Résumé de la thèse: "Explainability of Possibilistic and Fuzzy rule-based systems"

Ismaïl Baaj

Introduction

Aujourd'hui, les progrès de l'*Intelligence Artificielle* (IA) ont conduit à l'émergence de systèmes capables d'automatiser des processus complexes, en utilisant des modèles qui peuvent être difficiles à comprendre pour les humains [Fre14, Lip18]. Lorsque les humains utilisent ces systèmes d'IA, il est bien connu qu'ils veulent comprendre leurs comportements et leurs actions [HLD+19, TS81], car ils ont davantage confiance dans les systèmes qui peuvent expliquer leurs choix, leurs hypothèses et leurs raisonnements [GSC+19]. La capacité explicative des systèmes est devenue une exigence des utilisateurs pour l'acceptation de leur utilisation [ACM17], notamment dans les environnements à risque humain comme avec les véhicules autonomes ou en médecine [MAT+16]. Dans ce contexte, les lois ont récemment renforcé les droits des utilisateurs et leur permettent de revendiquer un droit d'explication dans l'IA [ACM17, Reg16]. En application de cette législation, le développement de la capacité explicative des systèmes d'IA apparaît aujourd'hui comme une nécessité. Cette nécessité est apparue en même temps que le récent regain d'intérêt pour l'Intelligence Artificielle eXplicable (abrégée XAI), un domaine de recherche qui vise à développer des systèmes d'IA capables d'expliquer leurs résultats d'une manière compréhensible pour les humains [ADRDS+20]. Si les premières approches visant à développer l'explicabilité des systèmes d'IA remontent aux années 70-80 [MHC+19], ce domaine a regagné en popularité avec le lancement récent d'un programme de la Defense Advanced Research Projects Agency (DARPA) qui tente d'apporter de la transparence à des modèles d'IA opaques tels que les réseaux neuronaux profonds [GA19]. Le développement de principes, de stratégies et de techniques d'interaction homme-machine pour générer des explications efficaces des résultats des systèmes d'IA (voir [ACMM21, ADRDS+20, Mil19, BC17]) répond aux attentes et aux besoins émergents des parties prenantes utilisant des systèmes d'IA [LOS+21].

Dans [DPU03], Dubois, Prade et Ughetto développent l'idée que les informations codées sur un ordinateur peuvent avoir un *accent négatif ou positif*. Les informations négatives peuvent être considérées comme des contraintes et correspondent à des énoncés qui excluent certaines situations parce qu'elles sont impossibles. Les informations positives modélisent des observations et correspondent à des énoncés qui décrivent ce qui est possible avec certitude parce que cela a été observé. Ces deux points de vue antagonistes sur l'information nous permettent de distinguer différents types de règles Si-Alors pour représenter les données et les connaissances, qui peuvent être modélisées de manière appropriée dans le cadre de la théorie des ensembles flous et de la théorie des possibilités.
Dans cette thèse, basée sur les travaux [DEGP07, DP20, DPU03, FP92], nous introduisons des paradigmes explicatifs pour deux systèmes d'IA :

- un système basé sur des règles possibilistes, où les règles possibilistes encodent des infor-

mations négatives et

- un système à base de règles à possibilité, qui encodent des informations positives.

Dans cette thèse, les capacités explicatives de ces systèmes sont développées pour les objectifs suivants :

1. l'établissement de points de rencontre entre les domaines de la *Représentation des connaissances et raisonnement* (KRR) et du *Machine Learning* (ML). Ils ont été récemment étudiés par [Ame19].

2. l'élaboration d'une *chaîne de traitement* pour XAI, qui a été proposée dans [BPO19], afin de pouvoir générer des explications en langage naturel des décisions des systèmes d'IA et de pouvoir évaluer ces explications.

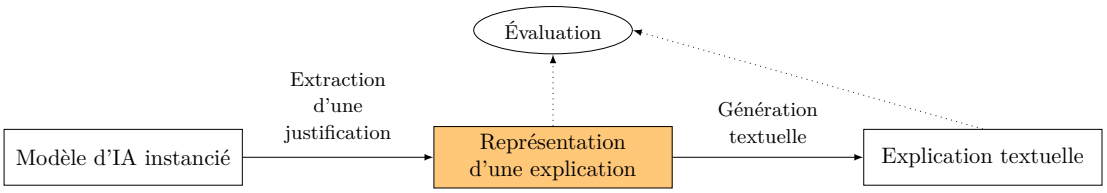

Figure 1: Chaîne de traitement proposée pour générer et évaluer les explications [BPO19]

Cette chaîne de traitement (Figure 1) comporte en son centre une *représentation d'une explication* et trois tâches :

- l'*extraction d'une justification* des résultats du système d'IA considéré, afin de former le contenu d'une explication [ADRDS+20, BC17, GMR+18],

- la *génération textuelle* d'une explication, avec des techniques de génération en langage naturel (NLG) [GK18, RD97],

- l'*évaluation* d'une explication, voir [DVK17, MZR21, TM11, vdWNCN21].

Les trois tâches sont séparées en processus distincts afin de permettre un développement spécifique de chacune d'entre elles et de dissocier les responsabilités. Dans cette thèse, afin d'élaborer cette chaîne de traitement, nous nous concentrons sur la tâche d'extraction d'une justification et la définition d'une représentation graphique d'une explication. À partir de cette représentation, on est en mesure de générer des explications en langage naturel et d'évaluer ces explications.

La thèse est structurée comme suit. Dans la partie A, on rappelle la théorie des possibilités, la théorie des ensembles flous et le cadre des graphes conceptuels. Les notions d'information positive et négative sont introduites, ainsi que les systèmes à base de règles floues et les systèmes à base de règles possibilistes. Enfin, nous faisons un bref rappel des approches explicatives de ces systèmes.
Dans la partie B, en accord avec notre premier objectif d'établir des liens entre KRR et ML, nous développons une interface possibiliste entre l'apprentissage et le raisonnement Si-Alors. Une telle interface est construite en généralisant le système d'équations min-max de Farreny et Prade [FP92], qui a été développé pour les systèmes à base de règles possibilistes. Dubois et Prade pensent que cette interface peut permettre le développement de méthodes d'apprentissage

possibilistes qui seraient cohérentes avec le raisonnement à base de règles [DP20].
Les parties C et D se concentrent sur l'élaboration de la chaîne de traitement. Dans la partie C, nous introduisons des méthodes pour justifier les résultats d'inférence des systèmes à base de règles possibilistes et floues. Nos méthodes nous permettent de former deux explications d'un résultat d'inférence d'un système à base de règles : sa justification et son caractère inattendu (un ensemble d'énoncés logiques qui ne sont pas impliqués dans la détermination du résultat considéré tout en étant liés à celui-ci). La notion de caractère inattendu s'inspire de celle donnée dans la théorie de la simplicité [Des17], où elle vise à capturer ce que les gens considèrent comme surprenant d'une situation donnée [SD15].
Dans la partie D, nous représentons ces deux explications par des graphes conceptuels [CM08]. Nous représentons également une explication qui est la combinaison d'une justification et du caractère inattendu d'un résultat d'inférence. Ces représentations nous permettent de voir graphiquement les résultats des multiples opérations analytiques effectuées pour générer des explications des décisions d'inférence. À partir de ces représentations, il est possible de produire des explications en langage naturel, en adaptant les systèmes NLG qui produisent du texte à partir des entrées du web sémantique [BACW14, GSNPB17].

1 Partie A

Dans le premier chapitre de la Partie A, nous présentons les outils nécessaires utilisés dans cette thèse. Nous commençons par rappeler la théorie des possibilités, qui est un cadre bien connu pour le traitement des informations incomplètes ou imprécises [DP88, DP15] introduit par Zadeh en 1978 [Zad78]. Nous nous concentrons sur le traitement possibiliste des systèmes à base de règles, qui a été développé dans les années 80 [FP86, FPW86].
Nous continuons ensuite en présentant la logique floue, qui a également été introduite par Zadeh [Zad65]. C'est une extension de la logique booléenne comprenant une notion de vérité partielle. Nous présentons les notions d'informations positives et négatives. Dans [DPU03], Dubois, Prade et Ughetto font la distinction entre les informations *négatives* et *positives* qui sont sous-jacentes à une règle. Cette distinction a également été revisitée par [JCDG08]. Une règle en logique classique est de la forme "si X est A alors Z est O" où $A \subseteq U$ et $O \subseteq V$ sont des sous-ensembles du domaine de la variable X et Z respectivement, qui relient les deux univers de discours U et V par leurs restrictions locales A et O. On peut interpréter une telle règle de deux points de vue différents, selon que l'on se concentre sur ses exemples ou ses contre-exemples :

- *Vue positive* : la règle est vue comme une condition de la forme "si X est A alors Z *peut être* O" et affirme que lorsque X prend sa valeur dans A, alors *toutes* les valeurs dans O sont admissibles pour Z. Les paires $(u, v) \in A \times O$ forment un ensemble d'exemples explicitement autorisés par la règle. Cette vue est l'interprétation conjonctive de la règle, en ne mettant l'accent que sur ses exemples.

- *Vue négative* : la règle est interprétée comme une contrainte de la forme "si X est A alors Z *doit être* O" et affirme de manière implicitement négative que les valeurs extérieures à O sont *exclues* lorsque X prend ses valeurs dans A. Les couples $(u, v) \in A \times \overline{O}$ forment l'ensemble des contre-exemples de la règle et sont explicitement interdits par la règle, tandis que les couples de l'ensemble $\overline{A \times \overline{O}}$ forment l'ensemble des couples de valeurs implicitement autorisés pour (X, Z). Comme nous avons :

$$\overline{A \times \overline{O}} = (\overline{A} \times V) \cup (U \times O) = (\overline{A} \times V) \cup (A \times O),$$

l'ensemble $\overline{A \times \overline{O}}$ est la réunion disjointe de l'ensemble des exemples $A \times O$ et de l'ensemble $\overline{A} \times V$ des paires de valeurs non engagées par la règle. Cette vue est l'interprétation implicative de la règle, en mettant l'accent uniquement sur ses contre-exemples (l'ensemble $A \times \overline{O}$) et correspond clairement à l'implication booléenne en logique classique.

Les notions d'informations positives et négaties nous permettent de rappeler les deux types de règles floues : les *règles floues conjonctives* qui encodent des informations positives et les *règles floues implicatives* qui encodent des informations négatives. Nous rappelons comment leur sémantique peut être définie dans le cadre de la théorie des possibilités.
Enfin, pour représenter les connaissances en termes de graphes, nous présentons le cadre des graphes conceptuels. Dans cette thèse, les graphes conceptuels nous permettront de donner une représentation graphique de certains de nos résultats.
Dans le second chapitre de la partie A, nous commençons par donner un aperçu des premiers développements des systèmes d'IA explicables, qui remontent aux années 70-80. Ensuite, nous examinons certaines approches pour développer les capacités explicatives des systèmes à base de règles floues composés de règles à possibilité. Enfin, nous étudions le système d'équations min-max de Farreny et Prade, qui a été proposé pour développer les capacités explicatives d'un système à base de règles possibilistes.

2 Partie B

Dans la Partie B, nous abordons un certain nombre de questions et/ou problèmes soulevés dans [DP20, FP92] par une étude approfondie du système d'équations min-max d'un système à base de règles possibilistes que nous avons rappelé dans la Partie A. Ce système d'équations décrit la distribution des possibilités de sortie et a été proposé pour effectuer une analyse de sensibilité [FP92]. Dans le cas de n règles possibilistes, nous donnons une construction canonique des matrices gouvernant le système d'équations. À partir de notre système d'équations généralisé, nous avons obtenu une formule explicite pour la distribution des possibilités de sortie et calculé les mesures de possibilité et de nécessité correspondantes. Nous avons donné une condition nécessaire et suffisante pour que la distribution des possibilités de sortie soit normalisée et déterminé, lorsque cela est possible, les solutions d'entrée minimales pour la normalisation. Nous avons défini un algorithme pour reconstruire le système d'équations lorsque nous supprimons une règle. Il produit le système d'équations associé au sous-ensemble de règles restant. Cet algorithme nous permet donc d'obtenir tous les sous-systèmes d'équations d'un système d'équations initial.
Nous avons ensuite étendu notre système d'équations au cas d'une cascade. Nous avons mis en évidence une relation entrée-sortie entre les systèmes d'équations associés à chaque ensemble de règles : elle relie le vecteur de sortie du premier système au vecteur d'entrée du second système. Il en résulte que le vecteur de sortie du second système s'obtient par des produits min-max imbriqués des matrices des deux systèmes d'équations. Enfin, nous avons montré qu'une telle cascade peut être représentée par un réseau de neurones min-max explicite.
Nous illustrons nos résultats par la construction du système d'équations associé à la cascade utilisée comme exemple dans [FP90, DP20] que nous représentons par un réseau de neurones min-max.

3 Partie C

Dans le premier chapitre de la Partie C, nous abordons l'explicabilité des résultats d'inférence d'un système à base de règles possibilistes. Nous nous appuyons sur l'approche de Farreny et Prade [FP92], qui a été rappelée dans le second chapitre de la partie A. En utilisant leur système d'équations min-max, les auteurs étudient deux buts explicatifs pour une valeur de l'attribut de sortie $u \in D_b$ (D_b est le domaine de l'attribut de sortie b), qui peuvent être formulés comme deux questions :

(i) Quelles sont les conditions pour que le degré de possibilité de u, qui est noté $\pi^*_{b(x)}(u)$, soit strictement supérieur ou inférieur à un $\tau \in [0, 1]$ donné ?

(ii) Quels sont les degrés des prémisses qui justifient $\pi^*_{b(x)}(u) = \tau$?

Pour ces deux questions, les paramètres des règles s_i et r_i sont fixés. Les auteurs de [FP92] donnent une condition suffisante pour obtenir $\pi^*_{b(x)}(u) > \tau$ pour une paire particulière (u,τ) de leur exemple. Pour la deuxième question, ils affirment qu'on peut lire directement les degrés de possibilité des prémisses impliquées dans le calcul du degré de possibilité d'une valeur de l'attribut de sortie. Leur affirmation est soutenue par une valeur u de l'attribut de sortie de leur exemple.
Dans le premier chapitre de la Partie C, nous abordons ces deux questions dans le cas général. Pour la première question, nous donnons les conditions nécessaires et suffisantes pour obtenir $\pi^*_{b(x)}(u) > \tau$ et $\pi^*_{b(x)}(u) < \tau$ selon les degrés des prémisses. Pour la deuxième question, nous donnons une condition nécessaire et suffisante qui permet de justifier $\pi^*_{b(x)}(u) = \tau$ par des degrés de prémisses. Cela permet d'extraire le sous-ensemble de prémisses dont les degrés sont impliqués dans le calcul de $\pi^*_{b(x)}(u)$. Cette extraction est effectuée en évaluant $\pi^*_{b(x)}(u)$ par rapport à un seuil fixé.
Nous définissons ensuite quatre fonctions de réduction des prémisses et nous les appliquons au sous-ensemble obtenu de prémisses liées à $\pi^*_{b(x)}(u)$. Cela nous conduit à former deux types d'explications de $\pi^*_{b(x)}(u)$:

- La *justification* de $\pi^*_{b(x)}(u)$, qui est formée en réduisant les prémisses sélectionnées à la structure responsable de leur degré de possibilité ou de nécessité. Deux fonctions de réduction des prémisses sont utilisées.
- Le *caractère inattendu* de $\pi^*_{b(x)}(u)$, qui est un ensemble d'expressions possibilistes possibles ou certaines liées au résultat d'inférence considéré dans le sens suivant : bien qu'il peut sembler y avoir une incompatibilité potentielle entre chacune des expressions possibilistes et le résultat d'inférence considéré, elles ne sont pas impliquées dans la détermination du résultat d'inférence. Elles sont extraites en appliquant les deux autres fonctions de réduction des prémisses.

Nos constructions sont illustrées par deux exemples.

Dans le second chapitre de la Partie C, nous étudions les capacités explicatives d'un système à base de règles floues composé de règles à possibilité (système d'inférence flou de Mamdani), où les prémisses des règles sont des conjonctions de propositions floues. Dans notre système, ces règles floues sont combinées de manière disjonctive. Nous nous concentrons sur l'explication sémantique des conclusions inférées d'un système Mamdani, sans envisager l'utilisation d'un

processus de défuzzification (un processus pour obtenir une valeur crisp à partir de l'ensemble flou agrégé de sortie) [VLK99].
À cette fin, nous fixons les notations utilisées pour les principaux objets d'une inférence d'un système de Mamdani : le degré d'activation d'une règle, la conclusion inférée et la distribution de possibilité de la variable des conclusions de la règle. Ensuite, nous énonçons le résultat principal : la conclusion inférée totale satisfait la sémantique α^*-possible au sens de Dubois-Prade [DP98], où α^* est le maximum des degrés d'activation des règles. Ce résultat est prouvé en deux étapes. Ensuite, nous justifions par un sous-ensemble pertinent de prémisses de règles, chacune des conclusions inférées de tout système de Mamdani, en évaluant le degré de la conclusion par rapport à un seuil fixé. Nous donnons un exemple d'un tel système qu'on utilise pour illustrer toutes les constructions qui suivent.
De façon similaire au cas d'un système de règles possibilistes (voir le premier chapitre de la Partie C), nous introduisons deux fonctions de réduction des prémisses. En les appliquant aux prémisses sélectionnées pour justifier une conclusion inférée, ces fonctions nous permettent de former deux types d'explications :

- La *justification* d'une conclusion, qui est un ensemble d'expressions de logique floue (conjonctions de propositions floues) suffisantes pour justifier, sémantiquement, la conclusion inférée. Il est formé en appliquant une fonction de réduction aux prémisses sélectionnées. Cette fonction de réduction est basée sur notre travail précédent, voir [BP19].

- Le *caractère inattendu* d'une conclusion, qui est un ensemble d'expressions de logique floue extraites en appliquant une autre fonction de réduction aux prémisses sélectionnées. Ces expressions de logique floue sont liées à la conclusion considérée dans le sens suivant : bien qu'il puisse sembler y avoir une incompatibilité potentielle entre chacune des expressions de logique floue et la conclusion considérée, elles ne sont pas impliquées dans la détermination de la conclusion inférée.

De telles explications sont formulées en effectuant des traitements sur les prémisses des règles, qui sont des conjonctions de propositions floues. Dans la Partie D, de telles explications seront représentées graphiquement par des graphes conceptuels. Dans ce but, nous montrerons que nous pouvons naturellement représenter les conjonctions de propositions floues par des graphes conceptuels.

4 Partie D

Dans le premier chapitre de la Partie D, nous élaborons un cadre pour représenter les explications en termes de graphes conceptuels. Nous commençons par définir les objets qui nous permettent de former une explication d'une décision d'inférence d'un système à base de règles. Nous affirmons qu'une explication est composée de :

- $m + 1$ déclarations ($m \geq 1$), où l'une est un *phénomène observé*. Les autres m énoncés sont liés à ce phénomène, et peuvent être soit des justifications du phénomène, soit des faits inattendus qui n'empêchent pas le phénomène de se produire.

- et d'un lien entre le phénomène et les autres m énoncés qui structure l'explication. Par exemple, un tel lien peut être désigné par "estJustifiéPar" ou "bienQue".

La *représentation d'une explication* en termes de graphes conceptuels est réalisée comme suit. Étant donné une explication, chacune de ses déclarations est représentée par un graphe conceptuel. Pour structurer l'explication, les graphes représentant les énoncés sont imbriqués dans

un graphe conceptuel racine qui contient un nœud de relation représentant le lien entre les énoncés. Le *graphe conceptuel imbriqué* qui en résulte est une représentation de l'explication.

Dans ce chapitre, nous commençons par donner le vocabulaire minimal (qui peut être considéré comme une représentation d'une petite ontologie) pour construire le graphe conceptuel racine R de la représentation. Ensuite, en supposant que le vocabulaire minimal est étendu afin de représenter les $m+1$ énoncés par les graphes conceptuels notés $D, N_1, N_2, \ldots, N_m$, nous donnons l'interprétation de chacun des graphes composant la représentation :

- D est graphe conceptuel qui est la représentation graphique d'un énoncé décrivant un phénomène observé,
- Pour $i = 1, 2, \ldots, m$, chaque N_i est un graphe conceptuel qui représente graphiquement une déclaration liée au phénomène représenté par D. Par exemple, il peut s'agir d'une déclaration qui justifie le phénomène ou d'une déclaration qui représente un fait inattendu.
- R structure l'explication en représentant le lien entre le phénomène observé D et les énoncés $N_1, N_2, \ldots, N_m$.

Nous terminons ce chapitre par la définition du graphe racine R et du graphe conceptuel imbriqué qui représente l'explication. Enfin, nous illustrons notre construction avec un exemple d'explication.

Dans le second chapitre de la Partie D, nous représentons graphiquement deux explications : la justification et le caractère inattendu du degré de possibilité $\pi^*_{b(x)}(u)$ d'une valeur u de l'attribut de sortie (voir Partie C). Pour représenter ces explications, nous nous appuyons sur notre cadre introduit au premier chapitre de la Partie D. Les graphes conceptuels qui en résultent sont des représentations visuelles des résultats de plusieurs opérations analytiques effectuées sur la base de règles possibilistes qui constituent des explications.

Nous commençons par spécifier, dans le contexte de l'explication d'une décision d'inférence possibiliste, les objets qui la composent. Pour les deux explications (justification et caractère inattendu), le phénomène observé est le degré de possibilité $\pi^*_{b(x)}(u)$ d'une valeur u de l'attribut de sortie. Les autres énoncés sont les expressions possibilistes possibles ou certaines capturées dans l'explication considérée.
Dans ce chapitre, nous introduisons la notion de *graphe conceptuel possibiliste*, qui est défini comme un graphe conceptuel où chaque nœud de concept est doté d'un degré et d'une sémantique. Un tel graphe nous permet de représenter graphiquement le phénomène observé ou chacune des expressions possibilistes possibles ou certaines capturées dans une justification de $\pi^*_{b(x)}(u)$ ou de son caractère inattendu.
Pour construire le vocabulaire permettant de représenter une explication, nous devons d'abord effectuer une étape de prétraitement basée sur une justification ou du caractère inattendu de $\pi^*_{b(x)}(u)$. Dans cette étape, nous définissons une *requête d'explication possibiliste*, qui est une structure pouvant capturer une justification de $\pi^*_{b(x)}(u)$ ou de son caractère inattendu et déterminer les énoncés d'explication. À partir d'une requête d'explication possibiliste, nous établissons une correspondance entre l'attribut et le sous-domaine de l'attribut sous-jacent à chacune

des propositions composant les expressions possibilistes de l'explication. Enfin, nous terminons cette section en définissant deux requêtes explicites d'explication possibiliste : une pour la justification de $\pi^*_{b(x)}(u)$ et l'autre pour son caractère inattendu.

Chaque requête d'explication possibiliste donne lieu à un vocabulaire. À partir de ce vocabulaire, on construit tous les graphes composant la représentation d'une explication. Ensuite, la représentation de l'explication est réalisée par emboîtement des graphes conceptuels possibilistes représentant les énoncés du graphe conceptuel racine, selon la définition présentée dans le premier chapitre de la Partie D.

Ensuite, nous étendons le cadre précédent afin de représenter une explication, qui est la combinaison de la justification et du caractère inattendu de $\pi^*_{b(x)}(u)$. Pour construire une telle représentation, nous combinons les deux requêtes d'explication possibilistes associées respectivement à la justification et au caractère inattendu en une nouvelle requête d'explication possibiliste. Ensuite, nous définissons un nouveau graphe racine pour structurer cette explication. Toutes nos constructions sont illustrées par les explications extraites des deux systèmes à base de règles possibilistes utilisés dans les exemples du premier chapitre de la partie C.

Dans le troisième chapitre de la Partie D, nous représentons graphiquement deux explications des résultats d'inférence d'un système de Mamdani (un système à base de règles floues composé de règles de possibilité) : la justification d'une conclusion et son caractère inattendu. Le contenu de ces explications est extrait à l'aide des méthodes introduites dans le second chapitre de la Partie C. Comme pour les explications des décisions d'inférence possibilistes, nous représentons ces explications par des graphes conceptuels en utilisant le cadre présenté dans le premier chapitre de la partie D. Les constructions sont similaires à celles des explications des décisions d'inférence possibiliste qui ont été représentées dans le chapitre précédent.

Nous commençons par définir les objets qui composent une explication d'une décision d'inférence floue. Pour les deux explications (justification et caractère inattendu), le phénomène observé est le degré flou $\alpha^*_{(Z,O_j)}$ de la conclusion considérée (Z, O_j). Selon l'explication représentée, les autres énoncés sont les expressions de logique floue (conjonction de propositions) qui composent soit la justification de la conclusion, soit son caractère inattendu.

Nous définissons un *graphe conceptuel flou* comme un graphe conceptuel où chaque nœud de concept est doté d'un degré flou et d'une sémantique. Chaque énoncé d'une explication sera représenté par un graphe conceptuel flou.

Ensuite, nous spécifions l'entrée de nos représentations, que nous appelons une *requête d'explication floue*. Une requête d'explication floue capture une justification ou un caractère inattendu, afin d'établir les énoncés de l'explication.

À partir d'une requête d'explication floue, nous définissons un vocabulaire qui étend celui du cadre (voir premier chapitre de la partie D). Avec ce vocabulaire, nous construisons des graphes conceptuels flous représentant les énoncés d'une explication et le graphe conceptuel racine de la représentation . En imbriquant les graphes conceptuels flous représentant les énoncés dans le graphe conceptuel racine, on obtient la représentation d'une explication par un graphe conceptuel imbriqué.

Nous représentons également une explication qui est une combinaison de la justification et du caractère inattendu d'une conclusion. Cette représentation est obtenue en combinant les deux requêtes d'explication floues associées respectivement à la justification d'une conclusion et à son caractère inattendu en une seule requête et en utilisant un nouveau graphe racine pour structurer l'explication.

Enfin, nos constructions sont illustrées par les explications des résultats d'inférence du système

d'inférence floue de Mamdani utilisé comme exemple dans le second chapitre de la partie D.

Conclusion et perspectives

Dans cette thèse, nous nous sommes concentrés sur deux objectifs XAI : l'établissement de points de rencontre entre KRR et ML et l'élaboration d'une chaîne de traitement pour la génération et l'évaluation d'explications AI (Figure 1). Nos paradigmes explicatifs ont été développés pour deux systèmes d'IA : un système à base de règles possibilistes, où les règles possibilistes encodent des informations négatives et un système à base de règles floues composé de règles de possibilité qui encodent des informations positives.

Dans la partie B, pour le premier objectif, nous avons introduit une interface possibiliste entre l'apprentissage et le raisonnement Si-Alors. L'interface a été définie en généralisant le système d'équations min-max de Farreny et Prade [FP92], qui a été proposé pour développer les capacités explicatives des systèmes à base de règles possibilistes. À partir du système d'équations généralisé, nous avons obtenu une formule explicite pour la distribution des possibilités de sortie, ce qui nous a permis de calculer les mesures de possibilité et de nécessité correspondantes. Nous avons donné une condition nécessaire et suffisante pour que la distribution des possibilités de sortie soit normalisée et déterminé, lorsque c'est possible, les solutions d'entrée minimales pour la normalisation. Nous avons défini un algorithme pour reconstruire le système d'équations lorsque nous supprimons une règle. Cet algorithme nous permet d'obtenir tous les sous-systèmes d'équations d'un système d'équations initial. Enfin, nous avons montré que le système d'équations associé à une cascade peut être représenté par un réseau de neurones min-max.
À partir de notre système d'équations généralisé, nous pouvons effectuer une analyse de sensibilité, en fixant les valeurs du vecteur d'entrée ou de sortie. Cette idée a été initialement suggérée par Farreny et Prade [FP92]. Pour établir qu'une base de règles possibilistes est cohérente [DP20], c'est-à-dire qu'étant donné que les distributions de possibilités des attributs d'entrée sont normalisées, la distribution de possibilités de sortie doit toujours être normalisée, nous pouvons rechercher des conditions générales portant sur les degrés des prémisses et les paramètres des règles. Enfin, pour développer des méthodes d'apprentissage possibilistes, il serait intéressant d'adapter, pour notre réseau de neurones, une méthode de descente de gradient min-max [BDR94, LQLC17, TL97]. Nous pouvons également envisager d'utiliser la méthode d'apprentissage de NEFLCASS [NK99], un système flou qui utilise l'inférence min-max commune et utilise un algorithme d'apprentissage heuristique. Une autre approche pour l'apprentissage des paramètres des règles serait d'utiliser le fait que le système d'équations a la forme d'un système d'équations de relations floues [San76]. Par conséquent, l'apprentissage peut être effectué en utilisant les algorithmes de résolution d'équations de relations floues, voir [Pee13]. Les méthodes d'approximation proposées dans [Cec00] peuvent également être utiles.

Dans la partie C, pour la chaîne de traitement, nous avons introduit des paradigmes explicatifs pour justifier les résultats d'inférence des systèmes à base de règles possibilistes et floues. Pour les deux types de systèmes à base de règles, nous avons développé une méthode pour sélectionner les prémisses de règles qui justifient un résultat d'inférence. Ensuite, nous avons défini des fonctions de réduction des prémisses pour les deux types de systèmes à base de règles. En les appliquant aux prémisses sélectionnées, cela nous a permis de former deux types d'explications d'un résultat d'inférence : sa justification et son caractère inattendu. Comme notre approche est basée sur un seuil qui a un impact majeur sur le contenu des explications, nous devons

trouver des moyens de le déterminer pour une base de règles. Il serait également important d'évaluer les explications afin de voir si elles conviennent aux utilisateurs [DVK17, MZR21]. Des protocoles d'évaluation ont été proposés pour les explications des résultats des systèmes à base de règles e.g., [BP19, vdWNCN21].

Dans la partie D, nous avons proposé une représentation graphique d'une explication. Tout d'abord, nous avons donné une méthode générale pour représenter les explications en termes de graphes conceptuels. Ensuite, nous l'avons étendue pour représenter les explications des décisions d'inférence possibilistes et floues. Pour chaque type de système à base de règles, nous avons représenté trois explications d'un résultat d'inférence : sa justification, son caractère inattendu et une combinaison de sa justification et de son caractère inattendu. Pour la représentation de ces explications, nous avons défini deux types de graphes : les graphes conceptuels possibilistes et les graphes conceptuels flous, où chaque nœud de concept est doté d'un degré et d'une sémantique associée. Pour les graphes que nous avons introduits, nous devons étendre les travaux sur les graphes conceptuels classiques [CM08] pour leur fournir un mécanisme d'interrogation et une interprétation logique.
La représentation peut être étendue à d'autres explications. Par exemple, nous pouvons l'étendre pour représenter les explications des résultats d'inférence d'une cascade (peut-être en imbriquant les représentations) ou les explications des résultats d'autres systèmes d'IA.
La représentation peut être utilisée par des systèmes de NLG pour produire des explications en langage naturel. Cela pourrait être fait en adaptant les systèmes NLG qui utilisent les entrées du web sémantique pour produire du texte [BACW14, GSNPB17]. Parmi eux, notons que le système FORGe [MDW19], qui a obtenu le meilleur score dans l'évaluation humaine du défi WebNLG [GSNPB17], est basé sur le transducteur de graphes MATE [BW10] qui utilise un graphe conceptuel en entrée.

References

[ACM17] US ACM. Public policy council.(2017). *Statement on algorithmic transparency and accountability*, 2017.

[ACMM21] José M Alonso, Ciro Castiello, Luis Magdalena, and Corrado Mencar. *Explainable fuzzy systems: Paving the way from interpretable fuzzy systems to explainable ai systems*. Springer, 2021.

[ADRDS+20] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[Ame19] Kay R Amel. From shallow to deep interactions between knowledge representation, reasoning and machine learning. In *Proceedings 13th International Conference Scala Uncertainity Mgmt (SUM 2019), Compiegne, LNCS*, pages 16–18, 2019.

[BACW14] Nadjet Bouayad-Agha, Gerard Casamayor, and Leo Wanner. Natural language generation in the context of the semantic web. *Semantic Web*, 5(6):493–513, 2014.

[BC17] Or Biran and Courtenay Cotton. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, volume 8, pages 8–13, 2017.

[BDR94] A Blanco, M Delgado, and I Requena. Solving fuzzy relational equations by max-min neural networks. In *Proceedings of 1994 IEEE 3rd International Fuzzy Systems Conference*, pages 1737–1742. IEEE, 1994.

[BP19] Ismaïl Baaj and Jean-Philippe Poli. Natural language generation of explanations of fuzzy inference decisions. In *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6. IEEE, 2019.

[BPO19] Ismaïl Baaj, Jean-Philippe Poli, and Wassila Ouerdane. Some insights towards a unified semantic representation of explanation for explainable artificial intelligence. In *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019)*, pages 14–19, 2019.

[BW10] Bernd Bohnet and Leo Wanner. Open soucre graph transducer interpreter and grammar development environment. In *LREC*, 2010.

[Cec00] Katarína Cechlárová. A note on unsolvable systems of max–min (fuzzy) equations. *Linear Algebra and its Applications*, 310(1-3):123–128, 2000.

[CM08] Michel Chein and Marie-Laure Mugnier. *Graph-based knowledge representation: computational foundations of conceptual graphs*. Springer Science & Business Media, 2008.

[DEGP07] Didier Dubois, Francesc Esteva, Lluís Godo, and Henri Prade. Fuzzy-set based logics-an history-oriented presentation of their main developments. *The Many Valued and Nonmonotonic Turn in Logic*, 8:325–449, 2007.

[Des17] Jean-Louis Dessalles. Conversational topic connectedness predicted by simplicity theory. In *CogSci*, 2017.

[DP88] D. Dubois and H. Prade. *Possibility theory: an approach to computerized processing of uncertainty*. Plenum Press, New York, 1988.

[DP98] Didier Dubois and Henri Prade. Possibility theory: qualitative and quantitative aspects. In *Quantified representation of uncertainty and imprecision*, pages 169–226. Springer, 1998.

[DP15] D. Dubois and H. Prade. Possibility theory and its applications: Where do we stand? In *Handbook of Computational Intelligence*, 2015.

[DP20] Didier Dubois and Henri Prade. From possibilistic rule-based systems to machine learning-a discussion paper. In *International Conference on Scalable Uncertainty Management*, pages 35–51. Springer, 2020.

[DPU03] Didier Dubois, Henri Prade, and Laurent Ughetto. A new perspective on reasoning with fuzzy rules. *International journal of intelligent systems*, 18(5):541–567, 2003.

[DVK17] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[FP86] Henri Farreny and Henri Prade. Default and inexact reasoning with possibility degrees. *IEEE transactions on systems, man, and cybernetics*, 16(2):270–276, 1986.

[FP90] Henri Farreny and Henri Prade. Explications de raisonnements dans l'incertain. *Revue d'intelligence artificielle*, 4(2):43–75, 1990.

[FP92] Henri Farreny and Henri Prade. *Positive and Negative Explanations of Uncertain Reasoning in the Framework of Possibility Theory*, page 319–333. John Wiley & Sons, Inc., USA, 1992.

[FPW86] Henri Farreny, Henri Prade, and E Wyss. Approximate reasoning in a rule-based expert system using possibility theory: A case study. In *IFIP Congress*, pages 407–414, 1986.

[Fre14] Alex A Freitas. Comprehensible classification models: a position paper. *ACM SIGKDD explorations newsletter*, 15(1):1–10, 2014.

[GA19] David Gunning and David Aha. Darpa's explainable artificial intelligence (xai) program. *AI Magazine*, 40(2):44–58, 2019.

[GK18] Albert Gatt and Emiel Krahmer. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170, 2018.

[GMR+18] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5), aug 2018.

[GSC+19] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. Xai—explainable artificial intelligence. *Science Robotics*, 4(37), 2019.

[GSNPB17] Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. The webnlg challenge: Generating text from rdf data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, 2017.

[HLD+19] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4):e1312, 2019.

[JCDG08] Hazaël Jones, Brigitte Charnomordic, Didier Dubois, and Serge Guillaume. Practical inference with systems of gradual implicative rules. *IEEE Transactions on Fuzzy Systems*, 17(1):61–78, 2008.

[Lip18] Zachary C Lipton. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018.

[LOS+21] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. What do we want from explainable artificial intelligence (xai)?–a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research. *Artificial Intelligence*, 296:103473, 2021.

[LQLC17] Long Li, Zhijun Qiao, Yan Liu, and Yuan Chen. A convergent smoothing algorithm for training max–min fuzzy neural networks. *Neurocomputing*, 260:404–410, 2017.

[MAT+16] Brent Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2):2053951716679679, 2016.

[MDW19] Simon Mille, Stamatia Dasiopoulou, and Leo Wanner. A portable grammar-based nlg system for verbalization of structured data. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, pages 1054–1056, 2019.

[MHC+19] Shane T Mueller, Robert R Hoffman, William J Clancey, Abagail K Emery, and Gary Klein. Explanation in human-ai systems: A literature meta-review synopsis of key ideas and publications and bibliography for explainable ai. Technical report, DARPA, 2019.

[Mil19] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

[MZR21] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 11(3-4):1–45, 2021.

[NK99] Detlef Nauck and Rudolf Kruse. Obtaining interpretable fuzzy classification rules from medical data. *Artificial intelligence in medicine*, 16(2):149–169, 1999.

[Pee13] Ketty Peeva. Resolution of fuzzy relational equations–method, algorithm and software with applications. *Information Sciences*, 234:44–63, 2013.

[RD97] Ehud Reiter and Robert Dale. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87, 1997.

[Reg16] General Data Protection Regulation. Regulation eu 2016/679 of the european parliament and of the council of 27 april 2016. *Official Journal of the European Union*, 2016.

[San76] Elie Sanchez. Resolution of composite fuzzy relation equations. *Information and control*, 30(1):38–48, 1976.

[SD15] Antoine Saillenfest and Jean-Louis Dessalles. Some probability judgments may rely on complexity assessments. In *CogSci*, 2015.

[TL97] Loo-Nin Teow and Kia-Fock Loe. An effective learning method for max-min neural networks. In *IJCAI*, pages 1134–1139. Citeseer, 1997.

[TM11] Nava Tintarev and Judith Masthoff. Designing and evaluating explanations for recommender systems. In *Recommender systems handbook*, pages 479–510. Springer, 2011.

[TS81] Randy L Teach and Edward H Shortliffe. An analysis of physician attitudes regarding computer-based clinical consultation systems. In *Use and impact of computers in clinical medicine*, pages 68–85. Springer, 1981.

[vdWNCN21] Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. Evaluating xai: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, 291:103404, 2021.

[VLK99] Werner Van Leekwijck and Etienne E Kerre. Defuzzification: criteria and classification. *Fuzzy sets and systems*, 108(2):159–178, 1999.

[Zad65] Lotfi A Zadeh. Fuzzy sets. *Information and Control*, 8(3):338–353, 1965.

[Zad78] Lotfi A Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy sets and systems*, 1(1):3–28, 1978.