

# Argumentation and Causal Models in Human-Machine Interaction: A Round Trip

Yann Munro<sup>1,\*†</sup>, Isabelle Bloch<sup>1,†</sup>, Mohamed Chetouani<sup>2,†</sup>, Marie-Jeanne Lesot<sup>1,†</sup> and Catherine Pelachaud<sup>3,†</sup>

<sup>1</sup>Sorbonne Université, CNRS, LIP6, Paris, France

<sup>2</sup>Sorbonne Université, CNRS, ISIR, Paris, France

<sup>3</sup>CNRS, Sorbonne Université, ISIR, Paris, France

## Abstract

In the field of explainable artificial intelligence (XAI), causal models and abstract argumentation frameworks constitute two formal approaches that provide definitions of the notion of explanation. These symbolic approaches rely on logical formalisms to reason by abduction or to search for causalities, from the formal modeling of a problem or a situation. They are designed to satisfy properties that have been established as necessary based on the study of human-human explanations. As a consequence they appear to be particularly interesting for human-machine interactions as well. In this paper, we show the equivalence between a particular type of causal models, that we call argumentative causal graphs (ACG), and abstract argumentation frameworks. We also propose a transformation between these two systems and look at how one definition of an explanation in the argumentation theory is transposed when moving to ACG. To illustrate our proposition, we use a very simplified version of a screening agent for COVID-19.

## Keywords

Causal models, Abstract argumentation frameworks, eXplainable Artificial Intelligence (XAI),

## 1. Introduction

In human-machine interaction, explainability is a very important property that helps improving the performance of the human-agent pair [1] as it increases the trust and the understanding that humans have in artificial intelligence systems. Many methods have been developed to contribute to the interpretability and explainability of artificial intelligence systems (XAI) and especially in this field of human-machine interaction [2]. Among them, symbolic approaches

---

AIC 2022, 8th International Workshop on Artificial Intelligence and Cognition, June 15–17, 2022, Örebro, Sweden

\*Corresponding author.

†These authors contributed equally.

✉ yann.munro@lip6.fr (Y. Munro); isabelle.bloch@sorbonne-universite.fr (I. Bloch);

mohamed.chetouani@sorbonne-universite.fr (M. Chetouani); marie-jeanne.lesot@lip6.fr (M. Lesot);

catherine.pelachaud@sorbonne-universite.fr (C. Pelachaud)

🌐 <https://lip6.fr/Yann.Munro> (Y. Munro); <https://webia.lip6.fr/~bloch/chaireIA.html> (I. Bloch);


<https://www.isir.upmc.fr/personnel/chetouani/> (M. Chetouani); <https://webia.lip6.fr/~lesot/> (M. Lesot);

<http://pages.isir.upmc.fr/~pelachaud/> (C. Pelachaud)

🆔 0000-0002-9155-6180 (Y. Munro); 0000-0002-6984-1532 (I. Bloch); 0000-0002-2920-4539 (M. Chetouani);

0000-0002-3604-6647 (M. Lesot); 0000-0003-1008-0799 (C. Pelachaud)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

rely on logical formalisms to reason by abduction or to search for causalities, from the formal modeling of a problem or a situation. It is in this type of approach that we are interested in in this paper.

To improve the way human-machine interactions are modeled in these frameworks, one way consists in drawing inspiration from human cognitive and social mechanisms, in particular the ones related to the explanation process, and deriving desirable properties and behaviors from them. In [3], Tim Miller, drawing on work in social and cognitive sciences, identifies essential characteristics that are needed when developing explainable artificial intelligence methods.

A first formal framework is based on the work of Joseph Halpern and Judea Pearl [4] on causality and in particular on causal models. This notion of causality is closely related to that of explanation. Indeed, explaining a fact is often associated with providing a cause, and therefore a definition of an explanation can be found in their work [5]. This framework was for instance implemented to generate explanations for an agent playing Starcraft II, a real-time strategy game [6]. This paper focuses on a special case of such models defined in Section 2.1, which we propose to call argumentative causal graphs (ACG).

Another framework proposing a definition of the notion of explanation is that of argumentation. Introduced by Phan Minh Dung in 1995 [7], the abstract argumentation framework (AAF) allows modeling the interactions between arguments coming from several entities or agents. Many XAI methods have been developed in this framework, see [8] for a survey.

After briefly presenting these two frameworks in Sections 2 and 3, we establish an equivalence between them through a transformation allowing us to go from argumentative graphs to argumentative causal graphs and vice versa (Sections 4 and 5, respectively). To the best of our knowledge, there is no work in this direction, which is why we propose transformations to link the two fields, which is the main contribution of the paper. The objective is not to present a new method nor a new framework, but instead to propose a method to move from one to the other and thus to allow for the exploitation of the interesting properties of each framework in the other one.

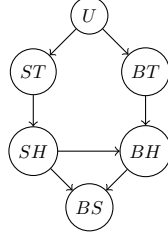
The paper illustrates the proposed principles on an example inspired by medical regulation assistants in the context of the health situation related to COVID-19. This is obviously a very simplified model of reality whose sole purpose is to illustrate our contributions and which is not intended to replace existing health instructions.

## 2. Causal Models

This section recalls the concepts of causal models defined by J. Halpern [9], a framework that leads to a definition of the notion of explanation. Including causality in the process of an explanation echoes some properties highlighted by works on explanation in social and cognitive sciences, as e.g. reported by T.Miller [3].

### 2.1. Definition

A causal model as introduced by J. Halpern [9] is a triplet  $M = (\mathcal{U}, \mathcal{V}, \mathcal{F})$  where  $\mathcal{U}$  is a set of exogenous variables, i.e. a set of variables whose values are independent of the model;  $\mathcal{V}$  is a set of endogenous variables;  $\mathcal{F}$  is a set of structural equations, one for each variable of



**Figure 1:** Causal graph associated with Example 1, inspired from [10].

$\mathcal{V}$ . They associate a value to each of the endogenous variables according to the values of the other variables. By associating each variable with a node and by drawing edges between these nodes to indicate the functional dependencies described by  $\mathcal{F}$ , the structural model  $M$  can be represented as a graph.

The equivalence discussed in Sections 4 and 5 focuses on a particular case of causal models which we propose to call Argumentative Causal Graphs (ACGs). These are triplets  $M = (\mathcal{U}, \mathcal{V}, \mathcal{F})$  for which (i) the variables are Boolean variables, and the structural equations are therefore written as logical formulas; (ii) these formulas do not contain disjunctions; (iii) the related graph is acyclic.

Let us introduce some additional useful notations:

Let  $M = (\mathcal{U}, \mathcal{V}, \mathcal{F})$  be a causal model. A **context**, written  $\mathbf{u}$ , is an assignment of the variables of  $\mathcal{U}$ . The pair  $(M, \mathbf{u})$  is called a **world**. Throughout the paper  $\mathcal{K}$  denotes a set of contexts.

Let  $\mathbf{X}$  be a set of variables of  $\mathcal{V}$ ,  $\mathbf{X} = \mathbf{x}$  denotes an assignment of the variables of  $\mathbf{X}$  with the values of  $\mathbf{x}$ .

Let  $\mathbf{u} \in \mathcal{K}$ ,  $(M, \mathbf{u}) \models \mathbf{X} = \mathbf{x}$  holds if  $\mathbf{X} = \mathbf{x}$  is the unique solution to the structural equations of  $\mathcal{F}$  in  $\mathbf{u}$ .

Let  $\mathbf{X} \in \mathcal{V}$  and  $\mathbf{x}$  be values of  $\mathbf{X}$ . We denote by  $\mathcal{K}_{\mathbf{X}=\mathbf{x}}$  the set of contexts  $\mathbf{u}'$  of  $\mathcal{K}$  such that  $(M, \mathbf{u}') \models \mathbf{X} = \mathbf{x}$ .

Let  $\mathbf{u} \in \mathcal{K}$ ,  $(M, \mathbf{u}) \models [\mathbf{X} = \mathbf{x}](\mathbf{Y} = \mathbf{y})$  holds if, in the world  $(M', \mathbf{u})$  defined as  $(M, \mathbf{u})$  in which the structural equations of  $\mathcal{F}$  determining the variables of  $\mathbf{X}$  are replaced by the assignment  $\mathbf{X} = \mathbf{x}$ , it holds that  $(M', \mathbf{u}) \models \mathbf{Y} = \mathbf{y}$ .

**Example 1.** (from [10]) – Suzy and Billy both throw a rock at a glass bottle. They are both perfectly accurate and are therefore sure to hit the bottle if they actually throw the rock. If either rock hits the bottle, it shatters. Suzy’s stone always hits first. To model the situation, the following variables are introduced:  $ST$  (respectively  $BT$ ) and  $SH$  (resp.  $BH$ ) represent “Suzy (resp. Billy) throws” and “Suzy (resp. Billy) hits”. Finally,  $BS$  refers to “the bottle shatters”. The set  $\mathcal{U}$  contains a unique variable summarizing the exogenous variables that represent factors outside the problem that influence whether Billy or Suzy throws the rock. The functions of  $\mathcal{F}$  complete the modeling of the problem. For example, the fact that Billy touches the bottle in the case (and only in this case) where he has thrown a stone and Suzy has not touched the bottle is represented as:  $BH = BT \wedge \neg SH$ . This results in the following causal model, illustrated in Figure 1:  $\mathcal{U} = \{U\}$ ;  $\mathcal{V} = \{ST, BT, SH, BH, BS\}$ ;  $\mathcal{F} = \{(SH = ST), (BH = BT \wedge \neg SH), (BS = SH \vee BH)\}$ .

## 2.2. Actual Cause

In this formalism, J. Halpern [9] then proposes to define the notion of cause as follows: given a formula  $\varphi$ , the assignment  $\mathbf{X} = \mathbf{x}$  is an **actual cause** of  $\varphi$  in the world  $(M, \mathbf{u})$  if the three following conditions are verified:

**AC1**  $(M, \mathbf{u}) \models (\mathbf{X} = \mathbf{x}) \wedge \varphi$ , i.e. both the cause and the consequence are true in the considered world.

**AC2** There exists a set  $\mathbf{W}$  of endogenous variables with values  $\mathbf{w}$  and a setting  $\mathbf{x}'$  for the variable  $\mathbf{X}$  such that if  $(M, \mathbf{u}) \models (\mathbf{W} = \mathbf{w})$  then

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

**AC3**  $\mathbf{X}$  is minimal: there is no strict subset of  $\mathbf{X}$  that verifies **AC1** and **AC2**. This condition aims to avoid having useless variables in the cause.

Condition **AC2** formalizes a counterfactual reasoning and checks whether, if the presumed cause  $\mathbf{X} = \mathbf{x}$  had not occurred (i.e. if  $\mathbf{X}$  had values  $\mathbf{x}' \neq \mathbf{x}$ ) **and** possibly other events had occurred (i.e.  $\mathbf{W} = \mathbf{w}$ ), the consequence would still occur.

**Example 1.** (continued) – Intuitively, one cause of the bottle shattering is the fact that Suzy threw the stone. Indeed, it was her stone that hit the bottle and thus broke it. However, if we ask the question: if Suzy had not thrown her rock, would the bottle have broken? The answer is ‘yes’ because Billy would have hit the bottle ( $BH = BT \wedge \neg SH$ ). Therefore the following counterfactual must be considered: if Suzy had not thrown her rock **knowing that** Billy did not hit the bottle, would the bottle have shattered? In this case, the answer is ‘no’, i.e. the fact that Suzy threw her stone is indeed an actual cause of the bottle shattering.

## 2.3. Sufficient Cause

Let  $\mathcal{K}$  be a set of contexts and  $\mathbf{u} \in \mathcal{K}$ . The assignment  $\mathbf{X} = \mathbf{x}$  is a **sufficient cause** of  $\varphi$  in the world  $(M, \mathbf{u})$  if the following four conditions are satisfied [9]:

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

**SC2** There exist a part of  $\mathbf{X}$ ,  $X = x$ , and another conjunction  $(\mathbf{Y} = \mathbf{y})$  (possibly empty) such that  $(X = x) \wedge (\mathbf{Y} = \mathbf{y})$  is an effective cause of  $\varphi$  in  $(M, \mathbf{u})$ , i.e. some part of  $\mathbf{X}$  is part of an actual cause in the considered world.

**SC3**  $(M, \mathbf{u}') \models [\mathbf{X} = \mathbf{x}] \varphi$  for all contexts  $\mathbf{u}' \in \mathcal{K}$ , i.e. if  $\mathbf{X} = \mathbf{x}$  then  $\varphi$  holds regardless of the context.

**SC4**  $\mathbf{X}$  is a minimal set that satisfies **SC1**, **SC2** and **SC3**.

**Remark 1.** There exists another version of the definition of sufficient cause proposed by T.Miller in [11], as an actual non-minimal cause, i.e. one that verifies only **AC1** and **AC2**. The major difference is in **SC3**. T.Miller’s view focuses only on the current context, in contrast to Halpern’s, who defines a sufficient cause over a set of given contexts. We choose here to consider Halpern’s definition because, among others, by weakening **SC3** a notion of explanatory power can be defined, which can be useful for comparing the generated explanations.

## 2.4. Explanation

When providing an explanation, it is important to consider the person to whom the explanation is dedicated. This person is called the explainee. For this reason, the search for actual cause and sufficient cause is constrained to a set of contexts  $\mathcal{K}$  determined by what the explainee considers as possible.

The assignment  $\mathbf{X} = \mathbf{x}$  is an **explanation** of  $\varphi$  relative to the set of contexts  $\mathcal{K}$  if the following three conditions are verified [9]:

**EX1**  $\mathbf{X} = \mathbf{x}$  is a sufficient cause for all contexts  $\mathbf{u}$  in  $\mathcal{K}$  that verify  $(\mathbf{X} = \mathbf{x}) \wedge \varphi$ .

**EX2**  $\mathbf{X}$  is minimal.

**EX3**  $\mathcal{K}_{(\mathbf{X}=\mathbf{x})\wedge\varphi} \neq \emptyset$ , i.e. at least one of the contexts considered as possible by the explainee is compatible with the explanation.

The explanation is said to be non-trivial if it also satisfies:

**EX4**  $(M, \mathbf{u}') \models \neg(\mathbf{X} = \mathbf{x})$  for some context  $\mathbf{u}' \in K_\varphi$ .

The set of contexts  $\mathcal{K}$  is determined by the explainee. Thus, it is possible that there are no sufficient causes in any context of  $\mathcal{K}$  (i.e.  $\mathcal{K}_{(\mathbf{X}=\mathbf{x})\wedge\varphi} = \emptyset$ ) and then, there is no possible explanation.

There exists a more general definition of explanation proposed by J. Halpern [9]. In particular, it addresses the problem mentioned above, taking into account the fact that the explainee does not have a perfect knowledge of the model, and that the explanation must thus bring some additional knowledge. For this purpose, not only an assignment  $\mathbf{X} = \mathbf{x}$  but also more complex assertions define explanations that allow the user to better understand the model: if no sufficient cause exists in the set of contexts  $\mathcal{K}$  considered by the explainee, then returning an additional formula may allow the latter to enlarge the set  $\mathcal{K}$  of possible contexts. That makes the definition richer and more compatible with human-human explanation.

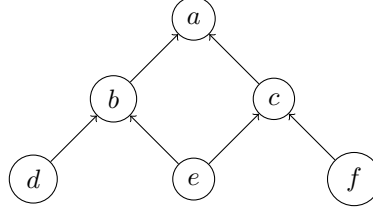
## 3. Abstract Argumentation Frameworks

This section briefly recalls P.M. Dung's [7] principles of Abstract Argumentation Frameworks (AAFs) as well as a definition of explanation [12] for this framework.

### 3.1. Definition

An AAF is a pair  $AF = (A, R)$  such that  $A$  is a finite set of arguments and  $R$  is a binary relation on  $A \times A$ , that is called the attack relation: an argument  $a \in A$  attacks  $b \in A$  if  $(a, b) \in R$ , also denoted  $R(a, b)$ . Since  $R$  is a binary relation with a finite support, an AAF can be represented as a graph.

This formalism does not impose any constraint on the internal structure of an argument, nor on the nature of an attack: an argument can simply be a statement in natural language. It can also be a formula defined in a certain language according to rules, as in the case of the ASPIC+ system [13].



**Figure 2:** Argumentative scenario modeling Example 2

**Example 2.** This example considers a simple scenario of an agent helping to screen for COVID-19. Let us imagine that a user, Billy, wakes up with some aches and pains. He decides to consult the agent. The agent asks a number of questions about his health. Indeed, having aches and pains is not enough to justify going for a PCR test, a self-test could be enough for example. The agent asks Billy to taste a condiment with a strong taste (salt, sugar, vinegar, etc.) to test whether he lost his sense of taste or not. Finally, the agent also checks whether he has been in close contact with someone who has COVID-19.

Their conversation can be represented by the following abstract argumentation framework, illustrated in Figure 2:  $A = \{a: \text{“A PCR test is necessary”}, b: \text{“No symptoms”}, c: \text{“Vaccinated”}, d: \text{“Aches and pains”}, e: \text{“Loss of taste”}, f: \text{“Close contact”}\}$  and  $R = \{(b, a), (c, a), (d, b), (e, b), (e, c), (f, c)\}$ .

In the case where Billy does not feel that he has any particular symptom or is vaccinated, a PCR test is not necessary. This is represented by the first two attack relations  $(b, a)$  and  $(c, a)$ . However, if he has aches and pains or a loss of taste, it is no longer possible to say that he does not have symptoms:  $(d, b), (e, b)$ . In the same way, if he is a close contact or no longer has any taste, the fact that he is vaccinated no longer justifies not going for a PCR test:  $(e, c), (f, c)$ . In particular, being vaccinated does not prevent one from getting COVID-19.

The graph shown in Figure 2 represents the case where Billy has aches and pains, lost taste and is close contact  $(d, e, f)$ . According to this graph,  $a$  is only attacked by non-accepted arguments (because they are attacked by non-attacked arguments) and can therefore be accepted. Thus, a PCR test must be performed.

### 3.2. Additional Definitions

Let  $Att_a^R$  denote **the set of direct attackers** of  $a$  for the attack relation  $R$ :

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

When only one attack relation is defined, the notation is simplified to  $Att_a$ .

A set of arguments  $S$  is **conflict-free** if there is no pair  $(a, b) \in S^2$  such that  $(a, b) \in R$ :  $\forall (a, b) \in S^2, (a, b) \notin R$ .

An argument  $a \in A$  is **acceptable** by a set  $S$  if  $S$  attacks all the attackers of  $a$ :

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

A set of argument  $S$  is said to be **admissible** if  $S$  is conflict-free and any element  $a$  of  $S$  is acceptable by  $S$ :

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

A set of arguments  $S$  is said to be **related admissible** if it is admissible and at least one of its arguments is attacked:

$$S \text{ is admissible and } \exists x \in S \text{ such that } Att_x \neq \emptyset.$$

Such an argument  $x$  is referred to as a **topic** of  $S$ .

**Example 2.** (continued) – Let us look for a related admissible set  $S_{ex}$  with  $a$  as a topic. Since  $a$  is attacked by  $b$ ,  $S_{ex}$  must contain an attacker of  $b$ . Let us take  $d$  for example:  $d$  is not attacked so it is acceptable by  $S_{ex}$ . Besides,  $a$  is also attacked by  $c$ . So we have to add an attacker of  $c$  to  $S_{ex}$ . Let us add for example  $e$ :  $e$  is not attacked, so it is also acceptable by  $S_{ex}$ . Finally, all attackers of  $a$  are attacked by an element of  $S_{ex}$ , so  $a$  is acceptable by  $S_{ex}$ . We have thus constructed  $S_{ex} = \{a, d, e\}$ . Considering all possibilities, the set of (**related**) admissible sets is:

$$S_{adm} = \{\{d\}, \{e\}, \{f\}, \{d, e\}, \{d, f\}, \{e, f\}, \{d, e, f\}, \\ \{a, d, e, f\}, \{a, d, f\}, \{a, e, f\}, \{a, d, e\}, \{a, e\}\}.$$

### 3.3. Explanations

In this AAF, Xiuyi Fan and Francesca Toni [12] propose the following definition for an explanation: let  $x \in A$  be an argument of  $A$ , an **explanation**  $S$  of  $x$  is a related admissible set with  $x$  as a topic. The explanation  $S$  is **compact** if it is minimal for the inclusion relation; it is **verbose** if it is maximal for the inclusion relation.

**Example 2.** (continued) – Argument  $a$  has two compact explanations: “a PCR test is needed” because Billy has “a loss of taste”, i.e.  $\{a, e\}$ , or because he has “aches and pains” and “is a close contact”, i.e.  $\{a, d, f\}$ . There also exists a verbose explanation: “loss of taste”, “aches and pains” and “contact cases”, i.e.  $\{a, d, e, f\}$ .

There are other definitions of explanation for argumentation systems (see e.g. [8] for a survey). However, in most cases, they require additional notions [14] and extend P.M. Dung’s framework [7]. For this reason, we do not consider them in this paper.

**Remark 2.** In the context of abstract argumentation, the objective is not to model the explainee. Instead, an AAF is rather a transcription of an exchange of arguments between several entities. The explanation thus serves to justify why an argument can be accepted by referring to the different arguments that are used to defend it: there is no notion of context. In particular, it is assumed that all the arguments and their interactions are known.

## 4. From AAF to ACG

This section and the following one present the main contribution of the paper, namely the equivalence between argumentative causal graphs (ACG) and abstract argumentation frameworks (AAF). This section presents a transformation of argumentative graphs into ACG. It also discusses how the notion of explanation is transported from AAF to ACG.

### 4.1. Proposed Transformation

Let  $AF = (A, R)$  be an AAF and its associated graph which is assumed to be acyclic.

For each argument  $a \in A$ , a Boolean variable  $X_a$  is created such that  $X_a = 1$  is read as “Argument  $a$  is accepted”. These variables constitute the set of endogenous variables. Moreover, for any unattacked argument  $a \in A$ , another Boolean variable  $\tilde{X}_a$  is created. These variables constitute the set of exogenous variables. Formally, let us define:

- $\mathcal{V} = \{X_a \mid a \in A\}$ ,
- $\mathcal{U} = \{\tilde{X}_a \mid (a \in A) \wedge (Att_a = \emptyset)\}$ ,
- $\mathcal{F} = \{F_{X_a} \mid X_a \in \mathcal{V}\}$  with:
  - ◊  $\forall a \in A$  such that  $Att_a \neq \emptyset$ ,  $F_{X_a} = \bigwedge_{b \in Att_a} \neg X_b$ ,
  - ◊  $\forall a \in A$  such that  $Att_a = \emptyset$ ,  $F_{X_a} = \tilde{X}_a$ .

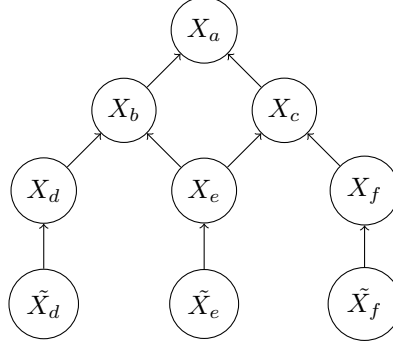
The triplet  $M = (\mathcal{U}, \mathcal{V}, \mathcal{F})$  is a causal model, acyclic and for which the structural equations in  $\mathcal{F}$  do not use disjunction. This model  $M$  is therefore an ACG.

**Remark 3.** *For each unattacked argument, we propose to build two variables, an endogenous one and an exogenous one. This duplication allows us to choose whether an unattacked argument is accepted or not through its exogenous representative  $\tilde{X}_a$  by initializing it to 0 or 1. Moreover, in the framework defined by J. Halpern and J. Pearl [4], only endogenous variables can be causes and thus explanations: through its endogenous representative, an unattacked argument can also be a cause.*

**Example 2.** *(continued) – The application of the proposed transformation leads to the construction of six endogenous variables:  $\mathcal{V} = \{X_a, X_b, X_c, X_d, X_e, X_f\}$ , and three exogenous variables, corresponding to the three unattacked arguments  $(d, e, f)$ :  $\mathcal{U} = \{\tilde{X}_d, \tilde{X}_e, \tilde{X}_f\}$ . In addition, the attack relations are transformed into structural equations. For example,  $a$  is attacked by  $b$  and  $c$ , so  $F_{X_a} = \neg X_b \wedge \neg X_c$ . With these transformations, we obtain the argumentative causal graph displayed in Figure 3.*

We call **default context** of the argumentation the unique context  $\mathbf{u}^*$  such that all exogenous variables are set to 1. It represents the situation described by the argumentative graph in which all unattacked arguments are accepted.





**Figure 3:** Argumentative causal graph created by applying the proposed transformation on the argumentative graph shown in Figure 2.

## 4.2. Back to Explanations

Causal models and abstract argumentation framework both have their own definition of the notion of explanation. This section shows that, using the proposed transformation, we can build an ACG counterpart explanation to any AAF explanation.

**Proposition 1.** *Let  $AF = (A, R)$  be an abstract argumentation framework whose associated graph is acyclic. Let  $a^* \in A$  be an argument such that there is an admissible set of which  $a^*$  is the topic. Let  $S$  be a compact explanation of  $a^*$ . Let  $M = (\mathcal{U}, \mathcal{V}, \mathcal{F})$  be the argumentative causal graph built from the transformation described in Section 4.1.*

*Let us consider:*

- $\varphi = (X_{a^*} = 1)$ ,
- $X_{arg} = S \setminus \{a^*\}$  and  $\mathbf{X} = \{X_a \mid a \in X_{arg}\}$ ,
- $\mathcal{K}$  a set of contexts that contains the default context  $\mathbf{u}^*$  ( $\mathbf{u}^* \in \mathcal{K}$ ).

*Then  $\mathbf{X} = \mathbf{1}$  is a non-minimal causal explanation of  $\varphi$  relative to  $\mathcal{K}$ , i.e.  $\mathbf{X} = \mathbf{1}$  satisfies **EX1** and **EX3** in  $\mathcal{K}$ .*

This proposition reintroduces the notion of explainee. Indeed, we create the set  $\mathcal{K}$  of contexts considered by the explainee. We only impose that  $\mathbf{u}^*$  belongs to  $\mathcal{K}$ . This assumption seems reasonable because it is the only context considered when working from a purely argumentative point of view.

**Proof 1.** *Let us prove that  $\mathbf{X} = \mathbf{1}$  satisfies **EX1** and **EX3** in  $\mathcal{K}$ .*

**(EX3)** *Let us first show that  $\mathbf{u}^*$  belongs to  $\mathcal{K}_{(\mathbf{X}=\mathbf{1}) \wedge \varphi}$ .*

*By assumption,  $\mathbf{u}^*$  belongs to  $\mathcal{K}$ .*

**(i)** *Let us prove by contradiction that  $\mathbf{u}^*$  belongs to  $\mathcal{K}_{\mathbf{X}=\mathbf{1}}$ .*

*Let us suppose that  $(M, \mathbf{u}^*) \models \neg(\mathbf{X} = \mathbf{1})$ . Then,  $\exists X_a \in \mathbf{X}$  such that  $X_a = 0$ . Now  $Att_a \neq \emptyset$ , and as  $F_{X_a} = (\bigwedge_{b \in Att_a} \neg X_b), \exists b \in Att_{X_a}$  such that  $X_b = 1$ .*

*However  $S$  is admissible thus  $\exists c \in Att_b \cap S$ .*

If  $Att_c = \emptyset$  then by definition of  $\mathbf{u}^*$ ,  $X_c = 1$ . It is not possible because  $X_b = 1$ . This leads to a contradiction.

Otherwise, as  $X_b = 1$  then  $X_c = 0$ . We can then recursively apply the same reasoning to  $X_c$ . As the graph is finite and acyclic, then it necessarily leads to a case where the defender ( $c$  here) is unattacked which leads to the same contradiction as before.

This shows that indeed  $\mathbf{u}^*$  is included  $\mathcal{K}_{\mathbf{X}=1}$ .

(ii) Let us prove that  $\mathbf{u}^*$  belongs to  $\mathcal{K}_\varphi$ .

As  $S$  is admissible with  $a^*$  as a topic and as the graph is acyclic,  $\forall b \in Att_{a^*}, \exists c \in X \cap Att_b$ . Now,  $(M, \mathbf{u}^*) \models \mathbf{X} = \mathbf{1}$ , so  $X_c = 1$  hence  $X_b = 0$ . Thus,  $\forall b \in Att_{a^*}, X_b = 0$  hence  $X_{a^*} = \bigwedge_{b \in Att_{a^*}} \neg 0 = 1$ .

Therefore  $\mathbf{u}^*$  belongs to  $\mathcal{K}_\varphi$ .

Thus,  $\mathbf{u}^*$  belongs to  $\mathcal{K}_{(\mathbf{X}=1) \wedge \varphi}$ , so this set is not empty, which shows that **EX3** is satisfied.

(EX1) Let us prove that  $\mathbf{X} = \mathbf{1}$  is a sufficient cause in  $\mathcal{K}$ , i.e. it verifies **SC1**, **SC2** and **SC3**, for all  $\mathbf{u} \in \mathcal{K}_{(\mathbf{X}=1) \wedge \varphi}$ . Let  $\mathbf{u} \in \mathcal{K}_{(\mathbf{X}=1) \wedge \varphi}$ .

**SC4** is a minimality condition which relates to the sufficient cause but which in the case of explanations is equivalent to **EX2** [9]. For that reason, we do not prove that  $\mathbf{X} = \mathbf{1}$  satisfies **SC4**.

1) **SC1** is verified by definition of  $\mathbf{u}$ .

2) Let us prove by contradiction that **SC3** is satisfied. Let  $\mathbf{u}'$  be a context such that  $(M, \mathbf{u}') \models [\mathbf{X} = \mathbf{1}] \neg \varphi$ .

As  $\neg \varphi$  holds (i.e.  $X_{a^*} = 0$ ), then according to  $\mathcal{F}$  for attacked argument ( $a^*$  is a topic of  $S$  so  $Att_{a^*} \neq \emptyset$ )  $\exists X_b \in \mathcal{V}$ , such that  $b \in Att_{a^*}$  and  $X_b = 1$ .

Now,  $S$  is admissible hence  $\exists c \in S \cap Att_b$ . Moreover, the graph is acyclic therefore  $c \neq a^*$  hence  $c \in \mathbf{X}$ . As  $\mathbf{X} = \mathbf{1}$ , it holds especially that  $X_c = 1$  and hence  $X_b = 0$  following  $F_{X_b}$ . This leads to a contradiction.

3) Finally, let us show that **SC2** is verified. To do so, we first build an actual cause of  $\varphi$  in  $\mathbf{u}$  and then show that this set does contain at least one element of  $\mathbf{X}$ .

(i) Let  $b \in Att_{a^*}$  and  $\mathbf{Z}_b = \bigcup_{c \in Att_b} \{X_c \mid (M, \mathbf{u}) \models (X_c = 1)\}$ .

As  $\mathbf{u} \in \mathcal{K}_{(\mathbf{X}=1) \wedge \varphi}$ , then  $(M, \mathbf{u}) \models X_{a^*} = 1$ , hence  $b \in Att_{a^*}$  leads to  $X_b = 0$ .

Now,  $S$  is admissible so in particular, as  $b \in Att_{a^*}$  and  $a^* \in S$ ,  $\exists c \in S \cap Att_b$ . As  $X_b = 0$  it holds that  $X_c = 1$ . Thus,  $X_c \in \mathbf{Z}_b$ . It follows that  $\mathbf{Z}_b$  is not empty.

Let us set  $\mathbf{Z} = \bigcup_{b \in Att_{a^*}} \mathbf{Z}_b$ .

$\mathbf{Z} = \mathbf{1}$  is not the set we are looking for to be an actual cause. Nevertheless, let us show that it satisfies **AC1** and **AC2** for  $\varphi = (X_a = 1)$  in the world  $(M, \mathbf{u})$ :

**AC1** is verified by construction of  $\mathbf{Z}_b$ .

**AC2**: By construction of  $\mathbf{Z}$ , if we force  $\mathbf{Z} = \mathbf{0}$ , then it holds that  $\forall b \in Att_{a^*}, \forall c \in Att_b, X_c = 0$ . Thus,  $F_{X_b} = \bigwedge_{c \in Att_b} \neg X_c = \bigwedge_{b \in Att_{a^*}} \neg 0 = 1$ . Therefore it holds that  $(M, \mathbf{u}) \models [\mathbf{Z} = \mathbf{0}] \neg \varphi$  so it follows that **AC2** is satisfied with  $\mathbf{W} = \emptyset$ .

Let us note  $\mathbf{Z}^m$  a minimal subset of  $\mathbf{Z}$  such that  $(\mathbf{Z}^m = \mathbf{1})$  verifies **AC1** and **AC2**. It is well defined and not empty because  $(\mathbf{Z} = \mathbf{1})$  satisfies **AC1** and **AC2**. Moreover,  $\mathbf{Z}^m$  satisfies **AC3** by definition of  $\mathbf{Z}^m$ . Therefore,  $(\mathbf{Z}^m = \mathbf{1})$  is an actual cause of  $\varphi$ .

(ii) Let us now show that we can build an actual cause  $(\mathbf{Z}' = \mathbf{z}')$  of  $\varphi$  such that  $\mathbf{Z}' \cap \mathbf{X} \neq \emptyset$ .

If  $\mathbf{Z}^m \cap \mathbf{X} \neq \emptyset$  then  $\mathbf{Z}' = \mathbf{Z}^m$  works.

Otherwise, i.e. if  $\mathbf{Z}^m \cap \mathbf{X} = \emptyset$ , then let  $b \in \text{Att}_{a^*}$  be an attacker of  $a^*$ :

$\exists X_c \in \mathbf{Z}^m$  such that  $c \in \text{Att}_b$ ,  $(M, \mathbf{u}) \models (X_c = 1)$  and  $X_c \notin \mathbf{X}$ .

As  $S$  is admissible and the graph is acyclic,  $\exists X_{c'} \in \mathbf{X}$  such that  $c' \in \text{Att}_b$ . Moreover,  $\mathbf{Z}^m \cap \mathbf{X} = \emptyset$ , hence  $X_{c'} \notin \mathbf{Z}^m$ . Finally, as  $\mathbf{u} \in \mathcal{K}$  we have  $(M, \mathbf{u}) \models (X_{c'} = 1)$ .

Let  $\mathbf{Z}^{m'} = (\mathbf{Z}^m \setminus \{X_c\}) \cup \{X_{c'}\}$ .  $\mathbf{Z}^{m'}$  also verifies **AC1** and **AC2**. As  $\mathbf{Z}^m$  is minimal by construction, then if  $\mathbf{Z}^{m'}$  is not minimal,  $\exists \mathbf{Z}' \subseteq \mathbf{Z}^{m'}$  such that  $\mathbf{Z}' \not\subseteq \mathbf{Z}^m$ . However,  $\mathbf{Z}^{m'} \setminus \mathbf{Z}^m = \{X_{c'}\}$  so  $X_{c'} \in \mathbf{Z}'$  and therefore we have  $\mathbf{Z}' \cap \mathbf{X} \neq \emptyset$ . Thus, we have built a set  $\mathbf{Z}'$  verifying **AC1** and **AC2**, minimal for the inclusion relation (**AC3**) and such that  $\mathbf{Z}' \cap \mathbf{X} \neq \emptyset$ . Therefore,  $\mathbf{Z}'$  satisfies **SC2**.

We have proved that whatever  $\mathbf{u} \in \mathcal{K}$ ,  $\mathbf{X} = \mathbf{1}$  satisfies **SC1**, **SC2**, **SC3**. Therefore, **EX1** is satisfied by  $\mathbf{X} = \mathbf{1}$ .

We proved that  $\mathbf{X} = \mathbf{1}$  satisfies **EX1** and **EX3**, hence  $\mathbf{X} = \mathbf{1}$  is an non-minimal causal explanation of  $\varphi$ .  $\square$

**Example 2.** (continued) – Let us illustrate this proposition with Example 2.

(i) We set  $a^* = a$ ,  $S = \{a, d, f\}$ ,  $\mathbf{X} = \{X_d, X_f\}$  and  $\varphi = (X_a = 1)$ .

By definition,  $\mathbf{u}^* = (X_d = 1, X_e = 1, X_f = 1)$ . It follows immediately that  $(M, \mathbf{u}^*) \models (X_d = 1 \wedge X_e = 1 \wedge X_f = 1)$ . In particular, it holds that  $(M, \mathbf{u}^*) \models (\mathbf{X} = \mathbf{1})$ . As a result,  $X_b = 0$  and  $X_c = 0$ , hence  $X_a = 1$ . Therefore,  $(M, \mathbf{u}^*) \models \varphi$ . Thus it holds that  $\mathbf{u}^* \in \mathcal{K}_{(\mathbf{X}=\mathbf{1}) \wedge \varphi}$  so **EX3** is indeed verified.

(ii) Let  $\mathcal{K}$  be a set of contexts, and  $\mathbf{u} \in \mathcal{K}_{(\mathbf{X}=\mathbf{1}) \wedge \varphi}$  ( $\mathcal{K}_{(\mathbf{X}=\mathbf{1}) \wedge \varphi}$  is not empty as it contains  $\mathbf{u}^*$ ). First, it holds that  $(M, \mathbf{u}) \models (\mathbf{X} = \mathbf{1}) \wedge \varphi$  since  $\mathbf{u} \in \mathcal{K}_{(\mathbf{X}=\mathbf{1}) \wedge \varphi}$ . Secondly, let  $\mathbf{u}' \in \mathcal{K}$ . In  $(M, \mathbf{u}')$  we have  $(M, \mathbf{u}') \models [\mathbf{X} = \mathbf{1}](X_b = 0 \wedge X_c = 0)$ , hence  $(M, \mathbf{u}') \models [\mathbf{X} = \mathbf{1}](X_a = 1)$ . Finally, we set  $\mathbf{Y} = \{X_e\}$  and  $X = \{X_d\}$ . As  $X = 0 \wedge \mathbf{Y} = \mathbf{0}$ , it holds that  $X_b = 1$ , hence  $X_a = 0$ . Thus,  $(M, \mathbf{u}) \models [X = 0 \wedge \mathbf{Y} = \mathbf{0}] \neg \varphi$ . This shows that  $\forall \mathbf{u} \in \mathcal{K}$ ,  $\mathbf{X} = \mathbf{1}$  is a sufficient cause of  $\varphi$ , i.e. **EX1** is satisfied.

Thus, we illustrate in this example the fact that a compact explanation in the AAF of Example 2 is indeed an non minimal causal explanation in its associated ACG.

## 5. From ACG to AAF

In this section, we propose the inverse transformation, from ACG to AAF, as well as a proof of equivalence between these two formal frameworks.

### 5.1. Proposed Inverse Transformation

Given an ACG  $M = (\mathcal{U}, \mathcal{V}, \mathcal{F})$  and a set of contexts  $\mathcal{K}$ , the proposed transformation builds the AAF  $(A', R')$  defined as:  $A' = \{a \mid X_a \in \mathcal{V}\}$ ,

For any couple  $(X_a, X_b) \in \mathcal{V}^2$  of endogenous variables, let  $\mathbf{Y} = \mathcal{V} \setminus \{X_a, X_b\}$ . If for any context  $\mathbf{u} \in \mathcal{K}$ ,  $(M, \mathbf{u}) \models [X_b = 1, \mathbf{Y} = \mathbf{0}](X_a = 0)$  then  $(b, a) \in R'$ .

**Example 2.** (continued) – For the case of the argumentative causal graph presented in Figure 3, we have  $\mathcal{V} = \{X_a, X_b, X_c, X_d, X_e, X_f\}$ , thus we set  $A' = \{a, b, c, d, e, f\}$ .

Let  $\mathcal{K}$  be a set of contexts. Let  $\mathbf{u} \in \mathcal{K}$  be a context of  $\mathcal{K}$ . We have  $F_{X_a} = \neg X_b \wedge \neg X_c$ . Thus,  $(M, \mathbf{u}) \models [X_v = 1](X_a = 0)$  with  $X_v \in \{X_b, X_c\}$ .

Therefore  $(M, \mathbf{u}) \models [X_v = 1, \mathbf{Y} = \mathbf{0}](X_a = 0)$  with  $X_v \in \{X_b, X_c\}$  and  $\mathbf{Y} = \mathcal{V} \setminus \{X_a, X_v\}$ . Thus  $(b, a) \in R'$  and  $(c, a) \in R'$ .

Applying the same reasoning to all structural equations of  $\mathcal{F}$  leads to  $\{(d, b), (e, b), (e, c), (f, c)\} \in R'$ .

Now let us consider  $\mathbf{Y} = \mathcal{V} \setminus \{X_a, X_v\}$  with  $v \in \{d, e, f\}$ .

$(M, \mathbf{u}) \models [X_v = 1, \mathbf{Y} = \mathbf{0}](X_a = 1)$  holds. Indeed, all structural equations have been replaced by  $F_X = 0$  except for  $X_a$  and  $X_v$ .

As  $X_v = 1$  and  $F_{X_a} = \neg X_b \wedge \neg X_c$ , then  $X_a = \neg 0 \wedge \neg 0 = 1$ .

Therefore  $(v, a) \notin R'$ .

As a result,  $R' = \{(b, a), (c, a), (d, b), (e, b), (e, c), (f, c)\}$ .

## 5.2. Equivalence between AAF and ACG

**Proposition 2.** Let  $AF = (A, R)$  be an abstract argumentation framework,  $M = (\mathcal{U}, \mathcal{V}, \mathcal{F})$  the argumentative causal graph generated from  $AF$  by the transformation presented in Section 4 and  $AF' = (A', R')$  the abstract argumentative framework resulting from the inverse transformation of  $M$ . Then:

$$AF = AF'.$$

**Proof 2.** Let  $AF = (A, R)$  be an AAF,  $M = (\mathcal{U}, \mathcal{V}, \mathcal{F})$  be the ACG built from  $AF$  and  $AF' = (A', R')$  the AAF resulting from  $M$ .  $A = A'$  by construction; let us prove that  $R = R'$  by double inclusion.

1. Let  $(a, b) \in A^2$  be two arguments of  $A$  such that  $R(b, a)$ . By definition,  $X_a = \neg X_b \wedge (\bigwedge_{c \in \text{Att}_a \setminus \{b\}} \neg X_c)$  and therefore  $X_a = 0$  if  $X_b = 1$ . Thus, for all contexts  $\mathbf{u}$ ,  $(M, \mathbf{u}) \models [X_b = 1, \mathbf{Y} = \mathbf{0}](X_a = 0)$  with  $\mathbf{Y} = \mathcal{V} \setminus \{X_a, X_b\}$ , therefore  $(b, a) \in R'$  and  $R \subseteq R'$ .
2. Let  $(a', b') \in A'^2$  be two arguments of  $A'$  such that  $(b', a') \in R'$ . Let  $\mathbf{Y} = \mathcal{V} \setminus \{X_{a'}, X_{b'}\}$ . By definition:

$$(M, \mathbf{u}) \models [X_{b'} = 1, \mathbf{Y} = \mathbf{0}](X_{a'} = 0)$$

Now  $A = A'$ , so  $\forall \alpha \in A$ ,  $X_\alpha = X_{\alpha'}$ . In particular it thus holds that,  $(M, \mathbf{u}) \models [X_{b'} = 1, \mathbf{Y} = \mathbf{0}](X_a = 0)$ , hence  $\text{Att}_a^R \neq \emptyset$ .

Moreover,  $F_{X_a} = \bigwedge_{z \in \text{Att}_a^R} \neg X_z$ . If  $b' \notin \text{Att}_a^R$  then with  $\mathbf{Y} = \mathbf{0}$ ,  $F_{X_a} = \bigwedge_{\beta \in \text{Att}_a^R} \neg 0 = 1$ , that contradicts the hypothesis.

It follows that  $b' \in \text{Att}_a^R$  and as a result,  $R' \subseteq R$ .

We proved that  $R \subseteq R'$  and  $R' \subseteq R$ , i.e.  $R = R'$ . □

## 6. Conclusion and Future Work

In this paper we established the equivalence between argumentative causal graphs and abstract argumentation frameworks. We also proposed explicit transformations to go from one to the other. This allows us to use all the work already done on both sides and select what we are looking for on each one opening new direction for enriching the representation of interactions: dynamic modeling offered by AAF and dynamic representation offered by ACG with the notion of context.

On the one hand, the notion of context present in causal models allows one to change the values of the variables as one wishes, and thus offers a dynamic framework. Moreover, it allows us to take into account the knowledge of the agents. Furthermore, the work of J. Pearl and J. Halpern [5] also introduces the notion of explanatory power and partial explanation, as well as a general definition which in addition provides knowledge of the model to the explainee. Thus, this framework offers a definition of an explanation that suits social and cognitive science point of view [3] quite well. However, all of the above requires to be able to create the causal graphs of such situations which is supposed to be given in [6] for example.

On the other hand, argumentation systems propose a more natural framework to model interaction situations, which can facilitate their implementation for systems interacting with humans. Thus, one approach could be to dynamically model an interaction with an AAF, to compute a result or an action and then to perform the transformation into ACG in order to generate explanations with the desired properties.

Ongoing works first aim at enriching the established equivalence in particular to allow using other relations between arguments, beyond the attack one, to model better complex interactions.

Finally, the objective of such frameworks is to propose explanations adapted to humans in order to increase their confidence in AI systems but also to facilitate human-machine interactions. Thus, another challenge of future work is to test these formal frameworks and the proposed transformation on more complete and complex examples of human-machine interaction and then to have these models subjectively evaluated by human users.

## References

- [1] D. C. Berry, D. E. Broadbent, Explanation and verbalization in a computer-assisted search task, *The Quarterly Journal of Experimental Psychology Section A* 39 (1987) 585–609. doi:10.1080/14640748708401804.
- [2] S. Wallkötter, S. Tulli, G. Castellano, A. Paiva, M. Chetouani, Explainable embodied agents through social cues: a review, *ACM Transactions on Human-Robot Interaction (THRI)* 10 (2021) 1–24.
- [3] T. Miller, Explanation in Artificial Intelligence: Insights from the Social Sciences, *Artificial Intelligence* 267 (2019) 1–38.
- [4] J. Y. Halpern, J. Pearl, Causes and Explanations: A Structural-Model Approach. Part I: Causes, *The British Journal for the Philosophy of Science* 56 (2005) 843–887.
- [5] J. Y. Halpern, J. Pearl, Causes and Explanations: A Structural-Model Approach. Part II: Explanations, *The British Journal for the Philosophy of Science* 56 (2005) 889–911.

- [6] P. Madumal, T. Miller, L. Sonenberg, F. Vetere, Explainable reinforcement learning through a causal lens, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 2493–2500.
- [7] P. M. Dung, On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games, *Artificial Intelligence* 77 (1995) 321–357. doi:10.1016/0004-3702(94)00041-X.
- [8] K. Čyras, A. Rago, E. Albin, P. Baroni, F. Toni, Argumentative XAI: A survey, in: Z.-H. Zhou (Ed.), Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, 2021, pp. 4392–4399. doi:10.24963/ijcai.2021/600.
- [9] J. Y. Halpern, *Actual Causality*, MIT Press, 2016.
- [10] L. A. Paul, N. Hall, *Causation: A User’s Guide*, Oxford University Press, 2013.
- [11] T. Miller, Contrastive explanation: A structural-model approach, *Knowledge Engineering Review* 36 (2021) E14. doi:https://doi.org/10.1017/S0269888921000102.
- [12] X. Fan, F. Toni, On Computing Explanations in Abstract Argumentation, in: ECAI, volume 263, 2014, pp. 1005–1006. doi:10.3233/978-1-61499-419-0-1005.
- [13] S. Modgil, H. Prakken, The ASPIC+ framework for structured argumentation: a tutorial, *Argument & Computation* 5 (2014) 31–62.
- [14] A. Borg, F. Bex, A basic framework for explanations in argumentation, *IEEE Intelligent Systems* 36 (2021) 25–35.