses efforts sur l'intelligibilité des sorties et l'explicabilité par extraits, comme premier enjeu stratégique de l'usage public de l'IA en général, et de l'AM en particulier. Notre question suivante sera donc : dans l'état de l'art actuel, qu'est-ce qui rend problématique l'explicabilité par extraits des procédures d'AM ?

4 L'intelligibilité de l'AM

En guise de piste de travail pour ce groupe que nous appelons de nos vœux, nous proposons dans le reste de cet article une approche originale des enjeux d'intelligibilité et d'explicabilité en AM. Nous limiterons tout d'abord la complexité du problème en nous restreignant au cas où les données de la procédure

16. [9], chapitre IV, section 5, article 40.

d'AM sont simples et intelligibles. Il s'agit bien entendu d'une idéalisation forte, puisque dans bien des cas l'AM est appliqué à des données complexes dont l'intelligibilité peut être délicate : une étude complète de l'intelligibilité de l'AM devrait donc embrasser le couple données-algorithmes. Mais nous nous intéressons ici à l'intelligibilité des procédures, et un jeu de données complexe peut poser des problèmes d'intelligibilité même pour des algorithmes simples, et ne relevant pas de l'AM : nous nous permettrons donc cette idéalisation. Nous nous restreindrons aussi au cas des algorithmes d'apprentissage supervisé. Nombre des conclusions que nous allons formuler pourraient aussi s'appliquer aux autres types d'apprentissage, mais leur discussion détaillée allongerait et obscurcirait trop la présentation.

Nous allons explorer une analogie entre les propriétés de compositionalité et d'élémentarité de la procédure, et la notion technique de segmentation de l'espace des données. Cette analogie nous donnera une formulation intuitive de certaines des raisons pour lesquelles certains algorithmes d'AM sont difficiles à comprendre et à expliquer. Dans l'esprit des remarques présentées dans la section précédente, nous formulons cette approche en termes de distinction entre intelligibilité de la procédure d'instanciation d'un modèle d'AM et intelligibilité de ses sorties, ainsi qu'en termes d'intelligibilité des critères caractérisant le segment d'appartenance d'un exemple dans l'espace des données, depuis sa représentation initiale jusqu'à sa représentation finale en tant que sortie.

4.1 Instanciation d'un modèle d'AM : quelques rappels

Commençons par quelques rappels concernant la procédure d'instanciation d'un modèle d'AM [17]. Le travail du concepteur d'un modèle d'AM se décompose en général en six étapes :

1. Formalisation mathématique du problème
2. Sélection d'un jeu de données adapté et définition d'un objectif final
3. Sélection d'une classe de modèles (régressions, arbres, réseaux de neurones, etc.) et des algorithmes d'optimisation de ce que l'on nomme paramètres et hyperparamètres du modèle
4. Instanciation d'un modèle par apprentissage sur les données
5. Production de sorties
6. Tests de performance le cas échéant, par exemple dans le cas des problèmes dits supervisés où l'on dispose d'une variable à prédire, dont l'historique des valeurs va servir, par comparaison avec les prédictions, de *ground truth*.

Détaillons les phases d'instanciation, de production de sorties et de tests de performance. Le jeu de données servant à la construction du modèle est généralement décomposé en trois parties :

17. Le lecteur familier de l'AM peut sauter sans dommages cette section, et se contenter de lire le paragraphe de conclusion.

— Le jeu d'entraînement : celui-ci sert à l'optimisation de ce que l'on nomme les paramètres du modèle, qui correspondent par exemple dans le cas d'un réseau de neurones aux poids des connexions une fois que le nombre de couches et de neurones par couche ont été fixés. Le concepteur peut choisir différents algorithmes d'optimisation des paramètres selon la classe de modèles choisie : descente de gradient (stochastique ou non) pour les régressions ou réseaux de neurones, algorithmes gloutons comme le scindage binaire pour les arbres de régression et de classification, ou encore optimisation séquentielle pour les *Support Vector Machines*. La précision atteignable dans la détermination d'une solution est fonction de l'algorithme d'optimisation choisi, car certains algorithmes comme la descente de gradient stochastique "s'agitent" autour du minimum local et ne retiennent donc qu'une solution sous-optimale.
— Le jeu de validation croisée : celui-ci sert à la métaoptimisation des variables considérées comme fixes lors de la phase d'entraînement, que l'on nomme hyperparamètres du modèle, et qui dans le cas d'un réseau de neurones correspondent au nombre de couches et de neurones par couche ou encore à des variables techniques telles que le « dropout ». Les hyperparamètres sont parfois fixés à l'issue d'une optimisation par exploration brute de toutes les possibilités, mais on recourt le plus souvent à des heuristiques à la justification plus ou moins bien fondée. Bien souvent, c'est cette dernière étape qui constitue tout le savoir-faire industriel d'une entreprise utilisant de l'AM, comme lorsqu'il s'agit d'optimiser un réseau profond pour l'analyse d'images.
— Le jeu de test : ce dernier ensemble, isolé dès le début de la procédure de construction d'un modèle, sert à en tester la performance sur des exemples qu'il n'a jamais vus, mesurant ainsi une capacité dite de généralisation. Dans le cas d'un problème de classification, quatre métriques de base de performance (vrais positifs, vrais négatifs, faux positifs, faux négatifs) sont regroupées dans ce que l'on nomme la matrice de confusion.

Dans le cas de l'apprentissage supervisé, l'instanciation d'un modèle est le processus d'apprentissage des corrélations entre variables explicatives et variable cible. Le modèle instancié est le modèle capable de produire des prédictions sur la base de l'apprentissage effectué.

On peut ainsi constater que la procédure d'instanciation d'un modèle d'AM est bien définie, et est parfaitement traçable via le code du programme et les divers mécanismes de suivi d'exécution de ce dernier. Cette compréhension de la procédure est fondée sur la bonne connaissance par les concepteurs des modèles d'AM de la nature et de la structure technique des données qu'ils manipulent, ainsi que du fonctionnement *in abstracto* des algorithmes employés. Par exemple, imaginons que la frontière séparant deux classes d'objets caractérisés par deux variables (x_1, x_2) est circulaire. Il est évident qu'une régression logistique de degré 1 n'est pas pertinente, car elle ne peut générer que des frontières linéaires. De même, un réseau de neurones peut approximer n'importe quelle frontière entre exemples de labels différents grâce au théorème d'approximation universelle [21], mais ce résultat ne dit rien sur l'efficacité algorithmique de l'apprentissage des

paramètres associés. Ainsi, le concepteur possède une intelligibilité claire des possibilités des algorithmes employés *en tant qu'algorithmes*, et des étapes de la procédure lui permettant d'instancier un modèle d'AM. Les heuristiques employées (en particulier lors du réglage des hyperparamètres), si elles peuvent générer un certain flou scientifique, ne constituent un problème qu'en vue de la production de sorties qui soient elles-mêmes intelligibles. D'un point de vue pratique, il est donc important de comprendre les obstacles sur lesquels peut buter le concepteur suite à une demande d'explication d'une sortie d'un modèle d'AM.

4.2 Quels obstacles à l'intelligibilité des sorties d'un modèle d'AM ? Le problème de la segmentation

Pour traiter ce problème, il nous faut introduire la notion de segmentation de l'espace des données. Raisonnons sur une sortie particulière d'un modèle d'AM supervisé : celle-ci correspond à un point dans l'espace de données final, lui-même fruit d'une possible transformation de l'espace de données initial par les étapes intermédiaires du calcul du modèle instancié. Le point possède un voisinage au sein duquel la sortie calculée prend une valeur unique [18] et dont les frontières sont déterminées par l'algorithme d'AM à partir des données d'entraînement. Ainsi, dans l'exemple de l'arbre de décision de la figure 4.2, trois segments sont établis dans l'espace des données, caractérisés par des intervalles de valeur de vitesse du vent v et d'humidité de l'air h. Les frontières de ces segments correspondent aux règles de transition entre les noeuds de l'arbre. À chaque segment est associée une température moyenne unique calculée à partir des exemples d'entraînement présents dans le segment. Un nouvel exemple pour lequel le modèle instancié doit fournir une sortie correspond à un point de l'espace (v, h), et est situé dans l'un des segments découpés par l'arbre. La valeur de la sortie calculée par l'arbre pour ce nouvel exemple correspondra à la température moyenne de son segment d'appartenance.

Deux cas de figures peuvent alors se présenter suite aux segmentations établies par un modèle dans l'espace des données : soit les frontières établies par l'algorithme sont explicites et formulées de manière intelligible, soit elles ne le sont pas. Ainsi, les arbres de décision sont considérés comme intelligibles justement à cause du caractère explicite des segmentations qu'ils établissent dans un espace de données initial non transformé. Dans le cas des réseaux de neurones ou des méthodes à noyaux, il est souvent difficile d'avoir une définition explicite des frontières définissant le voisinage d'un point, voisinage au sein duquel une sortie homogène est calculée.

C'est ici qu'on peut percevoir une analogie entre les propriétés de la segmentation des données d'une part, et les propriétés de compositionalité et d'élémentarité d'autre part (voir Fig. 4.2). Dans le cas de l'arbre de décision que nous

18. Par moyenne des valeurs des exemples d'entraînement dans le voisinage dans le cas d'une régression par exemple, ou par vote à la majorité dans le cas d'une classification.

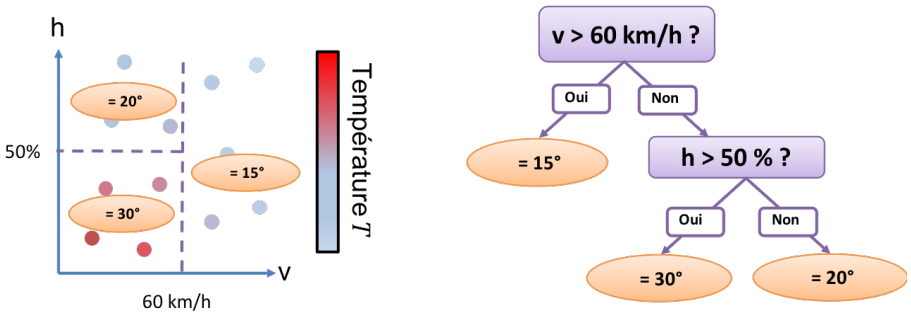

FIGURE 2 – Exemple de segmentation dans l'espace des données établie par un arbre de régression. v et h correspondent respectivement à la vitesse du vent et à l'humidité de l'air, tandis que la variable à prédire est la température T. Les frontières des segments (à gauche) sont en correspondance avec les règles d'évolution dans l'arbre (à droite), et sont optimisées à partir des données servant à l'instanciation du modèle.

venons de présenter, on dispose de variables explicites et simples, et de frontières bien définies : la procédure jouit des propriétés de compositionalité et d'élémentarité. D'autres modèles peuvent fournir une segmentation formalisable, mais dont le sens demeure opaque. C'est le cas des modèles dits *paramétriques*, où on dispose de relations mathématiquement bien définies entre variables, qui rendent le modèle intelligible pour l'expert. Mais leur sens peut être extrêmement difficile à expliquer en termes simples. Pour imaginer un exemple pédagogique, comment expliquer à un client que la banque lui refuse un crédit parce que le carré de son âge multiplié par son poids a dépassé un certain seuil ? Ce type de procédures conserve la compositionalité, mais il perd l'élémentarité, parce que les variables sur lesquelles sont basées ses décisions ne sont pas simples à expliquer. Enfin, il existe aussi des modèles pour lesquels la délimitation des frontières est difficile à définir, et leur sens demeure opaque même à l'expert. C'est le cas de certains calculs intermédiaires de réseaux de neurones profonds. La procédure perd alors à la fois la compositionalité et l'élémentarité : la décision finale de classification ne peut être divisée en sous-décisions, et les variables prises en compte par l'algorithme sont inintelligibles, voire impossibles à décrire en une formule mathématique. De là naît le sentiment d'opacité entourant certaines applications de l'AM, en particulier l'analyse d'images par des réseaux de neurones profonds, utilisée en médecine ou encore en conduite autonome. Cette opacité rend en pratique difficile toute correction de comportement « locale », c'est-à-dire qui n'ait pas recours à une modification de la procédure dans sa globalité, et notamment à une nouvelle instanciation du modèle modifiant par tâtonnement les heuristiques employées dans l'optimisation des hyperparamètres.

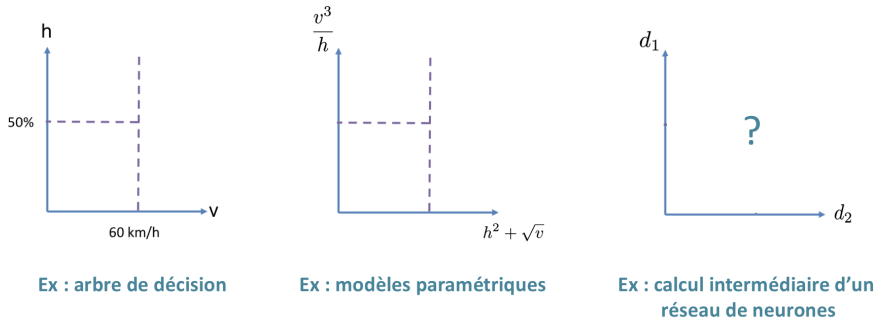

FIGURE 3 – Exemples de différences entre algorithmes d'AM en termes d'intelligibilité des segmentations établies dans l'espace des données : à gauche, une segmentation claire est établie par l'arbre de décision en fonction des variables d'entrée ; au milieu, un modèle paramétrique transforme de manière intelligible les variables d'entrée, la segmentation est alors formalisable via des formules mathématiques mais son sens peut être opaque pour une explication à un usager ; à droite, les transformations établies par un réseau de neurones dans l'espace des données n'aboutissent pas à une segmentation formalisable mathématiquement ou même tout simplement intelligible.

Il existe donc une analogie entre existence de segmentations à frontières explicites et compositionnalité d'une part, et intelligibilité des dimensions de l'espace de données de sortie et élémentarité d'autre part. Nous ne prétendons pas que les notions techniques épuisent les notions intuitives. Il est par exemple évident que la notion d'élémentarité ne se réduit pas l'intelligibilité des dimensions de l'espace des données de sortie. Cette analogie, si elle doit à coup sûr être raffinée, nous permet néanmoins de diagnostiquer de manière plus précise les difficultés à comprendre et à expliquer certains modèles d'AM dont les transformations intermédiaires de l'espace de données sont opaques.

Plusieurs techniques ont donc été développées afin de permettre à l'expert de vérifier les facteurs justifiant la sortie d'un modèle d'AM instancié. Par exemple, l'analyse classique de l'importance des variables [22], ou encore la méthode *Leave-One-Out Covariance* (LOOC) [23] qui consiste à relancer l'entraînement d'un modèle en remplaçant les valeurs d'une colonne par une valeur non signifiante, mesurent l'impact relatif de chaque variable sur les prédictions du modèle. De même, les algorithmes de réduction de dimension, comme l'analyse en composantes principales ou encore l'algorithme *t-distributed Stochastic Neighbour Embedding* (t-SNE) [24], permettent d'étudier les segmentations établies par un modèle sur des espaces de grande dimension en les projetant sur des espaces de dimension plus petite. Enfin, des techniques plus évoluées comme les *Local Interpretable Model-Agnostic Explanations* (LIME) [25] construisent un modèle intelligible localement proche d'un modèle difficile d'interprétation, en

se basant sur le prélèvement d'un échantillon d'observations autour d'un point quelconque, qui sert alors d'échantillon labellisé pour l'entraînement du modèle de substitution plus intelligible, comme un arbre de décision. Il faut noter que dans de tels cas, on établit localement une segmentation dont les frontières sont explicitables, et qui approxime la segmentation établie par le modèle original [26]. On peut d'ailleurs expliciter l'ensemble des frontières ou une sous-partie uniquement, selon le niveau de complétude pertinent et souhaité pour la description du segment.

Ces techniques sont très utilisées quelle que soit la classe de modèles, la difficulté de compréhension des sorties n'étant pas liée uniquement à l'algorithme, mais au couple données-algorithme, et en particulier à la transformation des données entre l'entrée et la sortie. Ainsi, ces techniques peuvent aider à comprendre les décisions prises par une forêt aléatoire qui utiliserait des variables socio-démographiques et personnelles pour prendre des décisions quant à l'octroi d'un crédit. Mais dans certains cas d'application comme l'analyse d'image, la transformation de l'espace de données entre l'exemple (à la représentation intelligible) fourni en entrée et la sortie calculée est opaque, et l'explicitation des caractéristiques (exactes ou approximatives) du segment où se situe la sortie ou un point intermédiaire du calcul pourra ne posséder aucun sens du point de vue de l'utilisateur humain (voir Fig. 4.2). Par exemple, dans le cas de la reconnaissance de chiffres sur une image de 28 pixels par 28 pixels en noir et blanc, un point de l'espace de données de départ va correspondre à une image parmi les 2^{784} images possibles, et un point de l'espace de sortie à un vecteur de longueur 10 dont les valeurs sont comprises entre 0 et 1. Chacun de ces deux points est donc parfaitement intelligible pour un humain suffisamment expert, mais les transformations intermédiaires de l'espace de données initial aboutissent à des représentations du point d'entrée qui sont peu compréhensibles, et ce même si les segments les contenant sont caractérisés par des frontières explicites. En effet, ces frontières seront elles-mêmes définies en référence à des dimensions difficilement intelligibles. Quant aux tentatives d'interprétation des résultats des couches intermédiaires d'un réseau de neurones comme des montées en abstraction dans la représentation des données initiales, elles ne sont basées que sur de vagues analogies avec les propriétés des systèmes biologiques, et aucune approche formelle permettant d'interpréter sans ambiguïté la signification des couches intermédiaires d'un réseau profond n'est établie à ce jour.

L'application de notre analogie aux polémiques de ces dernières années autour du fonctionnement des réseaux de neurones profond permet à la fois de l'illustrer et d'en raffiner l'analyse. La première concerne le système de classification d'images de Google qui a confondu en juillet 2015 deux Afro-américains avec des gorilles, et que Google n'a pu corriger qu'en retirant du jeu d'entraînement les images associées à des gorilles et autres primates [27]. La seconde concerne le premier accident fatal d'une voiture autonome, en mai 2016, suite auquel un audit a permis de démontrer que l'autopilote ne savait pas distinguer la partie blanche d'un semi-remorque du fond de ciel brillant [28]. L'analyse

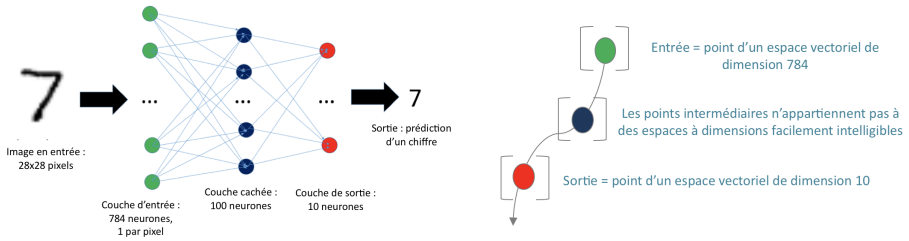

FIGURE 4 – Pour les réseaux de neurones, en particulier en apprentissage profond, seules les couches d'entrée et de sortie possèdent une interprétation claire

de ces accidents montre qu'il peut être possible de produire des explications simples et compréhensibles des sorties d'un réseau de neurones profond, et il est probable que dans de tels cas on n'ait pas eu besoin de comprendre les segmentations établies par le réseau, mais seulement d'analyser les ressemblances et dissemblances entre points voisins de l'espace des données. En quoi le défaut d'explicabilité par extrait est-il alors problématique ?

Dans le premier cas, l'audit a permis de comprendre les raisons des sorties, mais l'absence d'une interprétation claire des transformations intermédiaires des données a abouti un couplage fort entre calcul de la sortie et procédure d'instanciation du réseau, ce qui a empêché toute solution « locale » au problème, c'est-à-dire une solution ne passant pas par le réentraînement intégral du modèle. Dans le deuxième cas, un audit a également permis de comprendre l'origine du problème et d'exempter le constructeur de toute faute en considérant que le système a fait ce qui était prévu qu'il fasse dans une telle situation, contrairement au conducteur qui disposait de suffisamment de temps pour réagir et qui ne l'a pas fait, par manque de concentration sur la route comme exigé par le constructeur [19]. Mais de même que dans le premier cas, la résolution du problème par Tesla passera probablement par une adaptation du jeu de données servant à l'entraînement des modèles utilisés.

Le défaut d'explicabilité par extrait impose donc de considérer la procédure dans son ensemble, et le problème qui se pose avec certaines classes d'algorithmes d'AM n'est donc pas tant dans l'existence ou l'inexistence d'explications, que dans leur source. Pour comprendre les décisions erronées prises par les systèmes, il faut remonter jusqu'au processus d'entraînement sur les données, et pour résoudre les problèmes, il faut reprendre ce processus d'apprentissage. L'explication fournie n'est donc ni une explication des sorties par extraits, ni même une explication de l'intégralité de la procédure : elle consiste en une explication du processus d'engendrement de la procédure elle-même qu'est un modèle

19. Il s'agit là d'un cadeau juridique empoisonné pour les systèmes automatiques, car leur objectif ultime est bien que le conducteur puisse faire autre chose pendant que ces systèmes conduisent.

instancié. Pour reprendre notre analogie avec les prises de décision bureaucratiques, on se trouve dans une situation analogue à celle où l'on devrait expliquer une décision aberrante à un administré non par une explication par extraits ou par une explication de toute la procédure, qui sont ici inaccessibles, mais par une explication du processus politique qui a mené à la conception de la procédure elle-même. Si le processus d'instanciation d'un modèle est peut-être mieux compris et moins soumis au secret que les processus politiques, un tel état de faits est cependant problématique. De telles explications ne sont en effet nullement des justifications de la décision, qui permettent une contestation de leur légitimité ou une réaction adaptative de l'utilisateur : elles ne font qu'indiquer l'origine d'une aberration. On voit donc bien comment la perte de l'explicabilité par extraits entraîne une interaction totalement différente entre utilisateur et fournisseur. Cette interaction peut être privée des justifications que fournit l'explication par extraits, tout en comprenant des considérations sur l'origine des procédures dont l'explicabilité par extraits avait justement l'avantage de nous dispenser.

Cette dépendance de l'explicabilité des sorties au processus d'apprentissage est problématique à encore un autre titre. S'il est bien décrit d'un point de vue purement procédural, le processus d'apprentissage comprend une forte dimension heuristique, qui est une cause majeure de l'opacité de l'AM. En particulier, la représentativité de l'ensemble des données par rapport au problème considéré ne peut jamais être garantie, ce qui ajoute à la nature statistique des résultats de l'apprentissage une incertitude fondamentale souvent difficile à mesurer. Il s'agit là d'un problème complexe que nous n'aborderons pas dans cet article, et nous nous contenterons d'indiquer que selon qu'on ait affaire à un arbre de décision [29], à un processus gaussien [30], ou à un réseau profond [31] [32][33], les problèmes posés par la gestion de l'incertitude des sorties ne sont pas de même nature.

Ces considérations sur le manque d'intelligiblité et d'explicabilité de certaines méthodes d'AM nous aident à mieux en comprendre la source, mais le niveau le plus fondamental à analyser est celui du pourquoi du recours à ces méthodes : elles représentent tout simplement la meilleure solution possible pour certaines tâches. Pour conclure notre réflexion, nous devons donc nous pencher sur cette ultime question : qu'est-ce qui caractérise les problèmes pratiques pour lesquels les méthodes d'AM représentent la meilleure solution possible ?

4.3 Les sources structurelles de l'opacité : limites de la modélisation et heuristique de l'apprentissage

Il existe plusieurs pistes de réflexion sur la question du contexte d'utilisation des méthodes d'AM, qui ne sont pas nécessairement exclusives les unes des autres. Doshi-Velez et Kim [20] affirment que c'est lorsqu'il existe une incomplétude fondamentale dans la spécification du problème que l'AM est le plus utile, mais que c'est également dans ces situations là qu'il pose le plus de problèmes d'intelligibilité. Cette incomplétude est elle-même ancrée dans l'incapacité à fournir une paramétrisation complète du problème, d'où la nécessité d'une in-

teraction machine-données-concepteur humain pour modéliser le problème et interpréter les résultats en sortie, par nature incertains. La plupart des modèles d'AM visent alors à optimiser de manière heuristique une performance mesurée par un ensemble de métriques, comme les quatre métriques qui constituent la matrice de confusion évoquées précédemment pour le cas d'un classifieur.

Selon Lipton [1], les spécificités de l'usage de l'AM surgissent à une étape ultérieure de la résolution du problème, à savoir lorsque les métriques usuelles de performances de l'algorithme (prédiction et *ground truth*) ne sont pas suffisantes pour évaluer le travail de cet algorithme dans des circonstances d'usage réalistes. Cette caractérisation n'est pas incompatible avec celle de Doshi-Velez et Kim : l'un des problèmes cardinaux de l'AM est de comprendre comment l'optimisation d'une métrique permet de réaliser une tâche à la paramétrisation insaisissable.

Un exemple permet d'éclairer le lien entre paramétrisation incomplète du problème, insuffisance des métriques statistiques de performances, et exigence d'intelligibilité du modèle d'AM. Considérons un classifieur qui réussit à classer les chiens et les loups de manière très performante en utilisant le décor environnant : dans l'ensemble de données considéré, les loups sont plus souvent photographiés dans un décor comprenant de la neige en arrière-plan, ce qui n'est pas le cas pour les chiens (un exemple similaire peut être trouvé dans [25]). On considérera alors que ce modèle ne résout pas fondamentalement la tâche proposée, et la performance selon la métrique standard n'est pas suffisante pour détecter ce problème si les images utilisées pour tester la performance du modèle présentent le même biais. L'IA suroptimisée fonctionne comme un élève qui devine les réponses d'une question à choix multiples en analysant le ton de son professeur : la bonne réponse a été atteinte pour des raisons circonstancielles liées à un environnement artificiel d'exécution de la tâche, qui ne peuvent être généralisées. Le succès prédictif devrait au contraire être révélateur d'une véritable compréhension du problème. Pour obtenir une IA instanciée qui réalise vraiment la tâche voulue, et puisse voir son emploi généralisé, il ne suffit pas d'avoir juste : il faut avoir juste pour les bonnes raisons. L'intelligibilité des sorties —telle image est associée à un loup car le modèle repère un manteau de neige en arrière-plan— est donc un moyen de perfectionner la spécification du problème, et ce faisant d'adapter les métriques de performance ou le jeu de données utilisé.

Mais le problème du programmeur est précisément qu'il ne peut spécifier à l'avance tous les paramètres pertinents pour la tâche examinée. Il est impossible d'appliquer la méthodologie standard de la programmation "en V", et d'énumérer tous les problèmes potentiels et d'appliquer une batterie de tests unitaires, ou encore d'être certain que le jeu de données utilisé est représentatif de toutes les situations d'intérêt possibles. En l'absence d'une telle paramétrisation complète, la compréhension ne peut qu'évoluer par tâtonnements, chaque étape de la conception et de l'exécution d'un modèle d'AM visant à éliminer progressivement un bruit *dépendant du problème à résoudre* dont la forme est inconnue, et qui est considéré comme inessentiel pour la tâche à accomplir. Par exemple, dans un jeu de données de photos de chiens et de loups, les différences entre chiens

d'une photo à l'autre constituent un bruit à éliminer (de même pour les différences entre loups), tandis que les différences entre chiens et loups constituent un signal à utiliser pour un classifieur distinguant entre les deux espèces. Mais pour un classifieur de races de chiens, entraîné sur le sous-ensemble d'images constitué de chiens uniquement, les différences entre chiens constituent le signal que l'on cherche à exploiter.

Lorsque la paramétrisation d'un problème est complète, et que la forme de la relation entre paramètres est connue, on a recours à l'AM afin d'évaluer les paramètres à partir des données. Par exemple, dans le cas d'une régression linéaire modélisant la relation intensité - tension d'une résistance, on utilise les données expérimentales pour inférer une pente. Dans de tels cas, la forme de la loi d'association entrées-sorties vaut explication des décisions, sur le modèle des théories physiques, et permet un calcul relativement simple des frontières du voisinage du point dans l'espace de données représentant une sortie.

Mais le plus souvent, on a recours à l'AM justement à cause de l'absence d'une connaissance préalable des relations entre entrées et sorties. Ceci vaut *a priori* que la paramétrisation du problème soit complète ou non, et non uniquement dans le cas d'une paramétrisation incomplète, comme affirmé par Doshi-Velez et Kim, ainsi que Lipton. On cherche alors à construire empiriquement une fonction associant des entrées à des valeurs en sortie à partir des données dont on dispose. La forme de cette fonction n'étant pas connue, l'intelligibilité des sorties d'un modèle d'AM repose alors pour l'essentiel sur la production de justifications a posteriori du pourquoi de telle sortie à entrée fixée. Nous avons évoqué précédemment plusieurs techniques permettant de produire de telles justifications, comme l'importance des variables ou la méthode LIME, avec les limites évoquées quant à l'intelligibilité des segmentations réalisées par le modèle instancié.

L'emploi le plus stratégique de l'AM survient donc lorsqu'on l'emploie pour des problèmes pratiques d'une grande complexité, pour lesquels les modalités usuelles d'intelligibilité mathématique ont échoué. L'absence de description mathématique des relations entre entrées et sorties, et la difficulté qui en résulte à obtenir une formulation rigoureuse du problème à résoudre, est un facteur décisif rendant les méthodologies usuelles de programmation par preuves ou par tests unitaires inopérantes, et forçant le passage à une méthodologie heuristique. L'incapacité à obtenir une paramétrisation complète vient complexifier ce problème, mais n'est pas absolument nécessaire à son apparition. Dans une telle configuration méthodologique, l'opacité est constitutive d'une démarche vouée aux tâtonnements et à l'itération heuristique.

5 Conclusion

Notre travail a visé à raffiner les analyses existantes de la catégorie de transparence algorithmique, afin de faciliter son application aux enjeux cruciaux de l'AM. Nous avons notamment distingué quatre sens fondamentaux (loyauté, équité, explicabilité, intelligibilité), regroupés en deux familles (prescriptive et

épistémique).

Parce qu'elles sont essentielles pour comprendre les autres catégories, nous avons ensuite mis l'accent sur les propriétés épistémiques. À cause des problèmes d'intelligibilité propres à cette classe de modèles, il est notamment impossible d'appliquer les mêmes standards d'explicabilité à l'AM et aux algorithmes plus conventionnels dans la situation actuelle. Dans l'esprit du RGPD, nous appelons à la création d'un groupe de travail dédié à la transparence en IA, chargé de produire et de mettre à jour des standards d'explicabilité. Afin de contribuer à l'orientation du travail de ce groupe, nous avons distingué l'intelligibilité de la procédure et l'intelligibilité des sorties. La relative autonomie de ces deux intelligibilités est *de facto* une propriété essentielle des décisions bureaucratiques modernes, parce qu'elle rend possible ce que nous avons appelé l'*explicabilité par extraits* d'une sortie donnée. L'existence d'une forme d'explicabilité des décisions étant cruciale au fonctionnement administratif et légal de nos sociétés, l'enjeu premier pour l'usage de l'AM est de produire de l'intelligibilité et de l'explicabilité des sorties.

Nous avons par conséquent proposé une approche de l'intelligibilité de l'AM centrée sur celle des sorties, en tâchant de comprendre ce qui complexifie l'explicabilité par extraits des sorties des modèles d'AM. Nous avons exploré une analogie entre des propriétés des procédures bureaucratiques ordinaires, que nous avons appelées la compositionnalité et l'élémentarité, et des propriétés techniques des modèles d'AM formulées en termes de segmentation de l'espace des données. Nous avons montré la dépendance entre la compositionnalité des procédures et l'existence de segmentations explicites d'une part, et entre l'élémentarité des procédures et l'intelligibilité des dimensions de l'espace de sortie. Les réseaux de neurones profonds utilisés en analyse d'image s'avèrent être un exemple paradigmatique des difficultés associées à l'interprétation des sorties des modèles d'AM, car les segmentations qu'ils établissent dans l'espace des données peuvent être inintelligibiles, voire impossibles à formuler même pour un expert. C'est ce qui rend difficile le déboggage et la production d'explications des sorties libérées de la complexité de la procédure entière, et même de la complexité de l'engendrement de la procédure elle-même par le processus d'instanciation.

L'interrogation sur l'opacité des méthodes d'AM ne se réduit pas à l'examen des propriétés techniques de ces méthodes : elle porte aussi sur le contexte méthodologique qui nous pousse à les utiliser plutôt que d'autres méthodes a priori plus intelligibles. Nous avons montré que c'était l'incapacité à poser une modélisation formelle du problème traité qui constitue le facteur le plus fondamental, même lorsque l'on dispose d'une paramétrisation complète du problème. L'AM intervient de manière privilégiée dans des situations où les méthodes standards de modélisation et de programmation ont échoué, et où le processus nécessairement exploratoire de l'apprentissage sur les données vient combler le défaut d'intelligibilité créé par l'échec de ces méthodes.

Pour poursuivre ce travail, il serait souhaitable de lever notre hypothèse de simplicité des données et notre restriction à l'apprentissage supervisé. Une typologie fine des procédures bureaucratiques ordinaires, et une étude plus systématique des difficultés d'explication qu'elles posent, seraient également désirables.

Il sera également crucial pour le groupe de travail sur l'explicabilité de décider s'il doit exister une seule explication par procédure, ou si les explications peuvent varier en fonction du public ciblé.

Remerciements

Les auteurs tiennent à remercier Jean-Matthieu Schertzer et Nicolas Bousquet pour leurs relectures de l'article, ainsi que pour leurs commentaires pertinents qui ont permis de préciser certaines notions. Nous remercions également Henri Sahla pour ses remarques stimulantes.

Références

[1] Lipton, Zachary C. The Mythos of Interpretability. In *Proceedings of the 2016 ICML Workshop on Human Interpretability in Machine Learning*, 2016.

[2] Thomas Seiller. Why Complexity Theorists Should Care About Philosophy. In *ANR-DFG "Beyond Logic" Conference*, Cerisy-la-Salle.

[3] CERNA. Éthique de la recherche en apprentissage machine. Technical report, 2017.

[4] INRIA. TransAlgo : évaluer la responsabilité et la transparence des systèmes algorithmiques, Février 2017.

[5] Barocas Solon and Moritz Hardt. Fairness in Machine Learning. NIPS 2017 Tutorial.

[6] Executive Office of the President. Big Data : Seizing Opportunities, Preserving Values. Technical report, 2014.

[7] Camille Caldini. Google est-il antisémite ?

[8] Abdohalli, Benoush, Nasraoui, Olfa. Explainable Restricted Boltzmann Machines for Collaborative Filtering. In *Proceedings of the 2016 ICML Workshop on Human Interpretability in Machine Learning*, 2016.

[9] Parlement Européen. Règlement Général sur la Protection des Données. Règlement UE 2016/679 du Parlement Européen et du Conseil du 27 Avril 2016, 2016.

[10] Assemblée Nationale et Sénat. Loi 2016-1321 du 7 Octobre 2016 pour une République numérique. *Journal Officiel de la République Française*, 0235, 2016.

[11] Will Knight. The Dark Secret at the Heart of IA. *The MIT Technological Review*, 120(3), 2017.

[12] David Gunning. Explainable Artificial Intelligence (XAI).

[13] Goodman, Bryce, Flaxman, Seth. EU regulations on algorithmic decision-making and a "right to explanation". In *Proceedings of the 2016 ICML Workshop on Human Interpretability in Machine Learning*, 2016.

[14] Nick Condry. Meaningful Models : Utilizing Conceptual Structure to Improve Machine Learning Interpretability. In *Proceedings of the 2016 ICML Workshop on Human Interpretability in Machine Learning*, 2016.

[15] Hara, Satoshi, Hayashi, Kohei. Making Tree Ensembles Interpretable. In *Proceedings of the 2016 ICML Workshop on Human Interpretability in Machine Learning*, 2016.

[16] Krause, Josua, Perer, Adam, Bertini, Enrico. Using Visual Analytics to Interpret Predictive Machine Learning Models. In *Proceedings of the 2016 ICML Workshop on Human Interpretability in Machine Learning*, 2016.

[17] Egele, Manuel, Scholte, Theodoor, Kirda, Engin and Kruegel, Christopher. A survey on automated dynamic malware-analysis techniques and tools. *ACM. Comput. Surv.*, 44(2), 2012.

[18] Dhurandhar, Amit, Iyengar, Vijay, Luss, Ronny, Shanmugam, Karthikeyan. A Formal Framework to Characterize Interpretability of Procedures. In *Proceedings of the 2017 ICML Workshop on Human Interpretability in Machine Learning*, 2017.

[19] Adrian Weller. Challenges for Transparency. In *Proceedings of the 2017 ICML Workshop on Human Interpretability in Machine Learning*, 2017.

[20] F. Doshi-Velez and B. Kim. Towards A Rigorous Science of Interpretable Machine Learning. *ArXiv e-prints*, February 2017.

[21] Kurt Hornik, Maxwell B. Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2 :359–366, 1989.

[22] Leo Breiman. Random forests. *Machine Learning*, 45(1) :5–32, October 2001.

[23] Joseph P. Hoffbeck and David A. Landgrebe. Covariance matrix estimation and classification with limited training data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(7) :763–767, July 1996.

[24] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9 :2579–2605, 2008.

[25] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" : Explaining the predictions of any classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144, New York, NY, USA, 2016. ACM.

[26] Marco Tulio Ribeiro Marco ; Sameer Singh ; Carlos Guestrin. Introduction to Local Interpretable Model-Agnostic Explanations (LIME), 2016.

[27] James Vincent. Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech, 2018.

[28] National Transportation Safety Board Office of Public Affairs. Driver Errors, Overreliance on Automation, Lack of Safeguards, Led to Fatal Tesla Crash, 2017.

[29] Louis Wehenkel. On uncertainty measures used for decision tree induction. In *IPMU-96, Information Processing and Management of Uncertainty in Knowledge-Based Systems*, page 6. 1996.

[30] J. Lee, Y. Bahri, R. Novak, S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein. Deep Neural Networks as Gaussian Processes. *ArXiv e-prints*, 2017.

[31] Robin Senge, Stefan Bösner, Krzysztof Dembczynski, Jörg Haasenritter, Oliver Hirsch, Norbert Donner-Banzhoff, and Eyke Hüllermeier. Reliable classification : Learning classifiers that distinguish aleatoric and epistemic uncertainty. In *Information Sciences*, volume 255, pages 16–29. 2014.

[32] H. Wang and D.-Y. Yeung. Towards Bayesian Deep Learning : A Survey. *ArXiv e-prints*, April 2016.

[33] A. Kendall and Y. Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? *ArXiv e-prints*, March 2017.