Une terminologie pour une IA explicable contextualisée

Matthieu Bellucci, Nicolas Delestre, Nicolas Malandain, Cecilia Zanni-Merk

Une terminologie pour une IA explicable contextualisée

Matthieu Bellucci*, Nicolas Delestre*
Nicolas Malandain* Cecilia Zanni-Merk*

*Normandie Université, INSA Rouen, LITIS, St-Etienne du Rouvray 76800, France
matthieu.bellucci@insa-rouen.fr

Résumé. L'Intelligence Artificielle explicable a connu un gain de popularité ces dernières années, grâce aux nouvelles législations promouvant le « droit à l'explication ». De nombreuses méthodes ont été récemment développées pour aider à comprendre les modèles dits « boîtes noires », mais la définition de ce qu'est une explication n'est pas encore clairement définie. De plus, la communauté s'accorde à dire que de nombreux termes importants n'ont pas de définition consensuelle. Dans cet article, nous passons en revue la littérature et montrons qu'il existe un problème majeur concernant les définitions des termes comme explicabilité ou interprétabilité. Pour résoudre ce problème, nous proposons une terminologie qui prend en compte le contexte d'un système d'IA. Cette terminologie est compatible avec la majorité des définitions vues dans la littérature, de sorte qu'elle puisse servir de base pour les travaux futurs. Nous discutons également des métriques permettant d'évaluer certaines propriétés définies dans cette terminologie.

1 Introduction

L'Intelligence Artificielle (IA) est largement utilisée dans l'industrie et la vie quotidienne. Jusqu'à récemment, son utilisation n'était pas problématique car elle se limitait à des décisions sans conséquences, telles que des recommandations de films ou du ciblage publicitaire. Cependant, l'IA a fait son entrée dans des secteurs comme la banque, les assurances ou encore la médecine (Guidotti et al., 2018). Dans ces secteurs, il est essentiel de comprendre pourquoi et comment une IA a pris une certaine décision. Il est normal de se demander pourquoi un prêt bancaire a été refusé, ou de comprendre comment un diagnostic a été fait pour un patient. C'est ce droit à l'explication que défend, entre autres, la régulation européenne GDPR (2016). L'utilisation d'algorithmes d'IA complexes et en grandes dimensions rend la conception d'explications difficile, voire impossible. C'est ce besoin d'explicabilité qui a conduit à la popularisation du domaine de l'IA explicable (XAI), avec des projets tels que le programme XAI de la DARPA (Gunning, 2017). La figure 1 présente le cas d'utilisation le plus important de l'IA explicable selon ces auteurs : « lorsqu'en pratique, dans un certain contexte, un utilisateur doit comprendre, faire confiance et être responsable des conclusions tirées par un système d'IA ». Le modèle explicable et l'interface d'explication de cette architecture représentent ce que nous appelons un système intelligent explicable (XIS). Un modèle explicable associe ses sorties à un ensemble de règles (ou autre formalisme) pour fonder ses décisions. Une inter-

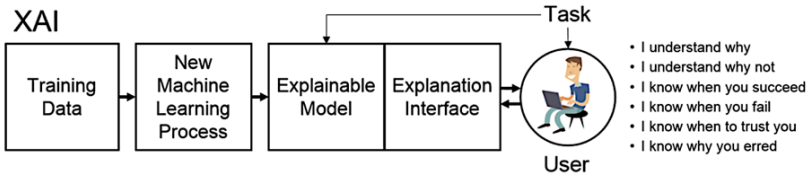

FIG. 1 – *Un exemple typique de cas d'utilisation du XAI, d'après Gunning (2017)*

face d'explication interagit avec un utilisateur pour lui fournir des explications, fondées sur les entrées, les sorties et les caractéristiques du modèle explicable.

Il n'existe pas encore de consensus sur la signification des termes utilisés, ce qui pose un problème pour évaluer la qualité et l'utilité des méthodes développées. De nombreux chercheurs ont abordé la question de l'absence d'une terminologie approuvée par la communauté, en soulignant les problèmes que cela engendre (Arrieta et al., 2020; Lipton, 2018; Adadi et Berrada, 2018). Ils soulignent également le manque d'indicateurs permettant d'évaluer les différents aspects de l'explicabilité d'un système. Cependant, à notre connaissance, il n'existe aucun article proposant une terminologie qui reprend les termes populaires de ce domaine et leur attribue des définitions non ambiguës. C'est ce que nous proposons dans cet article.

Dans la section 2, nous examinerons les termes observés dans la littérature et les définitions qui leur sont associées. Ensuite, dans la section 3, nous proposons une terminologie cohérente et non ambiguë fondées sur les définitions vues dans la section précédente, dont l'utilité sera illustrée par une étude de cas. Dans cette terminologie, nous proposerons certains concepts que nous qualifions de mesurable, car il est possible de déterminer différentes métriques mesurant la qualité ou le degré de présence de ces concepts. Nous discuterons donc de l'élaboration de ces métriques. Enfin, nous conclurons dans la section 4.

2 Analyse de la littérature

Nous avons étudié la littérature pour déterminer et comprendre la terminologie du XAI. Nous avons remarqué que de nombreux termes reviennent, mais sans consensus concernant leurs définitions (Adadi et Berrada, 2018; Doshi-Velez et Kim, 2017; Calegari et al., 2020). Parmi ces termes, explicabilité et interprétabilité sont les plus importants mais aussi les plus débattus. Pour quelques termes, la communauté du XAI est parvenue à un consensus sur leur sens général mais sans proposer de définition. Cette absence de terminologie claire constitue un problème pour la recherche dans ce domaine.

Nous avons étudié plus de 30 articles liés au XAI. Nous avons utilisé les mots-clés "XAI", "terminology", "taxonomy", "survey", "review", "explainability", "interpretability" dans le moteur de recherche Google Scholar, consulté entre octobre 2020 et janvier 2021 ; en se concentrant sur les études, taxonomies et revues, ainsi que les articles populaires proposant des méthodes d'explicabilité. Notre objectif était d'identifier les termes et leurs définitions associées qui reviennent dans cet état de l'art. Nous pensons que la combinaison de ces revues de la littérature donne une bonne vue d'ensemble de la terminologie du XAI. Le reste des articles nous a permis d'affiner nos définitions afin de les rendre compatibles avec les méthodes existantes.

Dans cette section, nous étudierons comment les termes récurrents sont utilisés et définis dans différents articles. Puis nous identifierons les points de convergence et divergence pour chacun d'entre eux. Enfin, nous déterminerons comment les termes sont liés entre eux.

2.1 Interprétabilité

Interprétabilité est le terme favori de la communauté scientifique pour décrire les techniques d'IA qui sont facilement compréhensibles (Adadi et Berrada, 2018). Ce terme est utilisé pour décrire des modèles ou algorithmes plutôt que des systèmes intelligents complets. La définition implicite que certains auteurs utilisent est qu'un modèle interprétable est un modèle facile à comprendre par la majorité des utilisateurs. Dans Lipton (2018), consacré à l'analyse de ce qu'est l'interprétabilité, l'auteur conclut que qualifier un modèle d'intrinsèquement interprétable est dénué de sens. Adadi et Berrada (2018) proposent la définition suivante : « Un système interprétable est un système qu'un utilisateur peut non seulement voir, mais aussi étudier et comprendre comment les entrées sont mathématiquement liées aux sorties ». Cette définition correspond bien à la définition implicite décrite plus tôt. Gilpin et al. (2018) proposent également une définition similaire. Lipton (2018) et Futia et Vetrò (2020) introduisent les notions de simulabilité et de décomposabilité. Ces notions peuvent être considérées comme des sous-ensembles d'interprétabilité. Ceci est cohérent avec la définition d'Adadi et Berrada (2018) : si nous pouvons décomposer ou simuler un modèle, cela implique que nous pouvons comprendre comment les entrées sont liées aux sorties.

Une autre composante qui apparaît dans de nombreuses définitions est l'effort cognitif nécessaire pour comprendre un modèle. Par exemple, Calegari et al. (2020) déclarent que « dans les algorithmes d'IA, l'interprétabilité correspond à l'effort cognitif requis par les utilisateurs humains pour donner un sens à la façon dont l'algorithme fonctionne, ou motiver le résultat produit ». Ribeiro et al. (2016) mentionnent que les limitations de l'utilisateur doivent être prises en compte et Gilpin et al. (2018) ajoutent que « le succès de cet objectif est lié à la cognition, aux connaissances et aux biais de l'utilisateur ». Cet effort cognitif demandé peut être mesuré par la complexité d'un modèle. La complexité d'un modèle est largement utilisée dans la littérature. Elle peut être définie comme une mesure de l'interprétabilité d'un modèle (Adadi et Berrada, 2018; Ribeiro et al., 2016; Guidotti et al., 2018). Des exemples de mesure de complexité pour différents modèles peuvent être trouvés dans différents articles (Gilpin et al., 2018; Futia et Vetrò, 2020; Ribeiro et al., 2016; Guidotti et al., 2018; Wu et al., 2018).

Nous pouvons conclure que l'interprétabilité fait référence à la capacité d'un objet à être compris et étudié par un utilisateur avec un effort cognitif raisonnable. Nous utilisons le terme objet plutôt que modèle, car comme Lipton (2018) l'a mentionné, l'entrée doit être compréhensible, tout comme le modèle. Par conséquent, nous pourrions également définir une entrée comme interprétable. L'interprétabilité englobe également différentes notions telles que la décomposabilité et la simulabilité.

2.2 Explicabilité et explications

Certains chercheurs utilisent explicabilité et interprétabilité comme synonymes, mais ne donnent pas de définition claire. Beaudouin et al. (2020) proposent la définition suivante de l'explicabilité : « la capacité, l'inclination ou l'aptitude à rendre clair ou compréhensible, ou à expliquer le sens d'un algorithme ». En d'autres termes, l'explicabilité est la capacité d'un système à générer des explications. Comme discuté dans une série de quatre articles par Hoffman,

Klein et al. (Hoffman et Klein, 2017; Hoffman et al., 2017; Klein, 2018; Hoffman et al., 2018), définir ce qu'est une explication est une tâche complexe. Expliquer quelque chose équivaut à répondre aux questions qu'un utilisateur peut se poser, pour être en mesure de comprendre ce qu'il observe. L'objectif est de lui fournir les informations pertinentes afin qu'il puisse raisonner par lui-même sur le fonctionnement d'un modèle, ou pourquoi un modèle a pris une certaine décision. Calegari et al. (2020) définissent une explication comme « une activité visant à rendre les détails pertinents d'un objet clairs ou faciles à comprendre pour un observateur ». Arrieta et al. (2020) et Guidotti et al. (2018) partagent la même définition, « une explication est une interface entre les humains et un décideur qui est à la fois une exacte représentation du décideur et compréhensible pour les humains ». Cette définition ajoute une dimension à ce que devrait être une explication. Elle doit être une exacte représentation, c'est-à-dire que l'explication doit être fondée sur les mécanismes du modèle et les entrées utilisées. Ribeiro et al. (2016) définissent une explication comme présentant « des indices textuels ou visuels qui permettent une compréhension qualitative de la relation entre les composants de l'instance et la prédiction du modèle ». Nous observons que cette définition est un sous-ensemble des précédentes. En effet, les indices textuels ou visuels peuvent être les détails pertinents dont nous parlions précédemment. Nous considérons que leur définition est adaptée à leur propre méthode, mais qu'elle manque de généralité. En conclusion, à partir de ces définitions, une explication est une interaction entre l'utilisateur et un autre agent dont le but est de fournir des détails pour répondre aux questions de l'utilisateur ou rendre un système ou une décision facile à comprendre.

La communauté du XAI semble s'accorder sur une taxonomie de méthodes de génération d'explications. Arya et al. (2019) proposent une taxonomie fondée sur des questions concernant ce qui est expliqué, comment cela est expliqué et à quel niveau. Cette taxonomie prend la forme d'un arbre de décision. Ils introduisent également l'idée qu'une explication peut être soit statique, soit interactive, ce qui est rarement abordé dans la littérature. Nous avons rencontré les notions d'explicabilité globale/locale et d'explications *post hoc*/directes dans la majorité des articles examinés.

Explications *post hoc*/directes Selon Futia et Vetrò (2020), les explications *post hoc* « ne cherchent pas à révéler comment un modèle fonctionne, mais elles se concentrent sur comment il s'est comporté et pourquoi ». Lipton (2018) dit que « ces interprétations peuvent expliquer les prédictions sans élucider les mécanismes des modèles ». Guidotti et al. (2018) les qualifient d' « approches de rétro-ingénierie ». Nous comprenons de ces définitions que les explications *post hoc* n'exploitent pas la logique du système. Arya et al. (2019) disent que les explications *post hoc* impliquent « des méthodes auxiliaires pour expliquer un modèle après son apprentissage ». Parmi les méthodes existantes, nous pouvons citer LIME (Ribeiro et al., 2016) ou SHAP (Lundberg et Lee, 2017). Les méthodes d'explication *post hoc* sont généralement agnostiques par rapport au modèle, elles peuvent donc être appliquées à n'importe quel modèle, qu'il soit interprétable ou non. Certains chercheurs argumentent que les méthodes d'explication opposées aux explications *post hoc* utilisent la nature « directement »interprétable du système qu'elles veulent expliquer (Arya et al., 2019; Arrieta et al., 2020; Calegari et al., 2020). Ce type d'explication est rarement défini explicitement dans la littérature et n'est donc pas associé à un terme spécifique. Nous proposons donc le terme d'explications directes pour désigner ces méthodes. Adadi et Berrada (2018) nomment ces méthodes « méthodes liées à la complexité », ce qui correspond à la conception de modèles intrinsèquement interprétables.

Cette notion d'utilisation de l'interprétabilité intrinsèque d'un modèle est également utilisée par Arya et al. (2019). Ces explications directes sont évidemment dépendantes du modèle, car celui-ci est utilisé comme source pour générer les explications.

Explications globales/locales L'explicabilité globale fait référence à l'explication du fonctionnement d'un modèle. Adadi et Berrada (2018) et Guidotti et al. (2018) disent que l'explicabilité globale « facilite la compréhension de toute la logique d'un modèle ». Hoffman et al. (2018) la décrivent comme l'explication de « la manière dont les catégories et les mécanismes conceptuels sont dérivables des instances et de leurs attributs ». L'explicabilité locale fait référence à l'explication d'une seule prédiction faite par une IA. Cette idée est partagée par de nombreux chercheurs (Arya et al., 2019; Adadi et Berrada, 2018; Doshi-Velez et Kim, 2017; Futia et Vetrò, 2020). Les explications locales sont généralement axées sur l'analyse de l'importance des caractéristiques dans les algorithmes d'apprentissage (comme LIME (Ribeiro et al., 2016)) et peuvent utiliser des mécanismes de raisonnement spécifiques pour répondre à des questions de type « que se passerait-il si ».

2.3 Transparence et boîtes noires

La transparence est un terme controversé. Pour certains auteurs, la transparence est étroitement liée à l'interprétabilité. Guidotti et al. (2018) discutent du problème des modèles dits « boîtes transparentes », qui, selon eux, « consistent à fournir directement un modèle qui est localement ou globalement interprétable ». Arya et al. (2019) disent qu' « un modèle directement interprétable est un modèle qui, par sa nature transparente, est compréhensible par la plupart des consommateurs ». Lipton (2018) avance que certains articles qualifient les modèles compréhensibles de transparents, tandis que les modèles incompréhensibles sont appelés boîtes noires. Selon les définitions de l'interprétabilité vues dans la section 2.1, un modèle compréhensible est interprétable. Par conséquent, la distinction entre transparent et interprétable n'est pas claire.

Beaudouin et al. (2020) proposent une définition différente : « La transparence correspond généralement à la mise à disposition d'informations sur le fonctionnement interne de l'algorithme, y compris la manière dont un système d'IA est développé, formé et déployé ». Ils ajoutent également que « la transparence ne signifie pas nécessairement que les informations sous-jacentes sont facilement compréhensibles ». Un modèle interprétable devient alors un modèle qu'un utilisateur peut comprendre, alors qu'un modèle transparent est un modèle qui fournit toutes les informations sur son développement. La transparence peut contribuer à fournir des explications, mais elle permet plutôt d'en garantir l'équité et l'impartialité.

Futia et Vetrò (2020) et Lipton (2018) distinguent trois niveaux de transparence. La transparence considérée au niveau du modèle entier est appelée simulatabilité, au niveau des composants individuels tels que les paramètres, appelée décomposabilité et au niveau de l'algorithme d'apprentissage, appelée transparence algorithmique. Nous soulignons que ces notions sont également liées à l'interprétabilité, comme indiqué dans la section 2.1.

Simulabilité Futia et Vetrò (2020) définissent la simulabilité comme « la vérification par une approche heuristique si un humain comprend le fonctionnement du modèle, et par conséquent si l'humain est capable de simuler le processus de décision ». Lipton (2018) propose une définition similaire, qui est également utilisée par Wu et al. (2018) : « L'homme doit être capable de prendre des données d'entrée ainsi que les paramètres du modèle et, en un temps raison-

nable, effectuer tous les calculs nécessaires pour produire une prédiction ». Arrieta et al. (2020) donnent une définition semblable et ajoutent que la complexité du modèle est très importante pour la simulabilité. Nous pouvons également déduire que la simulabilité d'un modèle dépend de la capacité cognitive de l'utilisateur, de la même manière que pour l'interprétabilité.

Décomposabilité Futia et Vetrò (2020), Arrieta et al. (2020) et Lipton (2018) sont tous d'accord sur la définition de la décomposabilité : « chaque partie du modèle - entrées, paramètres et calculs - admet une explication intuitive ».

Lipton (2018) définit les modèles boîtes noires comme des « modèles incompréhensibles ». Calegari et al. (2020) approfondissent la définition : « utilisé pour faire référence à des modèles où la connaissance n'est pas explicitement représentée, mais plutôt distribuée parmi des tenseurs de nombres réels, dont la complexité correspond rarement à nos capacités cognitives en tant qu'humains ». Nous retrouvons ici la notion de capacité cognitive, qui fait écho à la définition de l'interprétabilité.

3 Proposition de terminologie

Maintenant que nous avons étudié l'état actuel de la terminologie dans la littérature, nous proposons une terminologie qui utilise ces termes de manière non ambiguë. Les définitions proposées pour chaque terme sont présentées en *italique*. A la première apparition d'un terme défini de la terminologie dans cette section, le terme est présenté en **gras**.

Pour chaque terme correspondant à une propriété mesurable, nous discuterons de métriques pour évaluer celle-ci et des problèmes soulevés par ces métriques. Un problème majeur pour les métriques du XAI est le besoin d'avoir des avis des utilisateurs. Obtenir des avis représentatifs des utilisateurs est complexe, coûteux et souvent peu généralisable à différentes sortes de système. Il est envisageable de faire de tels sondages lorsque les utilisateurs d'un système sont connus et peu nombreux. Mais même dans ce cas de figure, ces mesures subjectives sont à manipuler avec précaution car le ressenti des utilisateurs peut varier dans le temps. Par exemple, un système peut être qualifié de boîte noire lorsqu'il est mis en place au sein d'un petit groupe de personnes. Puis, après un temps d'apprentissage et d'adaptation, les mêmes utilisateurs pourront le qualifier d'interprétable, car ils ont acquis les connaissances nécessaires pour le comprendre. C'est pourquoi, nous préférerons alors discuter de métriques objectives, qui ne demandent pas l'avis d'un utilisateur, lorsque cela est possible.

3.1 Terminologie générale

Dans cette section, nous définirons ce qu'est un XIS, ce qu'est l'**explicabilité** et l'**interprétabilité** et nous introduirons certaines notions que nous n'avons pas abordées dans la section 2. Nous discuterons également des métriques pour mesurer les concepts proposés.

Utilisateur Dans la littérature, la personne qui utilise un XIS est appelée un **utilisateur**, un observateur ou même un client. Nous proposons la définition suivante : *un utilisateur est un agent qui interagit avec un XIS*. Cette définition nous permet de considérer un humain ou un autre agent comme un utilisateur. Comme nous l'avons vu dans la littérature, l'explicabilité et l'interprétabilité doivent tenir compte de l'utilisateur afin de fournir une explication utile et pertinente (Adadi et Berrada, 2018; Futia et Vetrò, 2020; Ribeiro et al., 2016). De nombreuses définitions de cette terminologie dépendent de l'utilisateur et de sa tâche.

La création de mesures dans le domaine du XAI nécessite que ces mesures prennent en compte le contexte. Hoffman et al. (2018) parlent des différents indicateurs pour mesurer la qualité d'une explication. Dans leur conclusion, ils rappellent que selon le contexte ou l'application, la même valeur pour une mesure n'a pas la même interprétation. Par exemple, prenons une température de 50°C. S'il s'agit de la température extérieure, on dira qu'il fait chaud. Cependant, s'il s'agit de la température d'un four, on dira qu'il est froid. On remarque donc qu'au-delà du problème d'élaborer des mesures, il faut également se pencher sur l'interprétation de ces mesures, dans un contexte donné.

Explications et explicabilité A partir de notre analyse de la littérature, nous proposons la définition suivante : *une explication est le résultat d'une interaction entre un utilisateur et un explicateur dans le but de répondre aux questions de l'utilisateur.* Nous définissons l'explicabilité comme *la capacité d'un XIS à être expliqué à un utilisateur ou à fournir une explication*.

Dans la figure 1, l'interaction est représentée par les flèches entre l'interface d'explication et l'utilisateur. L'interface d'explication débute l'interaction en fournissant au moins la décision du modèle explicable. L'interface peut également ajouter une première explication à la décision, afin de répondre aux questions habituelles, ou bien elle peut attendre que l'utilisateur pose des questions. Le but d'une explication est de répondre aux questions de l'utilisateur sur le système et/ou sa décision. Une explication est considérée comme valide lorsque l'utilisateur n'a plus de questions. Cette notion d'interaction est également ce qui différencie l'explicabilité et l'interprétabilité.

Il est possible de mesurer différentes propriétés d'une explication. Par exemple, l'exactitude de celle-ci ou encore sa qualité. Hoffman et al. (2018) discutent des différents aspects qui font la qualité d'une explication et comment mesurer ces aspects. De nombreuses métriques proposées se fondent sur une évaluation par les utilisateurs, comme le mentionnent Arya et al. (2019). Des métriques plus spécifiques existent pour mesurer objectivement certaines propriétés d'explication, mais de manière générale, aucune ne fait consensus.

Interprétabilité Nous définissons l'interprétabilité comme *la capacité d'un objet ou d'un XIS à être observé, compris et étudié par un utilisateur, avec un effort cognitif raisonnable*. Il s'agit donc d'une propriété d'un XIS ou d'un objet. Ce terme désigne une propriété d'un système, mais compte tenu de son importance, nous avons décidé de le définir dans cette section. Pour qu'un objet soit interprétable par un utilisateur, ce dernier doit avoir accès à suffisamment d'informations pour pouvoir comprendre cet objet. Cette nécessité lie l'interprétabilité à la **transparence**. De nouveau, l'interprétabilité d'un objet dépend des connaissances de l'utilisateur, un même modèle peut être interprétable pour un utilisateur mais pas pour un autre.

Comme pour l'explicabilité, mesurer l'interprétabilité d'un objet ou d'un système dépend fortement de l'utilisateur cible. Carvalho et al. (2019) mentionnent la linéarité, la monotonicité et le niveau d'interactions entre les variables comme étant des indicateurs objectifs de l'interprétabilité. Dans la section 3.2.1, nous définissons des concepts qui peuvent être considérés comme des composants nécessaires de l'interprétabilité. C'est par le biais de ces composants que nous tentons de décomposer la mesure de l'interprétabilité en différents critères plus simples à mesurer. La **complexité**, la **décomposabilité** et la **simulabilité** qui seront définis plus tard permettront ainsi d'estimer l'interprétabilité d'un système.

Confiance Gunning (2017), dans le cadre du programme XAI de la DARPA, souligne que l'utilisateur final doit faire confiance aux conclusions tirées par un système d'IA. *Faire confian-*

ce, c'est croire que quelque chose est sûr et fiable. Si un utilisateur ne fait pas confiance à un XIS, il ne l'utilisera jamais, le rendant inutile. Pour qu'un utilisateur fasse confiance à un XIS, il faut lui fournir des informations complètes sur sa structure, sa fonction et son comportement. Cela se fait par le biais d'explications, de transparence et en fournissant les cas limites du système. Comme pour l'interprétabilité et l'explicabilité, la confiance est une notion très subjective, qui nécessite une évaluation par des humains. Hoffman et al. (2018) discutent plus en détails de mesurer la confiance d'un XIS.

3.2 Terminologie d'un système

De nombreuses notions que nous avons étudiées correspondent à des propriétés d'un XIS. Nous faisons une distinction entre un modèle et un système. Le modèle est l'objet qui calcule la prédiction à partir de l'entrée, ce qui signifie qu'il fait partie de ce système. Nous avons défini l'interprétabilité comme étant une propriété d'un objet ou d'un système, par conséquent, tout composant d'un système peut être interprétable en soi, sans garantir que le système entier soit interprétable. Nous divisons ces notions en deux parties, la première partie concerne les propriétés d'un XIS liées à son interprétabilité. La seconde partie concerne les concepts permettant d'accroître la confiance dans un XIS.

3.2.1 Vers des systèmes plus interprétables

Les notions définies dans cette section visent à comprendre les concepts sous-jacents de l'interprétabilité et les façons de la mesurer.

Boîte noire Nous considérons que *boîte noire est l'opposé d'interprétabilité*. On peut également parler d'opacité. Cette définition est directement liée à la définition de l'interprétabilité, ainsi le concept de boîte noire dépend de l'utilisateur. Certains modèles peuvent être jugés trop complexes pour être compris par quiconque. Les réseaux neuronaux profonds, par exemple, sont notoirement difficiles à comprendre et il est communément admis qu'ils sont des boîtes noires, quel que soit l'utilisateur.

La mesure du degré d'opacité est directement liée à la mesure du degré d'interprétabilité. Afin de déterminer si un système est boîte noire, on pourra fixer des seuils sur les indicateurs d'interprétabilité. Par exemple, si la complexité dépasse un certain seuil, on peut qualifier le modèle de boîte noire. Plus l'utilisateur est expérimenté, plus ce seuil sera élevé. Néanmoins, nous pensons qu'il est possible de déterminer des seuils standards pour certaines catégories d'utilisateurs.

Complexité La complexité est *la mesure de l'interprétabilité d'un objet, c'est-à-dire la mesure de la facilité avec laquelle un utilisateur peut simuler et/ou comprendre un objet*. Cette mesure est dépendante du modèle. De nombreux exemples sont donnés dans la littérature, par exemple la complexité d'un arbre de décision pourrait être le nombre de nœuds de l'arbre, ou la longueur maximale d'un chemin. La complexité d'une régression linéaire pourrait être le nombre de variables. D'autres exemples sont donnés dans Guidotti et al. (2018). Il s'agit d'une mesure objective de l'interprétabilité, qui ne demande pas de retour de l'utilisateur.

Simulabilité La simulabilité fait référence à *la capacité d'un modèle à être simulé ou reproduit par un utilisateur*. Pour mesurer la simulabilité, il faut définir des métriques permettant d'évaluer si un utilisateur peut simuler ou non le système. La simulabilité permet à un utilisateur d'effectuer de nombreuses expériences et manipulations du système par lui-même, il n'a

donc pas besoin de demander une interface d'explication pour obtenir des réponses. Certains auteurs soutiennent que si un modèle est simulable, alors le système qui utilise ce modèle est interprétable. Ces deux notions sont liées, mais le fait de pouvoir simuler un système ne signifie pas que l'utilisateur en comprend chaque composant. Cela nous conduit à la décomposabilité, qui permet de résoudre ce problème particulier.

Une mesure de la simulabilité est proposée par Slack et al. (2019). Il s'agit de mesurer le temps que met un utilisateur pour calculer la prédiction d'un modèle, en lui fournissant une représentation de ce modèle et une entrée. Nous pensons qu'il est possible d'automatiser cette tâche afin de ne plus avoir recours à l'humain. En effet, nous pouvons envisager de déterminer le temps de calcul de chaque opération en se basant sur une expérience avec des utilisateurs. Une fois ce temps de calcul connu pour chaque opération et chaque type d'utilisateur, nous pouvons déduire le temps de calcul pour l'ensemble du modèle. Cette intuition nécessite une exploration plus approfondie, mais nous pensons que la mesure de la simulabilité pourrait être automatisée.

Décomposabilité *Un système est décomposable si chacun de ses composants (entrées, paramètres et calculs) est interprétable*. Il est vital pour un utilisateur de pouvoir comprendre chaque composant s'il veut pouvoir comprendre comment une prédiction est faite. De même, la compréhension des paramètres et des calculs peut aider à comprendre et à simuler le modèle. Elle contribue à la confiance et à l'interprétabilité, de la même manière que la simulabilité. La combinaison de la simulabilité et de la décomposabilité conduit un système vers l'interprétabilité, bien que nous ne puissions pas conclure que leur combinaison garantit l'interprétabilité du système.

Mesurer la décomposabilité nécessite de déterminer l'interprétabilité de chaque composant d'un système. Les composants d'un système étant immuables, nous pensons qu'il serait donc possible de déterminer l'interprétabilité de chaque composant grâce à une seule étude, ce qui permettrait par la suite de déterminer la décomposabilité de tout système. De la même manière que pour la simulabilité, cette idée demande plus de réflexion afin de s'assurer de la qualité de cette analyse d'interprétabilité. Le problème de gérer les différents profils d'utilisateur se pose à nouveau.

Dans la section 3.1, nous discutions de la mesure de l'interprétabilité. Nous avancions qu'une faible complexité, un haut degré de décomposabilité et de simulabilité sont des propriétés nécessaires à l'interprétabilité. En décidant de seuils pour chacune de ces notions, nous pourrons alors déterminer si le système n'est pas interprétable. Malheureusement, nous pensons que d'autres conditions nécessaires à l'interprétabilité existent mais ne sont pas décrites dans cette terminologie. C'est pourquoi nous ne pouvons pas proposer de métrique pour mesurer l'interprétabilité d'un système.

3.2.2 Vers des systèmes dignes de confiance

Cette section contient des termes qui ne servent pas directement l'explicabilité ou l'interprétabilité, mais dont le but est de fournir les propriétés qu'un système devrait avoir pour gagner la confiance d'un utilisateur. Comme nous l'avons vu dans la section 3.1, la confiance est d'une importance capitale pour garantir l'utilisation d'un système.

Équité L'équité est *la capacité d'un système à éviter toute forme de discrimination injustifiée à chaque niveau du système*. Nous spécifions « discrimination injustifiée » car, comme discuté dans Corbett-Davies et Goel (2018), lorsqu'il existe une différence prouvée entre deux

groupes, il peut être discriminatoire de ne pas utiliser cette différence dans le système pour les deux groupes. La conception de systèmes équitables est extrêmement complexe, les solutions intuitives pour éviter les discriminations ne sont pas toujours valables, selon Corbett-Davies et Goel (2018). L'un des objectifs de l'explicabilité est de détecter quand un système discrimine un certain groupe ou utilise des variables qui ne devraient pas être pertinentes. Assurer l'équité d'un système contribue à la confiance accordée à ce système. Pour mesurer cette équité, Bellamy et al. (2018) proposent un logiciel permettant de mesurer de nombreux critères relatifs à l'équité, d'expliquer les résultats de ces mesures et de corriger automatiquement les biais et discriminations.

Transparence Comme nous l'avons vu dans la section 3.2.1, nous n'utilisons pas la transparence comme l'opposé de boîte noire. Nous préférons la définition plus globale de la transparence, proposée par Beaudouin et al. (2020). Ainsi, un système est transparent s'il fournit toutes les informations sur sa conception et son fonctionnement. La transparence ne garantit pas qu'un utilisateur sera en mesure de comprendre le système, mais garantit qu'il aura accès à toutes les informations concernant les données d'entraînement, la façon dont les données ont été traitées dans le cas où l'IA est fondé sur un apprentissage artificiel, les performances du système, etc. La transparence d'un système est une notion similaire aux programmes open-source, dans lesquels l'ensemble du code peut être vu et étudié par n'importe qui, ce qui empêche le concepteur d'utiliser ce programme à des fins malveillantes. Le fait de fournir ces informations est susceptible d'accroître la confiance dans un système, car tout utilisateur expérimenté peut vérifier que le système est correctement conçu et qu'il ne présente aucune faille ou parti pris susceptible d'altérer les prédictions et éventuellement d'entraîner une discrimination.

Compte tenu de notre définition de transparence, il est difficile de lui associer une métrique. Nous pouvons cependant définir une liste de critères à remplir pour chaque type de système. Par exemple, s'assurer de la disponibilité du code source ou encore la reproductibilité des résultats.

4 Conclusion

Nous avons introduit la notion de système intelligent explicable (XIS) qui est composé d'un modèle explicable et d'une interface d'explication. Nous avons ensuite étudié la littérature pour identifier les définitions des termes récurrents proposées par la communauté. Ensuite, nous avons proposé une terminologie pour une IA explicable contextualisée. Nous avons défini les concepts généraux du XAI, en particulier l'explicabilité et l'interprétabilité. Nous avons également souligné l'importance de l'utilisateur dans la conception des explications. Puis, nous avons défini des concepts liés à l'interprétabilité d'un XIS et le niveau de confiance qui lui est donné. Les premiers concepts aideront à déterminer l'interprétabilité d'un XIS, à travers une variété d'indicateurs différents, tels que la complexité, la simulabilité et la décomposabilité. Enfin, pour chaque concept mesurable défini, nous avons discuté des manières de les mesurer et des problèmes posés par l'élaboration de telles mesures. Parmi ces problèmes, l'intégration de l'utilisateur et son contexte dans les mesures est le plus important.

Cette terminologie permettra de mieux comprendre comment sont liés les termes du XAI et ainsi combiner certaines métriques pour évaluer des concepts de plus haut niveau comme l'explicabilité ou l'interprétabilité. De nombreux termes doivent encore être clairement définis, notamment les notions liées à l'effort cognitif requis par un utilisateur pour comprendre un

système. Nous devons identifier les facteurs qui entrent en jeu pour déterminer la capacité d'un utilisateur à comprendre ou non un XIS. Nous pensons que les connaissances de l'utilisateur (par exemple, son cursus ou ses expériences) ont un impact sur l'effort cognitif requis. D'autres facteurs relevant de la psychologie et de la sociologie pourraient également être pris en compte. Nous explorerons également les explications interactives qui n'ont pas encore été étudiées en profondeur par la communauté XAI. Nous espérons ainsi fournir des définitions nouvelles et améliorées qui tiendront compte des dernières avancées dans ce domaine.

Références

Adadi, A. et M. Berrada (2018). Peeking inside the black-box : a survey on explainable artificial intelligence (xai). *IEEE access 6*, 52138–52160.

Arrieta, A. B., N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al. (2020). Explainable artificial intelligence (xai) : Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion 58*, 82–115.

Arya, V., R. K. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilović, et al. (2019). One explanation does not fit all : A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv :1909.03012*.

Beaudouin, V., I. Bloch, D. Bounie, S. Clémençon, F. d'Alché Buc, J. Eagan, W. Maxwell, P. Mozharovskyi, et J. Parekh (2020). Flexible and context-specific ai explainability : a multidisciplinary approach. *Available at SSRN 3559477*.

Bellamy, R. K., K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, et al. (2018). Ai fairness 360 : An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv :1810.01943*.

Calegari, R., G. Ciatto, et A. Omicini (2020). On the integration of symbolic and sub-symbolic techniques for xai : A survey. *Intelligenza Artificiale 14*(1), 7–32.

Carvalho, D. V., E. M. Pereira, et J. S. Cardoso (2019). Machine learning interpretability : A survey on methods and metrics. *Electronics 8*(8), 832.

Corbett-Davies, S. et S. Goel (2018). The measure and mismeasure of fairness : A critical review of fair machine learning. *arXiv preprint arXiv :1808.00023*.

Doshi-Velez, F. et B. Kim (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv :1702.08608*.

Futia, G. et A. Vetrò (2020). On the integration of knowledge graphs into deep learning models for a more comprehensible ai—three challenges for future research. *Information 11*(2), 122.

GDPR (2016). General data protection regulation.

Gilpin, L. H., D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, et L. Kagal (2018). Explaining explanations : An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pp. 80–89. IEEE.

Guidotti, R., A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, et D. Pedreschi (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR) 51*(5), 1–42.

Gunning, D. (2017). Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web 2*(2).

Hoffman, R., T. Miller, S. T. Mueller, G. Klein, et W. J. Clancey (2018). Explaining explanation, part 4 : a deep dive on deep nets. *IEEE Intelligent Systems 33*(3), 87–95.

Hoffman, R. R. et G. Klein (2017). Explaining explanation, part 1 : theoretical foundations. *IEEE Intelligent Systems 32*(3), 68–73.

Hoffman, R. R., S. T. Mueller, et G. Klein (2017). Explaining explanation, part 2 : Empirical foundations. *IEEE Intelligent Systems 32*(4), 78–86.

Hoffman, R. R., S. T. Mueller, G. Klein, et J. Litman (2018). Metrics for explainable ai : Challenges and prospects. *arXiv preprint arXiv :1812.04608*.

Klein, G. (2018). Explaining explanation, part 3 : The causal landscape. *IEEE Intelligent Systems 33*(2), 83–88.

Lipton, Z. C. (2018). The mythos of model interpretability : In machine learning, the concept of interpretability is both important and slippery. *Queue 16*(3), 31–57.

Lundberg, S. et S.-I. Lee (2017). A unified approach to interpreting model predictions. *arXiv preprint arXiv :1705.07874*.

Ribeiro, M. T., S. Singh, et C. Guestrin (2016). " 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*, pp. 1135–1144.

Slack, D., S. A. Friedler, C. Scheidegger, et C. D. Roy (2019). Assessing the local interpretability of machine learning models. *arXiv preprint arXiv :1902.03501*.

Wu, M., M. Hughes, S. Parbhoo, M. Zazzi, V. Roth, et F. Doshi-Velez (2018). Beyond sparsity : Tree regularization of deep models for interpretability. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 32.

Summary

Explainable Artificial Intelligence has seen a surge in popularity in the past few years, thanks to new legislations that promote the "right to explanation". Many popular methods have been developed recently to help understand black-box models, but it is not clear yet how an explanation is defined. Furthermore, the community agrees to say that many important terms do not have commonly accepted definitions. In this paper, we review the literature and show that there is a major issue concerning the definitions of terms such as explainability or interpretability. To address this problem, we propose a terminology that takes into account the context of an AI system, i.e., its users, purposes or design. This terminology is compatible with the majority of the definitions encountered in the literature so that it can be a foundation for future works. We also discuss the metrics that evaluate some properties defined in this terminology.