Expliquer et justifier les systèmes de décisions algorithmiques

Clément Henin

N°d'ordre NNT : 2021LYSEI058

THESE de DOCTORAT DE L'UNIVERSITE DE LYON
opérée au sein de
INSA Lyon

Ecole Doctorale N° 512
École Doctorale Infomaths

Spécialité/ discipline de doctorat :

Informatique

Soutenue publiquement par :
Clément Henin

Expliquer et justifier les systèmes de décisions algorithmiques

---

Devant le jury composé de :

Solnon, Christine, Professeure, INSA Lyon, Présidente

Amer-Yahia, Sihem, Directrice de recherche, LIG, Rapporteure
Gambs, Sébastien, Professeur, UQAM, Rapporteur
Jacquelinet, Christian, Médecin Spécialiste, Agence de la Biomédecine, Examinateur
Mabi, Clément, Maître de conférence, UTC Compiègne, Examinateur

Le Métayer, Daniel, Directeur de recherche, Inria, Co-Directeur de thèse
Castelluccia, Claude, Directeur de recherche, Inria, Co-directeur de thèse

INSTITUT NATIONAL DE RECHERCHE EN SCIENCES ET TECHNOLOGIES DU NUMÉRIQUE

Résumé

Inria

Projet de thèse de doctorat

Expliquer et justifier les systèmes de décisions algorithmiques

par Clément HENIN

Dans un contexte favorable à la rationalisation des décisions par des objectifs mesurables et des méthodes quantitatives, le développement récent des technologies numériques a accéléré l'adoption des systèmes de décisions fondés sur un traitement algorithmique. De tels systèmes sont déjà présents dans de nombreux domaines et leur utilisation devrait encore s'accroître. Toutefois, si aucune mesure spécifique n'est prise, l'utilisation de tels algorithmes est porteuse de risques pour les citoyens et pour la société. Alors que plusieurs solutions, notamment juridiques et techniques, ont déjà été proposées par certains États et par la communauté de recherche *eXplainable Artififial Intelligence*, nous pensons que des efforts supplémentaires sont nécessaires pour répondre pleinement à tous les enjeux posés par ce type de systèmes.

Dans cette thèse, nous commençons par présenter les principaux travaux de recherche antérieurs qui composent le champ de recherche en *eXplainable Artififial Intelligence* en particulier les travaux visant à garantir l'intelligibilité des algorithmes de décision et notamment les méthodes d'explication en mode boîte noire, c'est-à-dire qui fonctionnent sans accéder au code de l'algorithme. Ces méthodes, en apparence diverses, partagent en fait une structure commune que nous identifions et que nous mettons à profit pour en établir une taxonomie. Par ailleurs, pour juger de l'intérêt des méthodes d'explication en général, il faut s'intéresser à leur évaluation. Cette question complexe reste un domaine de recherche actif dans lequel les études impliquant des utilisateurs humains ont une large part à jouer.

Ensuite, nous décrivons notre système interactif d'explication d'algorithme en boîte noire appelé *IBEX*. En s'appuyant sur les informations fournies par l'utilisateur, *IBEX* génère une explication adaptée au profil de la personne à ses besoins. L'utilisateur peut interagir avec le système d'explication à plusieurs niveaux en fonction de ses compétences et de son expertise. *IBEX* convient donc à la fois aux profanes (et dès la première utilisation), à des professionnels (qui pourraient être amenés à utiliser *IBEX* régulièrement) et à des experts en Intelligence Artificielle. Pour générer les explications, *IBEX* s'appuie sur un cadre générique commun pour les explications en boîte noire qui décompose notamment sur le processus d'explication en deux composantes distinctes. Cette décomposition et les combinaisons de choix qui en découlent permettent d'obtenir une variété d'explications qui peut être mise à

profit pour satisfaire au mieux la diversité des besoins. La démarche proposée dans IBEX a été testée dans le cadre d'une étude impliquant des utilisateurs (agents d'une autorité de régulation française) ayant des profils variés (juristes et ingénieurs).

Dans la suite, nous proposons une méthode originale permettant aux utilisateurs de systèmes algorithmiques d'en contester ou d'en justifier les décisions. Alors que la contestation occupe une place importante dans les textes de loi et que la question de leur justification par des normes extérieures est une préoccupation centrale des travaux de sciences sociales, il n'existe pas d'outils dédiés à ces objectifs spécifiques. En premier lieu, nous définissons les termes de justifications et de contestation. Une justification (respectivement une contestation) est un argument, soutenue par une preuve, qu'une décision est bonne (respectivement mauvaise) au sens d'un ensemble de normes extérieures prédéfinies. Justification et contestation sont vues ici comme des notions duales. Un dispositif technique, appelé *Algocate*, permet de rendre ces notions opérationnelles pour trois types de normes (règles, objectifs et référence) grâce à un système interactif permettant à l'utilisateur d'exprimer son point de vue sur le bien-fondé de la décision. Cette démarche est également testée auprès d'utilisateurs réels dans le cadre d'une étude impliquant des utilisateurs.

Les solutions proposées pour l'explication ou la justification de décisions algorithmiques doivent être confrontées au terrain. Dans le cadre de cette thèse, une collaboration de trois ans a été menée avec l'Agence de la Biomédecine. Elle a porté sur l'algorithme qui opère l'attribution des greffons cardiaque à l'échelle nationale, remplaçant un système dans lequel les médecins des centres de greffe avaient une grande autonomie. Après une phase d'analyse bibliographique et des entretiens menés sur le terrain dans des hôpitaux français, les principaux enjeux sociologiques et organisationnels autour de cet algorithme ont pu être identifiés. Suite à cela, un ensemble d'informations assurant l'explication et la justification du système a été mis à disposition des médecins des centres de greffe.

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations

ABM	Agence de la BioMédecine
AN	*Add random Noise*
BEM	Méthode d'Explication basée sur l'approche en Boîte noire
CEC	Circulation Extra-Corporelle
CF	ContreFactuelle
COMPAS	*Correctional Offender Management Profiling Alternative Sanctions*
DT	Arbre de décision
GAM	Modèles Additifs Généralisés
IA	Intelligence Artificielle
LA	*Approximation Linéaire locale*
ML	Apprentissage automatique
PC	Coefficients de Pearson
PD	Dépendance Partielle
PDP	*Partial Dependence Plot*
Pm	Permutation
RB	modèle à Base de Règles
RC	*Replace with Constant*
RGPD	Règlement Général sur la Protection des Données
SC	*Select Closest*
SDA	Système de Décisions Aglorithmique
SNAGC	Score National d'Attribution des Greffons Cardiaques
SVM	Séparateurs à Vaste Marge
XAI	*eXplainable Artificial Intelligence*

Chapitre 1

Expliquer et justifier les décisions algorithmiques

Dans un contexte favorable à la rationalisation des décisions par des objectifs mesurables et des méthodes quantitatives, le développement récent des technologies numériques a accéléré l'adoption des systèmes de décisions fondés sur un traitement algorithmique. De tels systèmes sont déjà présents dans de nombreux domaines et leur utilisation devrait encore s'accroître. Toutefois, si aucune mesure spécifique n'est prise, l'utilisation de tels algorithmes est porteuse de risques pour les citoyens et pour la société (Partie 1.1). Alors que plusieurs solutions, notamment juridiques et techniques, ont déjà été proposées par certains États et par la communauté de recherche *eXplainable Artififial Intelligence* (XAI), nous pensons que des efforts supplémentaires sont nécessaires pour répondre pleinement à tous les enjeux posés par ce type de systèmes (Partie 1.2). Nos contributions portent sur les aspects conceptuels (explications interactives, justifications et contestations), la mise en œuvre de ces concepts et des études de terrain visant notamment à assurer la pertinence globale de nos approches (Partie 1.3).

1.1 Contexte

La grande confiance que portent généralement les sociétés occidentales aux données et aux méthodes quantitatives, notamment pour la prise de décision, est un phénomène analysé de longue date par les sociologues (Partie 1.1.1). Les développements technologiques récents en traitement automatique des données, en particulier l'émergence du champ de l'Intelligence Artificielle (IA) (Partie 1.1.2), ont accéléré l'adoption de systèmes de décisions basés sur des algorithmes. De tels systèmes sont aujourd'hui présents dans de nombreux domaines et s'accompagnent de risques pour la société (Partie 1.1.3). Nous proposons dans cette partie d'illustrer notre propos à travers trois exemples de systèmes de décision présentés sous la forme d'encarts.

1.1.1 L'idéal de l'objectivité

Depuis longtemps, des machines nous remplacen et nous aident dans nos travaux manuels, aujourd'hui d'autres machines nous remplacent dans nos tâches intellectuelles. Cette évolution s'inscrit dans l'idéal de la "matérialité logique" dans lequel tout problème logique peut être rendu matériel et donc traité automatiquement par une machine physique adaptée. Cet idéal est loin d'être nouveau et avait conduit Blaise Pascal dès le XVII^{e} siècle à concevoir une machine à calculer capable de réaliser des opérations mathématiques simples. Aujourd'hui, ce ne sont plus seulement des calculs mathématiques qui sont confiés aux machines, mais des décisions complexes aux conséquences parfois considérables. Comment expliquer ce choix ? Qu'est-ce qui pousse nos sociétés à confier les décisions nous concernant à des automates ? Avant de décrire les dispositifs techniques qui rendent cela possible, il est utile de replacer l'avènement des algorithmes de décisions dans des tendances plus anciennes.

Puisqu'elles sont censées représenter le monde de manière transparente et objective, les données jouissent d'une grande légitimité dans le débat public et dans les processus de prise de décision. Les décisions fondées sur des données sont généralement réputées préférables aux décisions fondées sur le seul jugement humain. D'ailleurs, tout au long du XX^{e} siècle, la société a exigé de la part de plusieurs corps de professionnels de justifier leurs pratiques par

des mesures chiffrées [94]. Un exemple emblématique de ce phénomène est celui du corps médical qui a été forcé d'adopter (progressivement et notamment à la suite de différents scandales sanitaires au milieu du XX^{e} siècle) une évaluation systématique des traitements proposés aux patients par des mesures statistiques répondant à des exigences de méthodologie strictes. Dans cet exemple [1], c'est la nécessité de rendre des comptes (*accountability*) qui est le moteur du remplacement d'un savoir issu de l'expérience des professionnels par un savoir plus scientifique fondé sur des données brutes ("objectives") qui est censé conduire à plus de transparence. Depuis plus d'un siècle, ce rapport de force tend nettement au bénéfice de ce second type de savoir, surtout lorsque des décisions importantes sont en jeu.

Cette tendance a été largement documentée et commentée par de nombreux auteurs issus de différents champs des sciences sociales. Elle est par exemple abordée par Alain Desrosières qui analyse les mécanismes sociaux à l'œuvre dans l'établissement d'un argument fondé sur une mesure statistique (argument qui pourra participer à une décision) [33]. Elle est abordée également par Alain Supiot qui décrit comment des règles basées sur des objectifs mesurables se sont progressivement imposées dans les textes législatifs avec des conséquences souvent délétères [107]. Dans le champ du numérique, Dominique Cardon s'intéresse à l'utilisation accrue des algorithmes pour l'accès à l'information en ligne et en étudie les conséquences sociales [21]. Ces dispositifs techniques utilisant des statistiques, des règles basées sur des objectifs mesurables ou des méthodes algorithmiques pour rationaliser les décisions sont souvent critiqués par les sciences sociales : sous couvert d'une prétendue objectivité apportée par l'utilisation de données et d'analyses mathématiques parfois complexes, les raisons qui motivent et justifient une décision se retrouvent masquées et écartées de l'espace public du débat. En effet, bien qu'ils soient présentés comme objectifs, ces dispositifs sont en réalité conçus par des personnes porteuses de représentations auxquelles les systèmes de décisions résultants ne peuvent pas échapper.

La notion de "gouvernementalité", développé à l'origine par Michel Foucault, permet d'analyser les rapports de force inégaux observés dans de nombreux dispositifs numériques

1. Dans son livre *Trust in Numbers*, Theodore Porter décrit des situations similaires pour les comptables ou les ingénieurs des ponts et chaussées français [94].

[15]. Lorsque la décision est elle même algorithmique, cette gouvernementalité semble revêtir un caractère encore plus autoritaire car l'algorithme ne se contente plus de guider ou d'inciter l'utilisateur, puisqu'on lui confie directement la décision. Pour décrire cette évolution, la juriste Antoinette Rouvroy introduit la "gouvernementalité algorithmique" opérée par des systèmes "a-normatifs", c'est-à-dire des systèmes qui ne seraient justifiés par aucune norme [101]. Cela est bien sûr problématique s'agissant de systèmes prenant des décisions pouvant affecter des individus, voire la société dans son ensemble. En effet, lorsqu'une décision est prise par un professionnel, ce dernier doit la justifier par une norme extérieure, comme un texte de loi pour un juge ou le résultat d'une étude scientifique pour un médecin.

Les raisons de l'émergence des algorithmes ne sont pas seulement sociologiques. Les avancées technologiques en informatique et le développement des moyens de communication ont aussi contribué au développement des systèmes de décisions algorithmiques. La digitalisation croissante de nombreux aspects de nos sociétés a conduit à une explosion des volumes de données exploitables. Dans certaines situations, comme par exemple les systèmes de recommandations en ligne, le volume et la fréquence des données en jeu ne permettent pas d'envisager une décision humaine. Par ailleurs, alors que les décisions automatiques étaient jusqu'à récemment limitées à des traitements simples, l'utilisation d'algorithmes de plus en plus complexes a permis de diversifier et de généraliser leurs usages.

Exemple 1 : les recommandations de vidéos Youtube

Les systèmes de recommandation sont omniprésents sur internet. Ils concernent aussi bien des produits en vente sur des sites marchands (Amazon, Booking), les films et vidéos (Netflix, Youtube) que les articles de presse (Google News) ou les posts sur les réseaux sociaux (Facebook, Twitter). Même si chaque décision individuelle peut paraître anecdotique, ces systèmes peuvent avoir un impact considérable sur des secteurs économiques ou sur l'accès à l'information et donc sur le fonctionnement des démocraties. L'algorithme de recommandation de videos de Youtube est particulièrement intéressant car (1) il est massivement utilisé (Youtube étant le site le plus visité après le

moteur de recherche Google), (2) il a un impact avéré sur les choix des utilisateurs, et (3) il existe une controverse sur l'existence d'effets néfastes du système.

L'objectif de cet algorithme est de proposer en priorité les vidéos dont l'entreprise prévoit un temps de visionnage élevé afin de maximiser le temps passé sur la plateforme. Pour produire ces recommandations, Youtube s'appuie sur le contenu des vidéos et sur l'activité des utilisateurs (nombre de vues, de mentions "j'aime", d'abonnements aux chaînes, etc.). Le fonctionnement exact du système n'est pas dévoilé, mais l'entreprise précise que l'utilisation de réseaux de neurones profonds pour prédire le temps de visionnage a permis une amélioration "spectaculaire" de ses performances [29]. L'impact du système est avéré car les recommandations apparaissent comme l'une des principales sources de vues, comptant pour environ 30 % des vues c'est-à-dire quasiment autant que la première source (barre de recherche) [117].

Au premier abord, la légitimité de l'algorithme peut difficilement être remise en question. Youtube est un acteur privé fournissant un service à des utilisateurs qui consentent à l'utilisation de ce service dans les conditions spécifiées par Youtube. Pourtant, motivés par l'importance de la plateforme qui constitue une source d'information pour 28 % des américains [103], des journalistes s'interrogent sur le fonctionnement du système de recommandation. D'après leurs analyses, l'algorithme pousse les utilisateurs dans des "chambres d'écho " (ou bulles d'information) les confortant dans leurs avis et suggérant même des contenus similaires, mais plus controversés voire clivants [28, 100]. Certains accusent notamment ce phénomène des chambres d'écho d'avoir favorisé les contenus du candidat républicain Donald Trump au détriment de sa concurrente démocrate Hillary Clinton lors de l'élection présidentielle de 2016 aux États-Unis [76]. De son côté, l'entreprise critique les méthodes et les résultats de ces enquêtes en mettant en avant le fait que l'algorithme répond à une logique purement marchande.

1.1.2 Le développement de l'apprentissage automatique

Un algorithme d'IA est un algorithme ayant pour projet de simuler l'intelligence humaine. Plus précisément, on considère qu'un algorithme emploie une forme d'IA dès lors

que son comportement n'est pas explicitement implémenté par un développeur humain, mais que l'algorithme est capable d'une forme d'apprentissage à partir de données, il tombe dans ce cas dans le domaine de l'apprentissage automatique (ML [2]), ou par un système d'expérimentations, il appartient alors au domaine de l'apprentissage par renforcement. Dans le cas du ML, qui nous intéresse particulièrement dans cette thèse, l'algorithme est conçu pour reproduire une structure ou des motifs présents dans un jeu de données, appelées données d'entraînement. Au cours d'une phase d'entraînement, l'algorithme adapte un ensemble de paramètres, appelés *coefficients du modèle*, pour satisfaire au mieux un objectif défini par le développeur. Le plus souvent, il s'agit de prédire aussi fidèlement que possible un des attributs du jeu de données [3].

Le développement de l'IA a eu deux moteurs principaux. Le premier est la croissance exponentielle de la puissance de calcul des ordinateurs. La loi empirique de Gordon Moore postule que la densité de transistors sur les processeurs (et donc la puissance de calcul) double tous les deux ans [4]. Ainsi, des algorithmes effectuant des calculs de plus en plus complexes, et nécessitant notamment des phases d'apprentissage parfois très gourmandes, ont pu être employés. Le deuxième moteur de ce développement est l'explosion des volumes de données disponibles due notamment à la numérisation de nos sociétés. En effet, l'usage, aujourd'hui quasi-systématique, d'outils numériques dans les services administratifs, les communications, le commerce et l'industrie ont conduit à une multiplication des données accessibles ainsi qu'à une diversification des domaines d'utilisation de ces données.

Exemple 2 : le score COMPAS

Le score *Correctional Offender Management Profiling Alternative Sanctions* (COMPAS) est un algorithme estimant les risques de non-présentation à l'audience, de récidive ou de comportement violent pour les suspects ou les condamnés par la justice. Basé sur le

2. Sigle anglais de l'expression *Machine Learning*.
3. Il faut préciser qu'un autre type de ML existe et se distingue du ML *supervisé* décrit ici. Dans le cas du ML non supervisé, l'algorithme est conçu pour identifier des structures dans le jeu de données sans répondre à un objectif particulier. Ce type de ML est moins pertinent dans le contexte des décisions automatiques.
4. La possibilité d'utiliser les cartes graphiques dans le cas spécifiques de réseaux de neurones a encore accéléré ce phénomène.

ML, il estime les risques d'un profil à partir de 137 variables et des événements observés sur des profils similaires dans le passé. Présenté comme un système d'aide à la décision, son objectif est d'assister le magistrat par un recours à des données "objectives" sur des événements passés. Mis en place dans quelques états des États-Unis, il aurait déjà été utilisé plus d'un million de fois depuis 2000 [32].

Ce logiciel d'estimation des risques a été développé par l'entreprise Américaine Northpointe qui n'a pas rendu public les détails de fonctionnement de l'algorithme. Malgré cela, l'algorithme (vu donc comme une boîte noire) a été largement étudié et discuté sous de nombreux aspects. Sur le plan juridique d'abord, sa légitimité a été validée à deux reprises par la Cour suprême des États-Unis en 2015 et 2017. Par ailleurs, en 2016, une étude de l'association américaine Propublica compare les scores COMPAS proposés aux magistrats avec les événements réellement observés sur une période de 2 ans dans le comté de Broward. L'association montre une répartition inéquitable des cas de faux positifs et de faux négatifs entre les blancs et les non-blancs. Plus précisément, les personnes noires n'ayant pas récidivé ont été plus souvent classées à tort comme à haut risque (45 % contre 23 % pour les blancs) et les personnes blanches ayant récidivé ont été plus souvent classées à tort comme à faible risque (48 % contre 28 % pour les noirs) [11]. Cette enquête donne lieu à une controverse puisque la société Northpointe se défend de tout caractère discriminatoire en critiquant les résultats de Propublica et en pointant d'autres critères.

Enfin, la sociologue Angèle Christin, qui s'intéresse généralement à l'intégration d'algorithmes dans les organisations, a réalisé une étude sur leur utilisation dans les cours pénales [25]. Certaines de ses analyses conduisent à une critique plus profonde de l'utilisation des scores de risque. Parmi les fonctions sociales traditionnellement attribuées aux peines de prison (réparer un tort, prévenir d'autres crimes, permettre la réinsertion sociale et dissuader) le score de risque COMPAS, et les autres scores de manière générale, se focalisent principalement sur l'empêchement de nouveaux crimes (*incapacitation*). Cette situation est emblématique : sous couvert d'une prétendue objectivité, l'algorithme véhicule une vision politique de la décision du juge qui n'est pas

forcément partagée par tous les acteurs.

Aujourd'hui, ces algorithmes sont loin d'approcher l'intelligence humaine générale. Les algorithmes sont très spécialisés, nécessitent un grand nombre d'exemples pour apprendre à reproduire une structure (que les humains identifieraient très rapidement) et sont surtout incapables de curiosité, d'humour, d'empathie, etc. Toutefois, pour certains problèmes spécifiques, l'IA dépasse les meilleures intelligences humaines. Un exemple est la victoire en 2017 de l'algorithme AlphaGo contre le champion du monde du jeu de Go. Cet exemple a une valeur symbolique car le nombre de parties possibles de jeu de Go est tel que des capacités de simulation, même colossales, sont insuffisantes pour gagner contre un bon joueur... Il faut pour cela employer une forme d'intelligence.

Le développement de cette "intelligence" a été rendue possible par des progrès scientifiques en algorithmie. Les études historiques nous enseignent que le développement de l'IA est loin d'avoir suivi un cheminement linéaire. Au contraire, l'état du domaine est le résultat d'une succession d'hivers et de printemps et surtout le fruit d'une opposition entre différentes communautés [22]. Au fil des années, la communauté des "connexionistes" qui font le pari d'algorithmes inspirés du fonctionnement du cerveau humain semble s'imposer sur les autres. C'est l'écrasante victoire en 2012 du réseau de neurones profonds *AlexNet* dans la compétition de reconnaissance d'images *ImageNet Challenge* qui signe le début du regain de popularité des réseaux de neurones. Les prouesses se sont depuis étendues à beaucoup d'autres domaines : génération d'images, traitement automatique du langage naturel, reconnaissance vocale, simulation de systèmes chimiques et physiques complexes, etc. Les résultats obtenus par les réseaux de neurones ont souvent permis un gain significatif de performances par rapport aux autres types d'approches. Une contrepartie notable de ces performances est leur opacité de fonctionnement. En effet, les informations apprises par ce type d'algorithme prennent la forme de connexions entre neurones (souvent très nombreux) dont la lecture par une personne est tout sauf éclairante, en comparaison par exemple à la lecture d'un arbre de décision qui peut fournir une certaine intuition sur le fonctionnement de l'algorithme (à condition que sa taille soit raisonnable).

Les algorithmes reposant sur une forme d'apprentissage sont de plus en plus utilisés dans des processus de prise de décisions. En employant une technique d'apprentissage supervisé sur un jeu de données historiques contenant une variable déterminante pour la décision, l'algorithme est capable de fournir une prédiction pour cette variable. C'est la prédiction de cette variable, dont le risque d'une récidive (Exemple 2) ou le temps de visionnage d'une vidéo (Exemple 1) sont des exemples, qui est censée apporter un avantage notable au processus de décision.

1.1.3 Les systèmes de décisions algorithmiques

Un Système de Décisions Algorithmique (SDA) est un algorithme utilisé dans un processus de décision. Certains de ces systèmes prennent des décisions de manière autonome, par exemple celle de réduire ou d'augmenter la vitesse d'une véhicule, tandis que d'autres font simplement des suggestions qu'un utilisateur humain est libre de suivre ou de rejeter. Certains SDA ont un impact majeur sur la vie des individus concernés (Exemple 2 et Exemple 3), alors que d'autres prennent des décisions qui semblent insignifiantes pour chaque individu, mais peuvent en réalité avoir un impact considérable sur la société dans son ensemble (Exemple 1). Il nous paraît important de focaliser notre attention sur une finalité (prendre une décision) plutôt que sur une technique particulière (IA ou algorithme traditionnel). En effet, bien que l'IA a conduit à des changements majeurs dans ce domaine, nous avons aussi montré que les algorithmes de décision s'inscrivent également dans une histoire plus ancienne qu'il ne faut pas négliger.

L'utilisation des SDA est loin d'être anecdotique et de nombreuses entreprises et administrations en exploitent déjà [32]. Ils sont employés dans les domaines du transport (véhicules autonomes), de l'éducation (appariement élève/école), de la sécurité (surveillance vidéo), des ressources humaines (filtre de recrutement), de la médecine (diagnostics), etc. Dans ces différents secteurs, les SDA pourraient permettre d'améliorer globalement les décisions, en les rendant notamment plus rapides, reproductibles et économes, ce qui participerait à améliorer la qualité des services rendus. Finalement, ces systèmes pourraient se montrer plus

efficaces, transparents et responsables que les décideurs humains, à condition que les algorithmes eux-mêmes le soient.

En effet, les SDA s'accompagnent également de nombreux risques pour les individus (atteinte à la vie privée, à l'autonomie, risque de discrimination, etc.), pour les administrations (points de vulnérabilité, détournement par un État mal-intentionné, etc.) et pour la société (manipulation d'opinions, impact sur l'emploi, sur la justice, etc.) [40]. Ces risques peuvent être liés aux fonctionnalités intrinsèques du système, à des conséquences indirectes ou à des imperfections techniques (imprécisions, manque de robustesse, biais, etc.).

L'opacité des algorithmes est un facteur essentiel de ces risques. En effet, le décideur humain devrait pouvoir prendre une décision éclairée par le SDA, il devrait être "convaincu" par ce dernier que la décision est correcte et non pas se la voir imposer sans en comprendre les fondements. Si la décision est automatique, elle devrait pouvoir être facilement contrôlée par un responsable ou par une entité extérieure et *a minima* comprise par la personne concernée par la décision. De nombreux rapports, comme le rapport Villani [113] ou le rapport de la CNIL sur l'IA [27], mentionnent l'opacité comme l'un des problèmes centraux de ces systèmes.

L'opacité peut en fait avoir plusieurs sources. Elle peut être intentionnelle, dans le cas d'une entreprise souhaitant protéger sa propriété intellectuelle par exemple, ou la conséquence de certaines caractéristiques techniques des algorithmes (comme cela a été mentionné dans la Partie 1.1.2). Dans tous les cas, les utilisateurs d'un système opaque ne sont pas en mesure de comprendre le fonctionnement du système employé pour prendre les décisions. En plus des risques déjà mentionnés, l'opacité peut être un frein à l'adoption des SDA (alors même qu'ils pourraient avoir un effet positif), notamment pour les décisions critiques [112, 39]. Pour toutes ces raisons, des mesures doivent être prises pour limiter l'opacité et ses conséquences négatives.

Exemple 3 : attribution des greffons cardiaques

Les candidats à la greffe de cœur souffrent d'une insuffisance cardiaque mettant en jeu leur pronostic vital et ont épuisé toutes les alternatives thérapeutiques. L'opération

chirurgicale de transplantation pose en général peu de problèmes et la difficulté réside surtout dans l'obtention d'un greffon, car cela implique qu'une personne compatible et consentante se trouve en état de mort encéphalique. Même si les centres de greffe (côté receveur) ont globalement intérêt à collaborer afin d'obtenir de meilleurs appariements, le milieu reste compétitif puisque les médecins, désireux de soulager les souffrances de leur patient, souhaitent s'approprier la ressource, quitte à employer des stratégies illicites pour les favoriser au détriment des autres patients en attente [109]. Dans ce contexte, le système assurant l'attribution de cette ressource précieuse joue un rôle central dans la prise en charge. Depuis 2018, c'est le "Score Cœur", un algorithme développé par l'Agence de la Biomédecine, qui établit l'ordre de proposition des greffons, succédant à un système dans lequel les médecins transplanteurs avaient une grande autonomie dans le choix des patients à greffer. Pour calculer le score qui établit la priorité d'accès à un greffon, l'algorithme applique un ensemble de règles et de formules aux informations concernant le donneur et les candidats à la greffe. Ce traitement informatique combine des modèles prédictifs issus des données historiques de la greffe et des règles éditées "à la main" pour obtenir un compromis entre différents critères éthiques.

L'Agence de la Biomédecine dispose d'une légitimité administrative pour fixer les règles de l'attribution des greffons à l'échelle nationale. Même si les médecins critiquent le fonctionnement du score qui est jugé imprécis et inefficace, ce dernier est toutefois accepté car il permet d'éviter une situation de compétition entre centres de greffe dont l'impact sur les patients est jugé globalement délétère par l'Agence de la Biomédecine [54]. À la différence des exemples précédents, le Score Cœur a été développé par une administration. Une documentation publique fournit de nombreux détails sur le fonctionnement du système.

Pour justifier les décisions du système, quatre critères éthiques principaux sont avancés : l'évitement du décès sur la liste d'attente, l'équité d'accès à la greffe, l'évitement du décès après la greffe et des critères logistique de minimisation des transports de greffons (qui ont par ailleurs un impact sur les résultats de la greffe). Nous décrivons plus précisément cet algorithme dans le Chapitre 5 qui en traite spécifiquement.

1.2 Motivations

Plusieurs réponses d'ordre éthique, juridique et technique ont été proposées pour tenter de pallier les risques associés à l'émergence des SDA (Partie 1.2.1). Bien que ces réponses aillent dans le bon sens, nous les jugeons insuffisantes pour répondre à tous les enjeux (Partie 1.2.2).

1.2.1 Réponses existantes au problème de l'opacité

La mise en place des SDA et les problèmes liés à leur opacité ont suscité des réponses d'ordre éthique, juridique et technique. En premier lieu, de nombreux guides, livres blancs ou rapports formulent des recommandations pour ce type de systèmes. Ils sont rédigés par des associations, des organisations non gouvernementales, les pouvoirs publics ou des entreprises. En posant clairement les principaux enjeux éthiques, ces documents sont utiles pour alerter l'opinion publique et les décideurs. Parmi les thèmes abordés, les questions d'*accountability*, de discrimination, de transparence, du contrôle humain et de l'intelligibilité apparaissent comme les enjeux fondamentaux d'un développement éthique de ces systèmes [51].

Certaines de ces recommandations ont déjà été traduites en des termes juridiques. En Europe, l'article 22 du Règlement Général sur la Protection des Données (RGPD) fixe un cadre pour toute personne sujette à une décision "fondée exclusivement sur un traitement automatisé, y compris le profilage, produisant des effets juridiques la concernant ou l'affectant de manière significative" en établissant un "droit de la personne concernée d'obtenir une intervention humaine de la part du responsable du traitement, d'exprimer son point de vue et de contester la décision." Le considérant 71 préconise directement un droit "d'obtenir une explication quant à la décision" [5]. En France, la loi pour une République Numérique oblige les administrations qui prennent des décisions "sur le fondement d'un traitement algorithmique" à fournir "sous une forme intelligible" des informations sur le traitement effectué dont "les paramètres de traitement et, le cas échéant, leur pondération, appliqués à la situation de l'intéressé". Enfin, au Canada, l'article 6.2.3 de la Directive du 1er avril 2019 sur la

5. Ce droit est toutefois sujet à débat entre les spécialistes [114, 83].

prise de décision automatisée par les administrations prévoit une "obligation de fournir une explication significative aux personnes concernées sur la façon dont la décision a été prise et la raison pour laquelle elle a été prise". Il est intéressant de noter que cet article distingue la "façon" et la "raison" de la décision, ce qui fait écho à la distinction introduite dans cette thèse entre une explication et une justification. La Directive canadienne définit par ailleurs plusieurs niveaux d'explication en fonction de l'impact attendu de la décision.

Enfin, différentes réponses ont été apportées sur le plan technique. Pour répondre à la question des discriminations par exemple, la communauté scientifique a mené de nombreuses études sur l'équité des algorithmes proposant des méthodes pour quantifier les discriminations algorithmiques et les limiter. Pour répondre aux problèmes posés par l'opacité, qui nous intéressent particulièrement dans cette thèse, différentes recherches ont été menées afin de rendre intelligibles les traitements effectués par les algorithmes. Ces recherches se regroupent sous le sigle XAI pour *eXplainable Artificial Intelligence*. Ce domaine de recherche, à vocation pluri-disciplinaire, propose des méthodes spécifiques pour concevoir des algorithmes faciles à interpréter ou des méthodes pour expliquer les algorithmes qui ne le sont pas (Chapitre 2). Le champ de recherche est devenu très actif ces dernières années et certaines contributions ont permis des avancées significatives aussi bien sur le plan des méthodes [98], de la définition des objectifs [84] que de la prise en compte des impératifs juridiques [115].

Globalement, la démarche XAI consiste à fournir des informations pour comprendre le fonctionnement du SDA. Ces informations, ou explications, prennent la forme de graphiques, d'images ou de textes et indiquent par exemple les variables qui affectent le plus significativement la décision ou représentent sous une forme intelligible les traitements effectués par l'algorithme. Pour les calculer, certaines méthodes fonctionnent en "boîte blanche", c'est-à-dire en analysant le code du système ou les coefficients du modèle et d'autres en "boîte noire" en observant seulement les entrées et sorties de l'algorithme (sans accès au code du SDA).

Bien que ces réponses, éthiques, juridiques et techniques ont permis des avancées notables, elles restent globalement insuffisantes pour répondre à tous les défis posés par l'adoption massive des SDA.

1.2.2 Limites des réponses existantes

Au vu des enjeux et des risques associés à l'utilisation croissante des SDA, nous pensons que les réponses apportées sont actuellement insuffisantes. D'abord, malgré leur intérêt certain, les guides et rapports traitant de l'éthique de l'IA et des systèmes algorithmiques se contentent, pour la plupart, de formuler des recommandations très générales employant parfois des termes mal définis. En particulier, ils présentent rarement des solutions concrètes que les responsables de SDA ou les entités chargées de les contrôler pourraient mettre en œuvre. Par ailleurs, les chartes éthiques et autres codes de bonne conduite ne sont pas juridiquement contraignant. Il est donc légitime de se poser des questions sur leur effectivité.

Sur le plan juridique, nous considérons que les textes actuels vont dans la bonne direction, mais restent insuffisants à la fois dans leurs périmètres et dans leurs exigences. Par exemple, pour que les protections du RGPD s'appliquent, la décision doit être prise de manière autonome et produire des effets juridiques significatifs sur la personne concernée. Ce périmètre restreint exclut les décisions qui n'auraient pas d'impact significatif sur un individu particulier, mais en auraient sur la société dans son ensemble, comme certains systèmes de recommandation (Exemple 1). Aussi, le fait que la décision doive être fondée exclusivement sur le traitement algorithmique limite fortement cette obligation. De même, la loi pour une République Numérique et la directive Canadienne ne concernent que les administrations ce qui réduit considérablement leur impact. Par ailleurs, les exigences imposées en terme d'explication dans ces différents textes sont souvent floues [6] et semblent laisser la porte ouverte à des explications n'apportant pas de bénéfice majeur à leur destinataire. Un texte plus abouti devrait notamment préciser que les explications doivent être adaptées aux destinataires et contenir des informations appropriées à leurs objectifs, pour autant que ces derniers soient légitimes.

Enfin, sur le plan technique, les propositions actuelles du champ de recherche en XAI nous paraissent insuffisantes pour répondre à l'ensemble des enjeux. D'abord,rejoignant

6. Les décrets d'application de la loi pour une République Numérique sont toutefois à citer comme un bon exemple en la matière. Il prévoit notamment l'obligation de fournir : les paramètres de traitement et, le cas échéant, leur pondération, appliqués à la situation de l'intéressé ainsi que les opérations effectuées par le traitement.

l'alerte lancée par Tim Miller et ses co-auteurs [85], nous déplorons que les méthodes d'explications soient trop souvent conçues par des spécialistes en IA en fonction de leurs propres besoins. Nous argumentons que les explications devraient aussi être destinées aux profanes en IA, comme les personnes concernées par les décisions, certains professionnels du domaine (médecin ou juge) ainsi qu'à des personnes impliquées dans des activités d'audit d'algorithme, appartenant par exemple à une autorité de régulation, pouvant avoir ou non des connaissances technique en développement de SDA. Ces différentes catégories de personnes ont des besoins variés en explication. Ces besoins devraient être pris en compte pour formuler une explication adaptée à l'utilisateur, en terme notamment de complexité, de forme. En effet, la situation d'un médecin cherchant à s'assurer qu'il peut se fier au diagnostic proposé par un algorithme et la situation d'une personne subissant un refus et cherchant à obtenir une meilleure décision sont différentes et nécessitent des explications aux propriétés différentes. Pourtant, les méthodes d'explication de la littérature proposent généralement un seul type d'explication dont les objectifs et les destinataires attendus sont rarement précisés.

Par ailleurs, l'objectif de contestation occupe une place importante dans les textes de loi (ce droit apparaît par exemple dans le RGPD), mais les explications ne semblent pas parfaitement adaptées à cet objectif. Pour aborder la question de la contestation d'une décision, il faut employer une dialectique de justification. En effet, une décision est contestable dès lors qu'elle n'est pas justifiée et inversement une décision justifiée ne peut pas être contestée. Bien que la notion de justification apparaisse dans la littérature, elle est souvent confondue avec l'explication. Nous pensons pourtant que ces deux notions sont distinctes et mériteraient d'être clarifiées. Alors que le but d'une explication est de faire comprendre la décision, le but d'une justification est de convaincre que la décision est bonne (ou au contraire qu'elle est mauvaise dans le cas d'une contestation). Ces deux objets emploient des arguments de natures très différentes puisque l'explication se contente de décrire le fonctionnement de l'algorithme alors que la justification doit faire référence à une élément extérieur qui permet de porter un jugement sur le bien-fondé de la décision.

Enfin, alors que les études de terrain impliquant des utilisateurs devraient jouer un rôle central dans la conception des explications, il apparaît que ce type d'études est encore trop

rare [85]. Cela fait courir le risque de démarches hors-sol, décorrélées des besoins réels et ne répondant pas aux exigences nécessaires pour remplir les objectifs affichés. Les utilisateurs devraient être impliqués dans les recherches en XAI à deux niveaux : (a) des SDA réels devraient être étudiés afin de mieux comprendre les enjeux et d'en déduire les objectifs et contraintes spécifiques auxquels doivent répondre les explications, (b) les méthodes d'explications devraient être systématiquement testées par des utilisateurs humains sur des cas d'usage réels, ou dans le cadre d'expérimentations fictives.

Les limites des solutions existantes identifiées dans cette partie regroupent des critiques de différentes natures. D'une part, nous soutenons que les objectifs poursuivis par la communauté XAI sont en partie inadaptés aux enjeux réels tels qu'ils nous apparaissent au travers des textes de loi et des études de terrain. Une autre critique concerne les méthodes d'explication elles-mêmes souvent trop figées et incapables de s'adapter aux différents contextes dans lesquels elles sont susceptibles d'être employées. Enfin, la dernière critique concerne l'évaluation de ces méthodes qui n'implique pas suffisamment les utilisateurs. Les travaux présentés dans cette thèse tentent de répondre à ces critiques. Comme nous allons le voir, une telle ambition exige d'employer une démarche interdisciplinaire et notamment la prise en compte de la dimension sociologique du déploiement des SDA.

1.3 Démarche et contributions

Dans cette partie, nous décrivons la démarche générale adoptée pour pallier les limites identifiées dans les solutions existantes (Partie 1.3.1), puis nous présentons les différentes contributions apportées dans le cadre de cette thèse (Partie 1.3.2).

1.3.1 Démarche

Dans la suite de cette thèse, nous supposons que le code de l'algorithme n'est pas connu. Pour inférer des informations sur son fonctionnement, nous devons donc analyser les liens entre les entrées et les sorties du système. Bien que cette démarche en boîte noire limite les explications possibles, elle présente de nombreux avantages pratiques qui en justifient

l'adoption. En premier lieu, le mode boîte noire est une contrainte pour beaucoup de situations réelles dans lesquelles le code n'est pas disponible. Par exemple, le responsable du SDA peut refuser de le transmettre en vertu du droit à la propriété intellectuelle. De plus, en analysant directement les entrées et sorties du système, le mode boîte noire garantit que les explications proposées correspondent réellement au système utilisé, alors qu'une version altérée du code pourrait être transmise suite à une erreur ou intentionnellement afin de tromper le système d'explication. Enfin, cette démarche ne dépend pas du type d'algorithme considéré. Cela assure un large périmètre d'application ainsi qu'un gain considérable en efficacité puisque le développement d'une méthode suffit pour traiter un grand nombre de systèmes.

Dans notre démarche de recherche, le premier problème abordé est celui de l'interactivité des systèmes d'explications. Comme nous l'avons déjà mentionné, des contextes différents nécessitent des explications différentes. Prenons l'exemple d'un médecin cherchant à être convaincu par un outil d'aide au diagnostic et d'un auditeur d'une autorité de régulation vérifiant la légitimité d'un traitement. Alors que le premier cherche à connaître les caractéristiques prises en compte par l'algorithme dans le cas particulier de son patient pour s'assurer qu'elles sont pertinentes, le second s'intéressera plutôt au fonctionnement global du système, cherchant par exemple à savoir si le traitement est discriminant ou si l'une des données personnelles serait collectée sans être utilisée ce qui constituerait une violation du principe de minimisation du RGPD. Pour satisfaire cette variété de contextes, les utilisateurs doivent pouvoir interagir avec l'outil d'explication afin d'exprimer leurs besoins et d'obtenir l'explication adaptée. La conception d'un tel outil nécessite deux éléments :

— un système capable de générer une diversité d'explications,
— un protocole d'interaction permettant aux utilisateurs d'accéder à l'explication la plus adaptée à leur situation.

La réunion de ces deux éléments a donné lieu à l'outil IBEX (sigle de *Interactive Black-box EXplanations*) présenté dans le Chapitre 3.

L'intérêt ces outils d'explications est indéniable. En rendant intelligible le fonctionnement de l'algorithme, les explications permettent de détecter des comportements inattendus ou au

contraire de se rassurer sur la logique employée par le système. Cependant, lorsqu'il s'agit de contester une décision, les explications peuvent se montrer insuffisantes. Dans beaucoup de travaux en XAI, la notion de contestation (ou de justification) est confondue avec celle d'explication, nous soutenons ici l'idée qu'il s'agit de deux objectifs différents. En effet, alors que le but d'une explication est de faire comprendre la logique employée par l'algorithme, le but d'une contestation est de convaincre que la décision est mauvaise (le concept dual de la contestation étant la justification dont le but est de convaincre que la décision est bonne). L'explication est donc par nature intrinsèque, elle se limite à la description de la logique du SDA sous une forme intelligible. À l'inverse, la contestation (ou la justification) est extrinsèque puisqu'une norme extérieure doit être utilisée pour porter un jugement sur la décision. Par exemple, l'explication d'un refus de crédit pourrait se limiter à exposer le critère qui a été décisif pour l'algorithme (par exemple le fait que les mensualités dépasseraient le tiers des revenus), alors qu'une justification de ce refus devrait se référer à un principe supérieur pour convaincre que la décision est bonne (par exemple le fait que des mensualités trop élevées conduisent fréquemment à des défauts de remboursement, situations que la banque souhaite limiter). Pour permettre ces références à des normes, l'outil Algocate, présenté dans le Chapitre 4, a été conçu et implémenté. Sa pertinence a également été testée dans le cadre d'une étude impliquant des utilisateurs [59].

Les outils développés dans le cadre de cette thèse sont interactifs. Il nous paraît crucial que l'utilisateur soit maître des informations qu'il consulte et qu'il ne se retrouve pas dans la position d'un "destinataire passif de la sagesse du système" [91]. Une caractéristique importante d'IBEX et d'Algocate est qu'ils donnent à l'utilisateur une place centrale dans le choix des informations qu'il consulte.

Enfin, les travaux de cette thèse ont été accompagnés par une étude de terrain sur le système d'attribution des greffons cardiaques qui a permis de confronter les concepts à la réalité d'un algorithme de décision. Ce terrain s'est déroulé en trois phases : une phase d'analyse de l'algorithme (bibliographie et discussions avec les concepteurs), une phase d'entretiens avec les utilisateurs et une phase de développement d'un outil, largement inspiré d'IBEX et

d'*Algocate*, pour expliquer et justifier les décisions du système. Ce terrain a été particulièrement éclairant pour identifier certains enjeux généraux des SDA dont les manœuvres stratégiques employées par les acteurs. Pour y parvenir, des méthodologies propres aux sciences sociales ont été utilisées.

Nous espérons par cette thèse contribuer à combler un fossé entre plusieurs communautés. En effet, nous proposons d'introduire, dans le champ de l'informatique, des concepts inspirés des domaines de l'éthique. C'est dans cette communauté qu'apparaissent clairement les besoins de méthodes interactives ainsi que la particularité des justifications d'algorithmes et leur référence aux normes. Mais, à la différence d'un travail qui resterait purement théorique, cette thèse prend également le risque de se confronter à la réalité opérationnelle sur le "terrain" puisque les deux concepts introduits ont donné lieu à des outils qui ont été testés expérimentalement.

1.3.2 Contributions

Nos contributions s'organisent en 3 axes : les explications interactives, les justifications et l'étude de terrain sur le système d'attribution des greffons cardiaques.

Les explications interactives (Chapitre 3) :

Comme mentionné dans la Partie 1.2.2, la plupart des méthodes d'explication se focalisent sur un type d'explication, censé remplir une fonction spécifique déterminée par le concepteur de la méthode. Notre système d'explication interactif fonctionnant en boîte noire se présente sous la forme de plusieurs contributions. En premier lieu, un cadre générique pour décrire les méthodes d'explication en boîte noire a permis d'identifier de manière fine les composantes communes des méthodes d'explication : l'échantillonnage et la génération [60]. Les composantes des méthodes de la littérature sont décrites dans un formalisme unique qui facilite la comparaison des caractéristiques de chacune.

Au-delà des vertus comparatives, ce cadre générique peut être utilisé pour concevoir un système générique capable de fournir une grande diversité d'explications en exploitant différentes combinaisons des composantes d'échantillonnage et de génération [56]. Ce système

propose quatre stratégies d'échantillonnage et six méthodes de génération qui conduisent à un total de douze combinaisons possibles, chacune conduisant à une méthode d'explication ayant ses caractéristiques propres. En plus de ce choix, chaque composante peut être paramétrée par un ensemble d'options techniques qui peuvent être mises à profit pour modifier les propriétés de l'explication finale (par exemple sa simplicité ou sa généralité).

Pour que cette diversité d'explications puisse être utile à l'utilisateur final, nous avons conçu un protocole d'interaction lui permettant d'obtenir l'explication la plus appropriée [57]. Ce protocole d'interaction est constitué de trois niveaux allant du plus abstrait (adapté pour un utilisateur profane) au plus technique (adapté aux personnes ayant des connaissances avancées). Le niveau le plus abstrait est appelé le *contexte*. Il contient des informations de haut-niveau sur la demande d'explication (profil de la personne, objectif de l'explication et périmètre de l'explication). Le second niveau est celui des *exigences* ; elles définissent les propriétés attendues de l'explication (forme, simplicité, généralité, réalisme, actionnabilité et nature). Enfin, le dernier niveau fait directement référence aux options techniques du système générique d'explication. L'utilisateur peut interagir au niveau souhaité en fonction de ses connaissances et peut réagir à une explication, par exemple pour en obtenir une version plus complète. Notre système interactif d'explications a été évalué au cours d'une expérimentation impliquant des utilisateurs, agents d'une autorité française de régulation.

Les justifications d'algorithmes (Chapitre 4) :

Alors que la communauté de recherche en XAI se focalise sur la conception de méthodes pour expliquer le fonctionnement des algorithmes, de nombreux auteurs, issus notamment du domaine du droit, soulignent la nécessité de pouvoir contester les résultats d'un SDA. La contestatibilité, et son concept dual de justifiabilité, semblent en effet les conditions essentielles pour assurer l'*accountability* ("reddition de compte" ou "redevabilité" en français) et la légitimité des SDA. Pourtant, au sein de la communauté XAI, la distinction entre explication et justification est floue, les deux concepts se trouvant souvent confondus. Dans [58], nous proposons une clarification de la terminologie et argumentons de la nécessité de développer des outils spécifiques pour justifier et contester les SDA.

Dans [59], nous proposons une première version de notre outil de justification et de contestation. Il suppose que le SDA est fondé sur du ML et que son objectif est de reproduire les décisions d'un jeu de données historiques. L'outil considère que la décision est justifiée si elle ressemble aux décisions du jeu de données d'entraînement et qu'elle est contestable dans le cas contraire. L'approche est testée dans le cadre d'une étude impliquant des utilisateurs. Dans [55], nous élargissons l'approche à deux autres types de normes (règles et objectifs) et proposons un outil opérationnel nommé *Algocate*. Grâce au protocole d'interaction d'*Algocate*, l'utilisateur est invité à préciser ses raisons de penser que la décision est mauvaise (ou qu'elle est bonne) en mentionnant les caractéristiques des données d'entrée (son dossier par exemple) qui semblent aller dans ce sens. Ces raisons sont ensuite confrontées aux normes applicables pour ce SDA ce qui conduit à la génération de preuves allant soit dans le sens de la requête, soit dans le sens inverse. Notre système de justification et de contestation a été évalué au cours d'une expérimentation impliquant des utilisateurs, agents d'une autorité française de régulation.

Le système d'attribution des greffons cardiaques (Chapitre 5) :

Le système Score Cœur d'attribution des greffons cardiaques est un cas d'étude pertinent pour nos recherches. Il s'agit de décisions complexes, à fort impact et impliquant différents types d'utilisateurs pouvant avoir des intérêts divergents (concepteurs, médecins et patients). Pour comprendre finement les enjeux et l'impact de sa mise en place sur le secteur de la greffe cardiaque, nous avons mené une série d'entretiens semi-orientés (une méthodologie propre aux sciences sociales) dans quatre hôpitaux français. Alors que le Score Cœur les prive d'une partie de leur autonomie professionnelle et que l'attribution semble imparfaite, les médecins acceptent globalement ce système centralisé et automatique. En effet, il leur semble préférable à une attribution humaine qui était largement manipulée par les centres de greffe. Cette analyse a donné lieu à une publication dans une revue de sociologie spécialisée dans les technologies [54].

Dans une deuxième phase, nous avons développé un ensemble d'outils à destination des médecins et des concepteurs pour faciliter la compréhension du système. Ces outils, qui

prennent la forme d'une interface web, contiennent des pages de visualisation des données, des pages d'explications (du système dans son ensemble et de décisions individuelles) et des justifications des décisions. Après une analyse poussée des enjeux du système, ce cas d'étude met donc pleinement en œuvre les concepts proposés dans le cadre de cette thèse.

1.3.3 Liste des publications

Publications dans une revue internationale :

— *Beyond explainability : justifiability and contestability of Algorithmic Decision Systems*, C. Henin D. Le Métayer, AI & Society (Springer journal).

— *A framework to contest and justify algorithmic decisions*, C. Henin D. Le Métayer, AI and Ethics (Springer journal) : 1-14.

Publication dans une revue nationale (sociologie) :

— *Confier une décision vitale à une machine*, C. Henin, Réseaux 2021/1 : 187-213.

Publications dans un *workshop* (avec comité de lecture) :

— *A Multi-layered Approach for Tailored Black-box Explanations*, C. Henin D. Le Métayer, In Pattern Recognition. ICPR International Workshops and Challenges, 5-19. Cham : Springer International Publishing.

— *A Generic Framework for Black-box Explanations*, C. Henin D. Le Métayer, In 2020 IEEE International Conference on Big Data (Big Data) (pp. 3667-3676). IEEE.

— *Towards a Framework for Challenging ML-based Decisions*, C. Henin D. Le Métayer, 1st and 2nd International Workshops on Deceptive AI ECAI2020 and ECAI2021

— *Towards a Generic Framework for Black-box Explanation Methods*, C. Henin D. Le Métayer, IJCAI 2019 Workshop on Explainable Artificial Intelligence (XAI) (non archivé)

Chapitre 2

État de l'art

Dans ce chapitre, nous présentons les principaux travaux de recherche antérieurs à cette thèse qui composent le champ de recherche en XAI. La première partie (Partie 2.1) est un panorama des objectifs du champ de recherche et des enjeux autour de l'utilisation des SDA dans le contexte du développement de l'IA. Les Parties 2.1.2 et 2.1.3 se focalisent plus particulièrement sur les travaux visant à garantir l'intelligibilité des algorithmes. Dans la Partie 2.2, nous resserrons encore le périmètre sur les méthodes d'explication en boîte noire, c'est-à-dire qui fonctionnent sans accéder au code de l'algorithme. Ces méthodes, en apparence diverses (Partie 2.2.1), partagent en fait une structure commune que nous identifions (Partie 2.2.2) et mettons à profit pour en établir une taxonomie (Partie 2.2.3). Pour juger de l'intérêt de ces méthodes, il faut s'intéresser à leur évaluation (Partie 2.3). Cette question complexe reste un domaine de recherche actif (Partie 2.3.1) dans lequel les études impliquant des utilisateurs humains ont une large part à jouer (Partie 2.3.2). La dernière partie de ce chapitre est dédiée aux limites du champs de recherche (Partie 2.4). Certaines prennent la forme de pistes de recherche dont une partie est traitée pendant la thèse (Partie 2.4.1) et d'autres constituent des risques inhérents à l'approche XAI (Partie 2.4.2).

Les principales contributions de cette partie sont les suivantes :

— Une présentation des principaux enjeux de la recherche et XAI comprenant un ensemble de définitions et plusieurs revues thématiques de la littérature (méthodes en boîte blanche, méthodes en boîte noire et évaluation des explications).

— Un cadre générique pour décrire les méthodes d'explication en boîte noire et son application à 12 méthodes de la littérature. Ce cadre est notamment utile pour caractériser et comparer les méthodes existantes.

— Une identification de certaines limites du champ dont celles qui ont motivé les travaux de cette thèse.

2.1 Algorithmes éthiques, interprétables et explicables

Afin d'atténuer les risques accompagnant l'utilisation des SDA, de nombreuses recommandations sont formulées par les chercheurs, les pouvoirs publics et les entreprises comme l'*accountability* ou l'intelligibilité (Partie 2.1.1). Différents dispositifs techniques sont conçus pour tenter de les satisfaire. En se focalisant sur les enjeux d'intelligibilité, nous distinguons une branche de littérature visant à concevoir des algorithmes de décision interprétables (Partie 2.1.2) et une autre branche dont le but est d'obtenir des explications après la phase de conception du SDA (Partie 2.1.3).

2.1.1 Les enjeux éthiques de l'IA

Comme la plupart des transformations technologiques, le développement des SDA comporte des risques et des opportunités [40]. Dans des secteurs aussi divers que la santé, la justice, la sécurité, les transports ou l'énergie, l'utilisation d'algorithmes pour remplacer ou aider les humains promet des décisions plus fiables, équitables, efficaces, transparentes et responsables. Mais ces transformations peuvent aussi avoir des conséquences néfastes, qu'elles soient intentionnelles ou accidentelles, sur les individus et la société. Comme l'ont montré de nombreux auteurs, les algorithmes peuvent automatiser des biais discriminatoires en reproduisant les processus de décision humains [11, 19]. La généralisation de la collecte de données personnelles, souvent nécessaires pour concevoir et faire fonctionner les SDA, fait courir le risque d'une société de surveillance généralisée faisant peser avec force le poids des normes sociales sur les individus [24], sans parler du coût environemental colossal que représentent ces systèmes de stockage et de traitement. Aussi, le recours accru aux algorithmes

crée des points de vulnérabilité qui peuvent être exploités par des acteurs malveillants et qui rendent les conséquences d'éventuelles défaillances dramatiques (infrastructures nucléaires, manipulation de l'opinion publique [96], etc.).

Pour tenter de réduire ces risques, de nombreux auteurs (chercheurs, pouvoirs publics ou entreprises) formulent des *desiderata* pour un développement éthique de ces technologies [75, 32, 113, 40, 1]. Parmi les plus cités on retrouve les éléments suivants [1] :

— **Confidentialité** : Elle impose le respect de la vie privée dans la collecte et le traitement des données personnelles. Ce droit fondamental est notamment consacré dans le RGPD européen.

— **Sécurité et fiabilité** : Leur but général est d'assurer que le système remplit correctement sa fonction et ne peut pas être sujet à des défaillances ni victime d'une attaque ou d'un détournement de sa fonction d'origine [8, 96].

— ***Fairness*** : La *fairness* (ou équité) peut être définie comme l'absence d'utilisation injustifiée de certains attributs (genre, âge, religion, etc.) *i.e.* l'absence de biais discriminatoires [27]. Les discriminations algorithmiques ont reçu une attention particulière dans la littérature suite à la découverte de biais dans des systèmes existants [11, 19]. De nombreux outils, associés à différentes définitions de biais discriminatoires, sont aujourd'hui accessibles pour les identifier et les réduire [67, 93, 46]. Toutes les définitions n'étant pas compatibles, le choix de la mesure de *fairness* à adopter, qui revêt un caractère politique, est un préalable (non technique) à la mise en pratique de ces méthodes [17].

— ***Accountability*** : Reuben Binns [18] la définit comme l'obligation pour une partie de devoir rendre compte de ses décisions (c'est-à-dire de les justifier) à une autorité capable de lui infliger une sanction si la justification est jugée inadéquate. Alors que ce terme est l'un des plus employés, il est souvent mal défini dans les publications et prend un sens très général. Il s'agit avant tout d'un enjeu organisationnel et de

1. Un consensus clair sur la terminologie n'a pas encore été établi et d'autres auteurs font référence aux mêmes termes avec des définitions légèrement différentes. Dans la suite de cette thèse, nous utiliserons systématiquement les définitions proposées ci-dessous.

régulation, mais qui appelle également des développements technologiques [36].

— **Transparence** : Elle consiste à exposer (via la publication de code ou de documents) le fonctionnement des SDA. Cette mesure organisationnelle ou réglementaire, impliquant les personnes responsables du SDA permet de limiter l'opacité lorsque cette dernière est intentionnelle. La transparence est par exemple rendue obligatoire pour les administrations françaises dans la loi pour une république numérique. Bien que souhaitable dans de nombreuses situations, elle est parfois insuffisante pour répondre à tous les enjeux notamment lorsque le fonctionnement du système n'est pas intelligible [9].

— **Intelligibilité** : L'intelligibilité est la propriété d'un algorithme dont le fonctionnement est compréhensible [2, 48, 12]. Elle peut être globale si l'algorithme est compréhensible dans son ensemble, ou locale si les décisions individuelles sont compréhensibles. Elle peut s'adresser à différentes parties prenantes ayant ou non des connaissances techniques préalables. L'intelligibilité est une forme de "facilitateur" pour l'ensemble des autres *desiderata* [78]. Comme nous le verrons dans la suite, au moins deux stratégies peuvent être employées pour assurer l'intelligibilité : l'interprétabilité et l'explicabilité.

— **Contestabilité** : La contestabilité impose que les SDA ou leurs résultats puissent être contestés [91, 80]. Alors qu'elle est explicitement mentionnée dans certains textes de loi, dont le RGPD, peu d'outils ont pour objectif de l'assurer.

Les travaux de cette thèse se focalisent sur les deux derniers objectifs ainsi que sur la justification des SDA qui participe à l'*accountability*. En effet, la compréhension, la contestation et la justification des SDA constituent des caractéristiques essentielles pour les intéressés et sont par ailleurs des problématiques nouvelles dues à la complexification des traitements algorithmiques. La suite de cet état de l'art aborde en particulier les travaux sur l'interprétabilité et l'explicabilité, les travaux sur la contestabilité étant encore peu nombreux [55, 91, 80, 58].

2.1.2 Apprentissage automatique interprétable

Une première approche pour permettre l'intelligibilité est d'utiliser des algorithmes interprétables. Ces algorithmes, dont l'arbre de décision ou le modèle linéaire sont des exemples, peuvent être facilement compris par des humains sans avoir à employer une méthode spécifique. Les réseaux de neurones, dont l'information apprise lors de la phase d'apprentissage prend la forme de liens entre des neurones, sont à l'inverse difficiles à comprendre, même pour les spécialistes qui les conçoivent. C'est un exemple de modèle non interprétable. La nécessité de trouver un compromis entre l'inteprétabilité d'un algorithme et ses performances, généralement mesurées par la précision du modèle, est communément admise. Une représentation de ce compromis et de l'objectif de ce champ de recherche sont schématisés dans la Figure 2.1. On voit que les modèles les plus précis sont les moins interprétables et inversement. Pour améliorer cette situation, deux stratégies peuvent être employées : rendre compréhensibles les méthodes les plus performantes ou concevoir des méthodes à la fois performantes et intelligibles.

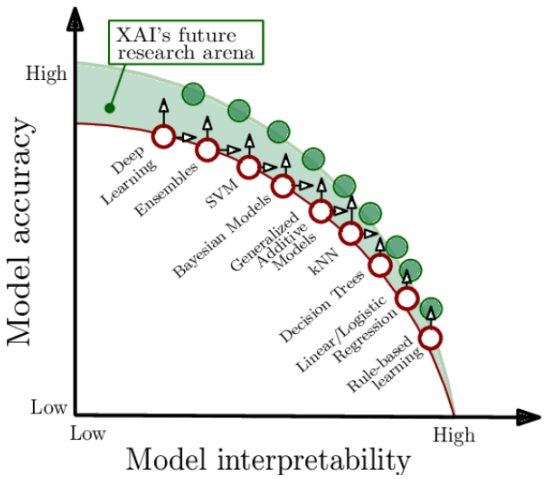


FIGURE 2.1 – Représentation schématique du compromis entre l'interprétabilité et les performances attendues. Les points rouges désignent l'état actuel du compromis et les points verts fixent l'objectif du champ de recherche. Figure extraite de [12].

Un courant de recherche, notamment en statistique et en optimisation tente de mieux formaliser les processus d'apprentissage des réseaux de neurones dont une grande partie

sont encore mal compris. Les bénéfices attendus de ces recherches sont nombreux car elles doivent permettre de rendre les processus à la fois plus fiables, en résolvant par exemple l'intriguant problème des attaques "adversariales" [108], plus performants en réduisant le nombre d'exemples et le temps nécessaires à l'apprentissage, et plus compréhensibles. À titre d'exemple, l'article [82] rend visible les représentations internes des réseaux de neurones en reconstituant des images à partir de l'information contenue dans les couches de neurones. Dans un registre plus théorique, la thèse de Marylou Gabrié utilise des modèles issus de la physique statistique pour tenter de mettre en lumière les mécanismes d'apprentissage à l'œuvre dans les réseaux de neurones [45]. L'impact de ces recherches pour les profanes est toutefois limité puisqu'il s'agit avant tout de permettre aux concepteurs des modèles de mieux comprendre les mécanismes statistiques à l'œuvre.

Une autre possibilité est de concevoir des algorithmes d'apprentissage à la fois performants et facile à interpréter. L'article de Himabindu Lakkaraju *et al.* est emblématique de cette démarche [71]. Pour que leur modèle à base de règles reste facile à comprendre, un terme mesurant l'interprétabilité est ajouté à la fonction objectif de l'apprentissage. Tao Lei adopte une démarche similaire en concevant un réseau de neurones capable de s'entraîner simultanément à prédire et à expliquer ses prédictions [74]. Dans l'exemple présenté, le réseau de neurones est capable de pointer les zones du texte qui ont influencé la prédiction. Un autre exemple de modèle interprétable est celui des modèles additifs généralisés (GAM) [52]. En imposant une indépendance stricte de l'action de chaque variable sur la prédiction finale grâce à une structure additive, les GAM ont en effet l'avantage d'avoir un fonctionnement très intuitif et facile à inspecter. Dans [23], les auteurs utilisent de tels modèles pour prédire, de manière intelligible, le risque d'aggravation pour les patients atteints de pneumonie [2].

2.1.3 Explications d'algorithmes

Que l'algorithme soit basé sur l'apprentissage inteprétable ou sur des méthodes traditionnelles, impliquant ou n'impliquant pas de phase d'apprentissage, il est souvent légitime

2. Le cas d'étude choisi fait écho à une étude des années 1990 qui avait conduit à adopter un modèle moins précis et interprétable plutôt qu'un réseau de neurones théoriquement très précis, mais opaque et jugé inacceptable pour des décisions à fort enjeu comme les diagnostics médicaux [7].

de chercher à obtenir des informations sur son fonctionnement. Ces informations sont généralement appelées des *explications d'algorithmes*. Une explication peut être *globale* si elle concerne l'algorithme dans son ensemble ou *locale* si elle concerne un résultat particulier. Pour obtenir ces informations, deux grandes classes de méthodes peuvent être employées. L'approche en "boîte blanche" suppose que le code de l'algorithme et les coefficients du modèle sont disponibles alors que l'approche en "boîte noire" ne repose pas sur cette hypothèse, la seule option consiste donc à fournir des entrées à l'algorithme et observer les sorties correspondantes.

Avec un accès au code, il est théoriquement possible de connaître l'ensemble des opérations exécutées par le programme. Ces opérations sont la cause de la décision et sont donc, au sens philosophique du terme, une explication de la décision. Par exemple, pour expliquer la sortie d'un arbre de décision, il suffit d'exposer l'ensemble des règles qui constituent l'arbre (explication globale) ou l'ensemble des règles appliquées dans ce cas particulier (explication locale). En pratique, sauf quelques exceptions comme l'arbre de décision, les opérations exécutées par les modèles sont si nombreuses et abstraites qu'elles n'apportent aucune information utile[3]. Le but des méthodes d'explication en boîte blanche est d'extraire une information compréhensible de cet ensemble d'opérations. Ces méthodes dépendent forcément de l'algorithme et du type de modèle utilisé.

La revue de littérature de Grégoire Montavon recense les grands principes des méthodes d'explication en boîte blanche appliquées aux réseaux de neurones profonds [89]. Les techniques de "maximisation d'activation"[4] permettent de visualiser sous forme d'image l'information encodée dans les coefficients des couches de neurones classe par classe. Afin de générer des explications locales, les méthodes de *layer-wise relevance propagation* peuvent être employées. Elles parcourent le réseau à l'envers (de la sortie vers l'entrée) pour identifier les entrées qui ont le plus pesé dans la classification finale. Les Séparateurs à Vaste Marge (en anglais *Support Vector Machine*, SVM) peuvent aussi être expliqués par des méthodes en boîte blanche. Glenn Fung *et al.* proposent d'approximer le SVM par un ensemble de règles

3. Par exemple, pour classer une seule image, le réseau de neurones ResNet doit effectuer près de 10^9 opérations mathématiques impliquant 5.10^7 coefficients [47]

4. Cet ensemble de techniques recherchent les points d'entrées de l'algorithme, qui peuvent être complètement fictifs ou plus réalistes selon les méthodes, conduisant à la valeur maximale d'une grandeur. La grandeur choisie est souvent la probabilité d'appartenance à une classe de sortie.

calculées à partir d'une analyse des coefficients du modèle [44]. Ces règles offrent une description globale du système et peuvent également être utilisées pour expliquer localement une sortie.

À la frontière boîte noire / boîte blanche, certaines méthodes proposent de fournir des explications pour n'importe quel modèle à condition que les dérivées de ce dernier soient accessibles. Dans [41], les auteurs font l'hypothèse d'un accès au gradient en tout point pour générer des cartes de saillance pour les classificateurs d'image. Dans [68], les auteurs supposent également un accès à la Hessienne (dérivée seconde) et un accès aux données d'entraînement. Leur méthode permet alors d'identifier avec un certain niveau de certitude l'exemple du jeu de données d'entraînement qui a eu le plus d'impact sur une classification. Les exemples "responsables" d'une classification peuvent ainsi être analysés afin de mieux comprendre le comportement du modèle et pour détecter d'éventuels exemples d'entraînement malicieux (c'est-à-dire délibérément conçus et introduits pour tromper le système [64]). Ces méthodes, qui nécessitent en pratique un accès en boîte blanche, ont toutefois l'avantage de ne pas dépendre d'un type de modèle trop spécifique.

Dans l'approche en boîte noire, puisqu'il n'est pas possible d'accéder au code ni aux coefficients du système, l'explication est issue d'une analyse des liens entre les entrées et les sorties. Une revue complète des méthodes d'explication en boîte noire, qui sont au centre des travaux de cette thèse, est présentée dans la partie suivante (Partie 2.2). À titre d'exemple, nous nous contentons ici de mentionner l'article de Marko Robnik-Šikonja publié en 2008 [99]. L'objectif des auteurs est d'estimer l'effet d'une variable A_i dans la classification de l'entrée x par une boîte noire f. L'idée générale, dont le principe se retrouve dans de nombreuses méthodes, est de calculer la différence $f(x) - f(x \setminus A_i)$ où $x \setminus A_i$ représente une version de x qui serait privée de l'information apportée par la variable A_i. Plus la différence est importante, plus l'impact de cette variable est significatif. L'estimation de $f(x \setminus A_i)$ n'est pas simple puisqu'il n'est pas possible, dans le cas général, de masquer l'information apportée par A_i. La moyenne suivante fournit une estimation possible de cette grandeur :

$$f(x \setminus A_i) = \sum_k P(A_i = a_k) f(x \leftarrow A_i = a_k), \tag{2.1}$$

avec $x \leftarrow A_i = a_k$ une version de l'entrée x dans laquelle on aurait remplacé la valeur de A_i par a_k. La somme se fait sur l'ensemble $\{a_k\}$, support de la variable A_i et est pondérée par la probabilité d'observer cette valeur $P(A_i = a_k)$.

Bien que cette méthode soit appréciable pour ses vertus pédagogiques, elle est simpliste et limitée puisqu'elle ne prend pas en compte les corrélations de A_i avec les autres variables et elle ne permet d'obtenir qu'un seul type d'explication. D'autres méthodes abordant ces problèmes spécifiques sont présentées dans la partie suivante.

2.2 Les méthodes d'explication en boîte noire

Dans cette partie, nous nous focalisons sur les méthodes d'explication basées sur l'approche en boîte noire (BEM) [5] de la littérature. Nous commençons par fournir une première intuition des traitements typiques de l'approche en boîte noire en décrivant une sélection de BEM de la littérature (Partie 2.2.1) [6]. Pour faciliter la comparaison et permettre une classification, nous proposons ensuite un cadre générique du fonctionnement de ces méthodes (Partie 2.2.2). Ce cadre est ensuite mis à profit pour élaborer une taxonomie des méthodes existantes (Partie 2.2.3).

2.2.1 Quelques exemples de méthodes

Une première manière de classer les méthodes d'explication concerne le périmètre visé : global ou local. Lorsque le modèle employé est complexe, il paraît illusoire d'en fournir une explication globale à la fois précise et compréhensible. Toutefois, une explication globale, même imprécise, peut avoir un intérêt pour saisir le fonctionnement "moyen" du modèle. Les méthodes globales s'appuient en général sur un jeu de données représentatif des variables d'entrée de l'algorithme. Elles peuvent présenter des explications sous forme de graphiques [69], de règles [72, 30] ou d'importances de variables [53, 62].

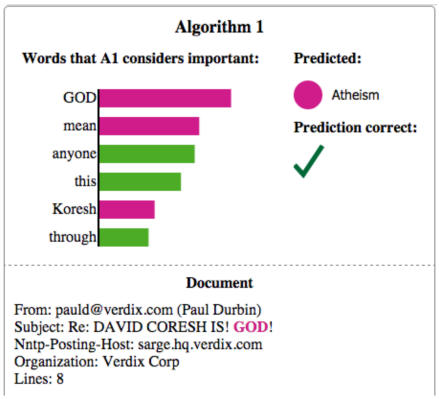
5. BEM est l'acronyme anglais de *Black-box Explanation Method*.
6. Pour une revue exhaustive des méthodes, le lecteur pourra s'orienter vers les publications suivantes : [48, 12, 2].

Même lorsque le modèle est complexe, il est possible de fournir des informations simples sur son fonctionnement au voisinage d'un point spécifique. Ces explications, appelées explications locales, sont notamment pratiques lorsque le destinataire ne s'intéresse qu'à un résultat en particulier et pas à l'algorithme dans son ensemble. Puisqu'aucune hypothèse ne peut être faite sur la boîte noire, obtenir une explication locale peut toutefois s'avérer être un défi technique. Par exemple, lorsque au moins deux variables interagissent de manière non additive dans le modèle, ce qui est le cas de l'immense majorité des méthodes modernes de ML, l'approche simpliste exposée dans la Partie 2.1.3 n'est plus suffisante. Ce problème est notamment soulevé dans un article écrit par les mêmes auteurs quelques années plus tard [118]. En effet, puisque l'influence de la variable A_i dépend de la valeur d'autres variables A_j, il faut faire varier simultanément A_i et toutes les autres variables A_j pour estimer sa contribution à la sortie de l'algorithme. Un résultat de théorie de jeu de Lloyd Shapley [102], permet ensuite aux auteurs d'isoler le rôle spécifique de la variable A_i dans la classification. D'autres outils d'explication s'inspirent également de cette méthodologie [31, 79] (voir Figure 2.2c).

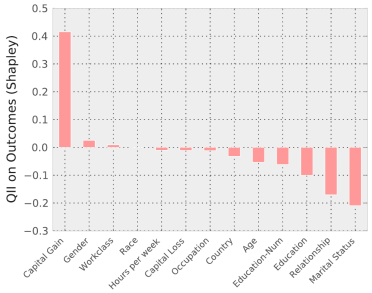
L'article de Marco Tulio Ribeiro qui décrit la méthode d'explication LIME (*Local Interpretable Model-agnostic Explanations*) fournit une méthode d'explication locale compréhensible [98]. La stratégie consiste à générer des versions perturbées d'images ou de textes en supprimant des pixels ou des mots de manière aléatoire. Ces versions perturbées sont ensuite classées par la boîte noire et un modèle linéaire pondéré est utilisé pour estimer la zone de l'image ou mot du texte qui est le plus souvent associée à une classe de sortie donnée (voir Figure 2.2b).

Certaines méthodes se focalisent sur l'obtention d'une explication ayant des caractéristiques spécifiques censées correspondre aux besoins du destinataire de l'explication. Ces caractéristiques varient d'un article à l'autre. Par exemple, la méthode Anchors [97], dont un exemple est présenté dans la Figure 2.2e, s'attache à trouver une explication sous forme de règles qui soit à la fois précise et générale (c'est-à-dire couvrant une grande zone de l'espace d'entrée) alors que la méthode BETA [72] s'attache à générer une explication sous forme de règles facile à interpréter (voir Figure 2.2a). Une autre option consiste à générer

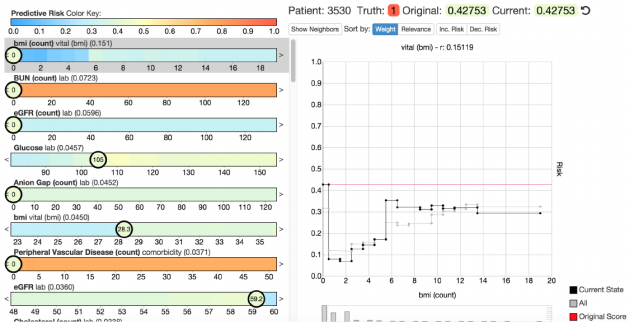
(A)



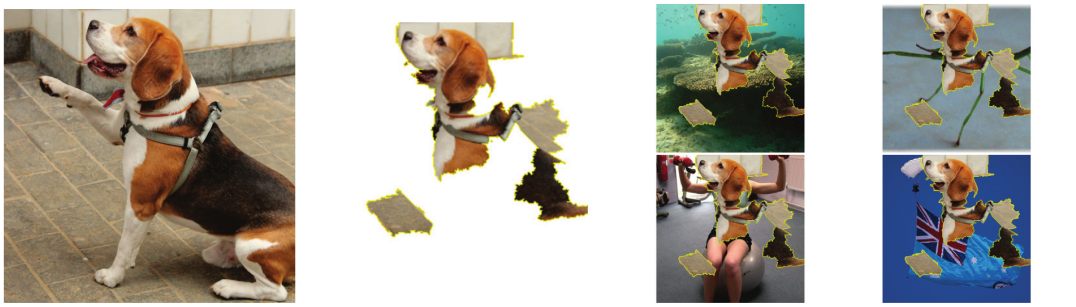
(B)



(C)



(D)



(a) Original image (b) Anchor for "beagle" (c) Images where Inception predicts $P(\text{beagle}) > 90\%$

(E)

FIGURE 2.2 – Exemples d'explications en boîte noire de la littérature. (A) Explication globale sous forme de règles [72]. (B) Importance de variables dans une classification de texte [98]. (C) Importance de variables dans une classification tabulaire [31]. (D) Système d'interaction visuel avec le modèle [69]. (e) Ensemble minimal de pixels suffisant à la prédiction de la classe "Beagle" (*Anchors*) et validation des *Anchors* par superposition sur des images aléatoires [97]. Ces figures sont extraites directement des articles mentionnés.

des exemples contrefactuels ou "contrefactuelles". Si l'explication concerne une entrée spécifique x, un exemple contrefactuel est l'entrée x' la plus proche de x obtenant un résultat différent de la boîte noire ($f(x) \neq f(x')$) [115]. Ce type d'explication est présenté comme particulièrement utile pour contester une décision algorithmique.

Dans la pratique, différents utilisateurs peuvent avoir différents besoins correspondant à des objectifs distincts pour les BEM. Il semble donc nécessaire de concevoir des BEM qui soient au minimum adaptatives et au mieux interactives pour être utiles dans une diversité de situations [81, 87]. Un développement prometteur dans cette direction consiste à mettre en œuvre une boîte à outils comprenant une variété de méthodes qui peuvent être sélectionnées par les utilisateurs en fonction de leurs besoins. Par exemple, AIX360 ("AI Explainability 360") [13] contient huit méthodes d'explication et permet à l'utilisateur de choisir en se basant sur un arbre de décision comprenant des questions telles que "comprendre les données ou le modèle ?" ou "modèle auto-explicatif ou explications post-hoc ?". Les outils proposés dans [69] contiennent également une forme d'interactivité par l'intermédiaire d'interfaces graphiques, comme celle de la Figure 2.2d.

Bien que ces méthodes paraissent avoir peu de points en communs, il est possible de les inscrire dans un cadre générique unique afin de les caractériser et de les comparer.

2.2.2 Un cadre générique pour décrire les BEM

L'objectif de cette partie est de définir un cadre générique commun pour décrire les BEM. Ce cadre est notamment mis à profit dans la partie suivante (Partie 2.2.3) pour présenter une taxonomie des méthodes existantes. Au-delà de son intérêt pour la classification, comprendre les BEM comme des instanciations d'un même cadre commun peut s'avérer utile pour définir une méthode d'explication générique et paramétrable (Partie 3.3).

Nous utilisons l'exemple simple d'un classificateur de spams pour décrire notre cadre générique. Il prend en entrée le texte d'un email et estime la probabilité que cet email soit un spam. Comme le code du classificateur n'est pas disponible, le système d'explication doit construire des emails, les soumettre au système et analyser leur classification pour inférer des informations sur le fonctionnement du système. Par exemple, pour évaluer le rôle de la

signature dans la classification d'un email, le système d'explication peut en créer différentes versions, avec et sans la signature, ou avec différents morceaux de texte dans la signature puis les fournir au classificateur de spams. Les classes attribuées à ces versions modifiées de l'email constituent alors l'information de base qui peut ensuite être transformée en une explication compréhensible par un humain.

Cet exemple simple met en évidence les deux principales composantes du cadre générique d'explication : (i) la sélection ou la création des entrées à soumettre à l'algorithme, appelée l'*échantillonnage* ; et (ii) l'analyse des liens entre les échantillons et leurs classifications pour générer le contenu de l'explication, appelée la *génération*. Si les données d'entrée sont représentées sous une forme difficilement intelligible, comme la valeur des pixels d'une image par exemple, une phase préliminaire doit s'ajouter pour permettre une représentation interprétable, comme cela est fait, entre autres, dans LIME [20].

Durant la phase d'échantillonnage, une base de données peut être utilisée pour estimer la distribution des variables d'entrée. Cette distribution est utile si l'on souhaite s'assurer que les explications reflètent le fonctionnement du modèle dans des conditions normales d'utilisation. Dans l'exemple des spams, il sera parfois inutile d'informer le destinataire de l'explication que tout email ayant une signature de plus de 2 000 caractères est systématiquement classé comme un spam si cette situation ne se présente presque jamais en pratique. Cette base de données est appelée *population* et elle est représentée par le symbole D. Par ailleurs, les explications peuvent concerner le modèle dans son ensemble ou seulement certaines entrées. Dans la suite, le *périmètre* de l'explication E décrit l'ensemble des points concernés par l'explication. Une explication globale est caractérisée par $E = D$ et une explication locale par $E = \{x_e\}$.

2.2.3 Taxonomie des méthodes

Les deux phases identifiées dans la partie précédente mettent en évidence la diversité des choix possibles dans la conception d'une BEM. Elles permettent aussi de classer les méthodes existantes en fonction de ces choix. Le Tableau 2.1 décrit comment ces choix se déclinent pour 12 BEM existantes de la littérature. Dans la suite, le contenu des choix

présentés dans le Tableau 2.1 est brièvement décrit, en considérant successivement les choix de conception pour l'échantillonnage (colonnes de gauche), la génération (colonnes du milieu) et le lien entre ces deux phases (colonne de droite).

Choix de conception pour l'échantillonnage :

- L'ensemble E décrit le périmètre de l'explication ($E = \{x_e\}$ si l'explication est locale et $E = D$ si elle est globale). Idéalement, le destinataire de l'explication devrait pouvoir se focaliser sur n'importe quel sous-ensemble de la population. Cependant, comme le montre le Tableau 2.1, les BEM de la littérature proposent la plupart du temps des explications globales ou locales, mais aucun choix intermédiaire. La seule exception est la méthode QII [31] qui permet d'expliquer la sortie de l'algorithme pour un groupe de données d'entrée.
- En général, les modèles se comportent différemment sur différents segments de la population. Par exemple, les systèmes de reconnaissance faciale sont moins précis dans la sous-population des femmes jeunes et plus précis dans celle des hommes blancs que dans la population générale [37]. Par conséquent, la prise en compte de la distribution des variables d'entrée grâce à la population est une caractéristique cruciale d'une BEM. Cette information est représentée par la colonne "Échantillonnage/D" du Tableau 2.1. Dans le Tableau 2.1, toutes les méthodes, sauf LIME [98] et LEMNA [50], prennent en compte la population D. Ces deux méthodes construisent des échantillons en masquant une partie des données d'entrée ce qui ne nécessite pas l'utilisation de D. L'impact de l'utilisation de la population sur les propriétés de l'explication générée est discutée plus en détail dans le Chapitre 3.
- La colonne "Échantillonnage/Type" du Tableau 2.1 fournit une information synthétique sur la stratégie employée pour l'échantillonnage. Une distinction est faite entre les échantillonnages par *sélection*, qui se caractérisent par le fait que les échantillons doivent appartenir à D et les échantillonnages par *perturbation* qui peuvent produire un élément quelconque de l'espace d'entrée. L'impact du type d'échantillonnage sur les propriétés de l'explication finale est détaillé dans le Chapitre 3.

Nom	Échantillonnage			Génération		Forme de l'explication	Itératif [b]
	E	D	Type [a]	Modèle	Objectifs		
Trepan [30]	Pop.	Pop.	P	Arbre de décision	Complexité, fidélité	Arbre de décision	I
BETA [72]	Pop.	Pop.	S	Modèle à base de règles	Interprétabilité, fidélité Équivocité	Règles globales	I
GoldenEye [53]	Pop.	Pop.	P	Valeur moyenne	Fidélité, Taille des interactions	Importance de variables	I
VIN [62]	Pop.	Pop.	P	Valeur moyenne	Projection ANOVA	Importance de variables	I
PDP ICE [69]	Pop.	Pop.	P	Valeur moyenne	NA	Graphe de dépendance par variable	NI
QII [31]	Indiv. ou groupe	Pop.	P	Valeur moyenne	NA	Importances de variables	NI
Anchors [97]	$\{x_e\}$	Pop.	P	Modèle à base de règles	Fidélité, complexité, Généralité	Règles locales	I
LIME [98]	$\{x_e\}$	$\emptyset$	P	Modèle linéaire	Fidélité, complexité	Importance de variables	NI
Shapley [118]	$\{x_e\}$	Pop.	P	Valeur moyenne	NA	Importance de variables	NI
LEMNA [50]	$\{x_e\}$	$\emptyset$	P	Mélange de modèles linéaires	Fidélité, complexité	Importance de variables	NI
Local gradient [16]	$\{x_e\}$	Pop.	S	Estimation par noyau	Fidélité	Gradient du modèle	NI
Contrefactuelles [115]	$\{x_e\}$	Pop.	S	Petites déviations	Distance à x_e	Exemple	I

TABLE 2.1 – Tableau comparatif des BEM. Les colonnes correspondent aux choix possibles dans le cadre générique. Les notations suivantes sont utilisées : (a) S : Sélection, P : Perturbation, ; (b) I : Itératif, NI : Non itératif.

.

Choix de conception pour la génération :

- Certaines BEM utilisent des méthodes d'apprentissage automatique interprétables pour obtenir une synthèse compréhensible de l'information contenue dans les échantillons. Par exemple, un modèle à base de règles peut être entraîné pour décrire les caractéristiques principales qui conduisent à la classification d'un email comme spam à partir des échantillons. Dans certains cas, la méthode utilise simplement une valeur moyenne pour fournir une information générale sur les échantillons. La colonne "Génération/Modèle" du Tableau 2.1 indique le type de modèle utilisé le cas échéant.
- Il est fréquent que des objectifs spécifiques soient imposés à la phase de génération afin d'obtenir des explications ayant les propriétés recherchées. Par exemple, dans le système à base de règles de l'exemple ci-dessus, un objectif de simplicité limitant le nombre total de règles pourrait être imposé au système. La colonne "Génération/Objectifs" montre les objectifs affichés de chaque BEM listés par ordre décroissant de priorité.

Choix généraux de conception :

- La colonne "Itération" du Tableau 2.1 permet de distinguer les méthodes utilisant un échantillonnage itératif (I) des méthodes non itératives (NI). Dans le mode itératif, les phases de génération et d'échantillonnage vont de pair et de nouveaux échantillons peuvent être obtenus en fonction des besoins de la génération. À l'inverse, dans le mode non itératif, les échantillons sont calculés en une seule étape. Intuitivement, le mode itératif peut conduire à des ensembles d'échantillons mieux adaptés aux besoins de la génération et obtenus de manière plus efficace. L'échantillonnage basé sur un algorithme génétique proposé dans LORE [49] est un exemple remarquable du potentiel des méthodes d'échantillonnage itératives.

En conclusion, cette partie fournit un cadre utile pour caractériser les méthodes d'explication en boîte noire. Cependant, au delà des méthodes employées pour générer les explications (qu'elles soient en boîte noire ou en boîte blanche), la question de leur évaluation doit se poser.

2.3 Évaluation des explications

La question de savoir si une explication remplit effectivement son rôle renvoie finalement à la question du but des explications et à la manière dont elles sont censées être utilisées (Partie 2.3.1). Bien que l'évaluation des explications reste une question de recherche encore ouverte et active, la nécessité d'impliquer des utilisateurs humains dans le processus est généralement admise (Partie 2.3.2).

2.3.1 Les enjeux de l'évaluation

Récemment, la question du but des explications dans le développement des SDA et des systèmes impliquant l'IA a suscité de nombreux travaux. Pour Finale Doshi-Velez [35], l'interprétabilité et les explications comblent un manque de formalisation et sont surtout utiles car elles servent d'autres propriétés souhaitables (*fairness*, fiabilité et *accountability*). Pour l'auteure, c'est d'ailleurs directement au regard de ces objectifs finaux que les explications

devraient être évaluées. Pour certains, les sciences humaines et sociales ont un rôle central à jouer dans la définition des enjeux et des propriétés souhaitables des explications. Dans sa revue de littérature, l'informaticien Tim Miller [84] regroupe les principaux essais de philosophie, de psychologie et de sciences cognitives en lien avec les explications interpersonnelles et pouvant avoir un intérêt pour les explications d'algorithmes[7]. L'article met notamment en avant l'importance du contexte et d'un certain niveau d'interaction dans l'explication. Selon l'auteur, il est nécessaire que les explications d'algorithmes prennent la forme d'un processus de communication plutôt que celle d'une explication statique. Une première manière de prendre en compte le contexte de l'explication est d'identifier les différents profils de destinataires d'explications [111, 116, 12, 88, 57]. Globalement, les auteurs semblent s'accorder sur au moins quatre catégories de destinataires d'explications : les experts techniques (spécialistes en IA ou en développement de SDA), les experts du domaine (professionnels ayant des connaissances sur les données), les profanes (personne n'ayant aucune connaissance spécifique comme la personne affectée par la décision par exemple) et les auditeurs extérieurs notamment chargés de contrôler la conformité du système. Chaque type de destinataire se distingue par son domaine d'expertise et l'objectif pour lequel il consulte une explication. En conséquence, la forme, le contenu et le mode d'interaction explicatif doivent avoir des caractéristiques adaptées à ses besoins spécifiques.

Cette contextualisation des explications doit également se répercuter sur la stratégie d'évaluation de leur pertinence. Pour Finale Doshi-Velez [35], trois niveaux d'évaluation doivent être considérés. Les évaluations fondées sur des applications (premier niveau) impliquent des personnes réelles réalisant des tâches réelles et mesurent directement les objectifs finaux de l'explication. Le deuxième niveau est celui des évaluations impliquant des humains réalisant une tâche simplifiée. Ce niveau, moins coûteux que le premier, est notamment utile pour tester la qualité de l'explication elle-même, dans le but par exemple d'en comparer différents types. Le troisième niveau n'implique pas d'utilisateur humain et s'appuie sur une mesure quantitative pour juger la qualité d'une explication (par exemple la

7. Dans la même veine, Shane T. Mueller [90] propose une méta-revue de littérature autour du concept d'explication pour les systèmes intelligents. Cet article regroupe les références pertinentes de sciences humaines et sociales incluant notamment des recherches sur la manière dont les humains formulent et acceptent des explications.

précision ou la "sparsité" d'un modèle connu comme interprétable). Dans un article récent [88], Sina Mohseni *et al.* synthétisent l'ensemble des recherches concernant l'évaluation des approches XAI et proposent un cadre et des recommandations à destination des concepteurs d'explications. Les auteurs distinguent notamment les huit objectifs possibles d'une approche XAI qui se répartissent entre différents types de destinataires [8]. Pour atteindre ces objectifs cinq types de mesure qui se répartissent dans les trois niveaux d'évaluation déjà cités sont proposés [9]. Enfin, à chacune de ces mesures correspondent cinq méthodes d'évaluation impliquant des utilisateurs ou des mesures quantitatives.

Les enjeux de l'évaluation des explications sont multiples et complexes. Si les mesures quantitatives sont indéniablement utiles afin de concevoir des méthodes d'explication efficaces, les études impliquant des utilisateurs humains, sur des tâches réelles ou simplifiées, restent la référence absolue dans ce domaine.

2.3.2 Études impliquant des utilisateurs

Alors même qu'elles apparaissent comme indispensables pour répondre aux enjeux des explications d'algorithmes, les évaluations par des utilisateurs humains sont encore peu répandues, notamment dans la communauté ML [85]. Nous présentons dans cette partie quelques exemples représentatifs des travaux visant à évaluer une méthode spécifique et d'études sur les utilisateurs visant à évaluer plus généralement une classe d'explications [10].

Les publications décrivant les méthodes d'explication incluent parfois des études sur les utilisateurs. Il s'agit le plus souvent d'études *ad hoc* impliquant typiquement quelques dizaines de personnes devant réaliser des tâches simplifiées (deuxième niveau d'évaluation dans la classification de Doshi-Velez). Les stratégies employées varient d'une étude à l'autre notamment afin de mesurer les objectifs visés par les auteurs. L'étude sur les utilisateurs présentée dans LIME [98] se focalise sur la capacité des utilisateurs à juger un modèle. Des

8. Les huit objectifs sont : transparence, confiance, limitation des biais, respect de la vie privée, visualisation et inspection, paramétrage du modèle, interprétabilité et débogage.
9. Ces mesures concernent : le modèle mental, l'utilité et la satisfaction des explications, la confiance, l'utilité dans l'exécution de tâches spécifiques et les mesures quantitatives.
10. Pour une revue exhaustive voir [88].

algorithmes basés sur l'apprentissage sont délibérément truqués via du sur-apprentissage [11] ou l'inclusion de variables trompeuses et les utilisateurs doivent tenter de détecter ces défauts à l'aide de LIME ou d'une méthode témoin. Dans *Anchors* [97], c'est la capacité à prédire la sortie du modèle pour une nouvelle entrée ainsi que le temps mis à réaliser cette tâche qui sont mesurés. Dans l'outil "*Interpretable Decision Set*" [71], c'est la facilité des utilisateurs à comprendre les règles du modèle qui est analysée. Nous notons que, dans chaque article, la méthode d'évaluation utilisée vise à confirmer une propriété mise en avant dans l'argumentaire de la méthode : la notion de confiance dans LIME, la précision et la généralité dans Anchors et le haut niveau d'interprétabilité dans "*Interpretable Decision Set*". Ainsi, les auteurs fixent à la fois l'objectif et la manière de le vérifier, souvent sans s'assurer que cet objectif est effectivement souhaitable dans des applications réelles. À l'inverse, la méthode d'évaluation proposée dans l'article de Josua Krause [69] prend la forme d'une étude de terrain avec des *datascientists*, utilisateurs finaux de l'outil de visualisation et d'explication (premier niveau d'évaluation selon la classification de Doshi-Velez). Bien que l'article ne présente aucune mesure quantitative, les retours des utilisateurs apportent une information riche sur l'utilité de la méthode dans des conditions réelles d'utilisation.

D'autres études, plutôt conduites au sein de la communauté des interfaces homme-machine, évaluent plus généralement l'impact de certaines propriétés d'explications comme la forme, le contenu ou la quantité d'information. Dans [70] par exemple, les auteurs étudient l'impact de l'exactitude et de l'exhaustivité des informations fournies par une explication. Les auteurs mettent notamment en avant que le caractère complet de l'explication est plus important que son exactitude, à condition que l'exactitude soit au-dessus d'un certain seuil. En deçà de ce seuil critique, la confiance est rompue et les utilisateurs rejettent en bloc les informations fournies.

Dans [95], ce sont les effets de l'interprétabilité et de la transparence sur trois propriétés (capacité à simuler le modèle, confiance et détection d'erreurs) qui sont analysés. Les auteurs comparent quatre conditions expérimentales : en boîte blanche ou en boîte noire avec un modèle simple ou un modèle complexe. Les utilisateurs doivent successivement tenter de

11. L'auteur entraîne délibérément un classificateur chien/loup à utiliser exclusivement la présence de neige dans le fond de l'image pour distinguer les photographies de ces deux types de canidés.

simuler la prédiction du modèle et se prononcer sur l'exactitude de sa sortie (après que la vraie sortie leur soit présentée). La confiance dans le modèle et la capacité à détecter des erreurs sont mesurées grâce à la deuxième tâche. Les résultats obtenus sont contre-intuitifs puisqu'ils tendent à montrer que les modèles simples et transparents ne suscitent pas plus la confiance des utilisateurs que les modèles complexes et opaques. La transparence du modèle aurait même tendance à réduire la capacité à détecter une erreur.

Dans [106], un protocole qui repose notamment sur des explications est mis en place pour encourager les retours d'utilisateurs afin d'aider les développeurs à améliorer les modèles de ML. Une comparaison de différentes formes d'explications montre que les explications sous forme de règles sont plus faciles à comprendre et globalement préférées par les utilisateurs. Le design de l'étude, réalisée avec des feuilles de papier et des stylos sous la forme d'entretiens individuels avec un expérimentateur, a également permis de mettre en avant l'importance de l'émotion ressentie par le destinataire de l'explication dans son acceptation.

Les études impliquant des utilisateurs apportent une information indispensable au développement d'explications répondant pleinement aux enjeux généraux. Au delà des modes d'évaluations déjà mentionnés, nous remarquons toutefois que les enjeux de l'explicabilité dans les organisations sont rarement étudiés alors qu'ils sont cruciaux pour établir les *desiderata* des méthodes. Il existe encore trop peu d'exemples de telles analyses qui seules peuvent assurer de l'adéquation entre les objectifs des méthodes d'explication et les enjeux réels en mobilisant des méthodes propres aux sciences sociales [38, 54].

2.4 Les limites de la littérature

L'XAI est un champ de recherche relativement récent et encore très dépendant des recherches menées en IA. À notre connaissance, il existe encore peu d'outils satisfaisants utilisables par l'industrie. Dans cette partie, nous abordons les limites du champ dont nous proposons des pistes d'amélioration dans la suite de la thèse (Partie 2.4.1), puis nous abordons les potentiels risques associés à l'utilisation de systèmes d'explication (Partie 2.4.2).

2.4.1 Limites des méthodes actuelles

Notre étude de l'état de l'art met en lumière certaines limites des méthodes actuelles dont une partie sont abordées au cours de cette thèse. En premier lieu, puisque les objectifs finaux des explications et la manière de les évaluer ne font pas encore l'objet d'un consensus et encore moins de standards, chaque auteur est libre de proposer ses propres objectifs, de concevoir une méthode d'explication qui semble y répondre et de proposer un protocole d'évaluation pour le vérifier. Ce mode de fonctionnement est problématique puisqu'il peut conduire à une divergence entre les besoins réels en explication et les méthodes développées. C'est un phénomène que certains décrivent comme le syndrome des "*inmates running the asylum*" [85]. Comme les auteurs de [85], nous pensons que les expérimentations de terrain, si possible portant sur des applications réelles comme celle présentée dans le Chapitre 5, peuvent participer à une meilleure définition des objectifs réels des explications. Par ailleurs, une meilleure formalisation des méthodes d'explication et plus généralement du processus d'explication peut être profitable pour rationaliser les différentes recherches. L'approche à plusieurs niveaux proposée dans le Chapitre 3 qui propose notamment une distinction entre le contexte, les propriétés et les paramètres techniques de l'explication semble être une proposition allant dans cette direction.

Même si les objectifs d'explication faisaient l'objet d'un consensus, la diversité d'utilisateurs et de situations est telle qu'un seul type d'explication ne peut pas tous les satisfaire. Une critique que nous pouvons formuler aux méthodes actuelles est leur caractère statique ou unilatéral ; elles ne permettent souvent pas d'adapter l'explication au contexte ni aux besoins spécifiques de l'utilisateur. Puisque chaque situation est unique (le même utilisateur dans le même contexte peut avoir des besoins différents en fonction de la situation), les méthodes d'explication devraient idéalement être interactives et donc laisser une grande liberté au destinataire de l'explication dans le choix des informations qu'il souhaite consulter. Les choix techniques employés dans le calcul des explications ne devraient donc pas être fixés de manière rigide, mais devraient au contraire pouvoir être modifiés en fonction du résultat des interactions avec le destinataire. Le système interactif d'explication proposé dans le Chapitre 3 va dans le sens d'un tel système en s'appuyant notamment sur les formalismes

déjà mentionnés.

Enfin, comme l'indique Doshi-Velez [35], les explications jouent avant tout un rôle de facilitateur pour d'autres objectifs comme l'*accountability* et la contestabilité. Pourtant, comme nous le montrons dans le Chapitre 4 les explications, dont le but est de faire comprendre la logique employée par le SDA, ne sont pas forcément adaptées à ces objectifs. En effet, il ne paraît pas légitime d'exiger d'une personne qu'elle comprenne le fonctionnement d'un SDA pour pouvoir le contester, surtout si cette personne est un profane sans connaissance spécifique sur les algorithmes. Bien que la nécessité de concevoir des SDA contestables est soutenue par un nombre croissant d'auteurs [91, 80], les recherches se focalisent quasi-exclusivement sur les méthodes d'explication. Le système de justification et de contestation présenté dans le Chapitre 4 est une première proposition allant dans ce sens.

2.4.2 Les risques liés aux explications

Alors que les explications semblent être une solution utile à certains problèmes posés par les systèmes automatiques de décision, leur mise en œuvre peut poser certains problèmes pratiques. Par exemple, comme souligné dans [26], les responsables de traitements qui pourraient se voir imposer de fournir des explications à leurs utilisateurs ont souvent intérêt à conserver une certaine forme d'opacité sur le fonctionnement de leur système. Cette divergence d'intérêt conduit l'auteur à redouter l'apparition de "*dark patterns*" de l'explication qui auraient pour but de tromper ou de décourager les utilisateurs.

Deux articles au moins s'intéressent déjà à la génération d'explications trompeuses, notamment capables de masquer le caractère discriminatoire d'un système de décision. L'existence de telles méthodes fait craindre que les méthodes d'explication puissent être détournées et intégrées dans des stratégies de "blanchiment éthique" (*fairwashing*). Dans [14], une stratégie d'énumération d'explications permet aux auteurs de sélectionner, parmi les explications ayant un niveau de précision comparable, l'explication qui ne laisse pas transparaître le caractère discriminant du système de décision. Dans le même registre, Dylan Slack *et al.* [105] montrent que la boîte noire utilisée pour prendre la décision pourrait être conçue de manière à détecter qu'elle est soumise à une analyse par un outil du type de LIME ou SHAP.

Tout comme les véhicules Volkswagen pouvaient tromper l'appareil de mesure lors des essais anti-pollution [12], la boîte noire pourrait modifier ses prédictions lorsqu'elle est soumise à une analyse de manière à masquer un comportement problématique que les responsables ne souhaiteraient pas dévoiler.

Bien que ces craintes peuvent sembler encore très théoriques, il existe déjà des exemples concrets de manipulation. Les travaux menés dans [10] analysent l'algorithme de publicité personnalisée du réseau social Facebook et plus précisément son système d'explication ("*Why am I seeing this [ad]?*"). Grâce à un travail minutieux de rétro-ingénierie du système de ciblage de Facebook, les auteurs ont pu montrer que les explications proposées sont souvent incomplètes (seule une partie des informations réellement utilisées pour le ciblage est mentionnée) et parfois trompeuses (parmi toutes les informations utilisées pour le ciblage, seules les moins sensibles sont présentées).

Sur un autre registre, des chercheurs ont montré que les explications peuvent être utilisées pour mener différents types d'attaques visant le SDA lui-même ou la confidentialité des données d'entraînement. Dans [104], les explications sont utilisées pour obtenir des informations confidentielles sur les données utilisées lors de l'entraînement d'un modèle (*membership attack*). Dans [86], c'est le modèle lui-même qui est entièrement reconstruit à partir des explications en quelques points. Bien que ces recherches concernent un type particulier d'explication (le gradient) rendant ce type d'attaques peu réalistes pour le moment, il faut noter que ces menaces pourront constituer un frein réel au déploiement des systèmes d'explications futurs.

12. Voir le scandale industriel du "Dieselgate" ou affaire Volkswagen.

Chapitre 3

Explications interactives

Dans ce chapitre, nous décrivons notre système interactif d'explication de SDA en boîte noire appelé *Interactive Black-box Explanations* (IBEX). En s'appuyant sur les informations fournies par l'utilisateur, IBEX génère une explication adaptée au profil de la personne à ses besoins. L'utilisateur peut interagir avec le système d'explication à plusieurs niveaux en fonction de ses compétences et de son expertise. IBEX convient donc à la fois aux profanes (et dès la première utilisation), à des professionnels (qui pourraient être amenés à utiliser IBEX régulièrement) et à des experts en IA. Pour générer les explications, IBEX s'appuie sur un cadre générique commun pour les explications en boîte noire qui décompose notamment le processus d'explication en deux composantes distinctes. Cette décomposition et les combinaisons de choix qui en découle permettent d'obtenir une variété d'explications qui peut être mise à profit pour satisfaire au mieux la diversité des besoins. IBEX a été validé dans le cadre d'une étude impliquant des utilisateurs (agents d'une autorité de régulation française) ayant des profils variés (juristes et ingénieurs).

Les contributions présentées dans ce chapitre sont les suivantes :

— un cadre théorique complet pour la conception de méthodes d'explication interactives basé sur des interactions à plusieurs niveaux et sur un système générique d'explication,

— un système d'interaction basé sur des interactions à plusieurs niveaux permettant à l'utilisateur d'obtenir une explication adaptée à ses besoins quel que soit son niveau d'expertise,

— un système générique d'explication en boîte noire paramétrable afin de pouvoir obtenir une grande variété d'explication,
— un ensemble de cas d'usage illustrant la démarche interactive,
— une étude impliquant des utilisateurs validant la pertinence de la méthode.

3.1 Démarche

La diversité des méthodes présentées dans le Chapitre 2 montre que la forme, le type d'utilisateur ciblé (ci-après "destinataire") et l'objectif des explications peuvent être très différents. L'un des principaux défis de l'XAI est de concevoir des méthodes d'explication qui puissent s'adapter à cette variété de situations. La nécessité de concevoir des explications interactives plutôt que statiques est défendue de manière convaincante par plusieurs auteurs [85, 81, 87, 2, 90], mais les développements dans cette direction restent rares dans le domaine de l'XAI. Une première manière de rendre interactif le processus d'explication est la démarche "boîte à outils" qui rend accessible une variété de méthodes pouvant être sélectionnées en fonction des besoins de l'utilisateur. Par exemple, AIX360 ("AI Explainability 360") [13] contient huit méthodes, regroupant des systèmes d'explication et des modèles interprétables, et guide les utilisateurs dans leur choix sur la base d'un arbre de décision comprenant des critères tels que "comprendre les données ou le modèle" ou "modèle auto-explicatif ou explications *post-hoc*". Toutefois, ces boîtes à outils ont un potentiel d'interaction limité puisqu'elles permettent seulement de passer d'une méthode à une autre alors même que ces dernières peuvent avoir des objectifs et des modes de fonctionnement très différents. Dans ce chapitre, nous proposons une nouvelle démarche pour rendre le processus d'explication plus interactif afin de mieux répondre aux besoins variés des utilisateurs, notamment les utilisateurs profanes. La démarche proposée se repose sur quatre éléments :

— Une formalisation du contexte d'explication (profil de l'utilisateur, objectif, etc.).
— La possibilité pour les destinataires d'exprimer leurs besoins en fonction de leur niveau d'expertise.

— Une méthode d'explication en boîte noire paramétrable capable de générer des explications satisfaisant certaines propriétés.

— Une correspondance entre les paramètres de la méthode d'explication et les besoins du destinataire.

Dans ce chapitre, nous présentons IBEX (pour "*Interactive Black-box Explanations*"), notre système interactif d'explication en boîte noire basé sur une démarche à trois niveaux ainsi qu'une implémentation de ce système du type "preuve de concept".

Notre démarche repose sur l'observation que, au-delà de leur diversité, les méthodes d'explication en boîte noire partagent certaines caractéristiques identifiées par le cadre générique présenté dans la partie 2.2.2 et ses deux composantes, l'*échantillonnage* et la *génération* qui ont inspiré l'architecture d'IBEX. L'architecture d'IBEX est générique dans le sens où de nombreuses BEM (y compris les méthodes existantes [60]) peuvent être décrites comme des instances spécifiques des différentes composantes. Ce cadre générique permet de travailler à un plus haut niveau d'abstraction que les systèmes de type boîte à outils. Les composantes peuvent être combinées et paramétrées de différentes manières ce qui conduit à une grande diversité d'explications. La richesse de cette combinatoire et la diversité des explications qui en résulte sont nécessaires pour répondre à la variété des besoins des utilisateurs.

Pour exploiter pleinement cette diversité, il faut parvenir à mettre en correspondance les besoins spécifiques des utilisateurs avec les explications les plus adaptées. Tous les utilisateurs, même les profanes doivent pouvoir efficacement interagir avec le système. Pour répondre à ces objectifs, nous proposons un protocole d'interaction à trois niveaux, appelés respectivement *contexte*, *exigences* et *options techniques* qui sont utilisés pour exprimer les besoins des destinataires de manière compréhensible :

1. Le *contexte* fournit des informations intelligibles par tous sur le profil du destinataire de l'explication et ses objectifs.
2. Les *exigences* caractérisent l'explication souhaitée, notamment sa forme, sa simplicité et sa généralité.
3. Les *options techniques* permettent de paramétrer le système générique d'explication ; elles incluent notamment l'étendue et le type d'échantillonnage à utiliser ainsi que les

objectifs et contraintes à appliquer pendant la phase de génération.

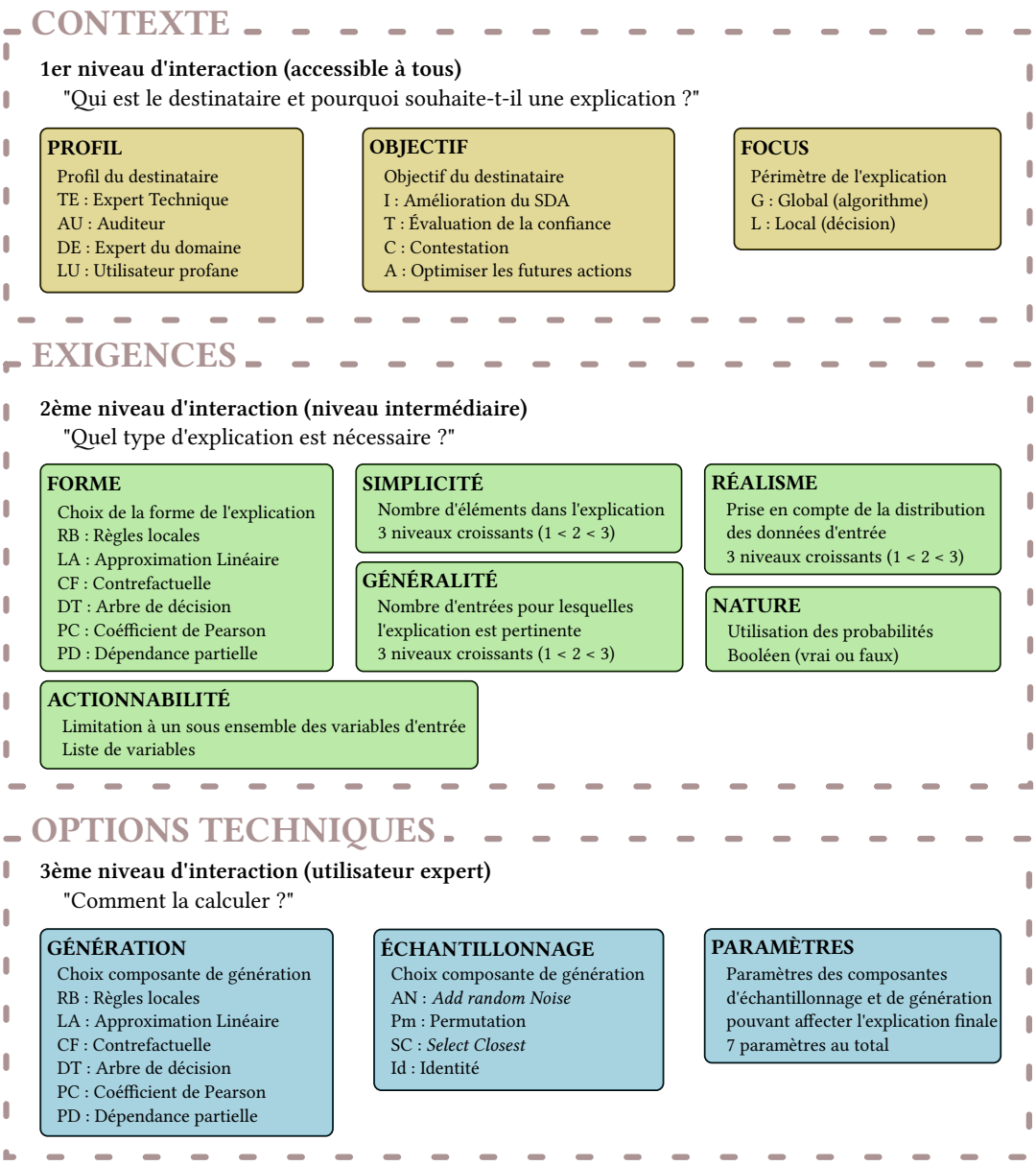


FIGURE 3.1 – Schéma récapitulatif des éléments constitutifs des trois niveaux d'interaction disponibles dans IBEX.

Ces trois niveaux et les différents éléments qui les constituent sont rappelés dans la Figure 3.1. La signification des différents éléments est détaillée plus précisément dans les parties suivantes. Schématiquement, ces trois niveaux pourraient être résumés de la manière

suivante : “Qui est le destinataire et pourquoi souhaite-t-il une explication ?”, “Quel type d’explication est nécessaire ?” et “Comment la calculer ?”. Les utilisateurs profanes devraient être en mesure d’exprimer leurs besoins au premier niveau, sans se soucier des exigences ni des options techniques. En revanche, un utilisateur expert, concepteur de SDA par exemple, pourra exprimer ses besoins directement sous la forme d’exigences ou même sous la forme d’options techniques. Quel que soit le niveau adopté par l’utilisateur pour exprimer ses besoins, ceux-ci doivent systématiquement être traduits en options techniques. Nous proposons dans ce chapitre, une méthode heuristique qui permet de traduire (1) les contextes en exigences et (2) les exigences en options techniques. Pour illustrer la correspondance entre les contextes et les exigences, un utilisateur profane dont le but est de contester une décision sera associé à des explications de préférence simples sous forme de contrefactuelles réalistes. En revanche, un utilisateur expert dont l’objectif est d’améliorer le SDA se verra suggérer, par défaut, des explications générales basées sur des règles.

Le rôle de ces correspondances est de permettre aux utilisateurs d’exprimer plus facilement leurs besoins et d’obtenir des explications appropriées. Cependant, si l’explication générée par le système ne répond pas à leurs attentes, les utilisateurs peuvent soit imposer un choix différent (par exemple, un utilisateur profane peut demander une explication sous forme de règles), soit interagir avec le système après la génération d’une première explication (par exemple, pour exprimer qu’ils souhaitent une explication plus simple ou plus générale).

Nous présentons les deux premiers niveaux (contexte et exigences) dans la partie 3.2 avant de décrire le système d’explication générique dans la partie 3.3. Dans la partie 3.4, nous décrivons les correspondances entre les différents niveaux : du contexte aux exigences, puis des exigences aux options techniques. Dans la partie 3.5, nous illustrons les différentes fonctionnalités d’IBEX à travers plusieurs études de cas.

3.2 Contexte et exigences

Cette partie décrit successivement les deux niveaux supérieurs d'interaction d'IBEX : le contexte (Partie 3.2.1) et les exigences (Partie 3.2.2). La correspondance entre ces niveaux est décrite dans la partie 3.4.

3.2.1 Contexte

Dans le protocole d'IBEX, le contexte représente le plus haut niveau d'interaction, c'est-à-dire le niveau le plus éloigné des options techniques. Il doit rendre possible à tous les utilisateurs, y compris les profanes, l'expression des besoins dans un vocabulaire simple et accessible. Un contexte est composé du *SDA* en question et de quatre éléments liés à la demande d'explication : le *profil*, l'*objectif*, le *focus* et le *point d'intérêt*. Ces éléments sont présentés ci-dessous :

— Le *SDA* comprend l'algorithme de décision et, si elles sont disponibles, les données d'apprentissage ou des données historiques d'utilisation.

— Le *profil* caractérise le profil du destinataire parmi l'une de ces quatre catégories $\{TE, AU, DE, LU\}$. TE représente les experts techniques, AU les auditeurs, DE les experts du domaine et LU les utilisateurs profanes. Les experts techniques sont les concepteurs, les développeurs, les testeurs, ou plus généralement toute personne ayant des connaissances sur la conception et la mise en œuvre de SDA. Les auditeurs sont également supposés avoir un haut niveau d'expertise dans la réalisation des tâches spécifiques d'audit ou d'évaluation de SDA. Les experts de domaine ne sont pas censés avoir une expertise sur le SDA lui-même ou sur la technologie utilisée, mais ils connaissent le domaine d'application. Les médecins, les juges ou les officiers de police sont des exemples d'experts de domaine. La dernière catégorie, celle des utilisateurs profanes, comprend ceux qui ne sont pas censés posséder de connaissances spécifiques. Il peut s'agir de personnes concernées par des décisions reposant sur le SDA ou de simples citoyens [1].

1. D'autres taxonomies de profils de destinataires ont déjà été proposées dans des travaux antérieurs, notamment dans [111] et [12]. Cette contribution reprend globalement ces propositions en les simplifiant pour des raisons pragmatiques de conception de l'outil.

— L'*objectif* caractérise l'objectif du destinataire de l'explication parmi quatre valeurs possibles $\{I, T, C, A\}$. I (*Improve*) représente l'amélioration du SDA, T (*Trust*) est l'évaluation du niveau de confiance pouvant être placé dans le système, C la Contestation d'une décision et A (*Action*) la prise d'action pour optimiser de futures interactions avec le système (par exemple pour augmenter les chances d'obtenir un crédit à la banque). L'amélioration du SDA comprend son test, l'évaluation de sa précision, la détection des biais et des erreurs et toute action visant à détecter les faiblesses potentielles. L'évaluation de la confiance présente des similitudes avec l'amélioration, mais fait l'hypothèse que le système a déjà été testé et l'objectif est donc plutôt de confirmer qu'il se comporte comme prévu. L'évaluation de la confiance comprend une variété d'objectifs liés à l'utilisation du SDA (éviter les mauvaises décisions [48], améliorer l'acceptation des résultats [48], augmenter la prévisibilité des résultats [84] et apprécier les limites du SDA [116]) ou liés aux connaissances apportées par l'algorithme dans son domaine d'application (causalité, transférabilité [78, 12]). La contestation d'une décision et l'optimisation des futures interactions sont deux réactions alternatives pour la personne affectée par une décision [115]. La deuxième vise à informer sur les actions qui peuvent avoir un impact sur le dossier de la personne (ses données d'entrée) et donc sur d'éventuelles décisions futures reposant sur le SDA.

— Le *Focus* caractérise le périmètre de l'explication. Il prend une valeur dans l'ensemble $\{G, L\}$. G représente une explication globale et L une explication locale. Une explication globale concerne le comportement du SDA pour l'ensemble des données d'entrée alors qu'une explication locale vise le comportement du SDA pour une entrée spécifique (ou autour d'une entrée spécifique).

— Le *Point d'intérêt* définit le point x de l'espace d'entrée qui est le point d'intérêt du destinataire lorsque le focus est local (dans le cas contraire, le point d'intérêt est vide).

La Figure 3.1 offre une vision synthétique du contexte. Certains de ces éléments peuvent être omis par l'utilisateur, le seul élément obligatoire étant le *SDA*. Une explication par défaut sera alors proposée et l'utilisateur pourra interagir à partir de celle-ci afin d'affiner l'expression de ses besoins (par exemple en demandant une explication plus simple ou plus

complète).

3.2.2 Exigences

Les exigences fournissent un niveau d'abstraction intermédiaire. Alors que le contexte caractérise la demande du destinataire, les exigences caractérisent les propriétés requises de l'explication. Elles peuvent être spécifiées directement par des utilisateurs experts ou par certains utilisateurs profanes, en fonction de leur niveau de compréhension. Les exigences sont constituées de sept éléments [2] : la *forme*, la *simplicité*, la *généralité*, le *point d'intérêt*, le *réalisme*, l'*actionnabilité* et la *nature*. En dehors du *réalisme*, qui est une contribution de cette thèse, l'introduction de chacun de ces éléments est motivée par des travaux antérieurs et des études expérimentales, comme mentionné ci-dessous :

— La *forme* correspond simplement à la forme finale de l'explication (règle, graphique, contrefactuelle, etc.) [48, 78]. Le rôle de la forme de l'explication sur son acceptation est analysé dans [106, 77]. Ici, la forme de l'explication doit être choisie parmi les propositions suivantes : les explications basées sur des règles (RB), les approximations linéaires locales (LA), les explications contrefactuelles (CF), les arbres de décision (DT), les coefficients de corrélation de Pearson (PC) et les diagrammes de dépendance partielle (PD). La liste complète des formes implémentées dans la version actuelle d'IBEX est fournie dans la partie 3.3.4.

— La *simplicité* caractérise le niveau de simplicité requis pour une explication. Il s'agit d'une exigence clé, car elle est généralement liée à l'intelligibilité de l'explication [48, 84, 72]. La version actuelle d'IBEX propose trois niveaux croissants de simplicité : $simplicité \in \{1, 2, 3\}$.

— La *généralité* caractérise le niveau de généralité requis pour les explications locales, c'est-à-dire la taille de l'ensemble des entrées pour lesquelles l'explication est pertinente ([84] p.44). Certains auteurs utilisent le mot "*cover*" pour désigner le même concept [72, 97]. La version actuelle d'IBEX considère trois niveaux croissants de généralité : $généralité \in \{1, 2, 3\}$. Il faut noter que la généralité est définie uniquement

2. En plus du SDA, tel que défini dans le contexte.

pour les explications locales puisque les explications globales couvrent, par définition, l'ensemble des données d'entrée.

— Le *point d'intérêt* a la même définition que celle présentée par la partie sur le contexte (Partie 3.2.1). Comme la généralité, le point d'intérêt est défini uniquement pour les explications locales.

— Le *réalisme* caractérise le niveau de réalisme requis pour une explication. Le terme "réalisme", introduit dans cette thèse, fait référence à la prise en compte de la distribution des données d'entrée lors du processus de production de l'explication. Les explications réalistes sont préférables pour les destinataires qui s'intéressent à l'utilisation réelle du SDA. En revanche, les personnes intéressées par la logique interne du SDA, indépendamment de son utilisation réelle, peuvent préférer une explication "théorique" qui ne dépend pas de la distribution des données d'entrée. Cette notion est illustrée à l'aide de l'exemple d'un système de crédit qui émettrait systématiquement le risque maximum lorsque le dossier de demande mentionne une fraude antérieure au crédit. Bien que cette caractéristique ait un impact énorme sur la décision, elle est rarement utilisée en pratique, car peu de demandeurs de crédit se trouvent dans cette situation. La démarche réaliste tient compte de la probabilité de rencontrer cette caractéristique lors du calcul de l'explication, et lui attribue donc une importance plutôt faible. À l'inverse, la démarche "théorique" ne tient compte que du modèle lui-même, et lui attribue donc une importance élevée. La version actuelle d'IBEX considère trois niveaux croissants de réalisme : $réalisme \in \{1, 2, 3\}$.

— L'*actionnabilité* exprime le fait que les explications pouvant mener à des actions doivent être privilégiées. Une explication actionnable est une explication impliquant uniquement des variables sur lesquelles l'utilisateur du SDA peut agir ([115] p.42). Par exemple, dans le fichier d'entrée d'un demandeur de prêt, la variable âge n'est pas actionnable alors que le nombre de prêts en cours l'est. La version actuelle d'IBEX considère deux options : $actionnabilité \in \{T, F\}$. La valeur T signifie que l'actionnabilité est une exigence. Dans ce cas, le destinataire doit fournir la liste des variables qu'il considère actionnables.

— la *nature* correspond à la possibilité de recourir aux probabilités dans les explications ([84] p.44). La version actuelle d'IBEX considère deux options : $nature \in \{T, F\}$. La valeur F signifie que les explications probabilistes ne sont pas souhaitées et la valeur T qu'elles sont acceptables.

Comme pour le contexte, les exigences peuvent être définies partiellement. En outre, elles peuvent être exprimées en termes de préférences plutôt que de contraintes. Par exemple, un expert technique peut caractériser la simplicité par $3 > 2 > 1$ pour exprimer le fait qu'il préfère les explications simples, mais peut également se satisfaire d'explications plus complexes. À l'inverse, un utilisateur profane peut préférer sélectionner uniquement la valeur 3 désignant la simplicité maximale. Dans la suite, une distinction est faite entre les "*exigences préférentielles*" et les "*exigences fortes*". Les exigences préférentielles doivent également être hiérarchisées (*i.e.* classées par ordre d'importance). Par exemple, un expert technique qui souhaite déboguer ou améliorer le SDA peut considérer que la généralité est plus importante que la simplicité (*généralité* > *simplicité*).

3.3 Système générique d'explication

Le contexte et les exigences présentés dans la partie précédente permettent aux destinataires d'exprimer leurs besoins sans avoir à considérer les moyens à mettre en œuvre pour obtenir l'explication appropriée. Nous décrivons maintenant le processus de production des explications répondant à ces besoins. Avant de définir les options techniques qui permettent d'adapter les explications aux différentes combinaisons de contextes et d'exigences, nous présentons l'architecture du système générique (Partie 3.3.1). L'implémentation des composantes d'échantillonnage et de génération, ainsi que leurs paramètres sont décrits respectivement dans les parties 3.3.2 et 3.3.3. Ces éléments permettent de définir les options techniques, dernier élément du protocole à trois niveaux d'IBEX, dans la partie 3.3.4. La correspondance entre le contexte, les exigences et les options techniques est décrite dans la partie 3.4.

3.3.1 Les deux composantes du système générique d'explication

Une grande variété de besoins peut être exprimée grâce au contexte et aux exigences décrits dans la partie précédente. Pour produire une gamme d'explications répondant à tous ces besoins, il faut concevoir un système d'explication finement paramétrable grâce à un ensemble suffisamment riche d'options techniques. Idéalement, ces options devraient pouvoir se combiner de manière à répondre indépendamment aux différents besoins des destinataires d'explications. L'architecture que nous proposons est dite générique au sens où de nombreuses BEM (y compris des méthodes existantes [60]) correspondent à des instances spécifiques de ce système.

Cette architecture repose sur une extension du cadre générique proposé dans la partie 2.2.2 et nous utilisons à nouveau l'exemple du classificateur de spams (SDA acceptant en entrée le texte d'un email et produisant en sortie la probabilité que cet email soit un spam) pour le présenter. Les deux composantes principales du système sont : (i) la sélection des données d'entrée à soumettre à l'algorithme, appelée l'*échantillonnage* ; et (ii) l'analyse des liens entre les échantillons et leurs classifications pour générer le contenu de l'explication, appelée la *génération*. Nous proposons une caractérisation formelle de la composante d'échantillonnage (Partie 3.3.2) et de la composante de génération (Partie 3.3.3) qui servent de base à la production d'explications répondant aux besoins de l'utilisateur exprimés sous forme de contextes ou d'exigences (Partie 3.3.4). Les principales notations utilisées dans cette partie et dans les suivantes sont regroupées dans le Tableau 3.1.

Nom	Description	Exemple
F	Modèle boîte noire	Classificateur de spams
X	Espace d'entrée de F	Ensemble des emails
Y	Espace de sortie de F	$[0, 1]$
E	Périmètre de l'explication	Email x_e
S	Ensemble des échantillons	Versions modifiées de x_e
Θ	Paramètres de l'échantillonnage	Partie de l'email modifiée
D	Jeu de données décrivant la population	Jeu de données d'entraînement de F

TABLE 3.1 – Principales notations

3.3.2 Échantillonnage

Le rôle de l'*échantillonnage* est de sélectionner les entrées appropriées (ou "échantillons") pour répondre à une question sur le modèle F. Cette phase doit prendre en compte un certain nombre de facteurs. Le premier aspect est le périmètre de l'explication appelé E [3]. Si la question porte sur une seule entrée x_e, alors $E = \{x_e\}$; si la question porte sur l'ensemble du modèle F, alors $E = D$ avec D un multiensemble [4] représentant la population des entrées possibles de F. Celui-ci est supposé être mis à disposition du système générique d'explication. Dans le cas général, E et D peuvent être n'importe quel (multi)sous-ensemble des entrées possibles. X, l'ensemble des valeurs d'entrée, peut être considéré comme le support du multiensemble D. Dans l'exemple du classificateur de spams, X est l'ensemble de tous les emails possibles (c'est-à-dire l'ensemble de tous les textes d'un certain format) et D est une base de données réelles d'emails dont dispose le système générique d'explication et qui peut être utilisée pour estimer la distribution des données d'entrée de F. Un exemple typique de D est le jeu de données d'entraînement ou le jeu de données de test utilisés pendant le processus d'apprentissage. D peut aussi être un ensemble de données historiques accumulées pendant l'utilisation du modèle. Lorsque le système générique d'explication ne dispose d'aucune information sur cette distribution, D est l'ensemble vide ($D = \emptyset$).

Le résultat de l'*échantillonnage* est un ensemble d'échantillons $S = \{x_1, ..., x_n\} \in X^n$. Par exemple, pour répondre à une question concernant l'impact de la signature sur la classification d'un email particulier x_e, une possibilité est de créer un échantillon en supprimant la signature de x_e. Cette stratégie ne nécessite aucune information sur la distribution réelle de la population et peut donc être appliquée même si $D = \emptyset$. Cependant, la réponse obtenue grâce à cet échantillon peut ne pas être suffisamment réaliste ou précise. Une stratégie plus élaborée consiste à remplacer la signature originale de x_e par les signatures d'emails réels. Cette stratégie nécessite des informations sur la distribution réelle de la population ($D \neq \emptyset$) afin de garantir que l'ensemble d'échantillons reflète la distribution réelle des emails. De

3. Le périmètre E est lié à l'élément focus du contexte. Des précisions sur cette correspondance sont apportées dans la partie 3.4.
4. D est un multiensemble, car il peut contenir plusieurs occurrences de la même entrée pour refléter la distribution des valeurs dans la population réelle.

manière générale, la procédure d'échantillonnage peut être définie de la manière suivante :

$$S = \{h_\theta(x_e, x_p) \mid (\theta, x_e, x_p) \in \Theta \times E \times D, Z(\theta, x_e, x_p)\} \tag{3.1}$$

avec :

$$h_\theta : E \times D \to X. \tag{3.2}$$

Θ est l'ensemble des paramètres de l'échantillonnage, Z est une fonction de filtrage et h_θ définit la manière dont les échantillons sont générés [5]. En bref, le paramètre θ permet de générer plusieurs échantillons pour chaque paire (x_e, x_p) tandis que Z permet de filtrer une sélection de paires (x_e, x_p). Dans l'exemple du filtre anti-spams, E contient un seul email ($E = \{x_e\}$). Les emails sont représentés par le contenu de leurs différentes parties (en-tête, corps, signature, ...) et $h_\theta(x_e, x_p)$ est une version modifiée de x_e obtenue en remplaçant une partie de x_e par la partie correspondante dans un élément $x_p \in D$. La partie à remplacer est spécifiée par θ. Par exemple, en prenant $\Theta = \{(SIG)\}$ et en supposant que D contient 1000 emails, la procédure d'échantillonnage génère 1000 versions perturbées de x_e avec des signatures extraites d'emails de D. Le rôle du paramètre θ est donc de personnaliser la fonction d'échantillonnage. Par exemple, pour comparer l'impact de l'en-tête et de la signature de l'email dans la classification, θ pourrait être utilisé pour spécifier la partie de l'email qui doit être remplacée (en-tête, signature ou les deux). Avec $\Theta = \{(HDR), (SIG), (HDR, SIG)\}$, la procédure d'échantillonnage générerait 3000 versions de x_e avec l'en-tête, la signature ou les deux remplacées par les parties correspondantes d'autres emails dans D. Une autre possibilité offerte par la définition (3.1) consiste à utiliser une fonction de filtrage Z pour ne sélectionner, par exemple, que les emails ayant beaucoup de mots en commun avec x_e.

Pour rendre la présentation plus concrète, nous présentons trois stratégies d'échantillonnage locales qui sont des instances de la définition (3.1). Ces stratégies ont été implémentées dans IBEX (Partie 3.5). Dans le premier exemple, appelé "*Select Closest*" (SC), la fonction de filtrage Z est utilisée pour sélectionner le sous-ensemble des élément de D les plus proches de x_e en comparant la distance $d(x_e, x_p)$ à un seuil prédéfini r. Dans ce cas, h^{close} sélectionne

5. Si Θ, D ou E sont vides, ils sont fixés à $\{0\}$ dans (3.1), sinon S serait également l'ensemble vide.

la valeur d'entrée sans la modifier (cf. Figure 3.2b).

$$S = \{h_\theta^{close}(x_e, x_p) = x_p \mid (\theta, x_e, x_p) \in \{r\} \times E \times D, d(x_e, x_p) < \theta\} \tag{3.3}$$

Dans la stratégie "*Select Closest*", les échantillons appartiennent à l'ensemble D ce qui est adapté pour l'obtention d'explications réalistes ($réalisme = 3$). Le nombre d'échantillons et leur proximité avec x_e peuvent être ajustés grâce au paramètre r.

La deuxième stratégie, appelée "*Permutation*" (Pm) combine deux entrées pour en obtenir une nouvelle. La fonction d'échantillonnage suivante :

$$\forall i, h_\theta^{perm}(x_e, x_p)[i] = x_e[i] \text{ si } i \in \theta \text{ sinon } x_p[i], \text{ avec } x_e \in E, x_p \in D \tag{3.4}$$

combine les variables de x_e avec les variables de x_p ($x[i]$ désigne la $i^{ième}$ variable de x) pour obtenir une nouvelle entrée. Le paramètre θ définit l'origine de chaque variable (cf. Figure 3.2c). Le paramètre θ est généré aléatoirement de telle sorte que chaque variable provienne de x_p avec une probabilité p, paramètre de l'échantillonnage. Le calcul des valeurs de Shapley dans [118] ou la génération des explications locales sous forme de règles dans *Anchors* [97] sont basés sur des stratégies d'échantillonnage de ce type. Dans les échantillons, chaque variable prise indépendamment est tirée de la distribution empirique de X et les variables incluses dans le même θ doivent également satisfaire les corrélations existant dans D. L'échantillonnage par permutation représente un niveau intermédiaire de réalisme ($réalisme = 2$).

Enfin, "*Add random Noise*" (AN), génère des échantillons en ajoutant une certaine quantité de bruit aléatoire à x_e. Les échantillons sont alors des versions bruitées de x_e, le bruit étant tiré d'une distribution normale de moyenne nulle (cf. Figure 3.2d).

$$h_\theta^{noise}(x_e, x_p) = x_e + \theta, \text{ avec } \theta \sim \mathcal{N}(0, \sigma^2) \tag{3.5}$$

L'échantillonnage AN n'utilise pas l'information contenue dans la population D et les variables ne sont pas corrélées. Le paramètre σ représente l'écart-type du bruit ajouté : une valeur faible de σ conduit à des échantillons proches de x_e tandis que des valeurs plus élevées

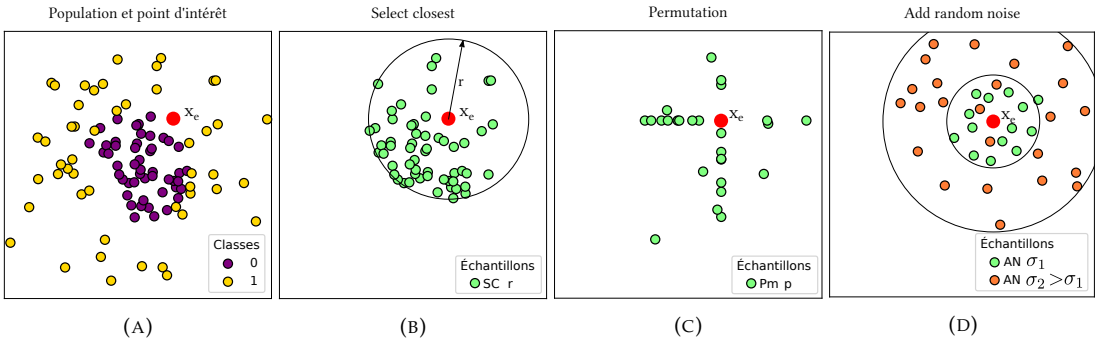


FIGURE 3.2 – Représentation schématique des processus d'échantillonnage locaux avec des variables continues en deux dimensions. Le point d'intérêt de l'explication (ou périmètre) est le point rouge x_e. (a) Population et point d'intérêt pour un problème de classification binaire. Les couleurs représentent les classes. (b) Échantillonnage "*Select closest*" avec un seuil r. Les échantillons sont les éléments de la population se situant à l'intérieur de la sphère de rayon r centrée sur le point d'intérêt. (c) Échantillonnage par "*permutation*" avec une probabilité p. Les échantillons sont des versions modifiées du point d'intérêt avec une ou deux variables issues de la distribution empirique. (d) Échantillonnage "*add random noise*" avec σ_1 et $\sigma_2 > \sigma_1$. Les échantillons sont des versions bruitées du point d'intérêt.

génèrent des échantillons dans un espace plus large (cf. les deux cercles de la Figure 3.2d). L'échantillonnage "*add random noise*" conduit à des explications qui ne sont pas soumises à des exigences de réalisme ($réalisme = 1$).

3.3.3 Génération

L'ensemble S d'échantillons et le modèle F (le SDA) sont les entrées du processus de génération d'explications. Ces deux éléments contiennent toute l'information nécessaire pour calculer l'explication, mais sous une forme inintelligible par l'utilisateur. L'objectif de la génération est d'extraire, à partir des échantillons et de leurs classifications, une explication compréhensible (et adaptée) pour l'utilisateur. Quelle que soit la forme finale de l'explication, le processus de génération peut être représenté comme le calcul d'un *proxy* du modèle F suivi par la construction d'une explication basée sur ce proxy quand le proxy lui-même n'est pas considéré comme une explication [6].

6. Un proxy prenant, par exemple, la forme de règles peut être directement fourni comme explication.

Dans l'exemple du classificateur de spams, une possibilité pour la composante *génération* est d'entraîner un modèle à base de règles mimant la classification des échantillons par le SDA. Par exemple, le modèle à base de règles pourrait prendre la forme suivante : "si la signature de l'email fait moins de 60 caractères, alors le classificateur considère qu'il s'agit d'un spam ; sinon, il considère l'email comme acceptable". Ces règles étant facilement interprétables, elles peuvent être utilisées directement comme explications. Dans d'autres situations, une étape supplémentaire doit être ajoutée pour obtenir l'explication à partir du modèle proxy. Cette phase peut retourner, par exemple, la ou les variables les plus importantes ou une version synthétique du proxy. Pour le classificateur de spams, l'explication pourrait par exemple être : "La longueur de la signature et le nombre de fautes de frappe sont les deux caractéristiques les plus importantes utilisées par le SDA pour décider si un email est un spam".

Le modèle proxy est désigné par f_w, une fonction du même type que le modèle F paramétré par w :

$$f_w : X \to Y \tag{3.6}$$

L'objectif de la génération est de trouver le meilleur proxy f_w pour répondre à la question du destinataire, ce qui revient à trouver les valeurs optimales de w. L'optimalité peut être définie formellement à l'aide de contraintes $o_i(w, S) \in \mathbb{B}$ et de critères $c_i(w, S) \in \mathbb{R}$ où $\mathbb{R}$ et $\mathbb{B}$ représentent respectivement l'ensemble des nombres réels et l'ensemble des booléens. L'objectif global prend la forme suivante :

$$\begin{aligned} w^* = \underset{w}{\operatorname{argmin}} \quad & \sum_i \lambda_i c_i(w, S) \\ \text{tel que} \quad & o_i(w, S) \end{aligned} \tag{3.7}$$

où $\lambda_i \in \mathbb{R}$ sont utilisés pour pondérer les critères.

Dans de nombreuses méthodes, l'objectif est de trouver les paramètres w tels que le proxy f_w soit le plus proche possible de F sur les éléments de S (échantillons). Généralement, trouver une bonne explication est une question de compromis. Un exemple typique est l'équilibre entre la précision et la complexité —souvent utilisée comme une estimation

de l'intelligibilité. Par exemple, une explication simple du classificateur de spams prenant la forme de règles qui prédirait correctement seulement 70 % des échantillons pourrait ne pas être acceptable ; à l'extrême opposé, une explication précise à 99 %, mais qui s'étalerait sur plusieurs pages de règles serait illisible et donc inutile pour le destinataire.

La distinction entre les critères et les contraintes offre une certaine souplesse dans la génération. Elle est d'ailleurs déjà utilisée implicitement par des méthodes de la littérature. Par exemple, *Anchors* [97], fixe une *contrainte* sur la précision du modèle à base de règles (la précision ne peut pas être inférieure à un seuil défini) et préconise que les explications soient très précises, tandis que BETA [72] fixe la précision du modèle à base de règles comme simple *critère* et préconise que les explications soient avant tout interprétables (contrainte).

Pour rendre la présentation plus concrète, nous présentons trois stratégies de génération locales, instances de la définition (3.7). Ces stratégies ont été implémentées dans IBEX (Partie 3.5). Le premier exemple est la génération d'un "modèle à base de règles" (RB) f_w où w représente l'ensemble des règles. Une possibilité est d'utiliser le nombre de règles comme critère et la précision du modèle comme contrainte, comme cela est fait dans [97]. Ce choix est représenté par la minimisation suivante :

$$\begin{aligned} w^* = \underset{w}{\text{argmin}} \quad & ||w|| \\ \text{tel que} \quad & \#\{x \in S, f_w(x) = F(x)\}/\#S > a \end{aligned} \tag{3.8}$$

avec $||w||$ le nombre de règles, # désignant le cardinal et a la précision minimale.

Le deuxième exemple emploie une "*Approximation linéaire locale*" (LA) du modèle, comme dans [98]. Dans ce cas, f_w est un modèle linéaire $f_w(x) = \sum_i w_i x[i]$ et la minimisation prend la forme suivante :

$$w^* = \underset{w}{\text{argmin}} \quad \lambda||w|| + \sum_{x \in S} (f_w(x) - F(x))^2 \tag{3.9}$$

ce qui équivaut à une régression de Lasso. Les coefficients linéaires obtenus fournissent des informations sur le comportement local du SDA. Plus précisément, en comparant leurs valeurs, il est possible d'estimer l'impact de la modification d'une variable sur la sortie du

modèle. Cette information peut facilement être convertie pour estimer l'importance relative des variables d'entrée sur la sortie du modèle [79].

Enfin, comme proposé dans [115], l'étape de génération peut être utilisée pour trouver une "*contrefactuelle*" (CF), qui peut être exprimée de la manière suivante :

$$\begin{aligned} w^* = \quad & \underset{w \in \{x - x_e, x \in S\}}{\operatorname{argmin}} \quad ||w|| \\ \text{tel que} \quad & f_w(x_e) \neq F(x_e) \end{aligned} \tag{3.10}$$

avec $f_w(x) = F(x + w)$ et $||w||$ désignant la distance entre $x + w$ et x. Suivant la définition de [115], une contrefactuelle est l'entrée la plus proche du point d'intérêt pour laquelle le SDA renvoie une sortie différente. Dans la formulation proposée, la différence entre le point d'intérêt et la contrefactuelle est appelée w ; elle doit être aussi faible que possible. L'équation (3.10) implique un critère de minimalité sur la norme de w, qui représente la distance entre la contrefactuelle et le point d'intérêt, et une contrainte sur la sortie du SDA pour cette contrefactuelle.

3.3.4 Les options techniques

Dans les parties précédentes, nous avons présenté les deux composantes de l'architecture du système générique d'explication, l'*échantillonnage* et la *génération*. L'architecture générique du système d'explication complet permet de le paramétrer de différentes manières. Les choix du type de composante (type d'échantillonnage et type de génération) et des paramètres constituent les *options techniques* qui peuvent être mises à profit pour générer une diversité d'explications. Dans cette partie nous résumons l'ensemble des options techniques disponibles pour les stratégies d'échantillonnage et de génération introduites dans les parties précédentes. La partie 3.4 se focalise sur la correspondance entre ces options et les niveaux d'interaction les plus abstraits.

Les composantes d'échantillonnage et de génération actuellement disponibles dans l'implémentation d'IBEX sont présentées de manière synthétisée avec leurs paramètres respectifs dans le Tableau 3.2. En considérant que, pour les explications locales, les phases d'échantillonnage et de génération sont indépendantes, il existe neuf combinaisons possibles et donc neuf instances différentes du système générique d'explication. Pour les explications globales, trois options supplémentaires sont possibles, ce qui donne un total de douze combinaisons. Le deuxième volet des options techniques concerne le choix de paramètres qui sont propres à chaque instance (cf. Tableau 3.2).

Nom	Composante	Focus	Param.	Description
Add random Noise (AN)	Échantillonnage	Local	σ	Ajout d'un bruit gaussien au point d'intérêt
Permutation (Pm)	Échantillonnage	Local	p	Permutation des variables du point d'intérêt et de la population
Select Closest (SC)	Échantillonnage	Local	r	Sous-population la plus proche du point d'intérêt
Identité (Id)	Échantillonnage	Global	$\emptyset$	Population
Replace with Constant (RC)	Échantillonnage	Global	α	Remplacement des valeurs d'une variable par une constante α
Modèle à base de règles (RB)	Génération	Local	a	Règles précises et simples
Approximation linéaire locale (LA)	Génération	Local	λ	Régression de Lasso
Contrefactuelle (CF)	Génération	Local	$\emptyset$	Entrée la plus proche classée différemment
Arbre de décision (DT)	Génération	Global	a_{DT}	Arbre de décision (échant. : Id)
Coefficients de Pearson (PC)	Génération	Global	$\emptyset$	Coefficients de corrélation entrées / sortie (échant. : Id)
Dépendance Partielle (PD)	Génération	Global	$n^{(i)}$	Valeur moyenne de la sortie sur la population pour chaque valeur de chaque variable (échant. : RC)

TABLE 3.2 – Options techniques : les différentes composantes et leurs paramètres respectifs. (i) n est le nombre d'intervalles à utiliser pour l'histogramme.

Pour illustrer la liberté offerte par les combinaisons des différentes instances d'échantillonnage et de génération, nous utilisons l'exemple des contrefactuelles (CF). Une contrefactuelle obtenue en utilisant l'échantillonnage réaliste *Select Closest* correspond à l'email de la population D le plus proche du point d'intérêt qui est classé différemment du point d'intérêt[7]. Ce type d'explication peut être utile pour un expert du domaine qui souhaite évaluer la confiance qu'il peut placer dans le SDA. En revanche, une contrefactuelle obtenue avec un échantillonnage non réaliste ne ressemble pas nécessairement à un email réel. Par exemple,

7. La notion de proximité entre emails peut par exemple reposer sur le nombre de mots communs.

si la stratégie d'échantillonnage consiste à ajouter des caractères aléatoires à l'email d'origine, les échantillons obtenus et donc l'explication finale ressemblent à des emails contenant de nombreuses fautes de frappe. Ce type de contrefactuelle convient mieux aux experts techniques qui s'intéressent au comportement "théorique" du SDA (même pour des entrées très éloignées de la distribution réelle).

Le Tableau 3.3 fournit un aperçu des paramètres des composantes d'échantillonnage et de génération. Ces paramètres associés à chaque composante permettent d'affiner la recherche d'explications. Par exemple, la valeur de σ (Définition (3.5)) a un impact sur la distance moyenne entre les échantillons et le point d'intérêt, grandeur appelée *étendue* de l'échantillonnage. Les explications obtenues avec de grandes valeurs de σ sont susceptibles d'être plus générales que les explications avec de petites valeurs de σ. De la même manière, la valeur du paramètre a (Définition 3.8) représente la précision minimale imposée lors de la recherche du modèle à base de règles. Ce paramètre peut être utilisé pour contrôler la simplicité de l'explication résultante.

Composante	Paramètre	Description
Add random noise	σ	Écart-type moyen du bruit
Permutation	p	Probabilité d'une inversion de valeur
Select closest	r	Distance à l'échantillon le plus éloigné
Modèle à base de règles	a	Précision minimale
Approximation linéaire locale	λ	Terme de régularisation (Lasso)
Arbre de décision	a_{DT}	Précision minimale
Graphe de dépendance partielle	n	Nombre d'intervalles dans l'histogramme

TABLE 3.3 – Liste des paramètres techniques

3.4 Du contexte aux explications

Dans les parties précédentes, nous avons proposé un protocole comportant trois niveaux d'interaction (contexte, exigences et options techniques). Chaque niveau est indépendant des autres afin que différents types d'utilisateurs, en fonction de leur expertise et de leur besoin, puissent utiliser celui qui leur convient le mieux sans avoir à connaître ou à comprendre les autres. Toutefois, quel que soit le niveau employé, les choix de l'utilisateur doivent être

traduits en options techniques nécessaires au paramétrage du système générique d'explication présenté dans la Partie 3.3.1. Dans cette partie, nous présentons les deux phases de ce processus : la traduction du contexte en exigences (Partie 3.4.1) et la traduction des exigences en options techniques (Partie 3.4.2). Ce processus aboutit à plusieurs possibilités pour l'ensemble des options techniques fournies au système générique d'explication. Chacune de ces possibilités correspond à une explication dont les propriétés exactes ne peuvent pas être connues à l'avance. La partie 3.4.3 décrit une étape finale d'évaluation *post-hoc* des explications pour sélectionner la réponse la plus adaptée. L'ensemble du processus est représenté schématiquement dans la Figure 3.3.

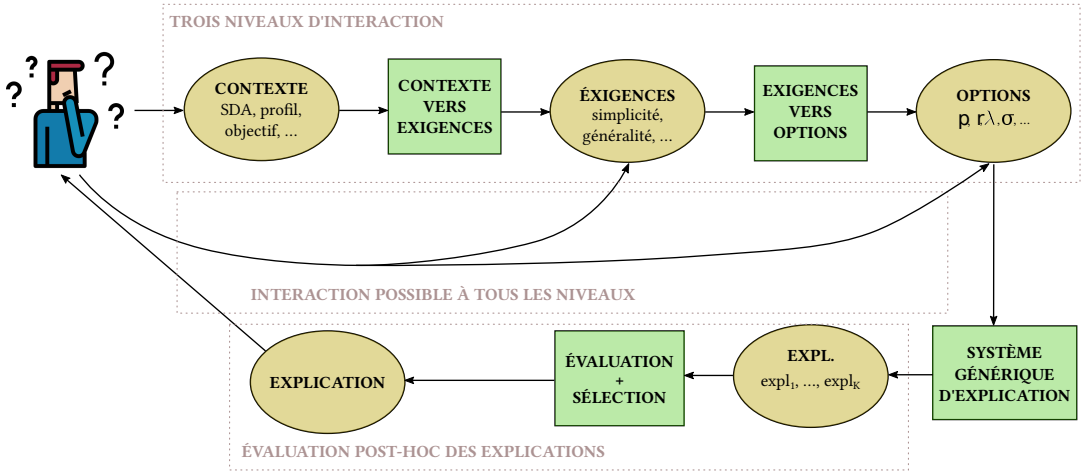


FIGURE 3.3 – Vue d'ensemble de la démarche à trois niveaux.

3.4.1 Du contexte aux exigences

Le contexte, défini dans la partie 3.2, est composé de quatre éléments[8] : *profil*, *objectif*, *focus* et *point d'intérêt*. Ces éléments doivent être convertis en exigences composées de sept éléments (Partie 3.2.2) : *forme*, *simplicité*, *généralité*, *point d'intérêt*, *réalisme*, *actionnabilité* et *nature*.

8. En plus du *SDA* qui se transmet simplement de niveau en niveau.

La partie 3.2 a introduit une distinction entre les *exigences fortes*, qui imposent des valeurs spécifiques (par exemple *généralité* = 1), et les *exigences préférentielles*, qui expriment une préférence plutôt qu'une contrainte (par exemple *généralité* : $3 > 2 > 1$). Lorsque plusieurs préférences sont exprimées, elles doivent être triées par ordre d'importance (par exemple *généralité* $>$ *simplicité*). D'un point de vue opérationnel, toutes les explications qui ne répondent pas aux exigences fortes sont d'abord filtrées, puis l'explication répondant au mieux aux exigences préférentielles parmi les explications restantes est proposée à l'utilisateur.

La première étape de la procédure de traduction consiste à utiliser l'élément *focus* du contexte pour filtrer les *formes* qui ne peuvent pas être utilisées. Par exemple, si *focus* = *G* (l'explication doit être globale), les explications contrefactuelles (*CF*) ne sont pas appropriées. Plus généralement, le *focus* associé à chaque forme est présenté dans le Tableau 3.2. Si *focus* = *L* (explication locale), alors l'élément *point d'intérêt* des exigences est obtenu directement à partir du contexte. Les autres éléments des exigences sont dérivés des éléments *profil* et *objectif* du contexte, comme présenté dans le Tableau 3.4.

En général, les explications simples sont préférées aux explications complexes ([84] p.44). La simplicité est exprimée comme une préférence de faible priorité, à moins que l'objectif ne soit l'évaluation de la confiance. L'exception notable à cette règle est le cas des utilisateurs profanes pour lesquels une exigence forte est nécessaire ($simplicité = 3$).

La généralité d'une explication (qui n'est pertinente que pour les explications locales) renforce la capacité du destinataire à comprendre les résultats du SDA pour des entrées similaires au point d'intérêt. Par conséquent, la généralité devrait être maximale ($généralité = 3$) lorsque l'objectif est d'évaluer la confiance qui peut être placée dans le modèle ([84] p.44). En revanche, lorsque l'objectif de l'utilisateur profane est de contester une décision ou d'optimiser de futures interactions, un niveau de généralité plus faible est mieux adapté. Dans la même situation, mais lorsque le destinataire est un expert du domaine ou un auditeur, un niveau intermédiaire de généralité est utilisé, car il peut vouloir comprendre la logique du SDA au-delà du point d'intérêt.

Un niveau élevé de réalisme favorise les explications qui trouvent une justification dans les données d'entraînement lorsque le SDA utilise du ML [73]. En fonction du contexte,

Expert technique		**Expert du domaine**		
*Improve**	*Trust*	*Trust**	*Contester*	*Action*
forme : RB >DT >LA >PD >PC >CF	forme : RB >DT >LA >PD >PC >CF	forme : RB >DT >LA >PD >PC >CF	forme : RB >DT >LA >PD >PC >CF	forme = CF
simplicité : 3 >2 >1	simplicité : 3 >2 >1	simplicité : 3 >2 >1	simplicité : 3 >2 >1	simplicité : 3 >2 >1
généralité : 3 >2 >1	généralité = 3	généralité = 3	généralité = 2	généralité : 1 >2 >3
réalisme = 1	réalisme : 3 >2 >1	réalisme = 3	réalisme = 1	réalisme = 2
actionnabilité = F	actionnabililté = F	actionnabilité = F	actionnabilité = F	actionnabilité = T
nature = T	nature = T	nature : T >F	nature : F >T	nature = F
general >forme >simple	simple >real >forme	simple >nat >forme	forme >nat >simple	simple >gen
Auditeur		**Profane**		
Trust	*Contester**	*Trust**	*Contester*	*Action*
forme : RB >DT >LA PD >PC >CF	forme : RB >DT >LA PD >PC >CF	forme : RB >DT >LA PD >PC >CF	forme : RB >DT >LA PD >PC >CF	forme = CF
simplicité : 3 >2 >1	simplicité : 3 >2 >1	simplicité = 3	simplicité = 3	simplicité = 3
généralité = 3	généralité = 2	généralité = 3	généralité = 1	généralité : 1 >2 >3
réalisme = 3	réalisme = 1	réalisme = 3	réalisme = 1	réalisme = 2
actionnabilité = F	actionnabilité = F	actionnabilité = F	actionnabilité = F	actionnabilité = T
nature : T >F	nature : T >F	nature : F >T	nature : F >T	nature = F
simple >nat. >forme	forme >nat. >simple	nat. >forme	forme >nat.	

TABLE 3.4 – Table de correspondance entre le contexte et les exigences. Les exigences fortes apparaissent en noir, les exigences préférentielles en vert. Les étoiles indiquent les objectifs par défaut utilisés lorsque l'utilisateur ne précise pas son objectif.

cela peut être un avantage ou un inconvénient. Les explications qui ne sont pas soutenues par des données d'entraînement permettent d'analyser des frontières de décision qui, bien qu'elles fassent partie du modèle, correspondent à des cas qui se retrouvent rarement dans les données réelles, comme mentionné dans l'exemple du système de crédit de la Partie 3.2.2. Lorsque l'objectif de l'explication est d'évaluer la confiance, les frontières de décision réellement utilisées doivent être la préoccupation principale, ce qui justifie l'échantillonnage réaliste proposé dans le Tableau 3.4. En revanche, les experts techniques peuvent vouloir étudier ces frontières de décision "théoriques" afin d'évaluer la robustesse du modèle pour des exemples hors de la distribution. Par ailleurs, un faible réalisme peut également être approprié pour remettre en question le SDA, car l'existence d'une frontière problématique (qui ne respecterait pas les lois anti-discrimination par exemple) pourrait être condamnable, même si elle n'est que rarement utilisée en pratique.

Enfin, pour optimiser de futures interactions avec le SDA, des contrefactuelles probables devraient être proposées en priorité. Cependant, imposer le réalisme maximal s'avérerait trop restrictif, car un tel niveau de réalisme suppose que la modification proposée conduise à une entrée qui a déjà été observée historiquement. Par conséquent, le niveau intermédiaire est utilisé. Aussi, les explications doivent se concentrer sur les variables actionnables afin de ne suggérer que des modifications faisables en pratique. Par conséquent, l'exigence *actionnabilité* est fixée à T lorsque l'objectif est *action* et à F autrement.

Comme l'ont montré des études antérieures ([84] p.44), l'utilisation de probabilités dans les explications n'est généralement pas éclairante, surtout lorsque l'explication est locale. Toutefois, certains profils, tels que les auditeurs et les experts techniques, peuvent être intéressés par une vision nuancée de la situation que l'usage des probabilités peut fournir.

Pour conclure, il faut souligner que les choix présentés dans le Tableau 3.4 peuvent être facilement ajustés dans l'implémentation d'IBEX. En effet, il paraît illusoire de proposer une correspondance qui s'appliquerait sans distinction à tout type d'applications. Chaque déploiement d'IBEX devra adapter au mieux cette proposition en fonction du contexte.

3.4.2 Des exigences aux options techniques

Nous décrivons maintenant la denière étape de l'interaction, c'est-à-dire la correspondance entre les six exigences et les options techniques du système[9] :

— **Forme :** l'élément forme est directement utilisé pour choisir l'instance de la composante de génération du système d'explication. Cependant, lorsque la *forme* est exprimée comme une exigence préférentielle, ce choix peut dépendre d'autres exigences, notamment la *nature*, comme cela est détaillé plus loin.

— **Simplicité :** la simplicité peut être contrôlée via la taille de l'explication (*i.e.* le nombre d'éléments qui la constituent) en utilisant les paramètres de la composante génération. Par exemple, les approximations linéaires locales utilisent le terme de pénalisation λ pour contrôler le nombre de coefficients non nuls : augmenter le poids de la pénalisation accroît donc la simplicité de l'explication résultante. La correspondance

9. Le *Point d'intérêt* se transmet sans modification des exigences aux options techniques.

pour les autres instances de la composante de génération est présentée dans le Tableau 3.5.

— **Généralité :** l'exigence de généralité a un impact sur l'étendue de l'échantillonnage, c'est-à-dire la distance entre le point d'intérêt et les échantillons. Une explication dérivée d'exemples proches du point d'intérêt (faible étendue) a peu de chances d'être générale. L'étendue de l'échantillonnage peut être contrôlée par les paramètres de l'échantillonnage. Par exemple, "*add random noise*" contrôle l'étendue avec le paramètre σ, écart-type moyen du bruit ajouté, comme le montre la Figure 3.2d. La correspondance des autres paramètres est présentée dans le Tableau 3.5.

— **Réalisme :** le réalisme d'une explication est entièrement déterminé par le choix de la composante d'échantillonnage. Les stratégies d'échantillonnage sont classées ci-dessous en fonction du niveau de ressemblance attendu entre les échantillons et les éléments de la population. La stratégie "*add random noise*", qui n'utilise pas la population, est associée à un faible niveau de réalisme ($réalisme = 1$), tandis que les stratégies "*select closest*" et "identité", qui génèrent des échantillons appartenant directement à la population, sont associées à un niveau élevé de réalisme ($réalisme = 3$). Les stratégies "Permutation" et "*replace with constant*" sont associées à un niveau intermédiaire ($réalisme = 2$).

— **Actionnabilité :** si l'utilisateur choisit de se concentrer sur les variables actionnables, les autres variables sont écartées afin de ne pas être prises en compte dans le calcul de l'explication finale.

— **Nature :** cette exigence limite les instances possibles pour la phase de génération en fonction de leur utilisation des probabilités. Par exemple, les explications sous forme de règles incluent une probabilité (celle qu'un échantillon soit correctement prédit), tandis que les explications contrefactuelles ne le font pas. Les composants impliquant des probabilités sont RB et DT.

Exigence	Composante	Param.	Effet
Simplicité	LA	λ	Simplicité élevée $\Longrightarrow$ λ élevé
	RB	a	Simplicité élevée $\Longrightarrow$ a faible
	DT	a_{DT}	Simplicité élevée $\Longrightarrow$ a_{DT} faible
	PD	n	Simplicité élevée $\Longrightarrow$ n faible
Généralité	AN	σ	Généralité élevée $\Longrightarrow$ σ élevé
	Pm	p	Généralité élevée $\Longrightarrow$ p élevé
	SC	r	Généralité élevée $\Longrightarrow$ r élevé

TABLE 3.5 – Impact des exigences sur les valeurs des paramètres d'explication

La procédure de correspondance présentée dans cette partie produit généralement plusieurs ensembles d'options techniques, en particulier en présence d'exigences préférentielles. La partie suivante présente une procédure *post-hoc* pour évaluer toutes les explications calculées et sélectionner, parmi ces candidates, celle qui répond le mieux aux exigences.

3.4.3 Evaluation *post-hoc* des explications

Comme aucune hypothèse ne peut être faite sur la régularité de la boîte noire, il n'est pas possible de prédire avec certitude les propriétés exactes d'une explication à partir des options techniques retenues pour les générer. Par exemple, si le destinataire souhaite une explication locale ayant un bon niveau de généralité et exprimée comme un modèle à base de règles, la forme et la distance de la frontière de décision la plus proche du point d'intérêt ont une influence sur le nombre de règles nécessaires pour atteindre une certaine précision. Pour surmonter ce problème et garantir que l'explication générée par IBEX répond aux exigences, il est nécessaire de recourir à une dernière étape d'évaluation. Cette étape consiste à calculer toutes les explications correspondant aux différents ensembles d'options techniques produits par l'étape précédente et à évaluer leurs propriétés de manière *post-hoc*.

L'évaluation des qualités des explications est une question de recherche complexe qui dépasse le cadre de ce chapitre. Ici, les explications sont évaluées simplement au regard de deux exigences (*simplicité* et *généralité*), les autres exigences ne nécessitant pas d'évaluation *post-hoc*, en suivant la méthode présentée ci-dessous. La simplicité est approximée par le nombre d'éléments de l'explication (par exemple le nombre de règles, le nombre de modifications dans une contrefactuelle, etc.). Il s'agit d'un indicateur de simplicité d'une explication [35]

qui présente des limites évidentes puisque la taille ne reflète pas toujours la simplicité. Cette définition de la simplicité a toutefois l'avantage d'être opérationnelle et facile à mettre en œuvre pour tout type d'explication ; elle est d'ailleurs fréquemment utilisée dans la littérature [42].

L'évaluation de la généralité repose sur une stratégie de test de l'explication sur les entrées de la population proches du point d'intérêt. Par exemple, pour une explication sous forme de règles, il suffit de vérifier que les règles prédisent la bonne sortie sur les proches voisins du point d'intérêt. Si l'explication n'est pas valide pour un nombre minimum d'entrées (les T_1 plus proches voisins), alors la généralité est de 1 ; si elle est valide pour les T_1 entrées les plus proches, mais pas pour les T_2 entrées ($T_2 > T_1$), alors la généralité est de 2 ; si elle est valide pour les entrées les plus proches T_2, alors la généralité est de 3 [10].

Enfin, il est important de souligner que la définition des besoins du destinataire (exprimés dans l'un des trois niveaux définis plus haut) n'est que la première étape de l'interaction. Lorsqu'une explication a été générée par IBEX sur la base de l'ensemble des options techniques résultant de cette étape initiale, le destinataire de l'explication peut réagir à cette explication avec une nouvelle demande. Cette demande peut se référer à l'explication initiale (par exemple, demander une explication "plus générale", ou "moins simple", ou une explication sous une forme différente) ou peut être entièrement nouvelle. En permettant aux destinataires d'interagir à plusieurs niveaux d'abstraction, IBEX leur donne la possibilité d'exprimer leurs besoins de manière très précise et de réagir aux explications précédentes.

3.5 Études de cas

Cette partie illustre le protocole d'interaction à trois niveaux et le système générique d'explication en présentant quelques exemples de l'application d'IBEX à plusieurs études de cas. IBEX met en œuvre la démarche présentée dans les parties précédentes et son code est disponible publiquement [11].

10. Les valeurs de T_1 et T_2 varient d'une application à une autre et peuvent être paramétrées.
11. https ://gitlab.inria.fr/chenin/ibex

IBEX permet des interactions à chaque niveau du protocole décrit dans la partie 3.4. Ainsi, la phase d'interaction commence par le choix du niveau (contexte, exigences ou options techniques) par l'utilisateur. Par défaut, l'interaction se fait au niveau "contexte", qui est le plus approprié pour les utilisateurs profanes. L'interaction initiale à ce niveau a pour but de caractériser les besoins du destinataire exprimés sous une forme simple (tels que définis dans la partie 3.2). En pratique, IBEX pose les questions suivantes :

1. Choisissez un ensemble de données (parmi les jeux de données proposés).
2. Êtes-vous intéressé par des explications globales (G) ou locales (L) ?
3. Quel est le point d'intérêt ? (pour les explications locales uniquement)
4. Comment voulez-vous être considéré par IBEX : comme un expert technique (TE), un utilisateur profane (LU), un expert du domaine (DE) ou un auditeur (AU) ?
5. Quel est l'objectif de votre demande d'explication : s'agit-il d'améliorer le SDA (I), d'évaluer la confiance que vous pouvez placer dans le SDA (T), de contester le SDA (C), ou d'optimiser de futures interactions avec le SDA (A) ?
6. Quelles sont les variables actionnables ? (pour l'objectif A uniquement)

L'utilisateur peut sauter toutes ces questions (sauf la première) s'il le souhaite et notamment s'il a un doute sur la réponse. Dans tous les cas, IBEX génère alors une première explication basée sur ce contexte (potentiellement partiel) et demande à l'utilisateur s'il souhaite poser d'autres questions. Si c'est le cas, l'utilisateur a deux options : il peut soit faire une nouvelle demande d'explication (nouvelle itération), soit réagir à l'explication précédente. Dans ce dernier cas, il doit interagir au niveau des exigences en demandant par exemple une explication plus simple, une explication plus générale ou une explication actionnable. Il peut également consulter les exigences de l'explication produite, et même les modifier pour en obtenir une nouvelle. Lorsque la nouvelle explication lui est proposée, l'utilisateur aura à nouveau la possibilité d'arrêter, de formuler une nouvelle question ou de réagir à l'explication actuelle.

Pour illustrer les avantages de la démarche en termes de polyvalence et d'interactivité, nous présentons l'utilisation d'IBEX dans les situations suivantes :

1. Un utilisateur profane demandant des explications sur un SDA (basé sur un réseau de neurones) fonctionnant à partir des données du *adult census dataset* [12] dans le but d'évaluer la confiance qui peut être placée dans le modèle (Partie 3.5.1).
2. Un expert du domaine demandant des explications sur un SDA (basé sur un réseau de neurones *long short term memory*) appliqué à un jeu de données textuelles dont le but est d'analyser les sentiments de tweets à propos de compagnies aériennes [13]. Son but est également d'évaluer la confiance qu'il peut accorder au SDA (Partie 3.5.2).
3. Un utilisateur profane demandant des explications sur un SDA (basé sur une forêt aléatoire) appliqué au jeu de données *german credit* [14] dans le but d'optimiser des interactions futures avec le SDA (Partie 3.5.3).
4. Un auditeur demandant des explications sur un SDA (basé sur un réseau de neurones) fonctionnant à partir des données du *adult census dataset* dans le but de contester le SDA (Partie 3.5.4).

Les trois jeux de données, accessibles en ligne, sont utilisés comme population. Leurs caractéristiques sont résumées dans le Tableau 3.6.

Jeu de données	Variables	Type	Sortie	Modèle
Adult census	Données démographiques (âge, niveau d'éducation, etc.)	Tabulaire	Revenus annuels ('<50k', '>=50k')	Réseau de neurones
German credit	Demande de crédit (montant, emploi actuel, etc.)	Tabulaire	Profil de risque ('*bad*', '*good*')	Forêt aléatoire
Twitter airline	Tweets sur des compagnies aériennes	Textuelles	Sentiment	LSTM

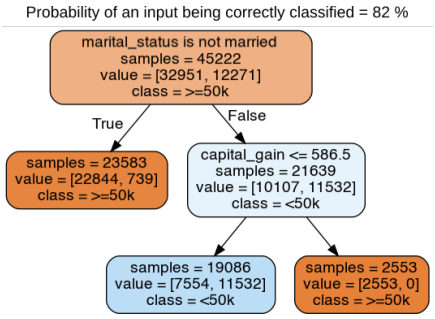
TABLE 3.6 – Jeux de données et modèles utilisés dans les cas d'étude

3.5.1 Utilisateur profane cherchant à renforcer la confiance

La première étude de cas concerne le jeu de données *adult census*. Ces données, extraites du recensement américain de 1994, contiennent des informations personnelles sur les citoyens américains, telles que leur âge, leur niveau d'éducation ou leur état civil. L'objectif du SDA est de prédire, sur la base de ces caractéristiques, si un individu gagne plus ou moins de 50 000 $ par an. Un utilisateur profane qui souhaite évaluer la confiance qu'il peut placer

12. https ://archive.ics.uci.edu/ml/datasets/Adult
13. https ://www.kaggle.com/crowdflower/twitter-airline-sentiment
14. https ://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

dans le SDA choisirait les réponses suivantes : *SDA=adult census*, *focus=G*, *profil=LU*, *objectif=T*. À partir de ce contexte, IBEX a généré l'explication présentée dans la Figure 3.4a. L'explication est simple, elle est composée d'un arbre de décision avec seulement deux nœuds et trois feuilles, ce qui est cohérent avec les choix présentés dans le Tableau 3.4.



(A)

```
EXIGENCES FORTES :
  actionnabilité = F
  simplicité = 3
  généralité = NA
  réalisme = 3
EXIGENCES PRÉFÉRENTIELLES :
  nature : F > T
  forme : DT > PD > PC
```

(B)

```
EXIGENCES FORTES :
  actionnabilité = F
  simplicité = 2
  généralité = NA
  réalisme = 3
  forme : DT
EXIGENCES PRÉFÉRENTIELLES :
  nature : NA
```

(C)

FIGURE 3.4 – (a) Explication générée par IBEX pour le jeu de données *adult census* à partir du contexte initial. IBEX a utilisé les exigences présentées dans la Figure 3.4b. (b) Exigences correspondant au contexte initial (G, LU, T). Les exigences préférentielles sont classées par ordre de priorité. (c) Exigences révisées après la demande "moins simple" de l'utilisateur.

Les exigences générées par IBEX pour ce contexte sont présentées dans la Figure 3.4b. L'exigence préférentielle $nature : F > T$ apparaît, ce qui signifie qu'une explication n'impliquant aucune probabilité aurait été préférée par l'utilisateur. Néanmoins, l'explication générée par IBEX est probabiliste. L'explication de ce choix d'IBEX est que les premières formes envisagées par IBEX (*PD* et *PC*) ont conduit à des explications considérées comme trop complexes pour satisfaire l'exigence forte $simplicité = 3$. Pour cette raison, à la suite de l'étape d'évaluation *post-hoc*, c'est l'explication sous forme d'un arbre de décision qui a été proposée.

L'utilisateur est presque satisfait de cette première explication, mais soupçonne la logique du SDA d'être en réalité plus complexe et l'explication proposée d'être trop simpliste. Grâce au système d'interaction, il peut demander une explication "moins simple". IBEX génère alors les exigences présentées dans la Figure 3.4c, puis l'explication présentée dans la Figure 3.5. Cette explication est effectivement moins simple que la précédente et elle fournit une description plus précise de la logique du SDA.

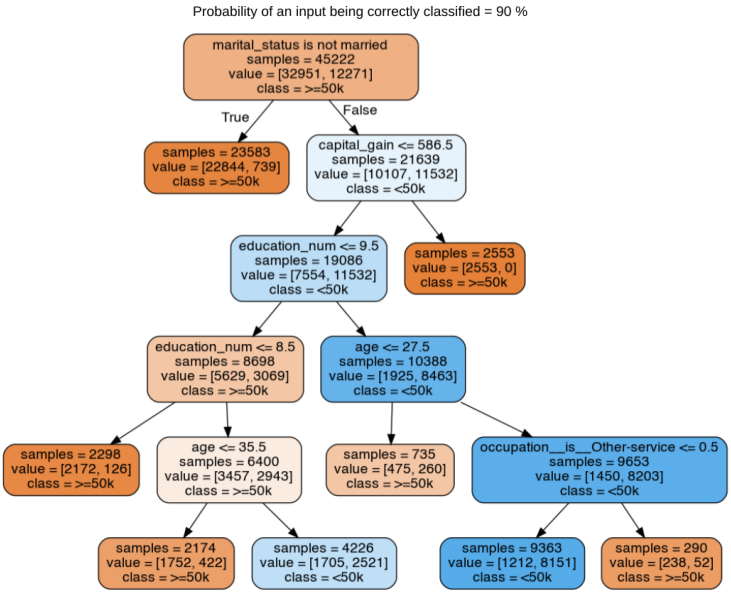


FIGURE 3.5 – Explication générée par IBEX pour le jeu de données *adult census* à partir des exigences révisées présentées dans la Figure 3.4c

3.5.2 Expert du domaine cherchant à renforcer la confiance

La deuxième étude de cas concerne les données *Twitter airline*. Ce jeu de données contient des tweets sur des compagnies aériennes et l'objectif du SDA est de les classer selon trois catégories d'émotion : négative, neutre ou positive. Les tweets négatifs sont censés exprimer des émotions négatives (colère, irritation, etc.), les tweets positifs sont censés exprimer des émotions positives (bonheur, gratitude, etc.) et les tweets neutres ne révèlent aucune ou peu d'émotion.

Supposons qu'un employé travaillant dans le service clientèle d'une entreprise souhaite mieux comprendre le SDA afin de l'utiliser plus efficacement. Cet employé utilise IBEX en tant qu'expert du domaine (*DE*), en demandant des explications sur des tweets spécifiques (L) dans le but d'évaluer la confiance qu'il peut placer dans le système (T).

Le Tableau 3.4 montre qu'IBEX associe ce contexte à des explications générales et réalistes. La Figure 3.6 présente des explications locales générées par IBEX suite à cette requête. Toutes les explications répondent à l'exigence de généralité : elles sont donc valables pour de nombreuses entrées (tweets) ce qui facilite la compréhension du modèle. Le fait que les

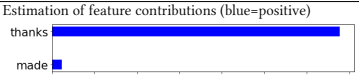
Tweet	Pred	Explanation
@AmericanAir although you have no control of the weather, you came through with a great customer service	Positive	IF the words "great" and "weather" appear THEN the tweet is positive with proba > 90 % among samples
@SouthwestAir Jason (108639) at Gate #3 in SAN made my afternoon!!! #southwestairlines #stellarservice #thanks!	Positive	Estimation of feature contributions (blue=positive) 
@AmericanAir all right, but can you give me an email to write to?	Neutral	IF the words "can" and "right" appear THEN the tweet is neutral with proba > 90 % among samples
@VirginAmerica You have any flights flying into Boston tomorrow? I need to be home and you Cancelled Flightled my flight and didn't do anything	Negative	IF the word "cancelled" appears THEN the tweet is negative with proba > 99 % among samples
@SouthwestAir What can we do to bring you back to Jackson, MS?! We miss you terribly around here. These other airlines are horrible!!	Negative	IF the word "horrible" appears THEN the tweet is negative with proba > 99 % among samples

FIGURE 3.6 – Exemple d'explications pour un expert du domaine. Toutes les explications sont générales et réalistes. Toutes, sauf une, utilisent la forme *RB* ($nature = T$), qui est préférée par les experts du domaine.

explications soient simples et pourtant précises tend à montrer que le modèle n'utilise pas de combinaisons complexes de mots et classe les tweets en fonction de la présence ou de l'absence d'un nombre limité de mots-clés (comme "*canceled*" ou "*thanks*"). Il est intéressant de noter qu'une des explications est sous la forme d'un graphique d'importance de variable (*LA*) alors que toutes les autres sont sous la forme de règles (*RB*). *A priori*, *RB* est la forme préférée pour ce contexte (car il implique des probabilités), mais, pour ce tweet spécifique, la seule explication qui correspondait à l'exigence de généralité était sous la forme *LA*.

3.5.3 Profane cherchant à optimiser de futures interactions avec le SDA

La troisième étude de cas concerne le jeu de données *german credit* qui contient des informations sur les crédits (montant, durée, but, etc.) et sur les demandeurs (type d'emploi, nombre de crédits en cours, etc.). Le SDA classe les demandes comme risquées ("*bad*") ou sûres ("*good*"). Ce cas d'étude concerne une personne dont la demande de crédit a été refusée et qui aimerait savoir comment l'améliorer pour obtenir un meilleur profil de risque à l'avenir. Cette requête concerne un utilisateur profane (*LU*) dont l'objectif est d'optimiser

des interactions futures (*A*) pour une entrée spécifique (*L*). D'après le Tableau 3.4, IBEX associe ce contexte à une contrefactuelle (*CF*) ayant une valeur de réalisme de *2* (la valeur *1* conduirait à des modifications peu probables de la demande tandis que la valeur *3* pourrait aboutir à une contrefactuelle trop éloignée de la valeur optimale). En outre, les modifications proposées par la contrefactuelle doivent être limitées aux caractéristiques actionnables (par exemple, la durée du crédit ou le nombre de crédits en cours), qui sont fournies par le destinataire. La contrefactuelle générée par IBEX à partir de ce contexte (cf. Tableau 3.7) suggère deux modifications du dossier actuel : la durée du crédit et la possession d'un téléphone.

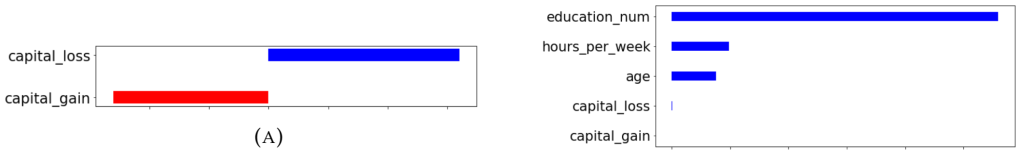
Variable actionnable	Montant du crédit	Durée	Crédits en cours	Type d'emploi	Possède un téléphone	Sortie
Demande actuelle	10722	47	1	*unskilled resident*	*yes*	***Bad***
Contrefactuelle	10722	36	1	*unskilled resident*	*none*	***Good***

TABLE 3.7 – Explication sous forme d'une contrefactuelle réaliste limitée aux variables actionnables.

3.5.4 Auditeur cherchant à contester le SDA

La dernière étude de cas concerne un auditeur (*AU)* qui souhaite contester (*C*) une décision spécifique (*L*) basée sur le jeu de données *adult census*. Une façon de contester un SDA est de montrer qu'il utilise des variables qui ne sont pas autorisées. Dans ce contexte, IBEX choisit un faible réalisme, car c'est la forme des frontières de décision, indépendamment de leur utilisation dans la pratique, qui importe dans ce cas. À titre de comparaison, la Figure 3.7 montre un exemple illustrant la différence entre les explications réalistes et non réalistes. L'explication à gauche de la Figure 3.7a a été générée par IBEX pour ce contexte (avec $réalisme = 1$) alors que l'explication de droite a été générée avec une valeur élevée de réalisme ($réalisme = 3$) [15]. Cette dernière prend également en compte la probabilité d'observer un changement dans les variables d'entrée pour estimer l'effet de ce changement sur la sortie du SDA. En revanche, l'explication à gauche décrit l'effet d'un changement sur la sortie sans tenir compte de la distribution réelle des données d'entrée.

15. L'échantillonnage correspondant à $réalisme = 3$ est "*select closest*" alors que celui correspondant à $réalisme = 1$ est "*add random noise*".



(A) (B)

FIGURE 3.7 – Deux explications de forme *LA* générées par IBEX pour la même entrée que le SDA classe comme '$< 50k$'. Les valeurs positives (barres bleues) indiquent un impact positif de la variable sur la sortie '$< 50k$' et les valeurs négatives (barres rouges) indiquent un impact négatif sur la sortie '$< 50k$'. (a) L'explication de gauche (générée avec réalisme=1) reflète la logique du SDA, sans tenir compte de la distribution réelle des données d'entrée. (b) L'explication de droite (générée avec réalisme=3) reflète le comportement du SDA sur des données réelles (basées sur la distribution de l'ensemble des données d'entrée).

L'explication de gauche montre que *capital_gain* et *capital_loss* sont les variables ayant le plus fort impact sur cette décision, tandis que l'explication de droite met l'accent sur *education_num* et *hours_per_week*. Premièrement, il faut noter qu'aucune de ces explications n'est plus "juste" que l'autre, elles reflètent le comportement du SDA sous des angles différents. Une analyse complémentaire de la population (ici le *adult census dataset*) montre que moins de 9% de toutes les entrées ont des gains en capital et qu'elles sont presque toutes classées '$> 50k$'. Cela justifie le fait que, même si *capital_gain* est un très bon indicateur de la classe '$> 50k$', il n'apparaît pas dans l'explication de droite, car il s'agit d'un événement rare. Le choix d'un faible niveau de réalisme a permis à IBEX d'identifier cet aspect du modèle qui pourrait être utilisé pour contester une décision, surtout si *capital_gain* n'est pas censé jouer un rôle pour ce type de décision. Ces explications peuvent également être utiles à un expert technique qui souhaite améliorer le SDA, comme le confirme le Tableau 3.4 (*TE* est également associé à *réalisme=1*).

3.6 Expérimentation impliquant des utilisateurs

Afin d'évaluer la démarche proposée dans IBEX, une étude impliquant des utilisateurs humains a été menée en collaboration avec la CNIL. Dans un contexte où la CNIL, acteur historique de la protection des données personnelles, cherche à développer des compétences

d'audit d'algorithmes, une version d'IBEX a été mise à disposition d'agents de la CNIL dans le cadre d'une mise en situation d'audit d'algorithme. L'étude résultant de cette collaboration (Partie 3.6.2) a produit différents apports à nos travaux de recherche :

— Certaines grandeurs quantitatives ont pu être mesurées directement lors de la réalisation de tâches fictives avec (ou sans) IBEX afin d'en mesurer la pertinence. Sur les cinq critères mesurés (dont la capacité à expliquer la sortie d'un SDA et le temps d'exécution de la tâche), les participants ayant accès à IBEX ont obtenu de meilleurs résultats que le groupe témoin (Partie 3.6.3).

— Des retours qualitatifs ont également pu être collectés lors d'une série d'entretiens menés avec des agents de CNIL ayant participé à l'étude et grâce à des commentaires libres collectés directement sur l'interface d'expérimentation. Ces retours permettent d'analyser le contexte de l'étude, d'identifier les besoins spécifiques attendus en termes d'explication pour réaliser des audits d'algorithmes et de mieux identifier les modalités d'utilisation d'IBEX par des utilisateurs réels (Partie 3.6.1).

3.6.1 Contexte

Depuis sa création en 1978, la CNIL est chargée de veiller à la protection des données personnelles contenues dans les fichiers et traitements informatiques ou papier, aussi bien publics que privés. À ce titre, elle exerce diverses activités liées à ce domaine dont l'accompagnement de la nouvelle réglementation (notamment européenne), du conseil auprès d'administrations envisageant des traitements de données à caractère personnel ainsi que des contrôles sur des entités publiques ou privées réalisant de tels traitements. À la suite d'un de ces contrôles, en cas de manquements constatés, la CNIL peut prononcer différentes mesures ou sanctions à l'égard du responsable de traitement. Afin de mener à bien cette mission de contrôle, les agents du service dédié ont développé un savoir-faire spécialisé impliquant la proche collaboration d'auditeurs des systèmes d'information (profil technique) avec des juristes. Ces contrôles peuvent s'effectuer en ligne, sur place ou sur pièces (à partir de documents) et peuvent concerner des éléments facilement vérifiables (par exemple l'oubli

d'une mention dans les conditions générales d'utilisation) ou plus complexe (par exemple la présence d'un biais discriminant dans un SDA).

Lorsque le contrôle implique l'analyse d'un SDA pouvant être complexe, les agents de la CNIL se limitent le plus souvent à des entretiens avec les équipes techniques ou commerciales et, dans certains cas, à une analyse rétrospective des décisions prises par le système. En particulier, les agents de la CNIL ne s'appuient pas systématiquement sur une analyse de code ni sur une analyse en boîte noire pour réaliser le contrôle. Les raisons de cette limitation sont avant tout pratiques. Les interviewés, agents de la CNIL au service des contrôles, ont pointé du doigt la faible durée des contrôles (généralement inférieure à 3 jours) et la complexité des systèmes d'information rencontrés impliquant souvent plusieurs infrastructures physiques louées à des prestataires. Toutefois, les interviewés précisent que, dans la plupart des contrôles, la constatation de l'infraction est simple et que les entretiens servent le plus souvent à confronter les responsables afin qu'ils reconnaissent le manquement.

Bien que les contrôles actuels semblent pouvoir se passer d'outils spécifiques à l'analyse des algorithmes, l'usage croissant de l'IA dans de nombreux services commerciaux et administratifs et l'évolution de la réglementation conduisent la CNIL à s'intéresser à de tels outils pour mener ses missions. En effet, dans le projet européen de règlement sur l'IA [16] il est prévu que les applications employant l'IA et considérées comme à haut risque devront pouvoir être auditées par un organisme extérieur comme la CNIL.

Ainsi, afin de renforcer ces capacités de contrôle pour s'adapter aux nouvelles pratiques et de préparer les missions d'audits algorithmiques, la CNIL s'intéresse aux outils d'explications d'algorithmes. Suite aux entretiens menés à l'issue de l'expérimentation, nous avons pu identifier certaines caractéristiques semblant correspondre à ces besoins en audit de la part d'un régulateur extérieur. En premier lieu, il apparaît que les explications en boîte noire semblent mieux adaptées que les explications en boîte blanche. En effet, même s'il sera probablement possible d'obtenir le code du système lors d'un audit, les analyses en boîte noire semblent plus efficaces puisqu'un même outil pourra être utilisé sur de nombreux systèmes et qu'il est difficile de prédire précisément comment un code sera exécuté une fois placé dans

16. https://ec.europa.eu/commission/presscorner/detail/fr/ip_21_1682 (consulté le 5/08/2021).

son environnement réel. Concernant le focus des informations fournies, les informations globales (concernant l'algorithme dans son ensemble) semblent en général mieux adaptées. Toutefois, certains juristes interrogés mettent aussi en garde sur d'éventuelles difficultés à convaincre un juge à l'aide d'un argument purement statistique. Enfin, les interviewés ont également souligné l'intérêt de disposer d'outils spécifiques qui permettraient de fournir une preuve directe d'un manquement à la loi. Par exemple, si un outil permettait d'assurer que l'algorithme n'utilise jamais une donnée particulière, la CNIL pourrait immédiatement constater un manquement au principe de minimisation de la collecte des données.

3.6.2 Déroulement

À cause de la crise sanitaire liée à l'épidémie de COVID-19, l'étude s'est déroulée entièrement en ligne par l'intermédiaire d'un système de visioconférence et d'une plate-forme d'expérimentation [17]. Avant de donner accès à la plate-forme, les participants ont assisté à une courte présentation du contexte et du déroulement de l'étude. Afin de ne pas influencer l'utilisation des outils, peu de détails ont été donnés sur la manière dont les explications devaient être consultées et interprétées. De cette manière, le caractère intuitif des outils a également pu être testé. L'expérimentation prend la forme d'une mise en situation au cours de laquelle les participants jouent le rôle d'agents de la CNIL devant réaliser un audit algorithmique sur une entreprise proposant un service basé sur un SDA.

Le cas d'usage concerne une entreprise proposant des crédits à la consommation pour l'achat d'un véhicule. Pour prendre les décisions d'octroi ou de refus, l'entreprise s'appuie sur un SDA employant 13 variables concernant le crédit demandé (pourcentage d'apport, montant, etc.), l'historique de crédits (autres crédits en cours, défauts, etc.) ainsi que l'âge du demandeur. En pratique, pour concevoir l'algorithme fictif audité par les participants, nous avons entraîné un arbre de décision avec un jeu de données en accès libre [18] auquel trois règles spéciales ont été ajoutées (les règles n'étaient pas affichées aux participants) :

17. La plate-forme sera accessible quelques mois après l'étude à l'adresse suivante : `algaudit.inrialpes.fr/`. Le code de la plate-forme est publiquement accessible `https://gitlab.inria.fr/chenin/algaudit`.
18. Vehicule Loan (Kaggle) : `https://www.kaggle.com/sneharshinde/ltfs-av-data`

— Règle 1 : tous les dossiers dont le demandeur est âgé de plus de 60 ans sont systématiquement refusés (règle volontairement discriminatoire),

— Règle 2 : les demandes ayant eu plus de deux crédits en défaut dans les 6 derniers mois sont systématiquement refusées,

— Règle 3 : les premiers crédits (nombre de crédits en cours nul) pour les faibles montants (prix du véhicule inférieur à 55 000) sont systématiquement acceptés.

Au total, 29 agents de la CNIL ont participé à l'expérimentation en ligne entre le 27 mai 2021 et le 9 juin 2021 au cours de 5 sessions. L'étude est composée de 3 étapes principales. Dans une première étape de mise en situation, le cas d'usage est présenté aux participants ainsi que 5 décisions de l'algorithme. À ce stade, les participants sont divisés en un groupe test ayant accès à IBEX et un groupe témoin, de taille similaire, qui a accès à une "calculette" permettant de simuler le SDA. Dans la deuxième phase, IBEX (ou la calculette) est mis à disposition sur deux exemples afin que les participants puissent se familiariser avec l'outil sans devoir suivre une consigne particulière. Dans la dernière phase, les participants réalisent 6 tâches, chacune correspondant à une décision du SDA tirée au hasard dans le jeu de données d'entraînement [19].

Une version simplifiée d'IBEX est proposée aux participants. Les interactions s'effectuent au niveau des exigences qui sont limitées à : la simplicité, le réalisme et la forme (voir la Figure 3.8). Au cours de chaque tâche, les participants répondent à trois questions à propos d'une décision spécifique qui leur est présentée :

— Quel a été le facteur déterminant de la décision (entrée à choix unique) ?

— D'autres facteurs ont-ils influencé la décision (entrée à choix multiples) ?

— Comment comprenez vous la décision ? S'agit-il d'un cas limite ou d'un cas simple (texte libre) ?

À partir des réponses à ces questions, nous proposons d'extraire 5 mesures quantitatives permettant d'évaluer l'utilité d'IBEX dans ce cas d'usage fictif, les performances du groupe test sont comparées à celle du groupe témoin ayant uniquement accès à la "calculette" :

19. À l'issue de ces trois phases, les participants entament une étude impliquant notre outil de justification présenté dans le Chapitre 4 sur le même cas d'usage. Cette partie est dédiée à l'expérimentation d'IBEX, la description de l'expérimentation d'Algocate étant détaillée dans la Partie 4.5.

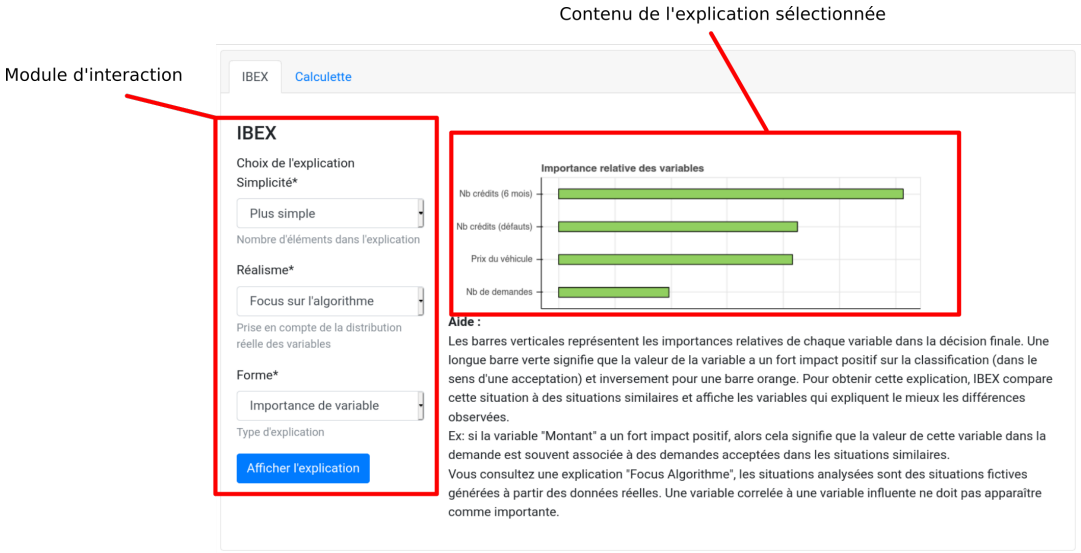


FIGURE 3.8 – Exemple de la représentation visuelle d'IBEX proposée sur la plate-forme en ligne pour l'expérimentation. Capture d'écran réalisée sur la plate-forme d'expérimentation en ligne.

1. Capacité à identifier le "facteur principal" de la décision,
2. Capacité à identifier les autres facteurs de la décision,
3. Capacité d'expliquer la décision sous forme d'un texte libre,
4. Capacité de distinguer les cas simples (correspondant à des feuilles de l'arbre de décision très déséquilibrées envers une classe) des cas limite (correspondant à des feuilles plus équilibrées),
5. Durée nécessaire pour réaliser une tâche.

Pour corriger les réponses, nous avons utilisé l'arbre de décision pour extraire le chemin de décision, c'est-à-dire l'ensemble des règles employées par le SDA. Nous avons considéré qu'un facteur était l'un des principaux pour cette décision s'il apparaît dans l'une des règles spéciales du SDA ou dans l'une des règles employées dans le chemin de décision de l'arbre à condition que celle-ci conduise à des nœuds de l'arbre ayant des classes majoritaires distinctes et compte pour au moins 1000 décisions. Ce dernier filtre permet d'exclure les règles ne modifiant pas la décision. Toutes les autres variables apparaissant dans le chemin de décision de l'arbre sont considérées comme des facteurs valides. Les explications sous forme

de texte ont été corrigées par l'auteur. Afin de ne pas influencer les résultats, les explications libres ont été corrigées sans que nous ne sachions s'il s'agissait d'une réponse du groupe témoin ou du groupe test. Une décision est considérée comme un "cas simple" lorsque l'une des règles spéciales est employée ou lorsque la feuille correspondante de l'arbre de décision se place dans le premier tiers des feuilles les plus déséquilibrées.

Pour les retours qualitatifs, nous avons mené une série de cinq entretiens semi-orientés (deux juristes et trois profils techniques) avec des participants à l'étude et nous avons posé des questions ouvertes sur l'interface d'expérimentation. L'analyse des retours collectés a été fait suivant une approche thématique prenant en compte la fréquence d'apparition des thèmes.

3.6.3 Résultats

En premier lieu, les résultats qualitatifs (entretiens et questions ouvertes) ont montré que les participants à l'étude ont jugé le contenu des explications globalement compréhensible et utile. Tous les participants qui ont eu accès à IBEX, sauf deux n'ayant pas répondu, ont déclaré que l'outil pourrait être utile pour un audit d'algorithme. L'interaction avec les trois exigences a été jugée simple et intuitive, même si la manière d'utiliser le réalisme de l'explication pour répondre aux questions a suscité quelques interrogations. Les participants ont par exemple apprécié la possibilité de changer la forme de l'explication afin d'avoir différents "angles d'approche" sur l'algorithme et pour confirmer les informations comprises avec les autres formes. Quelques pistes d'amélioration ont été proposées comme l'intégration d'explication globale (qui est possible en général avec IBEX, mais qui n'était pas disponible pour l'expérimentation) ou la mise à disposition d'informations conçues explicitement pour les contrôles.

Quantitativement, sur les cinq critères retenus, IBEX est plus performant que la calculette. Cette différence est significative pour trois critères (voir Figure 3.9). En particulier, IBEX permet efficacement de détecter le facteur principal dans 81 % des exercices (contre 66 % pour le groupe témoin) et permet d'identifier en moyenne 1,83 autres facteurs (contre 1,39 pour le groupe témoin). Ces résultats sont encourageants et valident l'utilité pratique d'IBEX et plus

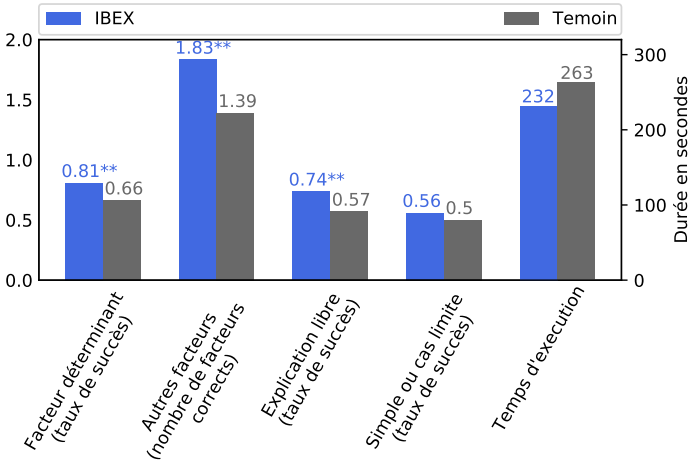


FIGURE 3.9 – Comparaison d'IBEX avec le groupe témoin (calculette). IBEX est plus performant sur les cinq critères retenus et la différence est significative (p-valeur < 0.05) sur les trois critères identifiés par le symbole "**".

généralement de la démarche en boîte noire. De plus, nous avons pu comparer la détection de la règle discriminante entre le groupe test et le groupe témoin. Seul un participant sur 14 du groupe IBEX n'a pas mentionné le caractère discriminant de l'algorithme contre trois sur 15 dans le groupe témoin. IBEX semble avoir procuré un avantage, mais un critère plus spécifique (par exemple une règle plus complexe mêlant plusieurs variables) aurait été plus éclairant.

De plus, les résultats décomposés par exigences permettent une analyse plus fine des propriétés de l'explication (Tableau 3.8). Comme nous pouvons le voir sur la ligne "Décompte", les explications sous forme d'importances relatives de variables (*LA*) semblent globalement préférées par les utilisateurs, mais les autres formes sont également largement consultées. D'après les informations recueillies lors des entretiens, les explications sous forme de graphique sont plus appréciées par les personnes ayant peu ou aucun bagage technique. Il est intéressant de noter que les explications sous forme de modèle à base de règle (*RBM*) sont significativement plus efficaces que les autres formes dans la détection du facteur principal. En contrepartie, elles conduisent à des durées d'exécution de la tâche significativement plus longues. Cette forme d'explication semble avoir été préférée par les profils techniques. Conformément à l'intuition, les explications les moins simples (*simplicité* = 1) conduisent à

	Forme			Simplicité		Réalisme		Moyenne
	LA	*CF*	*RBM*	*1*	*2*	*1*	*2*	
Décompte	184	70	87	239	102	249	92	
Facteur principal	0.81	0.90*	**0.91****	0.85	0.81	0.81	0.79	**0.81**
Autres facteurs	1.83	1.80	2.00	1.92	1.83	1.83	1.82	**1.83**
Explication libre	0.74	0.75	0.74	0.75	0.74	0.74	0.66	**0.74**
Simple ou limite	0.56	0.58	0.62	0.63*	0.56	0.56	0.55	**0.56**
Durée (secondes)	232	249	**290****	245	232	232	253	**232**

TABLE 3.8 – Cinq critères d'évaluation en fonction des exigences pour les tâches du groupe IBEX. La ligne "décompte" correspond aux exigences demandées par l'utilisateur (clic sur le bouton "Afficher l'explication"). La valeur des critères est mesurée sur les tâches ayant consulté au moins une fois l'explication correspondante. Si la différence avec les tâches n'ayant jamais utilisé cette exigence est significative (p-valeur < 0.05), la valeur est notée en gras avec le symbole **. Lorsque la différence n'est pas significative, mais montre une tendance (p-valeur < 0.1), la valeur est notée avec le symbole *.

des meilleurs résultats et une durée d'exécution plus longue. Toutefois, l'écart avec les autres explications n'est pas significatif.

En conclusion, ces résultats sont encourageants à plusieurs égards. Ils valident d'abord la possibilité pratique de concevoir un outil d'explication opérationnel à partir d'IBEX. De plus, les interactions sous forme d'exigences ont été bien comprises, rapidement, par des participants non-experts en IA et l'intérêt de ces interactions a pu être validé en entretiens et par l'usage intensif qu'en ont fait les participants. La comparaison des performances du groupe IBEX avec le groupe témoin fournit une preuve expérimentale forte de l'intérêt d'IBEX pour les critères sélectionnés.

Chapitre 4

Justifications et contestations

Dans ce chapitre, nous proposons une méthode originale permettant aux utilisateurs de systèmes algorithmiques d'en contester ou d'en justifier les décisions. Alors que la contestation des SDA occupe une place importante dans les textes de loi et que les justifications des SDA (ou plutôt leur absence) par des normes extérieures est une préoccupation centrale des travaux de sciences sociales, il n'existe pas d'outils dédiés à ces objectifs. En premier lieu, nous définissons les termes de justifications et de contestation. Une justification (respectivement une contestation) est un argument, soutenue par une preuve, qu'une décision est bonne (respectivement mauvaise) au sens d'un ensemble de normes extérieures prédéfinies. Justification et contestation sont vues ici comme des notions duales. Un dispositif technique, appelé *Algocate*, permet de rendre ces notions opérationnelles pour trois types de normes (règles, objectifs et référence) grâce à un système interactif permettant à l'utilisateur d'exprimer son point de vue sur le bien-fondé de la décision. Cette démarche est ensuite testée auprès d'utilisateurs réels dans le cadre d'une étude réalisée en collaboration avec la CNIL.

Les principales contributions de ce chapitre sont les suivantes :

— une définition générale des notions de contestation, justification, norme, preuve, affirmation et argument ;

— un cadre générique basé sur les définitions ci-dessus comprenant un protocole de contestation et de justification. Trois types de normes sont considérés dans ce document ; elles sont inspirées des trois principales théories morales (éthique de la vertu, conséquentialisme et éthique déontologique) ;

— un outil "preuve de concept" basé sur ce cadre et appelé *Algocate* [1] ;

— une expérimentation impliquant des utilisateurs pour évaluer la pertinence de l'outil.

4.1 Démarche

La possibilité de contester les résultats des SDA est une exigence essentielle lorsque ces derniers sont utilisés pour prendre des décisions ayant un fort impact sur les individus ou sur la société. C'est le cas, par exemple, des décisions prises par les juges, les professionnels de santé ou les banquiers. Cette nécessité est reconnue, dans une certaine mesure, par le RGPD, qui stipule qu'une personne "faisant l'objet d'une décision fondée exclusivement sur un traitement automatisé" a "le droit d'obtenir une intervention humaine de la part du responsable du traitement, d'exprimer son point de vue et de contester la décision" (article 22, paragraphe 3). Cependant, de nombreux obstacles peuvent limiter l'exercice de ce droit, le premier étant la difficulté pratique à comprendre les motifs d'une décision fondée sur les résultats d'un SDA. Pour que ce droit puisse être effectif, la contestabilité doit être rendue possible par des outils appropriés. De plus, comme l'indiquent Mulligan *et al.* la contestabilité favorise une implication critique, productive et responsable entre les utilisateurs et les algorithmes, entre les utilisateurs et les concepteurs de systèmes et, idéalement, entre les utilisateurs et les personnes soumises aux décisions (lorsqu'elles ne sont pas les utilisateurs) [91]. En effet, offrir des moyens de contester une décision peut être bénéfique pour :

— La fiabilité du SDA, en renforçant la capacité du décideur humain à détecter les résultats inappropriés.

— L'*accountability* du SDA, car les décisions peuvent être accompagnées de justifications.

— La légitimité du SDA, en améliorant l'autonomie du décideur humain. En effet, même si ce dernier n'a pas l'obligation formelle de se conformer à la suggestion du SDA, son autonomie est limitée s'il n'a pas la possibilité de la contester. Nous considérons en effet que la contestabilité est une condition essentielle pour préserver l'autonomie des décideurs humains.

1. Le nom *Algocate* est la concaténation des termes "*Algorithm*" et "*Advocate*".

Certains auteurs recommandent d'assurer la contestabilité dès la conception (*contestability by design*) et analysent les défis à relever pour la mettre en pratique [6, 61]. L'opacité est souvent mise en avant comme le premier obstacle à la contestation des décisions basées sur les SDA. En effet, il est difficile de contester les résultats d'un système lorsque les informations sur sa logique, son fonctionnement ou ses données d'entrée sont trop rares ou inintelligibles. La possibilité de produire des explications sur le fonctionnement d'un SDA facilite indéniablement la contestation. Toutefois, comme l'indiquent Mulligan *et al.* [91], fournir des explications n'est pas suffisant. Selon ces auteurs, la démarche réglementaire devrait permettre un dialogue équitable entre les professionnels et les systèmes d'aide à la décision ; les professionnels ne devraient pas être considérés comme des destinataires passifs de la sagesse du système dont ils dépendraient pour les remettre en question. Pour ces raisons, l'explicabilité seule est insuffisante et elle doit être élargie par des démarches réglementaires qui favorisent la contestabilité. Ce chapitre adopte la même position et présente une solution reposant sur les notions duales de contestation et de justification.

Dans la littérature, le terme justification est souvent employé dans un sens vague ou comme un synonyme d'explication. Pourtant, explications et justifications font référence à des concepts fondamentalement différents. Nous avançons les distinctions suivantes, qui sont cohérentes avec la caractérisation des justifications par T. Miller [84] :

— Le but d'une explication est de permettre à une personne (concepteur, utilisateur, personne affectée, etc.) de *comprendre* le fonctionnement du SDA. En revanche, l'objectif d'une justification est de convaincre que la décision est *bonne*. Par exemple, l'explication d'un refus de prêt peut être que le nombre de crédits en cours est trop élevé. Cette information permet de comprendre la logique du système en précisant la variable qui a été déterminante dans la décision. En revanche, elle ne dit rien sur le bien-fondé de la décision. Une justification de la même décision pourrait être que les demandes avec de nombreux crédits en cours ont une forte probabilité d'entraîner des défauts de remboursement, un risque que la banque souhaite réduire. Bien qu'elles s'appuient souvent l'une sur l'autre, les explications et les justifications ont des objectifs différents : une personne peut comprendre la logique qui conduit à un résultat particulier

sans être d'accord sur le fait que la décision qui en résulte est bonne ; à l'inverse, elle peut vouloir contester une décision (en étant convaincue qu'elle est mauvaise) sans comprendre la logique qui sous-tend l'algorithme.

— Les explications sont *descriptives* et *intrinsèques* dans le sens où elles ne dépendent que du système lui-même. En revanche, les justifications sont *normatives* et *extrinsèques* dans le sens où elles dépendent d'une référence (appelée *norme* dans la suite) permettant d'évaluer le bien-fondé des résultats. En effet, pour affirmer qu'un résultat est bon, il est nécessaire (1) de se référer à une définition (indépendante du SDA) de ce qu'est un bon résultat (la norme) et (2) de fournir la *preuve* que cette norme s'applique. Dans l'exemple ci-dessus, la norme est l'objectif de la banque de réduire les défauts de crédit. Une autre justification pourrait invoquer une norme de nature réglementaire comme un texte de loi qui interdirait l'octroi d'un nombre trop important de crédits.

La notion de norme est centrale dans la dialectique contestation-justification. En général, différents types de normes peuvent s'appliquer à un SDA. Ces normes peuvent avoir différentes sources de légitimité (juridique, éthique, sociale, économique, etc.) et peuvent être exprimées de différentes manières (par exemple, par la loi ou par la jurisprudence pour les normes juridiques). Lorsque plusieurs normes s'appliquent, elles peuvent être en tension, voire en contradiction [2]. Dans certains cas, il est possible de s'appuyer sur des règles de priorité pour établir la préséance d'une norme sur une autre (par exemple, le droit international prévaut généralement sur le droit national, la constitution prévaut sur le droit commun, qui prévaut sur les décrets, etc.) ; dans d'autres cas, de telles règles peuvent ne pas exister et les conflits entre les normes doivent être résolus par un décideur humain au cas par cas.

Techniquement parlant, les preuves peuvent prendre différentes formes en fonction du type de norme auquel elles se réfèrent. Par exemple, pour une décision fondée sur la jurisprudence, la preuve s'appuierait sur des décisions antérieures concernant des cas "similaires".

Les contestations et les justifications sont des notions duales : une contestation peut être

2. Dans le domaine de l'éthique, l'existence d'engagements incompatibles est parfois appelée "*moral overload*" [63].

considérée comme une *affirmation* selon laquelle une décision est *mauvaise*, soutenue par des preuves, tandis qu'une justification est une affirmation selon laquelle une décision est *bonne*, soutenue par des preuves. Dans les deux cas, les preuves font référence à une norme donnée.

Les contestations et les justifications peuvent prendre différentes formes. Par exemple, une affirmation pour justifier le refus d'une demande de prêt bancaire pourrait être : "Les demandeurs qui ont plus de trois prêts en cours sont généralement en défaut de paiement ; votre demande doit donc être rejetée". Une affirmation contestant la décision pourrait être la suivante : "Je sais que cette autre demande a été acceptée et qu'elle présente les mêmes attributs que la mienne, à l'exception du sexe, qui ne devrait pas être pris en compte ; ma demande devrait donc être acceptée elle aussi". Ici, le premier exemple fait référence à un objectif de minimisation des défauts et le second à une règle de non-discrimination, mais d'autres types de normes peuvent être utilisés.

Dans la suite de ce chapitre, les particularités des justifications (et des contestations) sont illustrées de manière informelle par un exemple (Partie 4.2), puis nous présentons un cadre général pour contester et justifier les décisions algorithmiques (Partie 4.3). Une implémentation de ce cadre, correspondant à des choix particuliers est présentée dans la Partie 4.4. Nous concluons ensuite par des résultats expérimentaux issus d'une étude utilisateurs menée avec la CNIL (Partie 4.5).

4.2 Présentation informelle du cadre

Dans cette partie, nous utilisons un hypothétique système de décision de crédit à la consommation pour illustrer le cadre générique de contestations et de justifications. Il est intéressant de noter qu'aucune information sur le fonctionnement du SDA ou sur les technologies utilisées pour le mettre en œuvre n'est nécessaire dans cet exemple. Les dossiers de demande de crédit sont composés d'informations sur le demandeur (âge, genre, type d'emploi et prêts en cours, etc.) et d'informations sur le crédit lui-même (prix du bien, pourcentage d'apport, etc.).

Les normes s'appliquant à cette étude de cas sont présentées, par ordre de priorité, dans la Figure 4.1. Elles proviennent de sources différentes : droits fondamentaux, règles commerciales et objectifs de la banque. Le système de justification permet d'élaborer des arguments à partir de ces normes, sans préjuger de leur légitimité. Plus précisément, le système de justification ne suppose pas que la légitimité d'une norme soit acceptée par toutes les parties. Par exemple, dans l'étude de cas considérée ici, la banque pourrait se référer à des règles commerciales internes qui ne sont pas connues ou acceptées par les clients. Dans une telle situation, si la banque fait référence à une norme qui n'est pas acceptée par le client, alors le protocole de contestation-justification requiert l'intervention d'un décideur humain (appelé ci-après "arbitre", même s'il ne s'agit pas nécessairement d'un arbitre au sens juridique du terme). L'avantage du protocole dans de tels cas est que la partie qui s'appuie sur une norme est obligée de la mentionner et de la soumettre à l'approbation de l'autre partie (ou de l'arbitre en dernier recours).

1. NORME 1 : Respect de la réglementation anti-discrimination.
   Le genre, l'origine ethnique et la religion ne doivent avoir aucune incidence sur les décisions du système.
2. NORME 2 : Règle métier pour l'acquisition de nouveaux clients.
   Afin de faciliter l'accès au marché de nouveaux clients, les personnes demandant un crédit pour la première fois sont acceptées, quelles que soient les caractéristiques de leur demande, à la seule condition que le prix du bien soit inférieur à 55 000 €.
3. NORME 3 : Objectif de minimisation des risques de défaut.
   Les risques de défaut de crédit doivent être minimisés. Une base historique de demandes de crédit est utilisée pour estimer les risques de défaut d'une nouvelle demande.

FIGURE 4.1 – Exemples de normes classées par ordre de priorité

La Figure 4.2 présente un exemple d'interaction entre un utilisateur et le système de justification. Les affirmations de l'utilisateur, traduites ici en langage naturel, sont désignées par les symboles *Ui* et les réponses du système par les symboles *Ai* :

— Dans U1, l'utilisateur conteste le rejet de sa demande de crédit et exprime les raisons pour lesquelles il pense que sa demande devrait être acceptée. L'utilisateur n'est pas tenu de produire des preuves à l'appui de sa contestation, car il peut ne pas être en mesure de le faire (il peut ne pas avoir accès aux données pertinentes par exemple).

— Dans A1, le système de justification fournit une preuve (application d'une norme à un cas particulier) de la contestation de l'utilisateur. Dans ce cas, la preuve est générée à partir de l'ensemble des données historiques du SDA.

— Dans A2, le système de justification fournit une justification de la décision de rejet soutenue par une preuve. La justification est une extension de la contestation incluant une condition supplémentaire sur le pourcentage d'apport. La preuve est à nouveau générée à partir de l'ensemble de données historiques du SDA. Elle montre que le pourcentage d'apport de la demande a généralement conduit à un fort taux de défaut dans le passé. Dans ce cas, la justification A2 est considérée comme plus forte que la contestation A1 en raison du très fort taux de défauts parmi les décisions historiques (87 % des crédits, soit 51 % de plus que la moyenne, alors que la différence pour la contestation A1 n'est que de 5 %), avec un nombre de décisions important (2 552).

— L'interaction peut alors continuer de différentes manières selon que l'utilisateur accepte ou non la norme utilisée dans A2. S'il ne l'accepte pas, il faut s'en remettre à l'arbitre pour trancher. Une autre possibilité est de contester à nouveau, mais sur une autre base. C'est l'option suivie dans la Figure 4.2 : U2 conteste la décision sur la base de la règle anti-discrimination en s'appuyant sur une demande similaire qui a été acceptée.

— A3 est la réponse produite par le système de justification après vérification de la validité de la contestation U2. Il s'agit de la dernière étape du protocole qui, dans ce cas, ne nécessite pas l'intervention de l'arbitre puisque les parties ont convenu d'une révision de la décision.

Les interactions présentées dans la Figure 4.2 montrent que le système fournit une aide à la fois aux utilisateurs qui contestent les décisions du SDA (par exemple, les individus affectés par les décisions, les régulateurs ou les décideurs humains qui ne sont pas sûrs de la suggestion du SDA) et aux utilisateurs qui veulent les justifier (généralement l'opérateur du SDA)[3]. Le système de justification est neutre, dans le sens où il est conçu pour trouver les meilleurs arguments (c'est-à-dire des affirmations soutenues par les preuves les plus fortes)

3. Dans la Figure 4.2, U1 et U2 sont des contestations tandis que A2 est une justification de la décision.

U1 Je pense que ma demande devrait être acceptée car le prix du bien que je souhaite acquérir est inférieur à 63 000 € et que j'ai plus de 24 ans.
A1 Dans l'historique des demandes, les 12 133 dossiers tels que prix du bien $\leq$ 63 000 € et âge $\geq$ 24 ans ont un taux moyen de défaut égal à 31 % (moyenne générale : 36 %).
A2 Toutefois, une analyse complémentaire par notre système montre que la NORME 3 tend à contredire votre affirmation. Pour arriver à cette conclusion, le système a considéré également le fait que les 2 552 dossiers tels que prix du bien $\leq$ 63 000 €, âge $\geq$ 24 ans et pourcentage d'apport $\leq$ 23 % (ce qui est votre cas) ont un taux moyen de défaut égal à 87 % (moyenne générale : 36 %).
U2 La demande de crédit *CA2* a été acceptée. Cette demande a les mêmes attributs que la mienne, à l'exception du genre. Par conséquent, ma demande devrait être acceptée également.
A3 La décision initiale enfreindrait en effet la NORME 1 (respect de la réglementation anti-discrimination). Votre demande doit donc être acceptée.

FIGURE 4.2 – Exemple d'interaction avec *Algocate* (issu d'une interaction réelle avec l'implémentation).

pour les deux parties. Cette neutralité est primordiale, étant donné le déséquilibre existant habituellement entre les individus qui sont affectés par les décisions et les concepteurs ou opérateurs du SDA.

4.3 Un cadre pour contester et justifier les décisions

Le but de cette partie est de définir plus précisément les concepts et le cadre générique présentés de manière informelle dans la partie précédente. Le système *Algocate*, qui est une implémentation particulière de ce cadre, est présenté dans la partie suivante (4.4). Le cadre repose principalement sur trois notions : les *affirmations*, les *normes* et les *preuves*. Le Tableau 4.1 fournit quelques exemples pour chacune de ces notions :

— Une affirmation définit ce qu'une décision devrait être (par exemple "acceptée" ou "rejetée"), selon une partie, et indique les raisons qu'a cette dernière de le penser (sous la forme de conditions sur les données).
— Une norme est une référence qui peut être utilisée par une partie pour soutenir une affirmation.
— Une preuve est l'application d'une norme à un cas particulier.

Le triplet <*affirmation*, *norme*, *preuve*> est appelé un *argument*. Un argument peut être soit une *justification* (si l'affirmation soutient la décision) soit une *contestation* (si l'affirmation

AFFIRMATION		
Absolue	$\forall x \in \mathbf{A}, x[\text{education}] = PhD \wedge x[\text{outstanding-loans}] \leq 2 \implies f(x) = 1$	Tous les docteurs ayant moins de deux crédits en cours devraient être acceptés
Relative	$\forall x \in \mathbf{A}, (\exists x' \in \mathbf{A}, f(x') = 1 \wedge (x'[\text{revenue}] \leq x[\text{revenue}] \wedge x'[\text{amount}] \geq x[\text{amount}])) \implies f(x) = 1$	Tous les cas pour lesquels il existe un cas accepté ayant des revenus plus faible et un montant plus élevé devraient être acceptés
NORME		
Règle (absolue)	$[Rule, \forall x \in \mathbf{A}, x[\text{outstanding-loans}] \geq 4 \implies f(x) = 0]$	Les crédits avec plus de 4 prêts en cours doivent être refusés
Règle (relative)	$[Rule, \forall x \in \mathbf{A}, (\exists x' \in \mathbf{A}, f(x') = 1 \wedge (x'[\text{revenue}] = x[\text{revenue}] \wedge x'[\text{education}] = x[\text{education}] \wedge x'[\text{age}] \geq x[\text{age}])) \implies f(x) = 1]$	Tous les cas pour lesquels il existe un cas accepté, plus âgé et avec les mêmes niveaux d'éducation et de revenus devraient être acceptés
Objectif	$[Obj, \Delta_O, nd]$ (avec $x[nd] = 1$ pour un non-défaut $x[nd] = 0$ sinon)	Le nombre de crédits acceptés et remboursés doit être maximisé (le nombre de défauts doit être minimisé)
Référence	$[Ref, \Delta_R, d]$ ($x[d]$ la décision de référence pour x)	Les décision du SDA doivent ressembler aux décisions prises antérieurement par les experts de la banque qui font référence
PREUVE		
Objectif	$[\Delta_O, 1200, 0.06]^{(i)}$ avec Δ_O données historiques avec $x[o]$ l'attribut objectif	Le taux de non-défaut pour les 1200 cas historiques qui correspondent à l'affirmation est de 93 % (moyenne globale : 87 %)
Référence	$[\Delta_R, 500, 0.28]^{(ii)}$ avec Δ_R données historiques et $x[d]$ décisions de référence	Parmi les 500 décisions d'experts correspondant à l'affirmation, 95 % ont été acceptés (moyenne globale : 67 %)

TABLE 4.1 – Exemples d'affirmations, de normes et de preuves. (i) La valeur 0.06 est le résultat de la différence 0.93 - 0.87. (ii) La valeur 0.28 est le résultat de la différence 0.95 - 0.67.

contredit la décision). Lorsque des arguments contradictoires sont émis, il faut pouvoir les comparer à l'aide d'une *relation de force* entre les arguments.

Dans ce qui suit, nous définissons successivement les affirmations (Partie 4.3.1), les normes (Partie 4.3.2), et les preuves (Partie 4.3.3) avant d'introduire la relation de force (Partie 4.3.4).

Dans la suite, nous utilisons les notations mathématiques suivantes. Les cas (entrées du SDA) sont caractérisés par des m-uplets $< a_1, \ldots, a_m >$ de m attributs. L'ensemble des valeurs possibles de l'attribut a_j est noté $\mathbf{A_j}$ et $\mathbf{A} = \mathbf{A_1} \times ... \times \mathbf{A_m}$ est l'ensemble de tous les cas possibles. La fonction de décision est appelée $f : \mathbf{A} \to \mathbb{B}$ avec $\mathbb{B} = \{0, 1\}$ (les décisions sont supposées binaires). La notation $x[j]$ désigne le $j^{ième}$ attribut de $x \in \mathbf{A}$. Par conséquent, $x[j] \in \mathbf{A_j}$. La notation est étendue aux ensembles S de cas avec $S[j] = \{x[j] | x \in S\}$. Par souci de lisibilité, le nom de la variable (par exemple "prix" ou "age") peut être utilisé à la place des indices dans les exemples. Par exemple, si le troisième attribut d'un cas x représente l'âge du demandeur, il y sera fait référence par la notation $x[age]$ (avec $age = 3$). Enfin, le cas considéré dans le protocole de contestation et de justification est appelé $x_s \in \mathbf{A}$.

4.3.1 Affirmations

Comme suggéré dans la partie précédente, les affirmations contiennent deux éléments d'information :

— La décision qui devrait résulter du SDA, selon l'émetteur de l'affirmation (1 pour "accepté" ou 0 pour "rejeté"),

— Les caractéristiques du cas qui plaident en faveur de cette décision, selon l'émetteur de l'affirmation.

Formellement, les affirmations sont donc définies par :

$$\forall x \in \mathbf{A}, C(x) \implies f(x) = \delta \tag{4.1}$$

avec $\delta \in \mathbb{B}$, la décision soutenue par l'émetteur et $C(x)$ la propriété censée justifier cette décision. Une affirmation, telle que décrite par l'équation (4.1) se lit donc de la manière suivante : tout cas x qui vérifie $C(x)$ devrait obtenir la décision δ. Une affirmation n'est pertinente pour un cas x_s que si $C(x_s)$ est vrai, ce qui sera supposé par la suite. Si $\delta \neq f(x_s)$, le but de l'affirmation est de contester la décision ; il s'agit alors d'une *affirmation contestataire*. Vice versa, si $\delta = f(x_s)$, l'objectif de la affirmation est de soutenir la décision ; il s'agit d'une *affirmation justificative*.

En général, $C(x)$ peut impliquer des comparaisons de x avec un nombre quelconque d'autres cas. Pour des raisons de simplicité, seules deux possibilités sont considérées ici : les conditions n'impliquant aucun autre cas et les conditions en impliquant un seul.

Le premier type d'affirmations, appelé *affirmations absolues*, ne fait référence à aucun autre cas. La forme générale de la condition $C(x)$ pour les affirmations absolues est la suivante :

$$C(x) = (x[i_1] \diamondsuit_1 v_1) \wedge \ldots \wedge (x[i_k] \diamondsuit_k v_k) \tag{4.2}$$

avec $i_p \in \{1, \ldots, m\}$, $\diamondsuit_p \in \{=, \leq, \geq\}$ et $v_p \in \mathbf{A_{i_p}}$ pour $p \in \{1, \ldots, k\}$. Les constantes v_p sont fournies par l'émetteur de l'affirmation. La première phrase (U1) de la Figure 4.2 et la première ligne du Tableau 4.1 sont des exemples d'affirmations absolues.

Le deuxième type d'affirmations, les *affirmations relatives*, impliquent une comparaison avec un autre cas x'. La deuxième phrase de l'utilisateur (U2) de la Figure 4.2 en est un exemple. Plus formellement, la condition $C(x)$ pour les affirmations relatives s'écrit de manière générale :

$$C(x) = \exists x' \in \mathbf{A}, f(x') = \delta \wedge (x'[i_1] \diamondsuit_1 x[i_1]) \wedge \ldots \wedge (x'[i_k] \diamondsuit_k x[i_k]) \tag{4.3}$$

avec $i_p \in \{1, \ldots, m\}$ et $\diamondsuit_p \in \{=, \leq, \geq\}$ pour $p \in \{1, \ldots, k\}$. Ce type d'affirmations correspond à une situation dans laquelle un autre cas x' est associé à une décision δ et la relation entre les deux cas semble justifier que la même décision soit prise pour x. La deuxième ligne du Tableau 4.1 fournit un exemple de condition exprimée dans cette syntaxe.

4.3.2 Normes

Comme suggéré plus haut, pour qu'une affirmation puisse être utilisée pour contester ou justifier une décision, elle doit être soutenue par une norme. Trois types de normes sont considérées dans notre système : les normes basées sur une règle, les normes basées sur un objectif et les normes basées sur une référence. Elles correspondent à trois stratégies communes pour soutenir une contestation (ou une justification). Il est intéressant de noter que ces trois modes reflètent également les trois principales théories morales (respectivement l'éthique déontologique, le conséquentialisme et l'éthique de la vertu). D'autres types de normes peuvent facilement être ajoutés au cadre, à condition que leur signification soit définie avec précision, comme nous proposons dans cette partie, et que leur force puisse être caractérisée comme nous le décrivons dans la Partie 4.3.4.

Normes basées sur une règle

La première catégorie de normes se caractérise par des règles fixes. Formellement parlant, ces règles peuvent être exprimées comme suit :

$$[Rule, Def]$$

avec *Rule* l'indicateur définissant le type de la norme (en l'occurrence "règle") et Def son contenu. Def est exprimé dans le même langage que les affirmations, c'est-à-dire par l'Équation (4.1) de la Partie 4.3.1, avec C définie par l'Équation (4.2) si la norme est exprimée de manière absolue ou l'Équation (4.3) si elle implique un second cas (règle relative). Les lignes 3 et 4 du Tableau 4.1 fournissent respectivement un exemple de règle absolue et un exemple de règle relative. Les normes NORME 1 et NORME 2 de la Figure 4.1 sont également des exemples de règle relative et de règle absolue.

Ce type de norme est courant en droit (règlements, directives, actes, règles contractuelles, etc.) et dans les affaires (règles sectorielles, règles d'entreprise, règles de procédure, etc.). En termes de théories morales, c'est aussi l'esprit de l'éthique déontologique qui repose sur des obligations morales (telles que "Tu ne mentiras point") que doit respecter tout agent rationnel [110] [4].

Normes basées sur un objectif

La deuxième façon de définir une norme consiste à recourir à un objectif mesurable qui peut être utilisé pour évaluer les décisions. En pratique, un objectif est exprimé comme un attribut qui devrait être maximisé.

Formellement, les normes basées sur un objectif peuvent être exprimées de la manière suivante :

$$[Obj, \Delta_O, ob]$$

où Obj est l'indicateur définissant le type de la norme (en l'occurrence "objectif"), Δ_O une base de données contenant les valeurs $\Delta_O[ob]$ de l'attribut objectif ob. La ligne 5 du Tableau 4.1 et la NORME 3 de la Figure 4.1 sont des exemples de normes basées sur un objectif.

Les normes basées sur un objectif sont courantes dans les entreprises et les organisations en général. Les décisions des juges par exemple sont soumises à un certain nombre d'objectifs dont certains peuvent être mesurables comme la prévention d'autres crimes ou la réinsertion sociale (cf. Exemple 2 du Chapitre 1). En termes de théories morales, elles peuvent être reliées au conséquentialisme, qui, comme le dit M. Timmons [110], explique

4. La principale représentante de cette école de pensée est la théorie morale de Kant.

le statut déontologique des actions (ou des décisions) en se référant aux valeurs des conséquences de ces actions (ou de ces décisions). Même si la portée des normes utilisables dans ce cadre générique ne se limite pas à des questions morales, comme le montrent les exemples, les normes basées sur un objectif suivent une démarche conséquentialiste en ce sens où, plutôt que de s'appuyer sur des règles fixes comme la première catégorie de normes, elles se concentrent sur l'évaluation de l'impact ou des conséquences des décisions.

Normes basées sur une référence

Le troisième type de norme est applicable lorsque des données de référence Δ_R sur les décisions passées sont disponibles. Le principe dans ce cas est qu'une décision est justifiée si elle est cohérente avec les décisions de référence. L'hypothèse implicite est que les données de référence sont légitimes, dans le sens où elles peuvent être utilisées comme un modèle "vertueux" pour les décisions futures.

Formellement, les normes basées sur une référence peuvent être exprimées de la manière suivante :

$$[Ref, \Delta_R, d]$$

où Ref est un indicateur définissant le type de la norme (en l'occurrence "référence"), Δ_R une base de données contenant les valeurs $\Delta_R[d]$ de l'attribut de décision d. La ligne 6 du Tableau 4.1 est un exemple de norme basée sur une référence.

Ce type de norme est courant en droit (utilisation de cas antérieurs ou jurisprudence). En termes de théories morales, elles peuvent être reliées à l'éthique de la vertu, qui est présentée comme suit par M. Timmons : nous prenons souvent d'autres personnes comme modèles de ce que nous aimerions être parce que nous pensons qu'ils possèdent certains traits de caractère admirables. En effet, les données de référence peuvent être vues comme un modèle de "bon comportement" (ici un modèle de "bonnes décisions"), le but du SDA étant de refléter le plus fidèlement possible le comportement de ce modèle. Lorsqu'un SDA repose sur un apprentissage automatique supervisé, l'ensemble de données d'apprentissage peut évidemment être utilisé comme données de référence.

4.3.3 Preuve

Dans un argument <*affirmation, norme, preuve*>, le composant *preuve* montre comment l'*affirmation* est soutenue par la *norme*. Les preuves peuvent prendre différentes formes selon le type de norme. La principale différence se situe entre les normes basées sur une règle, qui ne s'appuient pas sur des données et qui soutiennent ou contredisent une affirmation de manière binaire, et les autre normes (objectif ou référence) qui s'appuient sur des données et sont de nature quantitative. Dans ces derniers cas, une affirmation peut être plus ou moins soutenue par une norme.

Le cas des normes basées sur une règle est simple puisque ces normes sont exprimées dans le même langage que les affirmations. Une affirmation :

$$\forall x \in \mathbf{A}, C_s(x) \implies f(x) = \delta_s \tag{4.4}$$

est soutenue pas une règle :

$$\forall x \in \mathbf{A}, C_r(x) \implies f(x) = \delta_r \tag{4.5}$$

si et seulement si :

$$\delta_s = \delta_r \wedge \forall x \in \mathbf{A}, (C_s(x) \implies C_r(x)) \tag{4.6}$$

ce qui peut être vérifié à l'aide d'un simple système d'inférence comparant terme à terme les composantes $(x[i_j]\diamondsuit_j v_j)$ de C_s et C_r.

Les preuves impliquant des normes basées sur un objectif ou sur une référence reposent sur deux valeurs numériques appelées respectivement la "couverture" et l'"écart". Si l'affirmation est définie par :

$$\forall x \in \mathbf{A}, C(x) \implies f(x) = \delta \tag{4.7}$$

et si Δ est l'ensemble de données pertinentes (soit Δ_O soit Δ_R), il faut d'abord définir $\Delta_{|C}$, le sous-ensemble de cas correspondant à la condition de l'affirmation :

$$\Delta_{|C} = \{x \in \Delta | C(x)\} \tag{4.8}$$

La couverture $\gamma(\Delta, C)$, correspond au nombre de cas dans ce sous-ensemble :

$$\gamma(\Delta, C) = card(\Delta_{|C}) \tag{4.9}$$

avec $card(S)$ le cardinal d'un ensemble S.

Pour l'attribut correspondant à l'objectif ob, l'écart $\mu(\Delta, C, \delta)$ est défini de la manière suivante :

$$\mu(\Delta, C, \delta) = (2\delta - 1)(\overline{\Delta_{|C}[ob]} - \overline{\Delta[ob]}) \tag{4.10}$$

avec $\overline{S}$ la moyenne des valeurs de l'ensemble S. L'écart μ mesure la différence de moyennes entre le sous-ensemble caractérisé par C et l'ensemble de la population. Le facteur $(2\delta - 1)$ est justifié de la manière suivante. Si la différence est positive (objectif plus élevé dans $\Delta_{|C}$), alors la preuve soutient les affirmations telles que $\delta = 1$ et contredit les affirmations telles que $\delta = 0$. Vice-versa, si la différence est négative (objectif inférieur dans $\Delta_{|C}$) alors les preuves soutiennent les affirmations telles que $\delta = 0$ et contredisent les affirmations telles que $\delta = 1$. Dans tous les cas, un écart positif correspond donc à une preuve soutenant l'affirmation. Pour des raisons de lisibilité, il est préférable de montrer à l'utilisateur à la fois la moyenne du sous-ensemble $\Delta_{|C}[ob]$ et la moyenne de la population $\Delta[ob]$ comme cela est fait dans le Tableau 4.1. L'écart pour les normes basées sur une référence est défini de la même manière.

La définition complète de la preuve est la suivante :

$$[\Delta, \gamma(\Delta, C), \mu(\Delta, C, \delta)] \tag{4.11}$$

avec Δ le jeu de données pertinent (Δ_O ou Δ_R). Les deux dernières lignes du Tableau 4.1 montrent des exemples de preuves pour une norme basée sur un objectif et une norme basée sur une référence. Comme suggéré ci-dessus, les preuves basées sur des données peuvent soutenir les affirmations avec plus ou moins de force. Par exemple, dans la dernière ligne du Tableau 4.1, l'affirmation caractérise un sous-ensemble de 500 demandes avec un taux moyen d'acceptation supérieur de 28 % à la moyenne de la population (95 % - 67 % = 28 %).

Dans ce cas, l'argument est fort car la couverture (500) et l'écart (0,28) sont tous deux élevés, mais ce n'est pas le cas pour l'avant-dernière ligne du Tableau 4.1 dans laquelle l'écart n'est que de 6 % (0,06). La notion de force d'une preuve est discutée plus en détail dans la partie suivante.

Seuls les arguments *valables* sont considérés ici, c'est-à-dire les arguments $< S, N, E >$ tels que E soutient, dans une certaine mesure, S selon une norme N. Plus précisément, si N est une norme basée sur une règle, alors l'argument $< S, N, E >$ est valable seulement si la condition (4.6) est satisfaite. Si N est une norme basée sur des données, il est valable seulement si $\gamma(\Delta, C) \neq 0$ et $\mu(\Delta, C, \delta) > 0$.

4.3.4 Relation de force

Lorsque deux parties restent en désaccord sur une décision et que chaque partie fournit un argument pour soutenir sa position, il est important de pouvoir comparer ces arguments. Pour y parvenir, le système de contestation et de justification doit être doté d'une relation de préordre $\geqslant_a$ entre les arguments [5]. Cette relation, appelée "*relation de force*", repose sur des comparaisons des normes et des preuves qui composent les arguments :

$$[S, N, E] \geqslant_a [S', N', E'] \Leftrightarrow (N \geqslant_n N') \vee ((N = N') \wedge (E \geqslant_e E')) \tag{4.12}$$

En d'autres termes, un argument A est plus fort qu'un argument A' dans deux cas :

— A s'appuie sur une norme plus forte que A',

— A et A' s'appuient sur la même norme, mais la preuve de A est plus forte que celle de A'.

La relation de force $\geqslant_n$ entre les normes dépend essentiellement du domaine et doit être définie pour chaque application. À titre d'illustration, dans l'exemple de la Figure 4.1, la banque a spécifié une relation de force explicite entre trois normes (NORME 1 $\geqslant_n$ NORME 2 $\geqslant_n$ NORME 3). La première norme, qui est un droit fondamental, doit l'emporter sur toutes les autres. En revanche, la priorité de la NORME 2 sur la NORME 3 est un choix stratégique

5. Une relation de préordre est une relation binaire qui est réflexive et transitive.

de la banque. La relation de force n'est que partielle puisque certains arguments peuvent être incomparables.

La relation de force sur les preuves $\geqslant_e$, n'est pertinente que pour les preuves impliquant un jeu de données, puisque les preuves basées sur des règles sont de nature binaire. Une façon simple de la définir pour les normes basées sur des objectifs est la suivante :

$$[\Delta, \gamma, \mu] \geqslant_e [\Delta, \gamma', \mu'] \Leftrightarrow (\gamma \geq \gamma' \wedge \mu \geq \mu') \tag{4.13}$$

Les grandes couvertures conduisent à des arguments plus solides car elles améliorent la "force statistique" des preuves. La force des preuves croît également avec μ car des valeurs plus élevées de μ correspondent à un ensemble d'objectifs ou à des décisions de référence plus favorables. Cette définition de $\geqslant_e$ est intuitive, mais elle est conservatrice dans le sens où elle conduit à un grand nombre d'éléments de preuve incomparables. Une définition plus sophistiquée, mise en œuvre dans *Algocate*, est proposée dans la partie suivante.

4.4 "*Algocate*" : un système de contestation et de justification

Le cadre général introduit dans la partie précédente peut être rendu opérationnel de différentes manières pour construire un système de contestation et de justification. Cette partie décrit les principaux choix effectués dans la conception d'*Algocate*, une implémentation du type "preuve de concept". En premier lieu, nous présentons dans la Partie 4.4.1 les principales étapes du protocole et montrons les possibilités d'interactions avec le système. La relation de force utilisée dans *Algocate* est détaillée dans la Partie 4.4.2 puis son utilisation pour la génération d'affirmations dans la Partie 4.4.3.

4.4.1 Le protocole *Algocate*

Une session *Algocate* est toujours associée à une décision, qui est appelée *décision initiale* dans la suite. Pour reprendre la terminologie des explications, il s'agit donc ici de justifications "locales" d'algorithmes. Toutes les parties prenantes (concepteur, opérateur, décideur humain, personne affectée par la décision, auditeur, etc.) peuvent utiliser Algocate et

leurs motivations peuvent être variées. Il peut s'agir par exemple de trouver des arguments pour soutenir la décision initiale, pour la contester ou pour tester son bien-fondé. En cas de désaccord, l'intervention d'un arbitre humain pour trancher peut toujours être exigée, en accord avec la réglementation européenne. Enfin, Algocate fait l'hypothèse qu'un ensemble de normes ont été énoncées par les parties prenantes et/ou par des tiers indépendants (par exemple, des organismes de réglementation). Cependant, aucune hypothèse n'est faite sur l'exhaustivité de cet ensemble initial de normes ni sur leur acceptation par tous les utilisateurs. Comme indiqué ci-dessous, cet ensemble de normes peut évoluer, notamment pour tenir compte des verdicts de l'arbitre.

1. L'interaction avec *Algocate* commence par une affirmation initiale par un utilisateur (appelée *affirmation de l'utilisateur*). Si l'utilisateur a une norme spécifique à l'esprit pour appuyer son affirmation, il peut également la fournir, mais cette information n'est pas obligatoire. Dans ce qui suit, c'est la situation la plus courante (et la plus complexe) qui est considérée, celle dans laquelle l'utilisateur fournit seulement une affirmation.
2. *Algocate* analyse l'affirmation de l'utilisateur et recherche les normes et les preuves les plus appropriées pour générer les arguments valables les plus forts soutenant cette affirmation. Comme la relation de force n'est que partielle (cf. Partie 4.4.2), il est possible que certains arguments ne soient pas comparables. Dans ce cas, *Algocate* renvoie plusieurs arguments.
3. Dans l'étape suivante, Algocate tente de renforcer les arguments en cherchant de nouvelles affirmations. *Algocate* adopte une position neutre et considère à la fois les arguments qui soutiennent l'affirmation initiale et ceux qui la contredisent. *Algocate* renvoie les arguments valables les plus forts basés sur ces nouvelles affirmations (appelées *affirmations générées*). La génération des affirmations est décrite plus précisément dans la Partie 4.4.3.

Plusieurs options sont possibles pour l'utilisateur à ce stade. La situation la plus éclairante est le cas d'un utilisateur qui cherche à contester la décision initiale :

— Si *Algocate* a généré des preuves solides soutenant l'affirmation contestataire de l'utilisateur (à l'étape (2) ou l'étape (3)), alors celui-ci peut fournir ces preuves à l'arbitre pour demander une révision de la décision initiale.

— Si *Algocate* a généré des preuves solides allant à l'encontre de l'affirmation de l'utilisateur (à l'étape (3)), alors deux situations se présentent : soit l'utilisateur est convaincu que la décision initiale est légitime, soit il ne l'est pas. Dans le premier cas, le protocole s'arrête puisque le désaccord a été résolu. Dans le second cas, la première option pour l'utilisateur est d'essayer de contester la décision sur une base différente (avec une nouvelle affirmation), déclenchant une nouvelle itération du protocole. La deuxième option, s'il estime que l'argument généré par *Algocate* n'est pas légitime, est de le soumettre à l'arbitre. Cela peut être le cas, par exemple, si l'argument repose sur une norme jugée illégitime par l'utilisateur. Typiquement, l'utilisateur peut estimer que les données de référence ne représentent pas des décisions vertueuses (décisions discriminantes par exemple) ou tout simplement dénigrer le bien-fondé des règles invoquées. Dans ce cas, si l'arbitre confirme que la norme n'est pas acceptable, l'impact de son verdict va au-delà de la décision contestée par l'utilisateur : l'opérateur du SDA doit modifier le système pour corriger la situation (et l'ensemble des normes utilisées par *Algocate* doit être mis à jour en conséquence).

4.4.2 Relation de force pour les normes impliquant des données

La relation de force utilisée pour comparer les arguments doit répondre à deux critères : elle doit refléter l'intuition des parties et elle doit permettre de comparer le plus d'arguments possibles. Ces deux critères peuvent être en tension, comme le montre la Partie 4.3.4 qui introduit une relation simple et intuitive, mais qui laisse de nombreux arguments incomparables. Pour résoudre cette tension, *Algocate* s'appuie sur un score t mesurant la force des preuves basées sur des données :

$$[\Delta, \gamma, \mu] \succcurlyeq_e [\Delta, \gamma', \mu'] \Leftrightarrow t(\Delta, \gamma, \mu) \geq t(\Delta, \gamma', \mu') \tag{4.14}$$

$$t(\Delta, \gamma, \mu) = \mu\sqrt{\gamma} \tag{4.15}$$

Intuitivement, la preuve est considérée forte si l'écart μ est élevé, puisque cela signifie que la condition C a un fort impact sur la moyenne. Cependant, si la taille γ de $\Delta_{|C}$ est trop petite, l'écart observé pourrait être seulement dû au hasard. Ainsi, une preuve impliquant un ensemble de seulement deux cas doit être considérée plus faible qu'une preuve s'appuyant sur un sous-ensemble de 100 cas. Pour en tenir compte, il est courant de comparer la valeur moyenne de $\Delta_{|C}[o]$ avec l'espérance de la valeur moyenne d'un tirage aléatoire de même taille dans $\Delta[o]$. Selon la loi des grands nombres, l'écart-type moyen de ce sous-ensemble aléatoire est proportionnel à $1/\sqrt{\gamma}$. La valeur t mesure donc l'écart de moyenne observé en échelle d'écart-type moyen d'un tirage aléatoire [6]. La définition (4.15) rappelle le *test de Student* qui peut être facilement converti en *p-valeur* [7]. Même si la définition (4.15) est intuitive et semble posséder des propriétés statistiques raisonnables, d'autres choix sont possibles pour mesurer la force des preuves dans *Algocate* [8].

4.4.3 Génération d'affirmations

Le système de contestation et de justification ayant vocation à être neutre, il doit considérer indifféremment les arguments qui soutiennent l'affirmation de l'utilisateur et les arguments qui la contredisent. L'objectif de la procédure de génération d'affirmations d'*Algocate* est illustré par la Figure 4.3 pour une norme basée sur une référence. La partie gauche de la Figure représente les données de référence Δ_R, la décision initiale en rouge et l'affirmation de l'utilisateur (indiquant que les décisions dans la zone hachurée orange, qui représente $\Delta_{|C}$, devraient être 1). Cette affirmation est faiblement soutenue par la norme. La partie centrale de la Figure montre un argument plus fort soutenant la même conclusion (avec une condition supplémentaire représentée par la barre horizontale). Cependant, la partie droite de la figure fournit un argument encore plus fort (représenté par les deux barres verticales)

6. À un facteur multiplicatif près qui est de peu d'importance car t est employée ici comme une mesure relative.
7. La p-valeur peut être interprétée comme la probabilité que l'écart observé soit fortuit.
8. Par exemple, une définition reposant sur l'entropie de Shannon aurait également pu être considérée et aurait eu l'avantage de privilégier les affirmations conduisant à des groupes purs ce qui peut sembler convaincant. Toutefois, un désavantage notable de ce critère est qu'il ne fonctionne, à notre connaissance, que pour des mesures (références ou objectifs) binaires.

qui va à l'encontre de l'affirmation de l'utilisateur. Il s'agit en effet de l'argument le plus fort selon la relation de force $\geq_e$ définie dans la partie précédente. Visuellement, les tailles des ensembles sélectionnés sont à peu près similaires et la proportion de décisions négatives dans (c) est nettement plus élevée que la proportion de décisions positives dans (b).

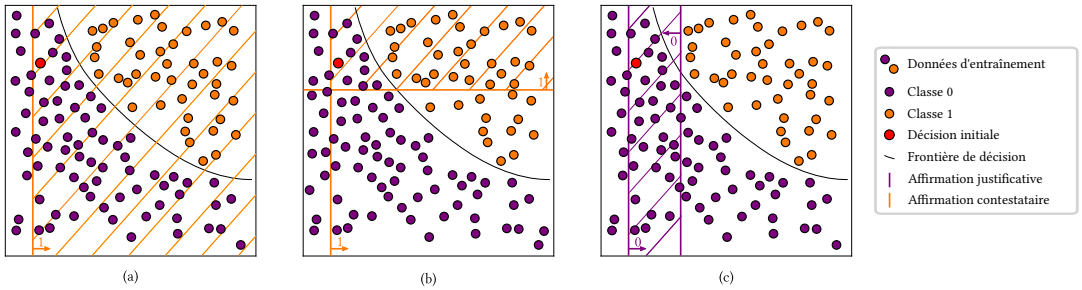


FIGURE 4.3 – Représentation schématique de la génération d'affirmations absolues pour une norme basée sur une référence. Les graphiques représentent les données de référence (données d'entraînement). Les deux classes sont représentées en orange et en violet. (a) L'affirmation initiale de l'utilisateur (ligne verticale orange) indique que toutes les décisions dans la zone hachurée en orange devraient être 1. Elle est faiblement soutenue par les données de référence (la zone hachurée orange contient de nombreux points violets). (b) Affirmation générée soutenant la position de l'utilisateur : toutes les décisions dans la zone hachurée orange devraient être 1. Une preuve forte soutient cette affirmation car le rapport orange/violet dans la zone hachurée est très élevé. (c) Affirmation générée contredisant l'affirmation de l'utilisateur : toutes les décisions dans la zone hachurée violette devraient être 0. Une preuve forte soutient cette affirmation car le rapport violet/orange est exactement égal à 1.

Plus généralement, le but de la procédure de génération d'affirmations est de trouver les conditions C^* et la décision δ (0 ou 1) telles que :

1. l'affirmation générée est cohérente avec la décision initiale x_s : $C^*(x_s)$ est vrai,
2. l'affirmation générée est un renforcement de l'affirmation initiale : $\forall x, C^*(x) \implies C(x)$,
3. la preuve soutenant cette affirmation est maximale : $t(\Delta, \gamma^*, \mu^*)$ est maximisé.

avec $\gamma* = \gamma(\Delta, C^*)$ et $\mu* = \mu(\Delta, C^*, \delta)$.

Dans la Figure 4.3, on voit que toutes les zones hachurées incluent la décision initiale (point rouge) et renforcent l'affirmation initiale (la ligne verticale dans (a) est réutilisée dans (b) et (c)). Aussi, il apparaît que les affirmations générées dans les figures (b) et (c)

contiennent un maximum de points corroborant l'affirmation (respectivement des points oranges et violets).

L'avantage de limiter la génération à des renforcements de l'affirmation de l'utilisateur est de garantir que les réponses du système prennent pleinement en compte les préoccupations de l'utilisateur. Cette personnalisation rend l'interaction plus constructive et plus pertinente. Bien sûr, il est toujours possible, pour n'importe quelle partie, de commencer une nouvelle interaction sur une base complètement différente (avec une affirmation différente), comme l'illustre U2 dans la Figure 4.2.

Techniquement, la procédure de génération considère toutes les normes possibles dans l'ordre décroissant de priorité. Chaque étape prend en entrée (1) la décision initiale x_s, (2) l'affirmation de l'utilisateur qui contient deux informations (C et δ) et (3) la base de données Δ correspondant à la norme. Pour les affirmations absolues, il produit un ensemble de triplets $\{(a_i^*, \diamondsuit_i^*, v_i^*), i = 1...K\}$ [9]. L'affirmation générée est obtenue en concaténant ces triplets à l'affirmation initiale, ce qui garantit la proriété de renforcement :

$$C^* = C \wedge a_1^* \diamondsuit_1^* v_1^* \wedge ... \wedge a_K^* \diamondsuit_K^* v_K^* \tag{4.16}$$

Avec cette définition de C^*, la couverture de l'affirmation générée est $\gamma(\Delta, C^*)$ (le nombre de points dans les zones hachurées de la Figure 4.3 (b) et (c)). L'objectif de la procédure de recherche est de trouver l'ensemble des triplets conduisant à la plus grande valeur de $t(\Delta, \gamma, \mu)$:

$$\max_{(a_i', \diamondsuit_i', v_i'), i=1...K} t(\Delta, \gamma', \mu') \tag{4.17}$$

Comme il ne semble pas exister de solution analytique générale à ce problème de maximisation, une stratégie de recherche exhaustive est employée dans *Algocate*. Par ailleurs, comme la complexité du problème est exponentielle en K, un algorithme glouton est utilisé pour trouver une solution approximative en un temps raisonnable. À chaque étape, tous les triplets possibles $(a_i', \diamondsuit_i', v_i')$ vérifiant la propriété $x_s[a_i'] \diamondsuit_i' v_i'$ (afin que l'affirmation soit cohérente avec la décision initiale) sont considérés et le triplet conduisant à la plus grande

9. Ensemble de paires $\{(a_i^*, \diamondsuit_i^*), i = 1...K\}$ pour les affirmations relatives. La génération des affirmations relatives n'est pas discutée ici car elle fonctionne de manière très similaire.

valeur de $t(\Delta, \gamma', \mu')$ est sélectionné. L'itération s'arrête lorsque la p-valeur associée au t-test du meilleur triplet est supérieure à un seuil donné [10].

4.5 Expérimentation impliquant des utilisateurs

Afin d'évaluer la démarche proposée dans ce chapitre, nous avons mené une étude impliquant des utilisateurs dans le cadre de la collaboration avec la CNIL. L'expérimentation a eu lieu conjointement avec celle d'IBEX présentée dans la Partie 3.6 dont le contexte est décrit dans la Partie 3.6.1. Dans la Partie 4.5.1, nous décrivons le déroulement de l'expérimentation d'Algocate et dans la Partie 4.5.2, nous détaillons les principaux résultats.

4.5.1 Déroulement

L'expérimentation d'Algocate s'est déroulée dans les mêmes conditions que celle d'IBEX, c'est-à-dire en ligne [11], après une courte présentation et sur la même mise en situation dans le domaine du crédit automobile. C'est le même algorithme qui est utilisé pour simuler le SDA d'octroi ou de refus de crédit (Partie 3.6.2).

Dans la mise en situation présentée aux utilisateurs, l'entreprise de crédit déclare que le SDA s'appuie sur les trois normes suivantes, classées par ordre de priorité, pour prendre les décisions :

1. NORME 1 : Règle métier d'exclusion des demandes ayant été deux fois en défaut dans les 6 derniers mois.

   Afin de minimiser les risques de défauts et de surendettement, les personnes ayant été au moins deux fois en défaut de crédit dans les 6 derniers mois sont systématiquement refusées.

2. NORME 2 : Règle métier pour l'acquisition de nouveaux clients.

   Afin de faciliter l'accès au marché de nouveaux clients, les personnes demandant un

10. L'implémentation actuelle d'*Algocate* utilise la valeur 0,2 qui peut facilement être ajustée en fonction des besoins de l'application.
11. La plate-forme sera accessible quelques mois après l'étude à l'adresse suivante : `algaudit.inrialpes.fr/`. Le code de la plate-forme est publiquement accessible `https://gitlab.inria.fr/chenin/algaudit`.

crédit pour la première fois sont acceptées, quelles que soient les caractéristiques de leur demande, à la seule condition que le prix du bien soit inférieur à 55 000 €.

3. NORME 3 : Objectif de minimisation des risques de défaut.
   Les risques de défaut de crédit doivent être minimisés. Une base historique de demandes de crédit est utilisée pour estimer les risques de défaut d'une nouvelle demande.

Les deux premières normes sont basées sur une règle, alors que la troisième est basée sur un objectif. La base de données historique mentionnée dans la troisième norme est la base de données d'entraînement de l'arbre de décision dont la variable cible est l'événement de défaut. Les participants ont accès à une version simplifiée d'Algocate permettant de faire des affirmations absolues (Figure 4.4) et de consulter les justifications ou contestations correspondantes (Figure 4.5).

FIGURE 4.4 – Interface d'Algocate (capture d'écran de l'expérimentation). Les participants peuvent sélectionner la décision attendue (δ) ainsi que l'ensemble des conditions qui constitue la condition C ($\Diamond_p$ et v_p). Par souci de lisibilité, seules six des treize variables sont présentées sur cette figure.

Avant de commencer les tâches de l'expérimentation, les participants peuvent tester Algocate sur deux exemples sans consigne particulière. Ensuite, six demandes de crédit, tirées

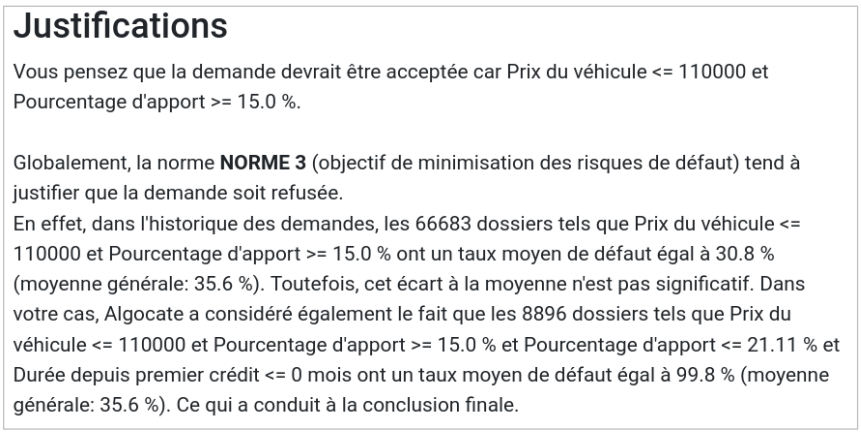


FIGURE 4.5 – Justification générée par Algocate (capture d'écran de l'expérimentation) en réponse à l'affirmation de la Figure 4.4. La justification prend la forme d'un texte généré automatiquement à partir de phrases modèle. Ici, la réponse d'Algocate va à l'encontre de l'affirmation de l'utilisateur.

au hasard parmi les données d'entraînement, sont présentées aux participants (données en entrée et décision du SDA). Pour chaque décision, les participants doivent répondre aux trois questions suivantes en ayant Algocate à leur disposition :

— La décision vous paraît-elle justifiée au regard des normes déclarées par la société (oui ou non) ?

— Indiquer le niveau de confiance de votre réponse (sur une échelle de 1 à 5).

— Justifier votre choix (sous forme de texte libre).

Puisque les normes considérées décrivent effectivement le fonctionnement du SDA, afin d'introduire artificiellement des décisions "injustifiées", une décision sur deux est volontairement inversée par la plate-forme d'expérimentation. Ainsi, avec une probabilité 1/2, les données d'entrée correspondant à une demande acceptée par le SDA sont associées à une décision de refus, ou inversement. Les réponses des participants ont été corrigées grâce au SDA, à l'arbre de décision et aux données d'entraînement. L'objectif de ces questions étant de mesurer la capacité des utilisateurs à détecter ces décisions "injustifiées" à l'aide d'Algocate.

4.5.2 Réception et résultats

Les retours qualitatifs obtenus sur Algocate offrent un regard contrasté sur l'outil. En premier lieu, l'immense majorité des utilisateurs a pu intégrer rapidement les notions de normes, de contestation et de justification. Une partie des répondants estime par ailleurs que l'utilisation de normes dans le cadre de contrôle ou d'audit algorithmique pourrait être éclairante puisqu'elle permettrait de confronter directement le responsable avec ses déclarations. Toutefois, les répondants ont globalement trouvé l'outil complexe à utiliser (notamment en comparaison d'IBEX) et ont souligné qu'en l'état, il n'était pas adapté aux contrôles et audits car les responsables de SDA n'explicitent généralement pas les normes comme l'exigerait Algocate. Globalement, l'outil a davantage été perçu comme un outil à destination des clients ou des utilisateurs finaux qu'à destination d'auditeurs.

Concernant la prise en main de l'outil, une partie des répondants a déclaré avoir eu des difficultés à comprendre l'interface d'affirmations. C'est en particulier le nombre de champs pouvant être remplis et l'utilisation des symboles ($\geq$ et $\leq$) qui semblent avoir gêné certains répondants. En revanche, les participants ont unanimement compris et globalement apprécié la forme des justifications et particulièrement l'usage des statistiques qui rendent l'argument compréhensible et neutre.

Enfin, les normes elles-mêmes ont été globalement bien comprises. C'est la norme sous forme d'objectif qui a soulevé le plus d'interrogation de la part des participants. Elle a été qualifiée de norme "floue" dans le sens où elle peut être utilisée pour soutenir et pour contester une même décision en fonction de l'affirmation choisie. Selon nous, ces réflexions sont saines puisqu'elles correspondent à une réalité dans laquelle certains sous-ensembles du jeu de données soutiennent un refus alors que d'autres soutiennent une acceptation pour une même décision.

Nous avons également extrait quelques critères quantitatifs de l'expérimentation. Au total, 29 agents de la CNIL ont participé à l'expérimentation et ont complété 165 tâches. D'abord, Algocate a été utilisé par les participants en moyenne 1,3 fois par tâche à réaliser. L'utilisation n'est toutefois pas uniforme puisque 56 tâches ont été faites sans Algocate. Les affirmations des participants sont le plus souvent cohérentes avec le fonctionnement

du SDA. Nous avons comparé les décisions soutenues par les participants avec les conditions sélectionnées pour les cinq variables les plus utilisées et, dans 86 % des cas, la condition proposée par le participant était cohérente avec le fonctionnement attendu du SDA. Par exemple, un participant soutenant que la demande doit être acceptée soulignait le fait que le pourcentage d'apport est élevé, facteur jouant effectivement en faveur de l'acceptation, et non pas l'inverse. C'est un élément rassurant qui tend à montrer que les grands principes de la dialectique de justification et de contestation ont été compris.

Concernant les performances des utilisateurs, ils ont été capables de détecter les décisions injustifiées dans 65 % des tâches. Puisqu'une partie importante des tâches a en fait été réalisée sans utiliser Algocate, nous avons pu décomposer les résultats entre deux groupes dans le Tableau 4.2. On voit que les participants qui ont utilisé Algocate ont obtenu de meilleurs résultats (67 % contre 61 %), toutefois la différence n'est pas significative. La confiance auto-déclarée (sur une échelle de Likert allant de 1 à 5) nous indique qu'Algocate améliore globalement la confiance des participants, notamment lorsque ces derniers ont obtenu la bonne réponse.

A utilisé Algocate \ **Réponse utilisateur**	Correcte	Incorrecte	%
Oui	73 (4,1)	36 (3,6)	**67 % (3,9)**
Non	34 (3,8)	22 (3,4)	**61 % (3,6)**
Total	**107**	**58**	**65 % (3,8)**

TABLE 4.2 – Résultats d'expérimentation : taux de succès pour détecter une décision injustifiée. Comparaison des utilisateurs ayant ou n'ayant pas utilisé Algocate. Le nombre entre parenthèses est la confiance auto-déclarée moyenne sur une échelle de Likert entre 1 et 5.

Afin de tester la pertinence des informations fournies par Algocate, nous avons comparé les réponses fournies par les utilisateurs avec les justifications qui leur avaient été proposées (Tableau 4.3). Premièrement, on voit que les participants ont suivi Algocate dans 82 % des cas, ce qui est un signal positif de la confiance accordée aux justifications. Il apparaît également qu'Algocate ne fournit pas systématiquement la bonne réponse. Cela n'est pas surprenant dans la mesure où les utilisateurs peuvent influer sur les justifications grâce aux

affirmations. Au total, avec les affirmations de nos participants, Algocate a indiqué une réponse correcte dans 74 % des cas ce qui est tout de même supérieur au taux de réussite des participants.

Réponse utilisateur / **Réponse Algocate**	Correcte	Incorrecte
Correcte	67	14
Incorrecte	6	22

TABLE 4.3 – Résultats d'expérimentation : comparaison des réponses d'Algocate et des participants.

Ces analyses nous enseignent que pour que le taux de succès dans cette tâche soit plus élevé, des améliorations pourraient être apportées sur les points suivants :

— La confiance des utilisateurs dans Algocate devrait être augmentée. En effet, si Algocate avait été systématiquement utilisé et suivi, le taux de succès serait probablement proche de 74 %.

— Afin de limiter les erreurs d'Algocate, il faudrait favoriser le dialogue contradictoire entre plusieurs parties aux objectifs opposés ce qui rapprocherait le système de son utilisation normale.

Pour résumer, cette expérimentation apporte de riches enseignements sur l'utilisation du protocole Algocate. D'abord, elle valide la possibilité d'implémenter le protocole Algocate et de le mettre à disposition d'utilisateurs non-spécialistes du domaine. Même si l'interface a été jugée complexe, les grands principes ont globalement été compris rapidement par les participants. Malheureusement, ces premiers résultats ne permettent pas d'apporter une preuve forte qu'Algocate est efficace en condition réelle pour détecter des décisions ne correspondant pas à un ensemble de normes fixées.

Ces premiers résultats ouvrent la voie à de nouvelles expérimentations qui prendraient la forme de jeux de rôle dans lequel les participants joueraient différentes parties prenantes (personne affectée par la décision, responsable de la décision et arbitre). En s'opposant sur des décisions individuelles, les participants seraient encouragés à utiliser pleinement l'outil

de justification et les affirmations provenant de chaque partie prenante offriraient des points de vue contradictoire plus proche de la réalité.

Chapitre 5

Étude de cas : l'algorithme d'attribution des greffons cardiaques

Les solutions proposées pour l'explication ou la justification de décisions reposant sur des SDA doivent être confrontées au terrain. Comme cela a été mentionné dans la Partie 2.3, l'évaluation des méthodes doit comprendre une validation par des utilisateurs qui prennent en compte tous les enjeux du SDA. Dans le cadre de cette thèse, une collaboration de trois ans a été menée avec l'Agence de la Biomédecine (ABM) sur le Score National d'Attribution des Greffons Cardiaques (SNAGC), aussi appelé communément "Score Cœur". Elle s'est déroulée en trois phases principales :

1. Une phase préliminaire qui a permis de caractériser le projet du Score Cœur grâce à des discussions avec ses concepteurs et une étude bibliographique,
2. Une phase d'enquête auprès des utilisateurs du Score Cœur dans quatre hôpitaux français qui a permis de compléter l'analyse bibliographique dans la caractérisation des enjeux [1],
3. Une phase de développement d'outils d'explication et de justification des décisions reposant sur le Score Cœur à destination des médecins et de l'ABM.

1. Une analyse de ces entretiens a donné lieu à une publication dans une revue de sociologie [54].

Ces trois phases ont conduit aux contributions suivantes :

— Une (ré)implémentation du Score Cœur à partir de la documentation permettant de valider la politique de transparence annoncée par l'ABM [2],

— Une analyse sociologique des enjeux du Score Cœur décrivant l'émergence, du SDA dans le champ de la greffe cardiaque ainsi que son impact sur les pratiques des professionnels [54],

— Le développement d'un outil d'explication et de justification sous la forme d'une interface web [3].

5.1 Le Score Cœur

Lancé en janvier 2018, le Score Cœur est un SDA pour l'attribution des greffons cardiaques. L'étude de cet algorithme est particulièrement intéressante dans le cadre d'une recherche sur les explications et justifications de SDA à plusieurs égards :

— Le terrain permet de mettre en lumière certains enjeux liés à l'utilisation de SDA (Partie 5.1.1). Pour identifier ces enjeux, des entretiens semi-orientés, suivant une méthodologie propre aux sciences sociales, ont été menés avec les principaux acteurs (concepteurs et utilisateurs du système),

— En étudiant le fonctionnement du score cœur et sa conception (Partie 5.1.2), il apparaît que l'algorithme embarque différents critères éthiques qui reflètent une certaine vision de l'attribution des greffons. Il est utile de pouvoir faire ressortir ces critères qui font écho aux normes pour la justification d'algorithmes proposées dans le Chapitre 4,

— Le contexte des explications est étudié sur un cas réel impliquant des acteurs exposés de manière différente au SDA ce qui permet d'identifier les différents utilisateurs et leurs besoins spécifiques (Partie 5.1.3).

2. https://gitlab.inria.fr/chenin/sc_explanations.
3. Le code est accessible publiquement en ligne (https://gitlab.inria.fr/chenin/explicoeur), une version d'essai du site est disponible en ligne sur des données générées (http://score-coeur.inrialpes.fr/).

5.1.1 Les enjeux du Score Cœur

Les candidats à une greffe de cœur souffrent d'une insuffisance cardiaque mettant en jeu leur pronostic vital et ont épuisé toutes les alternatives thérapeutiques. Compte tenu de la forte mortalité de ces patients et de leur qualité de vie dégradée, la greffe est considérée comme le traitement le plus adapté à leur situation. L'opération chirurgicale de transplantation pose en général peu de problèmes et la difficulté réside surtout dans l'obtention d'un greffon, car cela implique qu'une personne compatible et consentante se trouve en état de mort encéphalique. Au cours de l'année 2016, l'ABM a comptabilisé deux fois plus de candidats inscrits sur liste d'attente que de greffons disponibles [5], il s'agit donc d'une situation de "choix tragique" [20] dans laquelle un traitement est disponible, mais en quantité insuffisante pour répondre aux besoins de la population. Afin d'obtenir au juste moment un greffon compatible, les médecins transplanteurs des centres de greffe (côté receveur) ont intérêt à collaborer en mettant en place une liste nationale. Ce milieu reste pourtant compétitif puisque les médecins, désireux de soulager les souffrances de leurs patients, souhaitent s'approprier la ressource, quitte à employer des stratégies illicites au détriment des autres patients en attente [109].

Dans ce contexte, le système assurant l'attribution de cette ressource précieuse joue un rôle central dans la prise en charge. Depuis 2018, c'est le Score Cœur, un algorithme développé par une équipe de l'ABM, qui établit l'ordre de proposition des greffons[4]. Concrètement, l'algorithme du Score Cœur applique un ensemble de règles et de formules aux caractéristiques numérisées du donneur et des candidats pour calculer un score de priorité. Ce traitement informatique combine des modèles prédictifs issus des données historiques de la greffe et des règles éditées "à la main".

Avant l'instauration du Score Cœur en 2018, l'attribution des greffons est principalement locale (le greffon va dans le centre de greffe le plus proche du lieu de prélèvement) et les médecins transplanteurs ont une grande autonomie dans le choix des patients à greffer. En 2004, et jusqu'en 2018, un régime d'exception, appelé "super urgence", est mis en place pour

4. Il est important de noter qu'il s'agit bien d'un ordre de proposition ce qui veut dire que le médecin transplanteur peut toujours refuser le greffon. Toutefois, cet ordre a un impact majeur sur la prise en charge puisqu'il peut exclure certains types de candidats de la greffe.

permettre à certains candidats de profiter d'une priorité nationale d'accès. Tout candidat à la greffe recevant de la dobutamine —un médicament utilisé en cas de graves dysfonctions cardiaques— peut prétendre à ce statut de super urgence. Il est communément admis que ce système de priorité basé sur la dobutamine a été détourné par les médecins (au moins dans certains centres) afin d'accorder la priorité à leurs patients au détriment du reste de la liste[5]. D'ailleurs, plus de la moitié des greffés ont bénéficié du niveau maximum de priorité l'année ayant précédé la mise en place du Score Cœur [3]. Ce dispositif, ainsi que ces abus ont été critiqués pour des raisons médicales, d'efficacité et d'équité d'accès aux soins. Certains médecins, dont une partie se retrouve associée à l'équipe de conception, appellent dès 2011 à la mise en place d'un système d'attribution indépendant de la pratique médicale qui jugerait "objectivement" la priorité en se basant sur un score calculé automatiquement [34]. Dès le début, l'ambition est donc de normaliser l'attribution des greffons afin d'éviter les tactiques de contournement des règles par les acteurs locaux (centres de greffe).

Pour imposer ce dispositif au milieu de la greffe cardiaque, l'ABM dispose d'une légitimité administrative puisque l'État lui confie explicitement la régulation des activités de transplantation. L'objectif de l'ABM est donc de mener une politique de centralisation et de rationalisation des appariements entre greffons et candidats. Dès 2004, l'ABM met en place la base "Cristal", un système obligatoire de collecte des données relatives aux candidats et aux transplantations. Ces données jouent un rôle primordial dans la conception et le fonctionnement du Score Cœur comme nous le montrons dans la partie suivante. Un autre aspect notable du dispositif est la mise en place de réunions techniques organisées régulièrement à l'ABM pour que les centres de greffe puissent discuter des évolutions de l'algorithme. Christian Jacquelinet, l'un des concepteurs, insiste sur l'importance de ces réunions pour l'adhésion des centres de greffe au système et pour assurer que le Score Cœur s'adapte au mieux à la réalité de la prise en charge. Pour y participer, les médecins doivent être en mesure de comprendre et d'analyser les modalités de fonctionnement de l'algorithme afin de défendre les évolutions qu'ils souhaitent proposer. Comme nous le montrons dans la partie suivante, même si l'algorithme suit globalement des principes éthiques facilement énonçables, leur

5. Une faible dose de dobutamine peut être administrée, ce qui permet de profiter de la super urgence tout en limitant les risques d'effets secondaires.

mise en œuvre pratique peut s'avérer parfois complexe.

5.1.2 Fonctionnement du Score Cœur

L'attribution d'un traitement médical rare est un problème qui dépasse le cadre d'une réflexion purement clinique. Persad *et al.* [92] proposent trois catégories majeures de principes éthiques pour l'aborder.

(1) L'attribution *équitable* donne à chacun les mêmes chances d'accéder à la ressource. Le tirage au sort ou la règle du "premier arrivé, premier servi" en sont des exemples.

(2) La priorité au plus mal-loti ou *prioritarisme* accorde la priorité au patient qui souffre le plus, ou au patient le plus jeune qui a le moins profité de ses années de vie en bonne santé.

(3) La maximisation des retombées positives est un calcul *utilitariste* dont le but est de sauver le plus grand nombre de vies ou le plus grand nombre d'années de vie. Les patients qui ont le meilleur pronostic sont donc privilégiés.

Les auteurs précisent qu'aucun de ces principes n'est suffisant. Il n'est pas acceptable d'attribuer la ressource au patient qui souffre le plus si ce dernier est condamné à brève échéance (greffé ou pas), pas plus qu'il n'est acceptable de l'attribuer à celui qui en profitera le plus si ce dernier peut attendre alors qu'un autre patient est en situation d'urgence. Seule une combinaison de ces principes peut satisfaire nos exigences morales [6].

Le score cœur de l'ABM réalise un compromis entre ces trois critères en combinant quatre éléments (le risque de décès sur la liste d'attente, le risque de décès post-greffe et les appariements physiologique et géographique entre le donneur et le candidat) qui sont détaillés successivement [4] :

— Le "*Candidate Risk Score*", construit à partir des données Cristal, prédit, avec une précision jugée suffisante par les auteurs, le risque de mortalité des candidats sur liste d'attente (prioritarisme) tout en limitant la prise en compte des pratiques médicales que les médecins pourraient utiliser pour manipuler le système [65].

6. Ces principes peuvent être vus comme des normes servant à justifier les décisions de l'algorithme (Partie 4.3.2). Ils illustrent, sur un exemple concret, comment différentes normes peuvent être embarquées dès la conception d'un SDA.

— En 2018, un score prédictif du risque de décès post-greffe (utilitarisme), également construit à partir des données Cristal, est publié par l'ABM [66]. Cependant, la précision de ce score est jugée insuffisante pour envisager son implémentation complète dans le Score Cœur. Il est pris en compte en demi-mesure pour filtrer les candidats dont les chances de survie un an après la greffe sont inférieures à 50 % (alors que le taux moyen de survie s'établit autour de 85 %).

— Les règles de compatibilité pour l'appariement physiologique sont protéiformes et assurent différentes fonctions. Elles sont en premier lieu utilitaristes puisqu'elles évitent les greffes ayant peu ou aucune chance de succès (groupes sanguins ou morphologies incompatibles). L'appariement sur l'âge a un statut hybride puisqu'il vise à la fois à assurer l'utilitarisme (en greffant un cœur jeune sur un patient jeune, l'utilité de la ressource est maximisée) et le prioritarisme grâce à un appariement asymétrique (à différence d'âge égale, le patient jeune est privilégié). Enfin, l'appariement sur le groupe sanguin est conçu de manière à assurer l'équité vis à vis des candidats ayant un groupe sanguin rare qui ont statistiquement moins de chance de trouver un donneur compatible.

— L'appariement géographique permet d'éviter de transporter le greffon sur une longue distance. En effet, un transport long réduit les chances d'une greffe réussie (utilitarisme) et représente un certain coût que l'ABM souhaite maîtriser.

Le score final de l'ABM, qui combine ces différents éléments de manière complexe, prend la forme d'un indicateur compris entre 0 et 1152 "points" calculés pour l'ensemble des candidats inscrits sur la liste d'attente de greffe dès qu'un greffon est disponible. Le greffon est proposé par ordre décroissant de cet indicateur. Un système d'exception, appelé "composante expert", autorise les centres de greffe à demander une dérogation pour les patients souffrant de certaines pathologies spécifiques qui sont mal prises en compte par le "*Candidate Risk Score*". La demande prend la forme d'un dossier évalué par un comité d'experts extérieurs au centre du patient et peut conduire, si elle est acceptée, à l'obtention d'une prime de points qui accorde une priorité importante à celui qui en bénéficie.

Les informations contenues dans la documentation publique fournie par l'ABM sont

claires et détaillées. Nous avons notamment pu vérifier qu'elle était suffisamment précise pour permettre de coder entièrement le Score Cœur. Nous nous sommes livrés à cet exercice et avons pu valider, avec la collaboration de l'ABM, que les résultats obtenus dans les deux implémentations étaient les mêmes. Pourtant, malgré cette stricte transparence, comprendre comment ces différents critères éthiques sont combinés dans des cas particuliers reste un problème difficile qui nécessite au minimum un certain temps d'apprentissage.

5.1.3 Les enjeux de l'explication et de la justification

Pour compléter notre analyse bibliographique, nous avons mené une série de vingt-six entretiens conduits dans quatre centres hospitalo-universitaires français, dont un centre parisien, auprès de médecins transplanteurs et de personnels soignants. Ces entretiens, menés entre juin 2019 et décembre 2019, avaient pour objectif d'aborder la question de la compréhension de l'algorithme ainsi que d'étudier son utilisation. Ces entretiens ont permis, entre autres, d'identifier les enjeux de l'explication et de la justification dans le cas particulier de cet algorithme d'attribution.

La compréhension du Score Cœur est loin d'être uniforme. Au cours des entretiens, nous avons systématiquement demandé des explications sur l'algorithme ainsi que la liste des critères pris en compte afin d'estimer le niveau de connaissance des répondants en comparant les réponses aux informations de la documentation officielle [4]. Il apparaît que certaines variables sont bien connues alors que d'autres ne sont presque jamais citées. La prise en compte de la Circulation Extra-Corporelle (CEC)[7], par exemple, est connue de tous sauf un, alors que seules les deux secrétaires interrogées savent que les sexes du donneur et du receveur sont pris en compte. De même, moins de 20 % des répondants savent que la maladie initiale est prise en compte. Une première explication est que les utilisateurs de l'algorithme se focalisent sur les variables les plus importantes du système. Cela semble être corroboré par le propos de cette cardiologue : "C'est principalement celles que je regarde parce que je sais que c'est là que sont les points. C'est ça qui va pondérer mon patient et le classer."

7. La CEC consiste à remplacer temporairement le travail du cœur à l'aide d'une pompe extérieure. Il s'agit d'une intervention d'urgence aux effets secondaires importants visant à assurer une circulation sanguine minimale nécessaire au fonctionnement des autres organes vitaux.

Par ailleurs, alors que le score est composé de quatre éléments principaux (Partie 5.1.2), l'effort de compréhension se focalise en quasi-totalité sur le risque de décès sur liste d'attente. Pourtant, les variables d'appariements jouent un rôle au moins aussi important que les variables de gravité dans la sélection du candidat. Les médecins semblent focaliser leur attention là où ils ont un contrôle ou une impression de contrôle ; ils se désintéressent ainsi de l'appariement entre le donneur et le receveur, car cela ne constitue pas un levier qu'ils peuvent utiliser pour influencer les décisions. Par exemple, ce médecin se montre très pragmatique dans le choix des variables dont il a connaissance en déclarant : "Je connais les choses qui permettent de faire monter le score." La compréhension du Score Cœur s'intègre donc dans une dynamique de prise de contrôle par les médecins, qui reste toutefois très limitée en comparaison de l'autonomie que les médecins avaient dans le système antérieur. La mise en œuvre d'explications de cet algorithme doit prendre en compte cette dynamique pour ne pas encourager le contournement des règles car cela aurait des conséquences délétères à la prise en charge globale (Partie 5.1.1).

Une deuxième observation issue des entretiens est qu'une partie seulement des médecins s'intéresse au fonctionnement du Score Cœur. Environ la moitié des répondants connaît moins d'un tiers des critères pris en compte et beaucoup déclarent avoir des lacunes sur son fonctionnement. Certains montrent même un grand désintérêt pour le fonctionnement de l'algorithme et préfèrent se reposer sur leurs collègues plus investis dans le système. Dans les quatre centres interrogés, seules quelques personnes sont investies dans le score et jouent un rôle de référent. Il s'agit, le plus souvent, de cardiologues expérimentés situés à un niveau assez élevé dans la hiérarchie. Le coût d'apprentissage ne semble justifié que par des interactions régulières avec le système, ce qui conduit à une distribution inéquitable de la connaissance au sein des équipes. Par ailleurs, ceux qui concentrent la connaissance se retrouvent aussi en première ligne pour propager la confiance ou la défiance en l'algorithme au sein des équipes. Ces référents sont généralement les représentants des centres aux réunions techniques organisées par l'ABM pour décider des évolutions du Score Cœur.

5.2 Explicœur

Explicœur est un outil conçu pour faciliter la compréhension du Score Cœur par ses utilisateurs (médecins et ABM). Il regroupe différents outils pour visualiser, comprendre ou prédire les résultats du Score Cœur (Partie 5.2.1). Les outils proposés, qui constituent l'une des contributions de cette thèse, emploient des méthodes issues d'IBEX et d'*Algocate* ou des méthodes développées spécifiquement.

5.2.1 Visualiser, comprendre, prédire

À l'issue des deux phases d'analyse de l'algorithme et des entretiens décrits dans la Partie 5.1, plusieurs observations ont permis de concevoir un système adapté aux besoins des différents types d'utilisateurs. D'abord, malgré une documentation claire [4], il semble exister un coût d'apprentissage du Score Cœur ainsi que certaines difficultés pour mettre en application les informations générales de la documentation dans des cas particuliers. Pour que l'algorithme et les principes éthiques qui le motivent puissent être appréhendés plus simplement, il nous a paru pertinent de fournir en complément de la documentation les éléments suivants : des explications globales, des explications locales, des justifications locales ainsi que des représentations visuelles des données. Les concepteurs de l'algorithme, qui sont ici les responsables de la bonne utilisation globale des greffons, sont naturellement plus intéressés par les explications globales. À l'inverse, les médecins peuvent avoir de l'intérêt pour les informations locales (explication ou justification) et globales selon qu'ils souhaitent comprendre la situation d'un patient spécifique ou le fonctionnement général du Score Cœur (pour participer, par exemple, à une réunion technique de l'ABM). Les modalités de conception d'Explicœur doivent également prendre en compte le risque de l'utilisation des explications pour développer des tactiques de contournement. En accord avec l'ABM, nous avons donc limité l'accès à certaines informations qui pourraient favoriser de telles pratiques.

En plus de la complexité de l'algorithme, il existe une source supplémentaire d'incertitude sur la capacité d'un candidat à être bien classé : les caractéristiques du donneur et le

lieu de prélèvement pris en compte dans les fonctions d'appariement physiologique et géographique. Être capable d'estimer à l'avance, même de manière approximative, le nombre de points d'un candidat a une grande importance dans la pratique clinique, car cette estimation joue un rôle dans le choix de la stratégie de prise en charge. Un patient dont on sait qu'il va obtenir une proposition de greffon dans les dix prochains jours ne sera pas traité de la même manière qu'un patient dont on sait qu'il devra attendre plusieurs mois. Même si prédire avec certitude le délai d'obtention d'une proposition ne semble pas envisageable, des simulations peuvent apporter une estimation utile à la pratique clinique.

En résumé, trois types d'information sont proposés dans l'*Explicœur* avec pour objectifs de (1) *visualiser* synthétiquement les informations d'entrée de l'algorithme, (2) de *comprendre* les traitements réalisés par l'algorithme ainsi que les motivations sous-jacentes et (3) de mieux *prédire* les résultats de l'algorithme afin d'adapter la pratique clinique.

5.2.2 Les données Cristal

Les données Cristal relatives à la greffe cardiaque de l'année 2018 - 2019 ont été utilisées dans le cadre de notre collaboration avec l'ABM dans le but d'expérimenter et de concevoir ces différents outils. Ces données sont principalement structurées en trois types d'information [8] :

— Informations concernant le patient : date de naissance, groupe sanguin, poids, taille, maladie initiale, etc.

— Bilan périodique de suivi avant la greffe [9] : date du bilan, résultats de tests biologiques (NT-ProBNP, créatinine, bilirubine), traitements éventuels (CEC, assistance longue durée, dialyse), etc.

— Informations concernant le donneur : date du prélèvement, groupe sanguin, poids, taille, âge, etc.

8. Pour être exhaustif, il faudrait ajouter les informations sur les composantes experts déjà mentionnées ainsi que les contre-indications temporaires. Ces dernières sont décidées par les centres de greffe lorsqu'ils estiment que leur patient n'est pas en mesure de subir une transplantation (état de santé dégradé, absence).

9. Une mise à jour régulière des informations de ce bilan est nécessaire pour maintenir l'inscription du candidat sur liste d'attente.

Pour qu'un score soit calculé, il faut réunir toutes ces informations. Alors qu'un bilan ne peut être associé qu'à un seul patient, toutes les combinaisons de (patient, bilan) avec un donneur peuvent *a priori* être considérées. Par souci de réalisme, un filtrage sur les dates est réalisé. Ainsi, pour les patients inscrits sur la liste d'attente, c'est-à-dire ni greffés, ni décédés, une paire (bilan, donneur) est envisagée seulement si la condition suivante est vérifiée [10] :

$$\text{date du bilan} \leq \text{date du prélèvement} < \text{date du prochain bilan}$$

De plus, les bilans datés de trois mois ou plus avant le prélèvement sont également exclus. Ce critère est imposé par l'ABM afin d'assurer une mise à jour régulière des informations par les centres de greffe. Dans la suite, les triplets (patient, bilan, donneur) remplissant ces conditions sont appelés des "greffes envisagées".

Deux niveaux de décision peuvent être distingués : le niveau individuel comprenant la valeur du score pour une greffe envisagée et le niveau de l'attribution qui compare les scores de tous les candidats envisagés pour le même donneur. Les critères éthiques (mentionnés dans la Partie 5.1.2) se manifestent pleinement au second niveau car ces critères doivent comparer les caractéristiques des différents candidats. Cette distinction se répercute dans la conception des outils proposés dans la Partie 5.2.3. En effet, le niveau individuel fait plutôt référence au fonctionnement du Score Cœur et est donc mieux adapté à des explications alors que le niveau de l'attribution permet d'aborder pleinement les enjeux éthiques et est donc mieux adapté pour les justifications.

5.2.3 Présentation des outils

Explicœur comprend six outils destinés à apporter des informations complémentaires sur le fonctionnement du Score Cœur. Ces outils sont accessibles à travers une interface web [11]. Les six outils sont présentés et commentés successivement ci-dessous. Les données d'utilisation fournies par l'ABM dans le cadre de la collaboration sont utilisées comme

10. Par souci de simplicité, la condition prenant en compte les contre-indications temporaires n'est pas présentée ici.

11. Une version fonctionnant avec des données générées est accessible en ligne : `http://score-coeur.inrialpes.fr/` pendant quelques mois. Le code source de l'interface est publiquement accessible (urlhttps ://gitlab.inria.fr/chenin/explicoeur)

population au sens d'IBEX (voir Chapitre 3).

Partial Dependence Plot

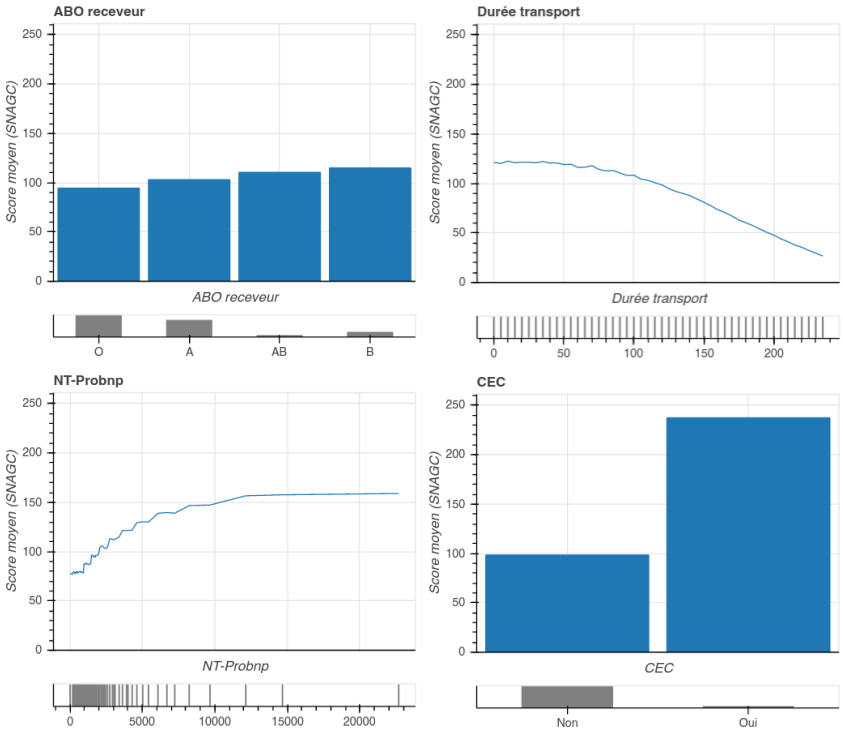


FIGURE 5.1 – Explication globale du Score Cœur sous la forme de *Partial Dependence Plot*. Par souci de concision, seuls quatre graphiques sont présentés. L'abscisse représente les valeurs possibles de la variable. L'ordonnée de la partie supérieure du graphique représente la valeur moyenne du score. La partie inférieure (en gris) représente la distribution de la variable.

Les *Partial Dependence Plot* (PDP) sont une représentation graphique de l'impact moyen de chaque variable sur la sortie de l'algorithme [43]. Il s'agit d'une explication globale dont un exemple est présenté dans la Figure 5.1 pour un sous-ensemble de variables. Le score moyen présenté en ordonnée correspond à la moyenne des scores qu'aurait la population si tous ses éléments partageaient la valeur en abscisse pour la variable considérée. Par exemple, pour calculer le score moyen affiché pour les candidats âgés de 50 ans, la valeur âge = 50 ans est attribuée artificiellement à toute la population et le score moyen de cette nouvelle population est calculé. Dans la Figure 5.1, la distribution de chaque variable est ajoutée en

dessous des valeurs moyennes.

Les PDP représentés dans la Figure 5.1, montrent que le groupe sanguin a en moyenne peu d'impact sur la valeur du Score Cœur. Les moyennes pour les groupes sanguins AB et B, qui sont les groupes rares (voir la distribution sous le graphique), sont légèrement supérieures aux autres groupes. Cela est cohérent avec l'avantage introduit délibérément afin de rétablir l'équité vis-à-vis de ces groupes qui sont naturellement défavorisés. Il apparaît par ailleurs que la présence d'une CEC a un fort impact sur la valeur du score, fait bien connu des praticiens.

Les PDP ont l'avantage de la simplicité puisque leur lecture est univoque. Ils représentent de manière compréhensible le comportement moyen du score sur toute la population. Toutefois, les connaissances qu'ils apportent peuvent être insuffisantes pour comprendre des situations spécifiques.

Arbre de décision

Comme nous le proposons dans le Chapitre 3, un arbre de décision peut être entraîné sur la population (greffes envisagées et leurs scores respectifs) afin d'obtenir une explication globale du système. Cet arbre de décision permet d'identifier les caractéristiques des greffes envisagées qui conduisent à des scores uniformes. En plus des règles affichées dans un arbre de décision traditionnel, des "info-bulles" ont été ajoutées à la demande de l'ABM pour obtenir des informations sur les sous-groupes identifiés.

Dans l'exemple utilisé dans la Figure 5.2, il apparaît que l'information la plus discriminante pour séparer les faibles scores des scores élevés est la valeur du NT-ProBNP [12]. Les seconds critères identifiés par l'arbre de décision sont la CEC et la présence d'une composante expert.

Explications locales

Comme nous l'avons montré dans la partie 3.3, des explications locales peuvent être générées via des opérations d'échantillonnage et de génération. Un exemple d'explication

12. Le NT-proBNP est un marqueur hormonal associé au risque cardio-vasculaire. Les valeurs élevées sont associées à un risque plus élevé.

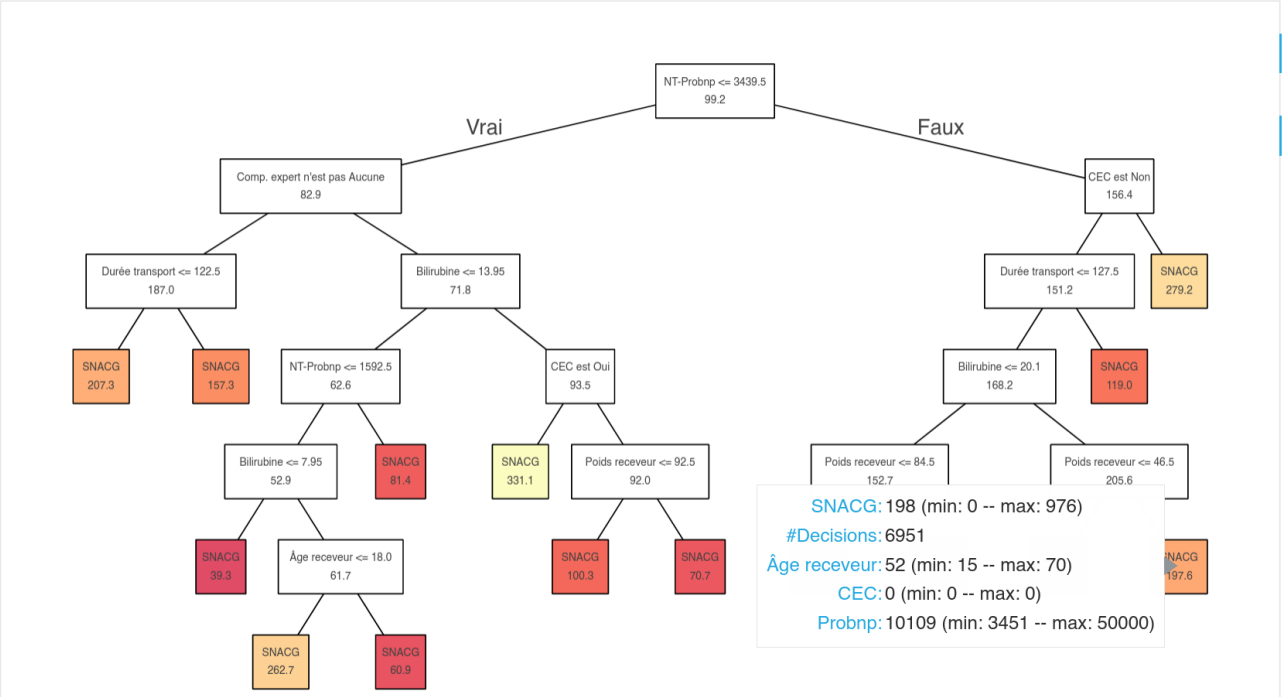


FIGURE 5.2 – Explication globale du Score Cœur sous forme d'un arbre de décision. L'info-bulle fait référence à la feuille en bas à droite de l'arbre.

locale fournie par Explicœur est présenté dans la Figure 5.3. Il contient des informations sur le score, le rang et les variables prises en compte par le Score Cœur dans la partie haute de la figure et propose des explications locales dans la partie basse. Ces explications prennent deux formes : importances relatives des variables et règles.

Pour obtenir ces explications, Explicœur met en œuvre un échantillonnage par permutation avec un filtrage des greffes envisagées ayant un score strictement positif. Ce filtrage permet de limiter la proportion des échantillons ayant un score nul dont une trop forte représentation rend l'étape de génération difficile. L'estimation des importances relatives de variables peut être faite facilement à partir de l'approximation linéaire locale (cf. Partie 3.3.3) et des échantillons ayant servi à la calculer. L'approximation linéaire locale fournit une estimation du gradient en un point donné. Par exemple, si l'approximation linéaire locale pour la variable "durée de transport" est négative, cela signifie que les échantillons ayant une durée de transport élevée ont en moyenne des scores plus faibles. Pour transformer

FIGURE 5.3 – Explication locale d'une décision individuelle. La partie haute rappelle le score et le rang du candidat ainsi que les variables considérées par l'algorithme. Suivent deux explications sous forme d'importances de variables et de règles. Dans l'exemple utilisé ici, le candidat est classé premier sur liste d'attente avec un score très élevé (1001). Les trois critères principaux pour expliquer ce score sont la présence d'une CEC, la compatibilité sanguine et la faible distance de transport (0 minute).

cette grandeur en importance relative de variable, elle est multipliée par l'écart entre la moyenne des échantillons et la valeur du point d'intérêt. Dans les importances de variables de la Figure 5.3, la valeur fortement positive pour la durée de transport signifie que la durée de transport pour ce greffon est plus faible que la moyenne des échantillons ce qui lui permet de "gagner des points" par rapport à la moyenne. Enfin, les explications sous forme de règles décrivent de manière simple les principales caractéristiques du point d'intérêt qui justifient sa valeur de score. Les explications sous forme de contrefactuelles (Partie 3.3.3), calculables avec IBEX, ont été retirées de la liste des explications locales car elles ont été

jugées de faible intérêt par les personnes interrogées lors des entretiens et car leur forme pouvait inciter à des tactiques de contournement des règles.

Justifications locales

Dans une première itération des justifications d'algorithme, nous avons proposé *Algocate* aux concepteurs du Score Cœur en se basant sur les normes affichées dans la documentation du système [4]. Bien que l'intérêt d'avoir à la fois des explications et des justifications a été reconnu, certains aspects de l'implémentation des concepts de justification et de contestation dans *Algocate* pouvaient porter à confusion. D'abord, la notion de contestation dans le cas spécifique du Score Cœur doit être adaptée. En effet, dans le système mis en place par l'ABM, c'est par l'intermédiaire des réunions techniques à l'ABM que les médecins peuvent "contester" ou critiquer l'algorithme. Or, ces réunions ne portent pas sur des décisions individuelles, mais sur les règles d'attribution, *i.e.* sur le choix et l'implémentation des normes applicables. Dans cette situation spécifique, la notion d'affirmation telle que définie dans la Partie 4.3.1 pour une décision individuelle semble moins adaptée. Par ailleurs, alors que le fonctionnement d'*Algocate* implique une analyse au niveau de la décision individuelle, la question qui soulève le plus de questionnement éthique est celle du choix d'un candidat parmi les autres candidats en attente. C'est donc sur le deuxième niveau de décision, relatif à l'ordre de proposition établi par le Score Cœur que doit porter la justification. Compte tenu de ces deux éléments, le calcul et la forme des justifications ont été adaptés par rapport au système *Algocate* tel qu'il est introduit dans la Partie 4.4.

Pour établir les normes de ce système de justification, nous nous sommes appuyés sur la documentation fournie par l'ABM [4] et sur la décomposition du Score Cœur en ses éléments constitutifs (Partie 5.1.2). Au total, Explicœur propose des justifications selon cinq normes qui sont présentées successivement :

— Risque de décès sur liste d'attente (ICAR) : *Candidate Risk Score* de l'ABM,

— Points de composante expert : nombre de points supplémentaires que confère une éventuelle composante expert au candidat,

— Survie post-greffe à un an : score prédictif du risque de décès post-greffe,

— Compatibilité avec le donneur : taux de perte des points après la prise en compte des différents filtrages de comptabilité physiologique,

— Réduction du transport du greffon : taux de perte des points après la prise en compte de l'appariement géographique.

À titre d'exemple, une justification de la décision précédente (Figure 5.3) est présentée dans la Figure 5.4. Pour chaque norme, la valeur correspondante pour la greffe envisagée est comparée à la valeur moyenne des autres candidats inscrits sur la liste d'attente de greffe le même jour en échelle d'écart-type. Ainsi, la valeur 1,1 pour la norme "Réduction du transport du greffon" signifie que le temps de transport du greffon est environ un écart-type en dessous de la moyenne des autres candidats en "compétition" sur ce même greffon : cela tend donc à justifier que le greffon lui soit attribué.

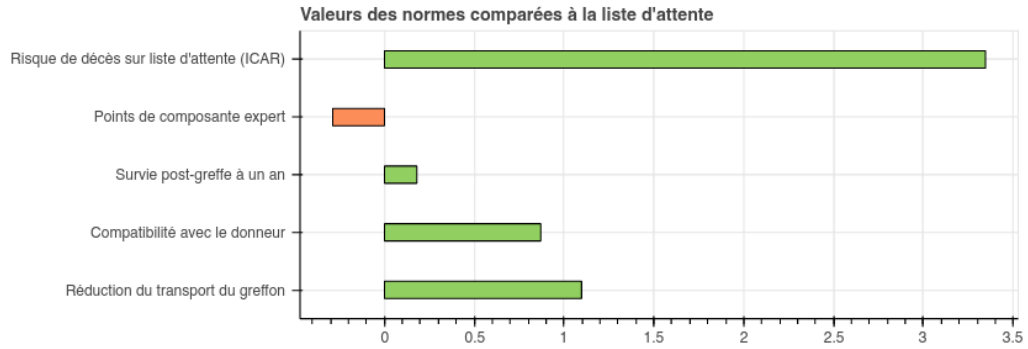


FIGURE 5.4 – Justification de la décision de la Figure 5.3 (présentée sur la même page dans l'interface web d'Explicoeur). Le risque de décès sur liste d'attente, la compatibilité avec le donneur et le temps de transport du greffon justifient que l'algorithme ait classé ce candidat en tête de la liste d'attente.

En comparaison d'Algocate, les justifications prennent ici une forme simplifiée tout en conservant la référence aux normes du système. L'exercice ici a consisté à adapter les grands principes de l'outil complet Algocate à la situation du score cœur et à ses enjeux. Par ailleurs, il est intéressant de noter que les justifications prennent, en un sens, la forme d'une explication boîte blanche puisque nous nous trouvons dans la situation particulière où les choix éthiques sont incorporés dans le SDA à travers les variables de la Figure 5.4.

Estimation des durées d'attente

L'un des buts affichés des travaux sur l'explicabilité est de permettre aux utilisateurs de mieux prédire les sorties d'un SDA. Comme cela a déjà été mentionné, le Score Cœur implique, en plus des difficultés "habituelles" de compréhension de l'algorithme, une autre source d'incertitude, celle qui est liée au hasard des greffons disponibles. Ces incertitudes ont des conséquences sur la prise en charge des patients puisque celle-ci dépend du temps d'attente avant d'obtenir une proposition. Pour réduire l'impact de cette incertitude, Explicœur inclut un système d'estimation des durées d'attente.

Le système d'estimation fonctionne par simulation aléatoire sur la liste d'attente à une date donnée. Il a l'avantage de prendre en compte à la fois les caractéristiques du candidat et les caractéristiques des autres candidats inscrits sur la liste au même moment. Les informations sur les candidats seuls ne sont pas suffisantes pour simuler un classement, car ce dernier dépend fortement des caractéristiques du donneur. Pour effectuer le classement, un donneur est tiré aléatoirement dans la base de données historiques. Nous faisons l'hypothèse que le candidat classé premier accepte la proposition et sort donc de la liste d'attente ; il est alors remplacé par un autre candidat tiré aléatoirement dans la base des données historiques. Les décès sur liste d'attente sont également simulés via une probabilité calculée sur la base du *Candidate Risk Score*. Les candidats dont le décès est simulé sont également remplacés. L'ensemble de ce processus (donneur aléatoire, simulation de greffe et simulation de décès sur liste) constitue une "journée" de greffe [13]. Ces journées sont répétées k fois afin de simuler le devenir des patients en liste d'attente pendant k jours. Ces k itérations constituent une simulation complète. Les simulations sont répétées un grand nombre de fois afin d'obtenir une dynamique moyenne.

L'ensemble de ces simulations permet d'estimer le temps d'attente d'un candidat en faisant l'hypothèse que les caractéristiques et la fréquence des greffons futurs sont proches de celles des greffons historiques. Explicœur affiche les résultats sous forme de probabilités d'obtenir une proposition dans les 3, 7, 14 ou 30 prochains jours (voir Figure 5.5). Par exemple, si dans 50 % des simulations le candidat est classé une fois en tête de liste pour l'un des 7 premiers greffons, alors il sera indiqué qu'il a 50 % de chance de recevoir une

13. Il y a un peu plus de 400 greffes par an en France : il est donc supposé qu'un greffon est disponible chaque jour en moyenne.

proposition dans les 7 prochains jours.

Id patient	Code équipe	Évènement	Durée transpo	Âge	Sexe	ABO	Taille	Poids	Jours depuis bila	NT-Probnp	Créatinine	Bilirubine	Prop. 3 jours	Prop. 7 jours	Prop. 14 jours	Prop. 30 jours
248767	PL7XA	DRG, CEC, XPCP	150	13	F	AB	180	71	6	9666	50	166	64 %	93 %	100 %	100 %
220238	DI3CA	XPCA	135	43	H	A	176	80	7	177	113	26	57 %	87 %	100 %	100 %
246647	MA4XE	DRG, XPCP1, SIA	215	1	F	A	88	13	33	10329	33	22	43 %	80 %	87 %	100 %
248472	BO5CB	XPCP2	140	6	F	A	110	15	78	3038.5	21	79	16 %	50 %	80 %	97 %
229212	MA4XE	XPCP2	115	8	F	O	122	20	30	5517	31	8	14 %	47 %	70 %	100 %
235142	LI1XA	XPCA	70	43	F	O	160	82	2	619	80	5	13 %	43 %	63 %	87 %
248797	PI7XA	DRG, CEC	65	57	H	O	191	84	1	3885	83	16	10 %	17 %	50 %	73 %
248771	PG7CA	XPCA, SIAV-BV	150	52	H	A	182	117	6	1497	83	9	10 %	10 %	63 %	97 %
242902	ST2XA	XPCP1	145	6	H	O	112	17	15	1203	25.5	2.9	10 %	57 %	90 %	100 %
231691	MA4CO	Dialyse	215	47	F	B	156	80	7	5873	NaN	26	10 %	20 %	36 %	67 %
246920	PI7XA		145	52	F	B	161	80	19	6814	163	28	7 %	7 %	17 %	47 %

FIGURE 5.5 – Présentation de l'estimation des durées d'attente. Les caractéristiques des candidats sont rappelées dans les colonnes de gauche et les probabilités d'obtenir une proposition sont affichées dans les colonnes de droite. Cette liste correspond au jour de décision présenté dans la Figure 5.3. Le système d'estimation estime le candidat à la $7^{ième}$ ligne, qui correspond au candidat des figures précédents, a 10 % de chances d'avoir une proposition dans les trois prochains jours.

5.2.4 Réception et conclusion

La durée réduite de la thèse et le contexte sanitaire de l'épidémie de COVID-19 n'ont pas permis d'effectuer une évaluation complète de l'outil Explicœur. Nous avons pu toutefois présenter l'outil à plusieurs reprises à l'équipe des scores d'attribution de l'ABM et à l'ensemble des centres de greffe dans le cadre d'une réunion du groupe technique. Si cela ne remplace pas une évaluation de l'outil, ces échanges ont été l'occasion d'obtenir un retour sur la réception et l'utilité perçue des méthodes d'explication sur un cas d'usage concret.

Comme prévu, l'ABM s'est montrée particulièrement intéressée par les explications globales du système. En particulier, les graphiques de *Partial Dependance Plot* (voir Figure 5.1 pour un exemple) ont été jugés utiles pour visualiser l'impact spécifique des différentes variables de manière compréhensible. La PDP pour le groupe sanguin (en haut et à gauche sur la Figure 5.1) pourrait notamment servir à justifier de l'équité de traitement au regard de ce critère. Plus généralement, ces explications globales qui sont calculées avec des méthodes indépendantes de l'ABM sont vues comme une façon de justifier les choix de conception à des acteurs extérieurs à l'ABM. L'explication a à la fois une valeur pédagogique et une valeur de normalisation de la représentation du fonctionnement de l'algorithme.

De leur côté, les médecins qui se sont exprimés sur l'Explicœur ont perçu un intérêt pour s'assurer d'une forme de contrôle sur le Score Cœur dans un contexte où la transparence sur ce type de système, déjà entérinée par la loi pour une république numérique, occupe une place grandissante dans le débat public. La volonté de comprendre semble s'intégrer dans une crainte de ne pas pouvoir répondre aux critiques extérieures au domaine de la greffe cardiaque. Par ailleurs, certains médecins affirment que l'outil peut s'avérer utile pour informer les patients et établir la stratégie de prise en charge de manière plus informée.

En conclusion, même si l'outil n'a pas pu être évalué formellement dans le cadre d'une étude scientifique, les premiers retours sont encourageants et tendent à montrer une bonne réception des informations mises à disposition.

Chapitre 6

Conclusion

6.1 Limites et perspectives

La durée réduite de la thèse nous a contraint à limiter nos recherches et à ne pas suivre toutes les pistes qui auraient pu ou dû être explorées. Par exemple, le rôle et l'impact des étapes d'échantillonnage et de génération sur les propriétés de l'explication finale mériteraient d'être plus détaillés. Des études complémentaires, comme celle proposée dans [57] sur le lien entre l'étendue de l'échantillonnage et la généralité de l'explication, devraient être faites pour rationaliser et rendre plus efficace la recherche d'explications avec un système d'évaluation post-hoc. Il faut reconnaître que la question des liens entre les options techniques et les propriétés de l'explication est complexe du fait de l'hypothèse boîte noire. Dans le système de contestation et de justification Algocate, la possibilité pour les différentes parties prenantes de proposer des normes devrait être testée pour en vérifier la pertinence. Cela impliquerait que plusieurs parties prenantes, aux objectifs divers, emploient des normes pouvant être discordantes pour justifier ou contester les sorties d'un SDA. Une telle étude utilisateur, qui prendrait la forme d'un véritable jeu de rôle, permettrait de valider complètement la démarche adoptée dans le Chapitre 4.

Des travaux futurs dans la continuation de cette thèse pourraient s'appliquer à définir des nouveaux types d'échantillonnage et de génération pour élargir encore le nombre d'explications accessibles avec IBEX. En particulier, des méthodes d'échantillonnage basées sur

des algorithmes génétiques pourraient être employées pour obtenir des ensembles d'échantillons ayant certaines propriétés statistiques de manière efficace. Ce type d'échantillonnage serait notamment adapté en haute dimension. Dans la continuation des travaux sur les justifications d'algorithmes, il pourrait être intéressant de concevoir des méthodes de justification globales pour connaître, par exemple, la norme qui a le plus d'impact sur l'ensemble des décisions du système ou pour obtenir des règles générales sur le caractère "éthique" de l'algorithme.

Par ailleurs, une hypothèse centrale des outils proposés est la disponibilité d'une base de données d'utilisation du SDA. Dans certaines situations, cette hypothèse n'est pas réaliste et les données doivent alors être recueillies, parfois sans le consentement du responsable de l'algorithme. Pour permettre une industrialisation des outils d'explication et de justification proposés, de futurs travaux pourraient tenter de les intégrer à des systèmes de collecte de données. Un tel mode de fonctionnement soulève bien sûr d'autres problèmes techniques comme la possibilité pour un algorithme d'employer des variables cachées, c'est-à-dire des variables qui ne sont pas connues par le système d'explication car elles n'apparaissent pas dans la base de données. Les explications devraient être adaptées pour permettre de détecter ces situations et les rendre explicites.

6.2 Retour sur nos contributions

Des algorithmes sont de plus en plus fréquemment utilisés pour prendre des décisions qui peuvent revêtir une forte dimension éthique et avoir un impact majeur sur la vie des individus (ex : Score Cœur ou COMPAS) ou sur la société dans son ensemble (ex : système de recommandation Youtube). Ces algorithmes sont parfois complexes et peuvent par exemple reposer sur un modèle utilisant du ML rendant leur fonctionnement difficile à analyser et à expliquer. Pour que l'usage de ces algorithmes reste légitime et accepté, un certain nombre d'exigences doivent leur être imposées comme la transparence, l'*accountability* ou l'intelligibilité.

De nombreuses solutions techniques sont proposées pour satisfaire ces exigences dont le

développement d'algorithmes interprétables ou de systèmes d'explication. Cependant, les outils existants ne permettent pas de relever tous les défis posés par l'utilisation croissante des SDA. En premier lieu, la majorité de ces outils sont à destination des spécialistes des SDA alors que les autres utilisateurs (profane, expert du domaine, auditeur, etc.) devraient être des destinataires privilégiés de ces explications et devraient même pouvoir contester les décisions sans avoir à comprendre le fonctionnement du SDA. Plus généralement, ils ne devraient pas être tributaires d'explications techniques et choisies de manière unilatérale pour exercer leur profession ou faire valoir leurs droits. En second lieu, les travaux de la communauté XAI nous semblent encore trop éloignés du terrain et mériteraient d'être d'avantage confrontés à des utilisateurs humains dans des situations réalistes.

Cette thèse défend l'idée que, pour que l'usage des SDA reste acceptable, il faut a minima que toutes les parties prenantes (concepteurs, utilisateurs, auditeurs) puissent consulter des informations adaptées à leurs compétences et à leurs besoins. Ces informations ne doivent pas se limiter à des répliques simplifiées d'explications conçues par les concepteurs de SDA dans le but d'en comprendre le fonctionnement interne. Elles doivent être adaptées à leurs destinataires et répondre aux multiples enjeux posés par les SDA. Nous montrons par ailleurs que la capacité à contester les décisions algorithmiques est indispensable à l'acceptation des systèmes et que cet objectif mobilise des méthodes spécifiques dont nous fixons les bases. Nous prenons le parti de travailler exclusivement en boîte noire ce qui assure l'indépendance vis-à-vis des techniques mobilisées par le SDA (règles, réseaux de neurones, SVM, etc.) et l'applicabilité même lorsque le code du SDA n'est pas disponible. Grâce à nos collaborations, nous proposons des travaux de recherche à la fois motivés par des cas d'usage réels et validés par des utilisateurs aux profils divers.

Comme nous le rappelons ci-dessus, notre étude de l'état de l'art montre que les solutions existantes sont insuffisantes. Notre thèse propose deux outils pour répondre aux défis de l'explicabilité. Plutôt que des améliorations incrémentales d'outils existants[1], IBEX et

1. D'autres auteurs ont appelé à plus d'interaction dans les processus d'explication d'algorithme [84, 90] et certaines contributions ont été faites dans ce sens comme la boîte à outil IBM 360 qui met à disposition un certain nombre de méthodes d'explication au choix de l'utilisateur [13]. À notre connaissance, il n'existe aucun autre cadre général reposant sur une décomposition fine des opérations de base des explications. La contestabilité des SDA a aussi fait l'objet d'un certain nombre de publications théoriques visant principalement à définir les *desiderata* [91, 80]. Les distinctions entre explication et justification, le lien entre les justifications et les contestations ainsi que la mise

Algocate sont des solutions novatrices s'appuyant sur un travail de conceptualisation des systèmes d'explication en boîte noire et des justifications. Nos contributions ne se limitent pas à des apports conceptuels puisque les deux outils ont donné lieu à des implémentations fonctionnelles sous forme de preuve de concept testées auprès d'utilisateurs.

Les études impliquant des utilisateurs humains, dont on sait l'importance dans ce domaine de recherche, constituent d'ailleurs une autre contribution majeure de cette thèse. La collaboration longue avec l'ABM sur Score Cœur a été une source d'inspiration pour les autres contributions de la thèse et l'occasion de confronter *in itinere* nos idées à un système automatisé réel prenant des décisions à fort impact. Les méthodes issues des sciences sociales ont permis de comprendre dans un niveau de détail généralement inaccessible dans le cadre d'une thèse d'informatique les modalités d'utilisation de l'algorithme et son impact sur les organisations hospitalières. Les expérimentations menées auprès des agents de la CNIL, en apparence plus traditionnelles, ont en fait la particularité de s'adresser à un public déjà professionnellement confronté à des questions d'explications d'algorithmes dans le cadre de leurs contrôles. A ce titre, elles constituent un argument fort de l'utilité des concepts introduits et du caractère opérationnel des outils.

En proposant des contributions aux niveaux conceptuel, opérationnel et appliqué, notre démarche se distingue des recherches traditionnelles menées à un seul niveau et dans une seule discipline. En effet, dès le début des travaux, il nous a semblé indispensable de placer les utilisateurs et les organisations impliquées dans les SDA au centre des reflexions. Ainsi, alors que la nécessité de concevoir des explications interactives est défendue par de nombreux auteurs du domaine de recherche, elle nous est apparue d'autant plus fortement au fil de nos rencontres avec les concepteurs du Score Cœur, les médecins utilisateurs, les juristes et les informaticiens impliqués dans les contrôles de la CNIL. Nous avons ainsi pris conscience que les utilisateurs pouvaient avoir des expertises et des objectifs très divers, mais également que chaque personne, indépendamment de son profil, avait ses goûts et ses préférences. Ceci nous a conduit à proposer notre démarche à plusieurs niveaux pour l'interaction car seul un système laissant une grande liberté à l'utilisateur pourrait être satisfaisant

à disposition d'un outil fonctionnel pour contester des décisions algorithmiques sont, également, des contributions de cette thèse.

et compréhensible pour une grande variété d'utilisateurs. Ainsi, riche de cette expérience de thèse interdisciplinaire, nous renouvelons l'appel de Tim Miller à faciliter les collaborations avec les sciences sociales en XAI.

Bibliographie

[1] URL : https://www.partnershiponai.org/research-lander/.

[2] Ashraf ABDUL et al. « Trends and Trajectories for Explainable, Accountable and Intelligible Systems : An HCI Research Agenda ». en. In : *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*. Montreal QC, Canada : ACM Press, 2018, p. 1-18. ISBN : 978-1-4503-5620-6. DOI : 10.1145/3173574.3174156. URL : http://dl.acm.org/citation.cfm?doid=3173574.3174156 (visité le 25/10/2018).

[3] ABM. *Agence de la biomédecine - Le rapport annuel médical et scientifique 2017*. 2017. URL : https://www.agence-biomedecine.fr/annexes/bilan2017/donnees/organes/03-coeur/synthese.htm.

[4] ABM. *Guide du Score Cœur*. version v2. 2018.

[5] ABM. *Site de l'Agence de la Biomédecine*. 2016. URL : https://www.agence-biomedecine.fr/annexes/bilan2016/donnees/organes/03-coeur/synthese.htm.

[6] Marco ALMADA. « Human intervention in automated decision-making : Toward the construction of contestable systems ». In : *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law - ICAIL '19*. ACM Press, 2019, 2–11. ISBN : 978-1-4503-6754-7. DOI : 10.1145/3322640.3326699. URL : http://dl.acm.org/citation.cfm?doid=3322640.3326699.

[7] Richard AMBROSINO et al. « The use of misclassification costs to learn rule-based decision support models for cost-effective hospital admission strategies. » In : *Proceedings of the Annual Symposium on Computer Application in Medical Care*. American Medical Informatics Association, 1995, p. 304.

[8] Dario AMODEI et al. « Concrete Problems in AI Safety ». In : *arXiv :1606.06565 [cs]* (2016). arXiv : 1606.06565. URL : http://arxiv.org/abs/1606.06565.

[9] Mike ANANNY et Kate CRAWFORD. « Seeing without knowing : Limitations of the transparency ideal and its application to algorithmic accountability ». In : *New Media & Society* 20.3 (2018), 973–989. ISSN : 1461-4448, 1461-7315. DOI : 10.1177/1461444816676645.

[10] Athanasios ANDREOU et al. « Investigating Ad Transparency Mechanisms in Social Media : A Case Study of Facebook's Explanations ». In : *Proceedings 2018 Network and Distributed System Security Symposium*. Internet Society, 2018. ISBN : 978-1-891562-49-5. DOI : 10.14722/ndss.2018.23191.

[11] Julia ANGWIN et al. « Machine Bias ». In : *ProPublica* (2016). URL : https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[12] Alejandro Barredo ARRIETA et al. « Explainable Artificial Intelligence (XAI) : Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI ». In : *arXiv :1910.10045 [cs]* (2019). arXiv : 1910.10045. URL : http://arxiv.org/abs/1910.10045.

[13] Vijay ARYA et al. « One Explanation Does Not Fit All : A Toolkit and Taxonomy of AI Explainability Techniques ». In : *arXiv :1909.03012 [cs, stat]* (2019). arXiv : 1909.03012. URL : http://arxiv.org/abs/1909.03012.

[14] Ulrich AÏVODJI et al. « Fairwashing : the risk of rationalization ». In : *International Conference on Machine Learning*. PMLR, 2019, 161–170.

[15] Romain BADOUARD, Clément MABI et Guillaume SIRE. « Beyond "Points of Control" : logics of digital governmentality ». In : *Internet Policy Review* 5.3 (2016). ISSN : 2197-6775. DOI : 10.14763/2016.3.433. URL : https://policyreview.info/node/433.

[16] David BAEHRENS et al. « How to Explain Individual Classification Decisions ». In : *J. Mach. Learn. Res.* 11 (2010), 1803–1831. ISSN : 1532-4435.

[17] Richard BERK et al. « Fairness in Criminal Justice Risk Assessments : The State of the Art ». In : *Sociological Methods & Research* (2018), p. 004912411878253. ISSN : 0049-1241, 1552-8294. DOI : 10.1177/0049124118782533.

[18] Reuben BINNS. « Algorithmic Accountability and Public Reason ». In : *Philosophy & Technology* 31.4 (2018), 543–556. ISSN : 2210-5433, 2210-5441. DOI : 10.1007/s13347-017-0263-5.

[19] Tolga BOLUKBASI et al. « Man is to Computer Programmer as Woman is to Homemaker ? Debiasing Word Embeddings ». In : *arXiv :1607.06520 [cs, stat]* (2016). arXiv : 1607.06520. URL : http://arxiv.org/abs/1607.06520.

[20] Guido CALABRESI et Philip BOBBITT. « Tragic choices ». In : (1978).

[21] Dominique CARDON. « Dans l'esprit du PageRank : Une enquête sur l'algorithme de Google ». In : *Réseaux* n° 177.1 (2013), p. 63. ISSN : 0751-7971, 1777-5809. DOI : 10.3917/res.177.0063.

[22] Dominique CARDON, Jean-Philippe COINTET et Antoine MAZIÈRES. « La revanche des neurones : L'invention des machines inductives et la controverse de l'intelligence artificielle ». In : *Réseaux* n° 211.5 (2018), p. 173. ISSN : 0751-7971, 1777-5809. DOI : 10.3917/res.211.0173.

[23] Rich CARUANA et al. « Intelligible Models for HealthCare : Predicting Pneumonia Risk and Hospital 30-day Readmission ». In : *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*. ACM Press, 2015, 1721–1730. ISBN : 978-1-4503-3664-2. DOI : 10.1145/2783258.2788613. URL : http://dl.acm.org/citation.cfm?doid=2783258.2788613.

[24] Angèle CHRISTIN. « What data can do : A typology of mechanisms ». In : *International Journal of Communication* 14 (2020), p. 20.

[25] Angèle CHRISTIN, Alex ROSENBLAT et Danah BOYD. « Courts and Predictive Algorithms ». In : (), p. 11.

[26] Michael CHROMIK et al. « Dark Patterns of Explainability, Transparency, and User Control for Intelligent Systems. » In : *IUI workshops*. T. 2327. 2019.

[27] CNIL. *Comment permettre à l'Homme de garder la main ?* 2017.

[28] CONSEIL-SUPÉRIEUR-AUDIOVISUEL. *Capacité à informer des algorithmes de recommandation. Une expérience sur le service YouTube*.

[29] Paul COVINGTON, Jay ADAMS et Emre SARGIN. « Deep Neural Networks for YouTube Recommendations ». In : *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, 2016, 191–198. ISBN : 978-1-4503-4035-9. DOI : 10.1145/2959100.2959190. URL : https://dl.acm.org/doi/10.1145/2959100.2959190.

[30] Mark W. CRAVEN et Jude W. SHAVLIK. « Extracting Tree-structured Representations of Trained Networks ». In : *Proceedings of the 8th International Conference on Neural Information Processing Systems*. NIPS'95. event-place : Denver, Colorado. MIT Press, 1995, 24–30. URL : http://dl.acm.org/citation.cfm?id=2998828.2998832.

[31] Anupam DATTA, Shayak SEN et Yair ZICK. « Algorithmic transparency via quantitative input influence : Theory and experiments with learning systems ». In : *2016 IEEE symposium on security and privacy (SP)*. IEEE, 2016, 598–617.

[32] Sonia DESMOULIN-CANSELIER et Daniel LE MÉTAYER. *Décider avec les algorithmes. Quelle place pour l'homme, quelle place pour le droit ?* Les sens du droit. Dalloz, 2020.

[33] Alain DESROSIÈRES. *La politique des grands nombres. Histoire de la raison statistique*. Poche/Sciences humaines et sociales. La Découverte, 2010. ISBN : 978-2-7071-6504-6. URL : https://www.cairn.info/la-politique-des-grands-nombres--9782707165046.htm.

[34] Richard DORENT, Eric EPAILLY et Laurent SEBBAG. « The effect of graft allocation system on outcomes in heart transplantation in France : Has the time come to take calculated survival benefit into account ? » In : *The Journal of Heart and Lung Transplantation* 30.11 (2011), 1299–1300. ISSN : 10532498. DOI : 10.1016/j.healun.2011.06.008.

[35] Finale DOSHI-VELEZ et Been KIM. « Towards A Rigorous Science of Interpretable Machine Learning ». In : *arXiv e-prints*, arXiv :1702.08608 (2017), arXiv :1702.08608. arXiv : 1702.08608 [stat.ML].

[36] Finale DOSHI-VELEZ et al. « Accountability of AI Under the Law : The Role of Explanation ». In : *arXiv :1711.01134 [cs, stat]* (2017). arXiv : 1711.01134. URL : http://arxiv.org/abs/1711.01134.

[37] Hachim EL KHIYARI et Harry WECHSLER. « Face verification subject to varying (age, ethnicity, and gender) demographics using deep learning ». In : *Journal of Biometrics and Biostatistics* 7.323 (2016), p. 11.

[38] Madeleine Clare ELISH. « The Stakes of Uncertainty : Developing and Integrating Machine Learning in Clinical Care ». In : *Ethnographic Praxis in Industry Conference Proceedings* 2018.1 (2018), 364–380. ISSN : 1559890X. DOI : 10.1111/1559-8918.2018.01213.

[39] Glyn ELWYN et al. « "Many miles to go ..." : a systematic review of the implementation of patient decision support interventions into routine clinical practice ». In : *BMC Medical Informatics and Decision Making* 13.S2 (2013). ISSN : 1472-6947. DOI : 10.1186/1472-6947-13-S2-S14. URL : https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-13-S2-S14.

[40] EUROPEANPARLIAMENT. *Understanding algorithmic decision-making : opportunities and challenges.* Publications Office, 2019. URL : https://data.europa.eu/doi/10.2861/536131.

[41] Ruth C. FONG et Andrea VEDALDI. « Interpretable Explanations of Black Boxes by Meaningful Perturbation ». In : *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, 3449–3457. ISBN : 978-1-5386-1032-9. DOI : 10.1109/ICCV.2017.371. URL : http://ieeexplore.ieee.org/document/8237633/.

[42] Alex A. FREITAS. « Comprehensible classification models : a position paper ». In : *ACM SIGKDD explorations newsletter* 15.1 (2014), 1–10.

[43] Jerome H. FRIEDMAN. « Greedy Function Approximation : A Gradient Boosting Machine ». In : *The Annals of Statistics* 29.5 (2001), 1189–1232.

[44] Glenn FUNG, Sathyakama SANDILYA et R. Bharat RAO. « Rule extraction from linear support vector machines ». In : *Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining - KDD '05*. ACM Press, 2005, p. 32. ISBN : 978-1-59593-135-1. DOI : 10.1145/1081870.1081878. URL : http://portal.acm.org/citation.cfm?doid=1081870.1081878.

[45] Marylou GABRIÉ. « Towards an understanding of neural networks : mean-field incursions ». Thèse de doct. 2019. URL : http://www.theses.fr/2019PSLEE035/document.

[46] Ahmad GHIZZAWI et al. « Fairank : An interactive system to explore fairness of ranking in online job marketplaces ». In : *22nd International Conference on Extending Database Technology (EDBT)*. 2019.

[47] Leilani H. GILPIN et al. « Explaining Explanations : An Approach to Evaluating Interpretability of Machine Learning ». In : *arXiv preprint arXiv :1806.00069* (2018).

[48] Riccardo GUIDOTTI et al. « A survey of methods for explaining black box models ». In : *ACM Computing Surveys (CSUR)* 51.5 (2018), p. 93.

[49] Riccardo GUIDOTTI et al. « Local Rule-Based Explanations of Black Box Decision Systems ». In : *arXiv :1805.10820 [cs]* (2018). arXiv : 1805.10820. URL : http://arxiv.org/abs/1805.10820.

[50] Wenbo GUO et al. « LEMNA : Explaining Deep Learning based Security Applications ». en. In : *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security - CCS '18*. Toronto, Canada : ACM Press, 2018, p. 364-379. ISBN : 978-1-4503-5693-0. DOI : 10.1145/3243734.3243792. URL : http://dl.acm.org/citation.cfm?doid=3243734.3243792 (visité le 05/11/2018).

[51] Thilo HAGENDORFF. « The Ethics of AI Ethics : An Evaluation of Guidelines ». In : *Minds and Machines* 30.1 (2020), 99–120. ISSN : 0924-6495, 1572-8641. DOI : 10.1007/s11023-020-09517-8.

[52] Trevor J HASTIE et Robert J TIBSHIRANI. *Generalized additive models*. T. 43. CRC press, 1990.

[53] Andreas HENELIUS et al. « A peek into the black box : exploring classifiers by randomization ». In : *Data Mining and Knowledge Discovery* 28.5–6 (2014), 1503–1529. ISSN : 1384-5810, 1573-756X. DOI : 10.1007/s10618-014-0368-8.

[54] Clément HENIN. « Confier une décision vitale à une machine ». In : *Reseaux* 1 (2021), 187–213.

[55] Clément HENIN et Daniel LE MÉTAYER. « A framework to contest and justify algorithmic decisions ». In : *AI and Ethics* (2021), 1–14.

[56] Clément HENIN et Daniel LE MÉTAYER. « A generic framework for black-box explanations ». In : *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020, 3667–3676.

[57] Clément HENIN et Daniel LE MÉTAYER. « A Multi-layered Approach for Tailored Black-Box Explanations ». In : *Pattern Recognition. ICPR International Workshops and Challenges*. Springer International Publishing, 2021, 5–19. ISBN : 978-3-030-68796-0.

[58] Clément HENIN et Daniel LE MÉTAYER. « Beyond explainability : justifiability and contestability of Algorithmic Decision Systems ». In : *AI & Society* (2021). DOI : 10.1007/s00146-021-01251-8.

[59] Clément HENIN et Daniel LE MÉTAYER. « Towards a framework for challenging ML-based decisions ». In : *1st and 2nd International Workshop on Deceptive AI@ ECAI2020 (to appear)*. 2020.

[60] Clément HENIN et Daniel LE MÉTAYER. « Towards a generic framework for black-box explanations of algorithmic decision systems ». In : *IJCAI 2019 Workshop on Explainable Artificial Intelligence (XAI)*. 2019.

[61] Tad HIRSCH et al. « Designing Contestability : Interaction Design, Machine Learning, and Mental Health ». In : *Proceedings of the 2017 Conference on Designing Interactive Systems*. DIS '17. Edinburgh, United Kingdom : Association for Computing Machinery, 2017, 95–99. ISBN : 9781450349222. DOI : 10.1145/3064663.3064703. URL : https://doi.org/10.1145/3064663.3064703.

[62] Giles HOOKER. « Discovering additive structure in black box functions ». In : *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04*. ACM Press, 2004, p. 575. DOI : 10.1145/1014052.1014122. URL : http://portal.acm.org/citation.cfm?doid=1014052.1014122.

[63] Jeroen van den HOVEN, gert-jan LOKHORST et Ibo POEL. « Engineering and the Problem of Moral Overload ». In : *Science and engineering ethics* 18 (mai 2011), p. 143-55. DOI : 10.1007/s11948-011-9277-z.

[64] Ling HUANG et al. « Adversarial machine learning ». In : *Proceedings of the 4th ACM workshop on Security and artificial intelligence*. 2011, 43–58.

[65] C. JASSERON et al. « Prediction of Waitlist Mortality in Adult Heart Transplant Candidates : The Candidate Risk Score ». In : *The Journal of Heart and Lung Transplantation* 36.4 (2017), S114. ISSN : 10532498. DOI : 10.1016/j.healun.2017.01.292.

[66] Carine JASSERON et al. « Optimization of heart allocation : The transplant risk score ». In : *American Journal of Transplantation* (2018). ISSN : 16006135. DOI : 10.1111/ajt.15201. URL : http://doi.wiley.com/10.1111/ajt.15201.

[67] Toshihiro KAMISHIMA et al. « Fairness-aware classifier with prejudice remover regularizer ». In : *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2012, 35–50.

[68] Pang Wei KOH et Percy LIANG. « Understanding black-box predictions via influence functions ». In : *International Conference on Machine Learning*. PMLR, 2017, 1885–1894.

[69] Josua KRAUSE, Adam PERER et Kenney NG. « Interacting with Predictions : Visual Inspection of Black-box Machine Learning Models ». en. In : *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*. Santa Clara, California, USA : ACM Press, 2016, p. 5686-5697. ISBN : 978-1-4503-3362-7. DOI : 10.1145/2858036.2858529. URL : http://dl.acm.org/citation.cfm?doid=2858036.2858529 (visité le 01/10/2018).

[70] Todd KULESZA et al. « Too much, too little, or just right ? Ways explanations impact end users' mental models ». In : *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. IEEE, 2013, 3–10. ISBN : 978-1-4799-0369-6. DOI : 10.1109/VLHCC.2013.6645235. URL : http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6645235.

[71] Himabindu LAKKARAJU, Stephen H. BACH et Jure LESKOVEC. « Interpretable Decision Sets : A Joint Framework for Description and Prediction ». In : *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*. ACM Press, 2016, 1675–1684. ISBN : 978-1-4503-4232-2. DOI : 10.1145/2939672.2939874. URL : http://dl.acm.org/citation.cfm?doid=2939672.2939874.

[72] Himabindu LAKKARAJU et al. « Interpretable & Explorable Approximations of Black Box Models ». In : *arXiv preprint arXiv :1707.01154* (2017).

[73] Thibault LAUGEL et al. « The Dangers of Post-hoc Interpretability : Unjustified Counterfactual Explanations ». In : *arXiv :1907.09294 [cs, stat]* (2019). arXiv : 1907.09294. URL : http://arxiv.org/abs/1907.09294.

[74] Tao LEI, Regina BARZILAY et Tommi JAAKKOLA. « Rationalizing Neural Predictions ». In : *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2016, 107–117. DOI : 10.18653/v1/D16-1011. URL : http://aclweb.org/anthology/D16-1011.

[75] Bruno LEPRI et al. « Fair, Transparent, and Accountable Algorithmic Decision-making Processes : The Premise, the Proposed Solutions, and the Open Challenges ». In : *Philosophy & Technology* 31.4 (2018), 611–627. ISSN : 2210-5433, 2210-5441. DOI : 10.1007/s13347-017-0279-x.

[76] Paul LEWIS. « Fiction is outperforming reality' : How YouTube's algorithm distorts truth ». In : *The Guardian* 2 (2018), p. 2018.

[77] Brian Y. LIM, Anind K. DEY et Daniel AVRAHAMI. « *Why and why not* explanations improve the intelligibility of context-aware intelligent systems ». en. In : *Proceedings*

of the 27th international conference on Human factors in computing systems - CHI 09. Boston, MA, USA : ACM Press, 2009, p. 2119. ISBN : 978-1-60558-246-7. DOI : 10.1145/1518701.1519023. URL : http://dl.acm.org/citation.cfm?doid=1518701.1519023 (visité le 29/10/2018).

[78] Zachary C. LIPTON. « The Mythos of Model Interpretability ». In : *arXiv :1606.03490 [cs, stat]* (2016). arXiv : 1606.03490. URL : http://arxiv.org/abs/1606.03490.

[79] Scott M LUNDBERG et Su-In LEE. « A Unified Approach to Interpreting Model Predictions ». In : *Advances in Neural Information Processing Systems 30*. Sous la dir. d'I. GUYON et al. Curran Associates, Inc., 2017, 4765–4774. URL : http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[80] Henrietta LYONS, Eduardo VELLOSO et Tim MILLER. « Conceptualising Contestability : Perspectives on Contesting Algorithmic Decisions ». In : *arXiv :2103.01774 [cs]* (2021). arXiv : 2103.01774. DOI : 10.1145/3449180. URL : http://arxiv.org/abs/2103.01774.

[81] Prashan MADUMAL et al. « A Grounded Interaction Protocol for Explainable Artificial Intelligence ». In : *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. AAMAS '19. Montreal QC, Canada : International Foundation for Autonomous Agents et Multiagent Systems, 2019, 1033–1041. ISBN : 9781450363099.

[82] Aravindh MAHENDRAN et Andrea VEDALDI. « Understanding deep image representations by inverting them ». In : *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015, 5188–5196. ISBN : 978-1-4673-6964-0. DOI : 10.1109/CVPR.2015.7299155. URL : http://ieeexplore.ieee.org/document/7299155/.

[83] Gianclaudio MALGIERI et Giovanni COMANDÉ. « Why a right to legibility of automated decision-making exists in the general data protection regulation ». In : *International Data Privacy Law* (2017).

[84] Tim MILLER. « Explanation in Artificial Intelligence : Insights from the Social Sciences ». In : *Artificial Intelligence* 267 (2017). DOI : 10.1016/j.artint.2018.07.007.

[85] Tim MILLER, Piers HOWE et Liz SONENBERG. « Explainable AI : Beware of inmates running the asylum ». In : *IJCAI-17 Workshop on Explainable AI (XAI)*. T. 36. 2017.

[86] Smitha MILLI et al. « Model Reconstruction from Model Explanations ». In : *arXiv :1807.05185 [cs, stat]* (2018). arXiv : 1807.05185. URL : http://arxiv.org/abs/1807.05185.

[87] Brent MITTELSTADT, Chris RUSSELL et Sandra WACHTER. « Explaining explanations in AI ». In : *Proceedings of the conference on fairness, accountability, and transparency*. 2019, 279–288.

[88] Sina MOHSENI, Niloofar ZAREI et Eric D. RAGAN. « A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems ». In : *arXiv :1811.11839 [cs]* (2020). arXiv : 1811.11839. URL : http://arxiv.org/abs/1811.11839.

[89] Grégoire MONTAVON, Wojciech SAMEK et Klaus-Robert MÜLLER. « Methods for Interpreting and Understanding Deep Neural Networks ». In : *Digital Signal Processing* 73 (2018). arXiv : 1706.07979, 1–15. ISSN : 10512004. DOI : 10.1016/j.dsp.2017.10.011.

[90] Shane T MUELLER et al. « Explanation in Human-AI Systems : A Literature Meta-Review Synopsis of Key Ideas and Publications and Bibliography for Explainable AI ». In : (), p. 204.

[91] Deirdre K. MULLIGAN, Daniel KLUTTZ et Nitin KOHLI. « Shaping Our Tools : Contestability as a Means to Promote Responsible Algorithmic Decision Making in the Professions ». In : *Available at SSRN 3311894* (2019). URL : https://ssrn.com/abstract=3311894.

[92] Govind PERSAD, Alan WERTHEIMER et Ezekiel J EMANUEL. « Principles for allocation of scarce medical interventions ». In : *The Lancet* 373.9661 (2009), 423–431. ISSN : 01406736. DOI : 10.1016/S0140-6736(09)60137-9.

[93] Geoff PLEISS et al. « On fairness and calibration ». In : *Advances in Neural Information Processing Systems*. 2017, 5680–5689.

[94] Theodore M PORTER. *Trust in numbers : The pursuit of objectivity in science and public life*. Princeton University Press, 1996.

[95] Forough POURSABZI-SANGDEH et al. « Manipulating and Measuring Model Interpretability ». In : *arXiv :1802.07810 [cs]* (2018). arXiv : 1802.07810. URL : http://arxiv.org/abs/1802.07810.

[96] Filipe N. RIBEIRO et al. « On Microtargeting Socially Divisive Ads : A Case Study of Russia-Linked Ad Campaigns on Facebook ». In : *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*. ACM Press, 2019, 140–149. ISBN : 978-1-4503-6125-5. DOI : 10.1145/3287560.3287580. URL : http://dl.acm.org/citation.cfm?doid=3287560.3287580.

[97] Marco Tulio RIBEIRO, Sameer SINGH et Carlos GUESTRIN. « Anchors : High-precision model-agnostic explanations ». In : *AAAI Conference on Artificial Intelligence*. 2018.

[98] Marco Tulio RIBEIRO, Sameer SINGH et 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*. ACM Press, 2016, 1135–1144. ISBN : 978-1-4503-4232-2. DOI : 10.1145/2939672.2939778. URL : http://dl.acm.org/citation.cfm?doid=2939672.2939778.

[99] Marko ROBNIK-ŠIKONJA et Igor KONONENKO. « Explaining classifications for individual instances ». In : *IEEE Transactions on Knowledge and Data Engineering* 20.5 (2008), 589–600.

[100] Kevin ROOSE. « The Making of a YouTube Radical ». In : *The New York Times* (2019). ISSN : 0362-4331. URL : https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html.

[101] Antoinette ROUVROY et Thomas BERNS. « Gouvernementalité algorithmique et perspectives d'émancipation : Le disparate comme condition d'individuation par la relation ? » In : *Réseaux* n° 177.1 (2013), p. 163. ISSN : 0751-7971, 1777-5809. DOI : 10.3917/res.177.0163.

[102] Lloyd S. SHAPLEY. « A value for n-person games ». In : *Annals of Mathematics Studies*. Cambridge University Press, 1953, 307–317.

[103] Elisa SHEARER et Elizabeth GRIECO. *Americans Are Wary of the Role Social Media Sites Play in Delivering the News*. 2019. URL : https://www.journalism.org/2019/10/02/americans-are-wary-of-the-role-social-media-sites-play-in-delivering-the-news/.

[104] Reza SHOKRI, Martin STROBEL et Yair ZICK. « Privacy Risks of Explaining Machine Learning Models ». In : *arXiv :1907.00164 [cs, stat]* (2019). arXiv : 1907.00164. URL : http://arxiv.org/abs/1907.00164.

[105] Dylan SLACK et al. « How can we fool LIME and SHAP ? Adversarial Attacks on Post hoc Explanation Methods ». In : *arXiv :1911.02508 [cs, stat]* (2019). arXiv : 1911.02508. URL : http://arxiv.org/abs/1911.02508.

[106] Simone STUMPF et al. « Toward harnessing user feedback for machine learning ». In : *Proceedings of the 12th international conference on Intelligent user interfaces*. 2007, 82–91.

[107] Alain SUPIOT. *La gouvernance par les nombres*. Fayard, 2015.

[108] Christian SZEGEDY et al. « Intriguing properties of neural networks ». In : *International Conference on Learning Representations*. 2014. URL : http://arxiv.org/abs/1312.6199.

[109] Pascal THIBAUT. « Allemagne : scandale de fraude au don d'organes ». In : *RFI* (2013). URL : http://www.rfi.fr/fr/europe/20130820-allemagne-scandale-fraude-don-organes-gottingen.

[110] Mark TIMMONS. *Moral Theory*. Rowman et Littlefield Publishers, 2013.

[111] Richard TOMSETT et al. « Interpretable to Whom ? A Role-based Model for Analyzing Interpretable Machine Learning Systems ». In : *CoRR* abs/1806.07552 (2018).

[112] Jean-Sébastien VAYRE. « Comment décrire les technologies d'apprentissage artificiel ? : Le cas des machines à prédire ». In : *Réseaux* n° 211.5 (2018), p. 69. ISSN : 0751-7971, 1777-5809. DOI : 10.3917/res.211.0069.

[113] Cédric VILLANI. « Donner un sens à l'intelligence artificielle ». In : (2018).

[114] Sandra WACHTER, Brent MITTELSTADT et Luciano FLORIDI. « Why a right to explanation of automated decision-making does not exist in the general data protection regulation ». In : *International Data Privacy Law* 7.2 (2017), 76–99.

[115] Sandra WACHTER, Brent MITTELSTADT et Chris RUSSELL. « Counterfactual Explanations Without Opening the Black Box : Automated Decisions and the GDPR ». In : *Harvard journal of law & technology* 31 (2018), 841–887.

[116] Adrian WELLER. « Challenges for Transparency ». In : *arXiv :1708.01870 [cs]* (2017). arXiv : 1708.01870. URL : http://arxiv.org/abs/1708.01870.

[117] Renjie ZHOU, Samamon KHEMMARAT et Lixin GAO. « The impact of YouTube recommendation system on video views ». In : *Proceedings of the 10th annual conference on Internet measurement - IMC '10*. ACM Press, 2010, p. 404. ISBN : 978-1-4503-0483-2. DOI : 10.1145/1879141.1879193. URL : http://portal.acm.org/citation.cfm?doid=1879141.1879193.

[118] Erik ŠTRUMBELJ et Igor KONONENKO. « Explaining prediction models and individual predictions with feature contributions ». In : *Knowledge and Information Systems* 41.3 (2014), 647–665. ISSN : 0219-1377, 0219-3116. DOI : 10.1007/s10115-013-0679-x.

FOLIO ADMINISTRATIF

THESE DE L'UNIVERSITE DE LYON OPEREE AU SEIN DE L'INSA LYON

NOM : Henin

DATE de SOUTENANCE : 13/10/2021

Prénoms : Clément

TITRE : Expliquer et justifier les systèmes de décisions algorithmiques

NATURE : Doctorat

Numéro d'ordre : 2021LYSEI058

Ecole doctorale : ED512

Spécialité : Informatique

RESUME : Dans un contexte favorable à la rationalisation des décisions par des données et des règles chiffrées, le développement récent des technologies numériques a accéléré l'adoption des systèmes de décisions fondés sur un traitement algorithmique. De tels systèmes sont déjà présents dans de nombreux domaines et leur utilisation devrait encore s'accroître. Pour certains types d'applications, l'opacité des systèmes peut être un frein, voire un obstacle rédhibitoire, à leur utilisation. Les explications d'algorithmes permettent d'atténuer ces obstacles, mais les méthodes existantes ne sont pas pleinement adaptées aux enjeux. En particulier, nous soutenons l'idée que les explications d'algorithmes doivent être adaptées à leurs destinataires, y compris les profanes en IA, et répondre aux multiples enjeux posés par les systèmes de décisions algorithmiques. Dans cette thèse, nous nous intéressons en particulier à la production d'explications en « boîte noire », c'est-à-dire sans accès au code du système de décision. L'avantage de cette démarche est de fournir des résultats qui peuvent s'appliquer à de nombreux systèmes, indépendamment de leur mode de fonctionnement. Notre première contribution est un système d'explications interactif, permettant à l'utilisateur de contrôler les propriétés de l'explication qui lui est fournie afin d'obtenir la plus adaptée à sa situation. La deuxième contribution propose une approche novatrice pour contester et justifier les résultats d'un algorithme. Ces approches théoriques ont donné lieu au développement de deux outils : IBEX et Algocate. Ces résultats théoriques sont confrontés au terrain au travers d'études utilisateurs, dont un travail mené sur l'algorithme du score cœur utilisé pour l'attribution des greffons cardiaques. Ce dernier combine des éléments sociologiques traitant notamment de l'appropriation par les acteurs de ce système de décision et le développement d'un outil adapté aux spécificités de ce système.

MOTS-CLÉS : Système de décisions algorithmique, explication d'algorithme, justification d'algorithme, contestabilité des algorithmes, attribution des greffons cardiaques,

Laboratoire de recherche : CITI Inria Rhônes-Alpes

Directeurs de thèse: Daniel Le Métayer Claude Castelluccia

Président de jury : Christine Solnon

Composition du jury : Sihem Amer-Yahia, Sébastien Gambs, Christian Jacquelinet, Clément Mabi, Christine Solnon, Claude Castelluccia, Daniel Le Métayer