De l'équivalence entre les modèles structurels causaux et les systèmes abstraits d'argumentation

Yann Munro, Isabelle Bloch, Mohamed Chetouani, Marie-Jeanne Lesot, Catherine Pelachaud

De l'équivalence entre les modèles structurels causaux et les systèmes abstraits d'argumentation

Y. Munro[1,2], I. Bloch[1], M. Chetouani[2], M-J. Lesot[1], C. Pelachaud[3]

[1] Sorbonne Université, CNRS, LIP6, Paris, France
[2] Sorbonne Université, CNRS, ISIR, Paris, France
[3] CNRS, Sorbonne Université, ISIR, Paris, France

prenom.nom@sorbonne-universite.fr

Résumé

Les modèles structurels causaux et les systèmes abstraits d'argumentation sont deux approches s'inscrivant dans les problématiques de l'intelligence artificielle explicable. Dans cet article, nous mettons en évidence une équivalence entre un cas particulier de ces modèles causaux, que nous appelons graphes causaux argumentatifs, et les systèmes abstraits d'argumentation. Nous proposons également une transformation permettant de passer d'une représentation à l'autre.

Mots-clés

Modèles structurels causaux, Systèmes abstraits d'argumentation, Explications en intelligence artificielle (XAI).

Abstract

In the field of explainable artificial intelligence, causal models and abstract argumentation frameworks are two formal approaches that provide a definition of an explanation. In this paper, we show the equivalence between a particular type of causal models, that we call argumentative causal graphs, and abstract argumentation frameworks. We also propose a transformation between these two systems.

Keywords

Causal models, Abstract argumentation frameworks, Explainable artificial intelligence (XAI).

1 Introduction

Il existe de nombreuses méthodes permettant de contribuer à l'interprétabilité et à l'explicabilité des systèmes d'intelligence artificielle (XAI) [5]. Des approches numériques ont pour objectif de fournir une explication en cherchant par exemple les corrélations entre les attributs d'entrée et de sortie. Des approches symboliques reposent sur des formalismes logiques pour raisonner par abduction ou rechercher des causalités, à partir de la modélisation formelle d'un problème ou d'une situation. C'est à ce type d'approche que nous nous intéressons dans cet article.

Pour améliorer la qualité de ces méthodes, une approche consiste à s'inspirer des mécanismes cognitifs et sociaux humains notamment ceux liés au processus d'explication et en déduire des propriétés et comportements intéressants. Dans [13], Tim Miller, en s'appuyant sur des travaux en sciences sociales et cognitives, dégage des caractéristiques essentielles qu'il conviendra de retrouver lors du développement de méthodes d'intelligence artificielle explicable.

Un premier cadre formel provient des travaux de Joseph Halpern et Judea Pearl [9] sur la causalité et notamment sur ce qu'ils appellent des modèles structurels causaux. Cette notion est en effet très intimement liée à celle d'explication : expliquer un fait revient souvent à fournir une cause et on retrouve donc logiquement dans leurs travaux une définition de cette notion [10]. Ce cadre a par exemple été mis en œuvre par Prashan Madumal et al. pour générer des explications pour un agent jouant à Starcraft II [12], un jeu de stratégie en temps réel. Cet article s'intéresse à un cas particulier de ces modèles que nous proposons d'appeler des graphes causaux argumentatifs.

Un autre cadre proposant une définition de la notion d'explication est celui de l'argumentation. Introduits par Phan Minh Dung en 1995 [6], les systèmes abstraits d'argumentation (*abstract argumentation framework*, AAF) permettent de modéliser les interactions entre des arguments provenant de plusieurs entités ou agents. De nombreuses méthodes de XAI ont déjà été développées dans ce cadre [17], que ce soit pour des problèmes modélisés initialement par des graphes argumentatifs ou bien pour des modèles qui à l'origine ne l'étaient pas.

Après avoir présenté brièvement ces deux cadres dans les sections 2 et 3, nous mettons en évidence une équivalence entre eux par l'intermédiaire d'une transformation permettant de passer des graphes argumentatifs aux graphes causaux argumentatifs et inversement (sections 4 et 5). A notre connaissance, il n'y a pas eu de travaux dans ce sens et c'est pourquoi nous proposons des transformations permettant de lier les deux champs, ce qui constitue la contribution principale de l'article. L'objectif n'est donc pas de présenter une nouvelle méthode ou un nouveau cadre mais bien de pouvoir passer de l'un à l'autre et donc permettre d'exploiter les propriétés intéressantes de chacun.

L'article illustre les principes proposés sur un exemple inspiré des assistants de régulation médicaux dans le cadre de la situation sanitaire liée au COVID-19 : il considère un

agent, humain ou autonome, dont l'objectif est de conseiller sur la nécessité de réaliser un test PCR [1]. Il s'agit évidemment d'un modèle très simplifié de la réalité dont l'unique but est d'illustrer nos contributions et qui n'a pas vocation à remplacer les consignes sanitaires existantes.

2 Modèles structurels causaux

Cette section rappelle les concepts définis par J. Halpern [8] qui conduisent à la définition de la notion d'explication dans le cadre des modèles structurels causaux.

2.1 Définition

Un modèle structurel causal tel qu'introduit par J. Halpern [8] est un triplet $M = (\mathcal{U}, \mathcal{V}, \mathcal{F})$ tel que :

- $\mathcal{U}$ est un ensemble des variables exogènes, c'est-à-dire un ensemble de variables dont les valeurs sont indépendantes du modèle ;
- $\mathcal{V}$ est un ensemble des variables endogènes ;
- $\mathcal{F}$ est l'ensemble des équations structurelles du modèle (une pour chaque variable de $\mathcal{V}$). Elles permettent d'associer une valeur à chacune des variables endogènes en fonction des valeurs des variables exogènes.

En associant à chaque variable un nœud et en traçant des arcs entre ces nœuds pour indiquer les dépendances fonctionnelles de $\mathcal{F}$, on obtient une représentation d'un modèle structurel M sous la forme d'un graphe.
L'équivalence discutée dans les sections 4 et 5 s'intéresse à un cas particulier de modèles structurels causaux que nous proposons d'appeler **graphes causaux argumentatifs** (GCA). Ce sont des triplets $M = (\mathcal{U}, \mathcal{V}, \mathcal{F})$ pour lesquels :

1. Les variables sont à valeur binaire. Les équations structurelles s'écrivent donc comme des formules logiques.
2. Ces formules ne contiennent pas de disjonction.
3. Le graphe associé est acyclique.

Dans la suite, les notations supplémentaires suivantes sont utilisées :

- Soit $M = (\mathcal{U}, \mathcal{V}, \mathcal{F})$, un modèle structurel causal. On appelle un **contexte**, noté $\mathbf{u}$, une affectation des variables de $\mathcal{U}$. La paire $(M, \mathbf{u})$ est appelée **monde**.
- Soit $\mathbf{X}$ un ensemble de variables de $\mathcal{V}$, on note $\mathbf{X} = \mathbf{x}$ une affectation des variables de $\mathbf{X}$ avec les valeurs de $\mathbf{x}$.
- Soit $\mathcal{K}$ un ensemble de contextes et $\mathbf{u} \in \mathcal{K}$, on note $(M, \mathbf{u}) \models \mathbf{X} = \mathbf{x}$ si $\mathbf{X} = \mathbf{x}$ est l'unique solution aux équations de $\mathcal{F}$ dans $\mathbf{u}$.
- Soit $\mathcal{K}$ un ensemble de contextes. Soit $\mathbf{X} \in \mathcal{V}$ et $\mathbf{x}$ des valeurs de $\mathbf{X}$. On note $\mathcal{K}_{\mathbf{X}=\mathbf{x}}$, l'ensemble des contextes $\mathbf{u}'$ de $\mathcal{K}$ tel que $(M, \mathbf{u}') \models \mathbf{X} = \mathbf{x}$.
- Soit $\mathcal{K}$ un ensemble de contextes et $\mathbf{u} \in \mathcal{K}$, la notation $(M, \mathbf{u}) \models [\mathbf{X} = \mathbf{x}](\mathbf{Y} = \mathbf{y})$ signifie que l'on se place dans le monde $(M, \mathbf{u})$ dans lequel les équations de $\mathcal{F}$ portant sur les variables de $\mathbf{X}$ sont remplacées par l'équation $\mathbf{X} = \mathbf{x}$.

1. *Reverse Transcriptase-Polymerase Chain Reaction*

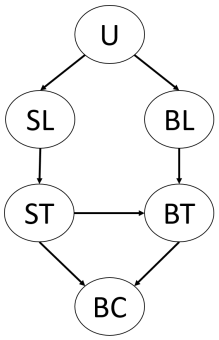

FIGURE 1 – Scénario causal, inspiré de [16].

Exemple 1. *Nous reprenons ici un exemple classique tiré de [16]. Suzy et Billy lancent tous les deux une pierre en direction d'une bouteille de verre. Ils sont tous les deux parfaitement précis et sont donc sûrs de toucher la bouteille s'ils lancent effectivement la pierre. Si l'une des pierres atteint la bouteille alors celle-ci se casse. La pierre de Suzy touche toujours la première.*
Pour modéliser la situation, on introduit les variables suivantes : SL (respectivement BL) et ST (resp. BT) représentent «Suzy (resp. Billy) lance» et «Suzy (resp. Billy) touche». Enfin, BC renvoie à «la bouteille se casse».
On introduit également un ensemble $\mathcal{U}$ de variables exogènes qui représentent des facteurs extérieurs au problème qui influencent le fait que Billy ou Suzy lancent la pierre.
Les fonctions de $\mathcal{F}$ complètent la modélisation du problème. Par exemple, le fait que Billy touche la bouteille dans le cas (et uniquement dans ce cas) où il a lancé une pierre et Suzy n'a pas touché la bouteille est représenté par la fonction suivante : $BT = BL \wedge \neg ST$.
On obtient le modèle structurel causal suivant, illustré figure 1 :
$\mathcal{U} = \{U\}$
$\mathcal{V} = \{SL, BL, SL, BT, BC\}$
$\mathcal{F} = \{(ST = SL), (BT = BL \wedge \neg ST), (BC = ST \vee BT)\}$

2.2 Cause effective

Dans le formalisme précédent des modèles structurels causaux, J. Halpern [8] propose ensuite de définir la notion de cause de la façon suivante.
L'affectation $\mathbf{X} = \mathbf{x}$ est une **cause effective** de φ dans le monde $(M, \mathbf{u})$ si les trois conditions suivantes sont vérifiées :

AC1 $(M, \mathbf{u}) \models (\mathbf{X} = \mathbf{x}) \wedge \varphi$, c'est-à-dire la cause et la conséquence sont toutes les deux vraies dans le monde considéré.

AC2 Il existe un ensemble $\mathbf{W}$ de variables endogènes avec des valeurs $\mathbf{w}$ et une configuration $\mathbf{x}'$ pour la variable $\mathbf{X}$ tels que si $(M, \mathbf{u}) \models (\mathbf{W} = \mathbf{w})$ alors :

$$(M, \mathbf{u}) \models [\mathbf{X} = \mathbf{x}', \mathbf{W} = \mathbf{w}]\neg\varphi$$

.

AC3 $\mathbf{X}$ est minimal : il n'existe pas de sous-ensemble de $\mathbf{X}$ qui satisfasse **AC1** et **AC2**. Cette dernière condition vise à éviter d'avoir des variables inutiles dans la cause.

Pour savoir qu'une chose est une conséquence d'une autre, il est possible de raisonner en se demandant : si la cause présumée ne s'était pas produite, est-ce que la conséquence se serait tout de même produite ? C'est ce que l'on appelle un scénario contrefactuel ou hypothétique. Si la réponse à la question précédente est « non », alors la cause présumée devient une cause effective.
La condition **AC2** renvoie à ce raisonnement sur les contrefactuels. Plus précisément, cette condition impose que s'il existe un scénario contrefactuel, c'est-à-dire un scénario dans lequel la cause présumée ne s'est pas produite ($\mathbf{X} = \mathbf{x}'$) **et** éventuellement d'autres évènements se sont quand même produits ($\mathbf{W} = \mathbf{w}$), tel que la conséquence à expliquer ne se produise pas, alors la cause présumée est bien une cause.

Exemple 1. *(suite) – Intuitivement, une cause de la bouteille qui se casse est le fait que Suzy ait lancé la pierre. En effet, c'est sa pierre qui a touché la bouteille et l'a donc cassée. Cependant, si on se pose la question : si Suzy n'avait pas lancé sa pierre, la bouteille se serait-elle cassée ? La réponse est oui car Billy aurait alors touché la bouteille* $(BT = BL \wedge \neg ST)$.
Il faut donc envisager le contrefactuel suivant : si Suzy n'avait pas lancé sa pierre ***et sachant que*** *Billy n'a pas touché la bouteille, la bouteille se serait-elle cassée ? Dans ce cas la réponse est effectivement non, c'est-à-dire que le fait que Suzy ait lancé sa pierre est bien une cause effective du fait que la bouteille se casse.*

2.3 Cause suffisante

Soit $\mathcal{K}$ un ensemble de contextes et $\mathbf{u} \in \mathcal{K}$. L'affectation $\mathbf{X} = \mathbf{x}$ est une **cause suffisante** de φ dans le monde $(M, \mathbf{u})$ si les quatre conditions suivantes sont vérifiées :

SC1 $(M, \mathbf{u}) \models (\mathbf{X} = \mathbf{x}) \wedge \varphi$.

SC2 Il existe une partie de $\mathbf{X}, X = x$, et une autre conjonction $(\mathbf{Y} = \mathbf{y})$ (éventuellement vide) telles que $(X = x) \wedge (\mathbf{Y} = \mathbf{y})$ est une cause effective de φ dans $(M, \mathbf{u})$, c'est-à-dire une partie de $\mathbf{X}$ est une partie d'une cause effective dans le monde considéré.

SC3 $(M, \mathbf{u}') \models [\mathbf{X} = \mathbf{x}]\varphi$ pour tous les contextes $\mathbf{u}' \in \mathcal{K}$, c'est-à-dire si l'on a $\mathbf{X} = \mathbf{x}$ alors on a φ quel que soit le contexte considéré.

SC4 $\mathbf{X}$ est minimal.

Remarque 1. *Il existe une deuxième version de la définition de cause suffisante proposée par T. Miller dans [14]. Il définit cette notion comme une cause effective non minimale, c'est-à-dire qui ne vérifie que* **AC1** *et* **AC2***. La différence majeure se situe dans* **SC3***. Le point de vue de T. Miller se concentre uniquement sur le contexte en cours, contrairement à J. Halpern qui définit une cause suffisante sur un ensemble de contextes donnés. On fait le choix ici de considérer plutôt la définition de J. Halpern notamment car en affaiblissant* **SC3** *on peut définir une notion de pouvoir explicatif utile pour comparer les explications générées.*

2.4 Explication

Lorsque l'on fournit une explication, il est important de tenir compte de la personne à qui est fournie cette explication. On appelle cette personne le destinataire de l'explication ou en anglais l'*explainee*. Pour cette raison, la recherche de cause effective et de cause suffisante va être contrainte à un ensemble de contextes $\mathcal{K}$ déterminé par ce que l'*explainee* considère comme possible.
L'affectation $\mathbf{X} = \mathbf{x}$ est une **explication** de φ relative à l'ensemble de contextes $\mathcal{K}$ si les trois conditions suivantes sont vérifiées :

EX1 $\mathbf{X} = \mathbf{x}$ est une cause suffisante pour tous les contextes $\mathbf{u}$ dans $\mathcal{K}$ qui vérifient $(\mathbf{X} = \mathbf{x}) \wedge \varphi$.

EX2 $\mathbf{X}$ est minimal.

EX3 $\mathcal{K}_{(\mathbf{X}=\mathbf{x})\wedge\varphi} \neq \emptyset$, c'est-à-dire les contextes considérés comme possibles par l'*explainee* sont compatibles avec l'explication.

L'explication est dite non triviale si elle vérifie en plus

EX4 $(M, \mathbf{u}') \models \neg(\mathbf{X} = \mathbf{x})$ pour certains contextes $\mathbf{u}' \in K_\varphi$.

L'ensemble de contextes $\mathcal{K}$ est déterminé par l'*explainee*. Ainsi, il est possible que cet ensemble soit trop restrictif, c'est-à-dire que les contextes considérés ne soient pas compatibles avec les explications. En effet, s'il n'existe pas de causes suffisantes dans au moins un contexte de $\mathcal{K}$ (c'est-à-dire $\mathcal{K}_{(\mathbf{X}=\mathbf{x})\wedge\varphi} = \emptyset$) alors, il n'y a pas d'explication possible.
Il existe une définition plus générale de la notion d'explication proposée par J. Halpern [8]. Elle permet notamment de remédier au problème mentionné ci-dessus. En effet, dans celle-ci, on prend également en compte le fait que le destinataire de l'explication n'a pas une connaissance parfaite du modèle, et donc l'explication doit apporter une connaissance supplémentaire. Pour cela, on renvoie non seulement une affectation mais également des formules permettant à ce destinataire de mieux comprendre le modèle. Ainsi, si aucune cause suffisante n'existe dans l'ensemble de contextes $\mathcal{K}$ considéré par l'*explainee*, alors renvoyer une formule en plus peut permettre à ce dernier d'élargir l'ensemble $\mathcal{K}$ des contextes possibles.

3 Système abstrait d'argumentation

Cette section rappelle brièvement les principes des systèmes abstraits d'argumentation de P.M. Dung [6] ainsi qu'une définition d'explication [7] pour ce cadre.

3.1 Définition

Un système abstrait d'argumentation est un couple $AF = (A, R)$ telle que :

— A est un ensemble d'arguments,

— R est une relation binaire sur $A \times A$.

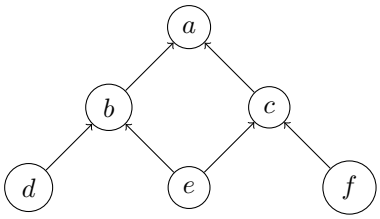

FIGURE 2 – Scénario argumentatif.

On appelle R la relation d'attaque et on dit qu'un argument $a \in A$ attaque $b \in A$ si $(a, b) \in R$ et on écrit $R(a, b)$. Comme R est une relation binaire à support fini, on peut naturellement représenter un système abstrait d'argumentation sous la forme d'un graphe associé.
Ce formalisme n'impose rien quant à la structure interne d'un argument, ni sur la nature d'une attaque. Ainsi, un argument peut simplement être un énoncé en langage naturel. Il peut également s'agir d'une formule définie dans un certain langage selon des règles, comme dans le cas du système ASPIC+ [15].

Exemple 2. *Cet exemple considère un scénario simple d'aide au dépistage du COVID-19 par un agent. Imaginons que l'utilisateur n'a pas l'impression d'avoir de symptômes particuliers, il s'est juste réveillé avec quelques courbatures. Il décide alors de consulter cet agent. Celui-ci pose un certain nombre de questions à l'utilisateur sur son état de santé. En effet, avoir des courbatures n'est pas suffisant pour justifier d'aller faire un test PCR, un auto-test pourrait par exemple suffire. L'agent lui demande de goûter un condiment au goût prononcé (sel, sucre, vinaigre ...) afin de tester son goût. Enfin, il faut également vérifier s'il est cas contact.*
Leur conversation peut être représentée par le système abstrait d'argumentation suivant, illustré figure 2 :

- $A = \{a$ *: «Test PCR nécessaire»,* b *: «Aucun symptôme»,* c *: «Parcours vaccinal complet»,* d *: «Courbatures»,* e *: «Perte du goût»,* f *: «Je suis cas contact»}*
- $R = \{(b, a), (c, a), (d, b), (e, b), (e, c), (f, c)\}$

Dans le cas où l'utilisateur n'a pas l'impression d'avoir de symptômes particuliers et est vacciné, un test PCR n'est pas nécessaire. Cela est représenté par les deux premières relations d'attaque (b, a) *et* (c, a)*. Toutefois, s'il a des courbatures ou une perte de goût, il n'est plus possible de dire qu'il n'a plus de symptômes (*$((d, b), (e, b))$*. De même, s'il est cas contact ou bien s'il n'a plus de goût, le fait d'avoir un parcours vaccinal complet ne justifie plus de ne pas aller faire de test PCR (*$((e, c), (f, c))$*. En particulier, être vacciné n'empêche pas d'attraper le COVID-19.*
Le graphe présenté en figure 2 représente le cas où l'utilisateur a des courbatures, perdu le goût et est cas contact (d, e, f)*. D'après ce graphe,* a *n'est attaqué que par des arguments non acceptés (car attaqué par des arguments non attaqués) et peut donc être accepté. Ainsi, il faut réaliser un test PCR.*

3.2 Quelques définitions supplémentaires

— On note Att_a^R **l'ensemble des attaquants directs** de a pour la relation R :

$$Att_a^R = \{b \in A \mid R(b, a)\}$$

Quand une seule relation d'attaque est définie, on note simplement Att_a.

— Un ensemble S est **sans conflit** s'il n'y pas d'arguments $(a, b) \in S^2$ tel que $(a, b) \in R$:

$$\forall (a, b) \in S^2, (a, b) \notin R$$

— Un argument $a \in A$ est **acceptable** par un ensemble S si S attaque tous les attaquants de a :

$$\forall b \in Att_a, \exists c \in S \cap Att_b$$

— Un ensemble S sans conflit et tel que tous ses éléments sont acceptables par S est dit **admissible** :

$$\forall (a, b) \in S^2, (a, b) \notin R$$
$$\text{et } \forall a \in S, \forall b \in Att_a, \exists c \in S \cap Att_b$$

— Un ensemble S est dit **admissible lié** s'il est admissible et si au moins un de ses arguments est attaqué :

$$S \text{ est admissible et } \exists x \in S \text{ tel que } Att_x \neq \emptyset.$$

Un tel x est alors appelé un **sujet** de S.

Exemple 2. *(suite) – Dans le cas de l'*AF *défini précédemment, cherchons un ensemble admissible lié* S_{ex} *de sujet* a*. Comme* a *est attaqué par* b*, il faut que* S_{ex} *contienne un attaquant de* b*. Prenons* d *par exemple. Ensuite* d *n'est pas attaqué donc il est acceptable par* S_{ex}*. De plus,* a *est également attaqué par* c*. Il faut donc ajouter un attaquant de* c *à* S_{ex}*. Ajoutons par exemple* e*. L'argument* e *est non attaqué, il est donc lui aussi acceptable par* S_{ex}*. Enfin, tous les attaquants de* a *sont attaqués par un élément de* S_{ex}*,* a *est donc acceptable par* S_{ex}*. On a ainsi construit* $S_{ex} = \{a, d, e\}$*.*
*De la même manière, l'ensemble des ensembles admissibles (**liés** de sujet* a*) est :*
$S_{adm} = \{\{d\}, \{e\}, \{f\}, \{d, e\}, \{d, f\}, \{e, f\}, \{d, e, f\},$
***{a,d,e,f}**,* ***{a,d,f}**,* ***{a,e,f}**,* ***{a,d,e}**,* ***{a,e}**$\}$.

3.3 Explications

Dans cette section, nous reprenons la définition d'explication donnée par X. Fan et F. Toni dans [7].
Soit $x \in A$, une **explication** S de x est un ensemble admissible lié de sujet x.
Une explication de x est dite **compacte** si elle est *minimale* au sens de l'inclusion.
Une explication de x est dite **verbeuse** si elle est *maximale* au sens de l'inclusion.

Exemple 2. *(suite) – L'argument* a *possède ici deux explications compactes : «un test PCR est nécessaire» car l'humain a «une perte de goût», soit* $\{a, e\}$*, ou car il a «des courbatures» et est «cas contact», soit* $\{a, d, f\}$*.*
Il y a également une explication verbeuse : «perte de goût, des courbatures et cas contact», $\{a, d, e, f\}$*.*

Il existe d'autres définitions de la notion d'explication pour les systèmes d'argumentation. Cependant, dans la plupart des cas, celles-ci nécessitent des notions supplémentaires [2] et sortent du cadre des AAF défini par P.M. Dung [6]. Pour cette raison, nous ne les considérons pas dans le cadre de cet article.

Remarque 2. *Dans le cadre de l'argumentation abstraite, l'objectif n'est pas de modéliser le destinataire de l'explication. Il s'agit plutôt d'une retranscription d'un échange d'arguments entre plusieurs entités. L'explication sert ainsi à justifier pourquoi un argument peut être accepté en renvoyant les différents arguments qui sont intervenus pour défendre ce dernier. Ici, il n'est pas question de contexte. En particulier, il est supposé que chaque entité connaît l'intégralité des arguments et comment ils interagissent.*

Les deux sections suivantes présentent la contribution principale de l'article, à savoir l'équivalence entre les GCA et les systèmes abstraits d'argumentation.

4 Passage des AAF aux GCA

Cette section présente une transformation des graphes argumentatifs en GCA. Nous nous intéresserons aussi à comment la notion d'explication se transporte des AAF aux GCA.

4.1 Transformation proposée

On considère un couple $AF = (A, R)$ et son graphe associé que l'on suppose acyclique.
On associe à chaque argument a une variable booléenne X_a telle que $X_a = 1$ se lit comme « l'argument a est accepté ». Ces variables constituent l'ensemble des variables endogènes.
De plus, pour tous les arguments a non attaqués, on crée une variable booléenne supplémentaire $\tilde{X_a}$. Ces variables forment l'ensemble des variables exogènes.
Formellement, posons :

- $\mathcal{V} = \{X_a \mid a \in A\}$,
- $\mathcal{U} = \{\tilde{X_a} \mid (a \in A) \wedge (Att_a = \emptyset)\}$,
- $\mathcal{F} = \{F_{X_a} \mid X_a \in \mathcal{V}\}$ avec :
  - $\diamond$ $\forall a \in A$ tel que $Att_a \neq \emptyset$, $F_{X_a} = \bigwedge_{b \in Att_a} \neg X_b$,
  - $\diamond$ $\forall a \in A$ tel que $Att_a = \emptyset$, $F_{X_a} = \tilde{X_a}$.

Le triplet $M = (\mathcal{U}, \mathcal{V}, \mathcal{F})$ est un modèle structurel causal, acyclique et dont les équations structurelles $\mathcal{F}$ n'utilisent pas de disjonctions. Ce modèle M est donc bien un GCA.
Pour chaque argument non attaqué, nous avons proposé de créer deux variables, une endogène et une exogène. Ce dédoublement permet de choisir si un argument non attaqué est accepté ou non par l'intermédiaire de son représentant exogène $\tilde{X_a}$ en l'initialisant à 0 ou à 1. De plus, dans le cadre défini par J. Halpern et J. Pearl [9], seules les variables endogènes peuvent être des causes et donc des explications. Ainsi, avec son représentant endogène, un argument non attaqué pourra lui aussi être une cause.

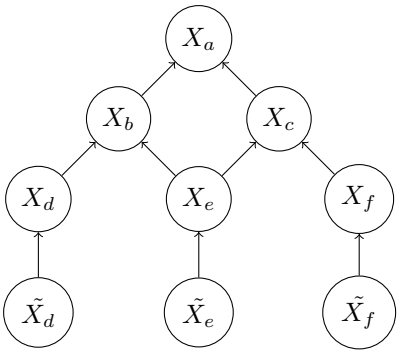

FIGURE 3 – Graphe causal argumentatif issu de la transformation du graphe argumentatif présenté en figure 2.

Exemple 2. *(suite) – L'application de la transformation décrite dans cette section à l'exemple 2 conduit à la construction de six variables endogènes : $\mathcal{V} = \{X_a, X_b, X_c, X_d, X_e, X_f\}$, et de trois variables exogènes, correspondant aux trois arguments non attaqués (d, e, f) : $\mathcal{U} = \{\tilde{X_d}, \tilde{X_e}, \tilde{X_f}\}$.*
Enfin, on transforme les relations d'attaque en équations structurelles. Par exemple, a est attaqué par b et c donc $F_{X_a} = \neg X_b \wedge \neg X_c$.
Avec ces transformations, on obtient le graphe causal argumentatif présenté en figure 3.

On appelle **contexte par défaut** de l'argumentation l'unique contexte $\mathbf{u}^*$ tel que toutes les variables exogènes valent 1. Il représente la situation décrite par le graphe argumentatif dans lequel tous les arguments non attaqués sont acceptés.

4.2 Retour sur les explications

Ces deux formalismes possèdent chacun leur propre définition de la notion d'explication. Avec la transformation que nous avons proposée, il est intéressant de voir si ces définitions sont compatibles ou non.

Proposition 1. *Soit $AF = (A, R)$ un système abstrait d'argumentation, dont le graphe est supposé acyclique. Soit $a^* \in A$ tel qu'il existe un ensemble admissible dont il est le sujet. Soit S une explication compacte de a^*. Soit $M = (\mathcal{U}, \mathcal{V}, \mathcal{F})$ le graphe causal argumentatif issu de la transformation décrite ci-dessus.*
On définit :

- $\varphi = (X_{a^*} = 1)$,
- $X_{arg} = S \setminus \{a^*\}$ *et* $\mathbf{X} = \{X_a \mid a \in X_{arg}\}$,
- $\mathcal{K}$ *l'ensemble des contextes considérés comme possibles par l'*explainee*. On fait l'hypothèse que le contexte par défaut* $\mathbf{u}^* \in \mathcal{K}$.

Alors $\mathbf{X} = \mathbf{1}$ est une explication, au sens causal, non minimale de φ relative à $\mathcal{K}$, c'est-à-dire $\mathbf{X} = \mathbf{1}$ vérifie **EX1** *et* **EX3** *dans $\mathcal{K}$.*

Dans cette proposition, nous avons réintroduit la notion de destinataire de l'explication. En effet, $\mathcal{K}$ représente l'ensemble des contextes considérés par l'*explainee*. Nous im-

posons seulement que $\mathbf{u}^*$ est inclus dans $\mathcal{K}$. Cette hypothèse semble raisonnable car il s'agit de l'unique contexte considéré lorsque l'on travaille d'un point de vue purement argumentatif.

Démonstration. Montrons que $\mathbf{X} = \mathbf{1}$ vérifie **EX1** et **EX3** dans $\mathcal{K}$.
(EX3) Montrons d'abord que u^* est inclus dans $\mathcal{K}_{(\mathbf{X}=\mathbf{1})\wedge\varphi}$.
Par hypothèse, $\mathbf{u}^*$ est inclus dans $\mathcal{K}$.
(i) Montrons par l'absurde que $\mathbf{u}^*$ est inclus dans $\mathcal{K}_{(\mathbf{X}=\mathbf{1})}$.
Supposons que $(M, \mathbf{u}^*) \models \neg(\mathbf{X} = \mathbf{1})$. Alors, $\exists X_a \in \mathbf{X}$ tel que $X_a = 0$. Or $Att_a \neq \emptyset$, donc comme $F_{X_a} = (\bigwedge_{b\in Att_a} \neg X_b), \exists b \in Att_{X_a}$ tel que $X_b = 1$.
Or S est admissible donc $\exists c \in Att_b \cap S$.
Si $Att_c = \emptyset$ alors $X_c = 1$ par définition de u^*. C'est impossible car $X_b = 1$. On arrive donc à une contradiction.
Sinon, comme $X_b = 1$ alors $X_c = 0$. Or $Att_c \neq \emptyset$, donc comme $F_{X_c} = (\bigwedge_{d\in Att_c} \neg X_d), \exists d \in Att_{X_c}$ tel que $X_d = 1$.
Avec S admissible, $\exists e \in Att_d \cap S$.
Si $Att_e = \emptyset$ alors $X_e = 1$ par définition de $\mathbf{u}^*$. C'est impossible car $X_d = 1$. On arrive donc à une contradiction.
Sinon, on peut encore une fois répéter ce raisonnement jusqu'à ce que $Att_e = \emptyset$ car le graphe est fini et acyclique.
Donc $\mathbf{u}^*$ est inclus dans $\mathcal{K}_{\mathbf{X}=\mathbf{1}}$.
(ii) Montrons maintenant que $\mathbf{u}^*$ est inclus dans $\mathcal{K}_{X_{a^*}}$.
Comme S est admissible de sujet a^* et que le graphe est acyclique, alors $\forall b \in Att_{a^*}, \exists c \in X \cap Att_b$. Or $\mathbf{X} = \mathbf{1}$, donc $X_c = 1$ et de fait $X_b = 0$. Ainsi, $\forall b \in Att_{a^*}, X_b = 0$ et donc $X_{a^*} = \bigwedge_{b\in Att_{a^*}} \neg 0 = 1$.
Donc $\mathbf{u}^*$ est inclus dans $\mathcal{K}_{X_{a^*}}$.
Ainsi, $\mathbf{u}^*$ est inclus dans $\mathcal{K}_{(\mathbf{X}=\mathbf{1})\wedge\varphi}$, et donc cet ensemble est non vide et **EX3** est satisfait.
(**EX1**) Montrons maintenant que $\mathbf{X} = \mathbf{1}$ est une cause suffisante dans $\mathcal{K}$, c'est-à-dire qu'il vérifie **SC1**, **SC2** et **SC3**, pour tout $\mathbf{u} \in \mathcal{K}_{(\mathbf{X}=\mathbf{1})\wedge\varphi}$.
SC4 est une condition de minimalité qui porte sur la cause suffisante mais qui dans le cas des explications est équivalente à **EX2** [8]. Pour cette raison, on ne démontre pas que $\mathbf{X} = \mathbf{1}$ vérifie **SC4**.
Soit $\mathbf{u} \in \mathcal{K}_{(\mathbf{X}=\mathbf{1})\wedge\varphi}$.
1) SC1 est vérifié par définition de $\mathbf{u}$.
2) Montrons par l'absurde que **SC3** est vérifié. Soit $\mathbf{u}'$ un contexte tel que $(M, \mathbf{u}') \models [\mathbf{X} = \mathbf{1}]\neg\varphi$.
Comme on a $\neg\varphi$ (c'est-à-dire $X_{a^*} = 0$), d'après $\mathcal{F}$ pour les arguments attaqués (a^* est un sujet de S et donc $Att_{a^*} \neq \emptyset$) $\exists X_b \in \mathcal{V}$, tel que $b \in Att_{a^*}$ et $X_b = 1$.
Or S est admissible donc $\exists c \in S$ tel que $R(c, b)$. De plus, le graphe est acyclique donc $c \neq a^*$ et donc $c \in X$. Comme $\mathbf{X} = \mathbf{1}$, on a en particulier $X_c = 1$ et donc $X_b = 0$ d'après F_{X_b}. On arrive à une *contradiction*.
3) Enfin montrons que **SC2** est bien vérifié :

- **(i)** On construit d'abord une cause effective de φ dans $\mathbf{u}$.
- **(ii)** On montre ensuite que cet ensemble contient bien au moins un élément de $\mathbf{X}$.

(i) Soit $b \in Att_{a^*}$, posons
$\mathbf{Z_b} = \bigcup_{c\in Att_b} \{X_c \mid (M, \mathbf{u}) \models (X_c = 1)\}$.
Comme $\mathbf{u} \in \mathcal{K}_{(\mathbf{X}=\mathbf{1})\wedge\varphi}$, alors $X_{a^*} = 1$ et donc $X_b = 0$.
Or S est admissible donc en particulier, $\forall\alpha \in S, \forall\beta \in Att_\alpha, \exists\gamma \in S \cap Att_\beta$.
Comme $b \in Att_{a^*}$ et $a^* \in S$, $Att_b \neq \emptyset$. De plus, $X_b = 0$ donc $\exists X_c \in Att_b$ tel que $X_c = 1$. Ainsi, $X_c \in Z_b$ donc $\mathbf{Z_b}$ est non vide.
Posons ensuite $\mathbf{Z} = \bigcup_{b\in Att_a} \mathbf{Z_b}$.
$\mathbf{Z}$ n'est pas le candidat pour être une cause effective. Toutefois, montrons qu'il vérifie **AC1** et **AC2** :

- **AC1** est vérifié par construction de $\mathbf{Z_b}$.
- Par construction de $\mathbf{Z}$, si on impose $\mathbf{Z} = \mathbf{0}$, alors on a $\forall b \in Att_{a^*}, \forall c \in Att_b, X_c = 0$. Or $F_{X_b} = \bigwedge_{c\in Att_b} \neg X_c = \bigwedge_{b\in Att_a} \neg 0 = 1$.
  On a donc bien $(M, \mathbf{u}) \models [\mathbf{Z} = \mathbf{0}]\neg\varphi$ et donc **AC2** est vérifié avec $W = \emptyset$.

Notons $\mathbf{Z^m}$ un sous-ensemble minimal de $\mathbf{Z}$ tel que $(\mathbf{Z^m} = \mathbf{1})$ vérifie **AC1** et **AC2**. Il est bien défini et est non vide car $(\mathbf{Z} = \mathbf{1})$ vérifie **AC1** et **AC2**. De plus, $\mathbf{Z^m}$ vérifie **AC3** par définition. On a donc construit $\mathbf{Z^m}$ tel que $(\mathbf{Z^m} = \mathbf{1})$ vérifie **AC1**, **AC2** et **AC3** c'est-à-dire que $(\mathbf{Z^m} = \mathbf{1})$ est une cause effective de φ.
(ii) Prouvons maintenant que l'on peut construire une cause effective $(\mathbf{Z}' = \mathbf{z}')$ de φ telle que $\mathbf{Z}' \cap \mathbf{X} \neq \emptyset$.
Si $\mathbf{Z^m} \cap \mathbf{X} \neq \emptyset$ alors $\mathbf{Z}' = \mathbf{Z^m}$ convient.
Sinon, c'est-à-dire si $\mathbf{Z^m} \cap \mathbf{X} = \emptyset$, soit $b \in Att_{a^*}$:

- $\exists X_c \in \mathbf{Z^m}$ tel que $c \in Att_b$, $(M, \mathbf{u}) \models (X_c = 1)$ et $X_c \notin \mathbf{X}$.
- Comme S est admissible et le graphe est acyclique, $\exists X_{c'} \in \mathbf{X}$ tel que $c' \in Att_b$. De plus, $\mathbf{Z^m} \cap \mathbf{X} = \emptyset$, donc $X_{c'} \notin \mathbf{Z^m}$. Enfin, comme $\mathbf{u} \in \mathcal{K}$ on a $(M, \mathbf{u}) \models (X_{c'} = 1)$.

Posons $\mathbf{Z^{m\prime}} = (\mathbf{Z^m} \setminus \{X_c\}) \cup \{X_{c'}\}$. $\mathbf{Z^{m\prime}}$ vérifie aussi **AC1** et **AC2**. Comme $\mathbf{Z^m}$ est minimal par construction, alors si $\mathbf{Z^{m\prime}}$ ne l'est pas, $\exists \mathbf{Z}' \subseteq \mathbf{Z^{m\prime}}$ tel que $\mathbf{Z}' \nsubseteq \mathbf{Z^m}$. Or $\mathbf{Z^{m\prime}} \setminus \mathbf{Z^m} = \{X_{c'}\}$ donc $X_{c'} \in \mathbf{Z}'$ et on a donc $\mathbf{Z}' \cap \mathbf{X} \neq \emptyset$.
On a donc construit un ensemble $\mathbf{Z}'$ vérifiant **AC1** et **AC2**, minimal pour l'inclusion (**AC3**) et tel que $\mathbf{Z}' \cap \mathbf{X} \neq \emptyset$ on a donc vérifié **SC2**.
On a montré que $\mathbf{X} = \mathbf{1}$ vérifie **EX1** et **EX3**, donc $\mathbf{X} = \mathbf{1}$ est une explication (au sens causal) non minimale de φ. □

Exemple 2. *(suite) – Reprenons l'exemple 2 et sa transformation associée présentée en figure 3.*
(i) *Posons* $a^* = a$, $S = \{a, d, f\}$, $\mathbf{X} = \{X_d, X_f\}$ *et* $\varphi = (X_a = 1)$.
On a $\mathbf{u}^* = (\tilde{X_d} = 1, \tilde{X_e} = 1, \tilde{X_f} = 1)$. *Donc* $(M, \mathbf{u}^*) \models (X_d = 1, X_e = 1, X_f = 1)$. *En particulier, on a bien* $(M, \mathbf{u}^*) \models (\mathbf{X} = \mathbf{1})$.
On a également $X_b = 0$ *et* $X_c = 0$ *et enfin* $X_a = 1$. *On a bien* $(M, \mathbf{u}^*) \models \varphi$.
On a donc bien $\mathbf{u}^* \in \mathcal{K}_{(\mathbf{X}=\mathbf{1})\wedge\varphi}$ *et en particulier* ***EX3*** *est bien vérifié.*

(ii) *Soit $\mathcal{K}$ un ensemble de contextes, et $\mathbf{u} \in \mathcal{K}_{(\mathbf{X}=\mathbf{1})\wedge\varphi}$ ($\mathcal{K}_{(\mathbf{X}=\mathbf{1})\wedge\varphi}$ est non vide car $\mathbf{u}^* \in \mathcal{K}_{(\mathbf{X}=\mathbf{1})\wedge\varphi}$).*
On a tout d'abord $(M,\mathbf{u}) \models (\mathbf{X}=\mathbf{1})\wedge\varphi$ car $\mathbf{u} \in \mathcal{K}_{(\mathbf{X}=\mathbf{1})\wedge\varphi}$.
Soit $\mathbf{u}' \in \mathcal{K}$. On a $(M,\mathbf{u}') \models [\mathbf{X}=\mathbf{1}](X_b=0 \wedge X_c=0)$ et donc $(M,\mathbf{u}') \models [\mathbf{X}=\mathbf{1}](X_a=1)$.
Posons $\mathbf{Y}=\{X_e\}$ et $X=\{X_d\}$. Avec $X=0\wedge\mathbf{Y}=\mathbf{0}$, on a $X_b=1$ et donc $X_a=0$. Ainsi, on a $(M,\mathbf{u}) \models [X=0\wedge\mathbf{Y}=\mathbf{0}]\neg\varphi$.
*On a montré que $\forall \mathbf{u}\in\mathcal{K}$, $\mathbf{X}=\mathbf{1}$ est une cause suffisante de φ, c'est-à-dire que **EX1** est bien vérifié.*

5 Passage des GCA aux AAF

Dans cette section, nous proposons la transformation réciproque ainsi qu'une démonstration de l'équivalence entre ces deux cadres formels.

5.1 Transformation proposée

On considère un graphe causal argumentatif de triplet $M=(\mathcal{U},\mathcal{V},\mathcal{F})$. On adopte la transformation suivante :

— $A'=\{a \mid X_a \in \mathcal{V}\}$,

— Pour la construction de la relation d'attaque : soit $(X_a,X_b)\in\mathcal{V}^2$, on pose $\mathbf{Y}=\mathcal{V}\setminus\{X_a,X_b\}$. Si on a $(M,\mathbf{u}) \models [X_b=1,\mathbf{Y}=\mathbf{0}](X_a=0)$ alors $(b,a)\in R'$.

Exemple 2. *(suite) – Reprenons le graphe causal argumentatif présenté en figure 3.*
On a $\mathcal{V}=\{X_a,X_b,X_c,X_d,X_e,X_f\}$. On pose donc $A'=\{a,b,c,d,e,f\}$.
Soit $\mathcal{K}$ un ensemble de contextes. Soit $\mathbf{u}\in\mathcal{K}$.
On a $F_{X_a}=\neg X_b\wedge\neg X_c$. En particulier, $(M,\mathbf{u}) \models [X_v=1](X_a=0)$ avec $X_v\in\{X_b,X_c\}$.
Cela reste vrai en imposant en plus $\mathbf{Y}=\mathbf{0}$ avec $\mathbf{Y}$ construit comme dans la transformation. Donc $(b,a)\in R'$ et $(c,a)\in R'$.
En appliquant le même raisonnement avec toutes les équations structurelles de $\mathcal{F}$ on a $\{(d,b),(e,b),(e,c),(f,c)\}\in R'$.
Posons $\mathbf{Y}=\mathcal{V}\setminus\{X_a,X_v\}$ avec $v\in\{d,e,f\}$.
On a $(M,\mathbf{u}) \models [X_v=1,\mathbf{Y}=0](X_a=1)$. En effet, toutes les équations structurelles ont été remplacées par $F_X=0$ sauf pour X_a et X_v : $X_v=1$ et $F_{X_a}=\neg X_b\wedge\neg X_c$ donc $X_a=\neg 0\wedge\neg 0=1$.
Donc $(v,a)\notin R'$.
On a donc $R'=\{(b,a),(c,a),(d,b),(e,b),(e,c),(f,c)\}$.

5.2 Équivalence entre AAF et GCA

Proposition 2. *Soit $AF=(A,R)$ un système abstrait d'argumentation, $M=(\mathcal{U},\mathcal{V},\mathcal{F})$ le graphe causal argumentatif associé à AF par la transformation décrite dans la section 4 et $AF'=(A',R')$ le système d'argumentation associé à M avec la transformation ci-dessus. Alors :*

$$AF=AF'.$$

Démonstration. Soit $AF=(A,R)$ un système abstrait d'argumentation, $M=(\mathcal{U},\mathcal{V},\mathcal{F})$ un graphe causal argumentatif associé à AF et $AF'=(A',R')$ le système d'argumentation associé à M.
Montrons que $AF=AF'$ c'est-à-dire $A=A'$ et $R=R'$.
• On a par construction $A=A'$.
• Montrons que $R=R'$ par double inclusion.

1. Soit $(a,b)\in A^2$ tel que $R(b,a)$.
   On a par définition $X_a=\neg X_b\wedge(\bigwedge_{c\in Att_a\wedge c\neq b}\neg X_c)$.
   En particulier, si $X_b=1$ alors on a :
   $$X_a=0\wedge(\bigwedge_{c\in Att_a\wedge c\neq b}\neg X_c)=0.$$
   Ainsi, pour tout contexte $\mathbf{u}$, on a $(M,\mathbf{u}) \models [X_b=1,\mathbf{Y}=\mathbf{0}](X_a=0)$ avec $\mathbf{Y}=\mathcal{V}\setminus\{X_a,X_b\}$, donc $(b,a)\in R'$ et $R\subseteq R'$.
2. Soit $(a',b')\in A'^2$ tel que $b'\in Att^{R'}_{a'}$.
   Soit $\mathbf{Y}=\mathcal{V}\setminus\{X_{a'},X_{b'}\}$. On a par définition
   $$(M,\mathbf{u}) \models [X_{b'}=1,\mathbf{Y}=\mathbf{0}](X_{a'}=0)$$
   Or $A=A'$, donc $\forall\alpha\in A$, $X_\alpha=X_{\alpha'}$. En particulier $(M,\mathbf{u}) \models [X_{b'}=1,\mathbf{Y}=\mathbf{0}](X_a=0)$ et donc $Att^R_a\neq\emptyset$.
   De plus, $F_{X_a}=\bigwedge_{z\in Att^R_a}\neg X_z$. Si $b'\notin Att^R_a$ alors avec $\mathbf{Y}=\mathbf{0}$ on a $F_{X_{a'}}=\bigwedge_{\beta\in Att^R_a}\neg 0=1$, ce qui contredit l'hypothèse.
   On a donc $b'\in Att^R_a$ et $R'\subseteq R$.

On a donc montré par double inclusion que $R=R'$. □

6 Conclusion et perspectives

Nous avons mis en évidence dans cet article l'équivalence qui existe entre les graphes causaux argumentatifs et les systèmes abstraits d'argumentation. Nous avons également proposé une transformation permettant de passer de l'un à l'autre. Cela permet de pouvoir utiliser le meilleur des deux mondes.
D'une part, la notion de contexte présente dans les modèles structurels causaux permet de faire varier les valeurs des arguments et offre donc un cadre dynamique. De plus, elle permet de tenir compte des connaissances des agents. Les travaux de J. Pearl et J. Halpern [10] introduisent également la notion de pouvoir explicatif et d'explication partielle, ainsi qu'une définition générale permettant en plus de donner une connaissance du modèle à l'*explainee*. D'autre part, les systèmes d'argumentation proposent un cadre plus naturel pour modéliser les situations d'interaction, pouvant faciliter sa mise en pratique pour des systèmes en interactions avec des humains. Ainsi, une démarche pourrait être de modéliser dynamiquement une interaction avec un AAF, de calculer un résultat ou une action puis d'effectuer la transformation en GCA afin de générer des explications aux propriétés voulues.

Toutefois, se limiter à la notion d'attaque entre arguments peut conduire à des situations un peu étranges dans lesquelles deux arguments n'interagissent pas entre eux alors qu'ils semblent clairement liés. Une première solution consiste à prendre la négation d'un de ces arguments. Une autre, beaucoup moins maladroite, est d'ajouter une relation binaire supplémentaire comme par exemple la relation de support [4]. Il serait alors intéressant de pouvoir intégrer ce genre de relation dans les équations structurelles des GCA. Cela nécessite de choisir un critère de décision en cas d'attaque et de support [4] : si un argument est attaqué par un argument non attaqué et supporté par un argument non attaqué, l'argument est-il accepté, non accepté, indéterminé ?

Les travaux de G. Brewka et al. [3] proposent une généralisation des AAF de P.M. Dung [6] appelée *Abstract Dialectical Framework* (ADF). Ce cadre formel remplace les relations d'attaques par des conditions d'acceptabilité des arguments, souvent sous la forme de formules logiques. On obtient donc une représentation avec des variables à valeurs binaires ou ternaires (arguments acceptés, refusés, ou indécis) dont la valeur est régie par des formules logiques. Cela fait évidemment écho à notre transformation des AAF en GCA mais plus généralement aux modèles structurels causaux de J. Halpern et J. Pearl [9]. Il serait donc intéressant d'explorer ce que chacune des approches propose et ce qu'il est possible de faire dans la continuation de ce que nous proposons avec les GCA.

Il existe également une formulation floue pour ces deux cadres [1, 11] qui apporte des outils permettant une représentation plus humaine des interactions, avec par exemple l'ajout d'un degré d'attaque et de support. L'étude des ces cadres et leur comportement par rapport à la transformation que nous proposons est une piste à venir pour étendre notre approche.

Enfin, l'objectif de tels cadres est de proposer des explications adaptées aux humains afin d'augmenter la confiance de ces derniers envers les systèmes d'IA mais également de faciliter les interactions entre humain et machine. Ainsi, un autre enjeu des travaux à venir consiste à tester ces cadres formels et la transformation proposée sur un exemple plus complet et complexe d'interaction entre humain et machine puis de faire évaluer subjectivement ces modèles par des utilisateurs humains.

Références

[1] Isabelle Bloch and Marie-Jeanne Lesot. Vers une formulation floue des explications par contraste. In *Rencontres Francophones sur la Logique Floue et ses Applications (LFA)*, 2021.

[2] AnneMarie Borg and Floris Bex. A basic framework for explanations in argumentation. *IEEE Intelligent Systems*, 36(2) :25–35, 2021.

[3] Gerhard Brewka, Stefan Ellmauthaler, Hannes Strass, Johannes P Wallner, and Stefan Woltran. Abstract dialectical frameworks. an overview. *IfCoLog Journal of Logics and their Applications*, 4(8) :2263–2317, 2017.

[4] Claudette Cayrol and Marie-Christine Lagasquie-Schiex. On the Acceptability of Arguments in Bipolar Argumentation Frameworks. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 378–389. Springer, 2005.

[5] Christophe Denis and Franck Varenne. Interprétabilité et explicabilité pour l'apprentissage machine : entre modèles descriptifs, modèles prédictifs et modèles causaux. Une nécessaire clarification épistémologique. In *Conférence Nationale en Intelligence Artificielle (CNIA)*, pages 60–68, 2019.

[6] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77 :321–357, 1995.

[7] Xiuyi Fan and Francesca Toni. On Computing Explanations in Abstract Argumentation. In *ECAI*, volume 263, pages 1005–1006, 2014.

[8] Joseph Y. Halpern. *Actual Causality*. MIT Press, 2016.

[9] Joseph Y. Halpern and Judea Pearl. Causes and Explanations : A Structural-Model Approach. Part I : Causes. *The British Journal for the Philosophy of Science*, 56(4) :843–887, 2005.

[10] Joseph Y. Halpern and Judea Pearl. Causes and Explanations : A Structural-Model Approach. Part II : Explanations. *The British Journal for the Philosophy of Science*, 56(4) :889–911, 2005.

[11] Jeroen Janssen, Martine De Cock, and Dirk Vermeir. Fuzzy argumentation frameworks. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU)*, pages 513–520, 2008.

[12] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. Explainable Reinforcement Learning Through a Causal Lens. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

[13] Tim Miller. Explanation in Artificial Intelligence : Insights from the Social Sciences. *Artificial Intelligence*, 267 :1–38, 2019.

[14] Tim Miller. Contrastive explanation : A structural-model approach. *Knowledge Engineering Review*, 36 :E14, 2021.

[15] Sanjay Modgil and Henry Prakken. The ASPIC+ framework for structured argumentation : a tutorial. *Argument & Computation*, 5(1) :31–62, 2014.

[16] Laurie Ann Paul and Ned Hall. *Causation : A User's Guide*. Oxford University Press, 2013.

[17] Kristijonas Čyras, Antonio Rago, Emanuele Albini, Pietro Baroni, and Francesca Toni. Argumentative XAI : A survey. In Zhi-Hua Zhou, editor, *Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4392–4399, 2021.