Proceedings of the 19th International Conference on Principles of Knowledge Representation and Reasoning
Special Session on KR and Machine Learning

Explaining Causal Models with Argumentation: the Case of Bi-variate Reinforcement

Antonio Rago[1] , **Pietro Baroni**[2] and **Francesca Toni**[1]
[1]Department of Computing, Imperial College London, UK
[2]Dipartimento di Ingegneria dell'Informazione, Università degli Studi di Brescia, Italy
{a.rago, ft}@imperial.ac.uk, pietro.baroni@unibs.it

Abstract

Causal models are playing an increasingly important role in machine learning, particularly in the realm of explainable AI. We introduce a conceptualisation for generating argumentation frameworks (AFs) from causal models for the purpose of forging explanations for the models' outputs. The conceptualisation is based on reinterpreting desirable properties of semantics of AFs as *explanation moulds*, which are means for characterising the relations in the causal model argumentatively. We demonstrate our methodology by reinterpreting the property of *bi-variate reinforcement* as an explanation mould to forge *bipolar AFs* as explanations for the outputs of causal models. We perform a theoretical evaluation of these argumentative explanations, examining whether they satisfy a range of desirable explanatory and argumentative properties.

1 Introduction

The field of explainable AI (XAI) has in recent years become a major focal point of the efforts of researchers, with a wide variety of models for explanation being proposed (see e.g. (Guidotti et al. 2019) for an overview). More recently, incorporating a causal perspective into explanations has been explored by some, e.g. (Schwab and Karlen 2019; Madumal et al. 2020). The link between causes and explanations has long been studied (Halpern and Pearl 2001); indeed, the two have even been equated (under a broad sense of the concept of "cause") (Woodward 1997). Causal reasoning is, in fact, how humans explain to one another (de Graaf and Malle 2017), and so mimicking such a trend lends credence to the hypothesis that machines should do likewise when their explanations target humans. Further, research from the social sciences (Miller 2019) has indicated the value of causal links, particularly in the form of counterfactual reasoning, within explanations, and that the importance of such information surpasses that of probabilities or statistical relationships for users. In this paper we outline a methodology for obtaining explanations from *causal models* (Pearl 1999), based on *(computational) argumentation* (see (Atkinson et al. 2017; Baroni et al. 2018) for recent overviews).

Argumentation has received increasing attention in recent years as a means for providing explanations of the outputs of a number of AI models (see (Vassiliades, Bassiliades, and Patkos 2021; Cyras et al. 2021) for recent overviews on argumentative XAI), e.g. for recommender systems (Teze, Godo, and Simari 2018), neural classifiers (Dejl et al. 2021), Bayesian networks (Timmer et al. 2015) and PageRank (Albini et al. 2020a). *Argumentative explanations* have also been advocated in the social sciences (Antaki and Leudar 1992; Miller 2019), and several works focus on the power of argumentation to provide a bridge between explained models and users, validated by user studies (Madumal et al. 2019; Rago et al. 2020). While argumentative explanations are wide-ranging in their application, their links with causal models have remained largely unexplored to date.

In this paper, we introduce a conceptualisation for generating *argumentation frameworks* (AFs) from causal models for the purpose of forging explanations for the models' outputs. Like (Albini et al. 2020b; Albini et al. 2021), we focus on explaining by *relations* – rather than by features, as is more conventional (e.g. for feature attribution methods such as (Lundberg and Lee 2017)). Our method is based on a reinterpretation of properties of argumentation semantics from the literature as *explanation moulds*, i.e. means for characterising argumentative relations (§3). Here, we focus on reinterpreting the property of *bi-variate reinforcement* (Amgoud and Ben-Naim 2018) as a basis for extracting bipolar AFs (Cayrol and Lagasquie-Schiex 2005) which may be used as explanations for the outputs of causal models. We provide a theoretical assessment of these explanations (§4), demonstrating how they satisfy desirable properties from both explanatory and argumentative viewpoints.

2 Background

Here, we provide the core background on causal models and computational argumentation, on which our method relies.

Causal models. A *causal model* (Pearl 1999) is a triple $\langle U, V, E\rangle$, where: U is a (finite) set of *exogenous variables*, i.e. variables whose values are determined by external factors (outside the causal model); V is a (finite) set of *endogenous variables*, i.e. variables whose values are determined by internal factors, namely by (the values of some of the) variables in $U \cup V$; each variable may take any value in its associated *domain*; we refer to the domain of $W_i \in U \cup V$ as $\mathcal{D}(W_i)$; E is a (finite) set of *structural equations* that, for each endogenous variable $V_i \in V$, define V_i's values as a function f_{V_i} of the values of V_i's *parents* $PA(V_i) \subseteq U \cup V \setminus \{V_i\}$. We use the term *binary* *causal model* to refer to any causal model $\langle U, V, E\rangle$ such that $\forall W_i \in U \cup V$, $\mathcal{D}(W_i) = \{0, 1\}$ (where 0 stands for "false" and 1 stands for "true", where suitable), and the term *gradual causal model* to refer to any causal model $\langle U, V, E\rangle$ such that $\forall W_i \in U \cup V$, $\mathcal{D}(W_i)$ is equipped with a partial order (we refer to this partial order as $\leq$; as usual, $a < b$ stands for $a \leq b$ and $b \nleq a$). Note that a binary causal model is a gradual causal model (with $0 < 1$).

Given a causal model $\langle U, V, E\rangle$ where $U = \{U_1, \ldots, U_i\}$, we denote with $\mathcal{U} = \mathcal{D}(U_1) \times \ldots \times \mathcal{D}(U_i)$ the set of all possible combinations of values of the exogenous variables (called *inputs*). Abusing notation, we refer to the value of any $W_i \in U \cup V$ given $\mathbf{u} \in \mathcal{U}$ as $f_{W_i}[\mathbf{u}]$: if W_i is an exogenous variable, $f_{W_i}[\mathbf{u}]$ is its assigned value in $\mathbf{u}$; if W_i is an endogenous variable, it is the value dictated by the structural equations in the causal model. We use the *do* operator (Pearl 2012) to indicate *interventions*, i.e., for any variable $V_i \in V$ and value thereof $v_i \in \mathcal{D}(V_i)$, $do(V_i = v_i)$ implies that f_{V_i} is replaced by the constant function v_i. We use the notation $set(W_i = w_i)$, for $w_i \in \mathcal{D}(W_i)$, to indicate that $do(W_i = w_i)$ if $W_i \in V$ or that W_i is assigned w_i if $W_i \in U$.

Argumentation. An *argumentation framework* (AF) is any tuple $\langle \mathcal{A}, \mathcal{R}_1, \ldots, \mathcal{R}_l\rangle$ with $\mathcal{A}$ a set (of *arguments*), and $\mathcal{R}_i \subseteq \mathcal{A} \times \mathcal{A}$, for $i \in \{1, \ldots, l\}$, (binary and directed) *dialectical relations* between arguments (Gabbay 2016; Baroni et al. 2017). In the abstract argumentation (Dung 1995) tradition, arguments in these AFs are unspecified *abstract* entities that can correspond to different concrete instances in different settings. Several specific choices of dialectical relations can be made, giving rise to specific AFs, including *bipolar AFs* (BFs, our focus in this paper) (Cayrol and Lagasquie-Schiex 2005), with $l = 2$ and $\mathcal{R}_1$ and $\mathcal{R}_2$ dialectical relations of *attack* and *support*, respectively, referred to later as $\mathcal{R}_-$ and $\mathcal{R}_+$. Given a BF $\langle \mathcal{A}, \mathcal{R}_-, \mathcal{R}_+\rangle$, for $\alpha_1 \in \mathcal{A}$, we will use the notation $\mathcal{R}_-(\alpha_1) = \{\alpha_2 | (\alpha_2, \alpha_1) \in \mathcal{R}_-\}$ and $\mathcal{R}_+(\alpha_1) = \{\alpha_2 | (\alpha_2, \alpha_1) \in \mathcal{R}_+\}$. The meaning of BFs (including the intended dialectical role of the relations) may be given in terms of *gradual semantics* (e.g. see (Baroni et al. 2017; Baroni, Rago, and Toni 2018)), defined by means of mappings $\sigma : \mathcal{A} \to \mathbb{V}$, with $\mathbb{V}$ a given set of *values* of interest for evaluating arguments. The choice of gradual semantics for BFs may be guided by *properties* that σ should satisfy (e.g. as in (Baroni, Rago, and Toni 2018)). We will utilise, in §3, a variant of the property of *bi-variate reinforcement* for BFs from (Amgoud and Ben-Naim 2018). We will also use, in §4, the following notions of coherence from (Cayrol and Lagasquie-Schiex 2005). Let a *path* from $\alpha_x \in \mathcal{A}$ to $\alpha_y \in \mathcal{A}$ via a relation $\mathcal{R} \subseteq \mathcal{R}_- \cup \mathcal{R}_+$ be a sequence of arguments $\alpha_1, \ldots, \alpha_n$, $n \geq 1$, such that $\alpha_1 = \alpha_x$, $\alpha_n = \alpha_y$, and for each i, $1 \leq i < n$, $(\alpha_n, \alpha_{n+1}) \in \mathcal{R}$. Then, a set of arguments $S \subseteq \mathcal{A}$ is *internally coherent* iff $\forall \alpha_x, \alpha_y \in S$, $\nexists$ a path $\alpha_1, \ldots, \alpha_n$ from α_x to α_y via $\mathcal{R}_- \cup \mathcal{R}_+$ such that $(\alpha_{n-1}, \alpha_n) \in \mathcal{R}_-$ and for $1 \leq i < n-1$, $(\alpha_i, \alpha_{i+1}) \in \mathcal{R}_+$, nor such that $(\alpha_1, \alpha_2) \in \mathcal{R}_-$ and for $2 \leq i < n$, $(\alpha_i, \alpha_{i+1}) \in \mathcal{R}_+$. S is said to be *externally coherent* iff $\forall \alpha_x, \alpha_y \in S$, $\nexists \alpha_z \in \mathcal{A}$ such that $\exists$ a path from α_x to α_z via $\mathcal{R}_+$ and $\exists$ a path $\alpha_1, \ldots, \alpha_n$ from α_y to α_z via $\mathcal{R}_- \cup \mathcal{R}_+$ such that $(\alpha_{n-1}, \alpha_n) \in \mathcal{R}_-$ and for $1 \leq i < n-1$, $(\alpha_i, \alpha_{i+1}) \in \mathcal{R}_+$, or such that $(\alpha_1, \alpha_2) \in \mathcal{R}_-$ and for $2 \leq i < n$, $(\alpha_i, \alpha_{i+1}) \in \mathcal{R}_+$.

i.

U_1	U_2	V_1	V_2
1 margherita	1 pineapple	0 ∼Italian	0 ∼enter
1 margherita	0 ∼pineapple	1 Italian	1 enter
0 margarita	1 pineapple	0 ∼Italian	0 ∼enter
0 margarita	0 ∼pineapple	0 ∼Italian	0 ∼enter

ii. **iii.**

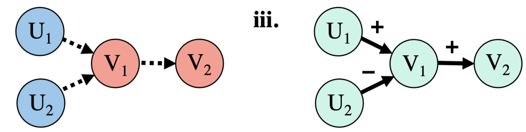

Figure 1: Toy example. (i) Combinations of values (1 or 0) resulting from the structural equations (the assignment of 1 to U_1 may be read as *"margherita" is spelt correctly on the menu* – simply given as 'margherita' in the table, the assignment of U_2 to 0 may be read as *there is no pineapple on the pizzas* – simply given as '∼pineapple' in the table, etc.). (ii) Graphical representation of the influence graph, with (exogenous/endogenous) variables in the causal model indicated by (blue/red respectively) nodes and influences represented by dashed arrows. (iii) RX for $\mathbf{u} \in \mathcal{U}$ such that $f_{U_1}[\mathbf{u}] = 1$ and $f_{U_2}[\mathbf{u}] = 0$.

3 From Causal Models to Explanation Moulds and Argumentative Explanations

We see the task of obtaining *explanations* for causal models' assignments of values to variables as a two-step process: first we define *moulds* characterising the core ingredients of explanations; then we use these moulds to obtain, automatically, (instances of) AFs as argumentative explanations. Moulds and explanations are defined in terms of *influences* between variables in the causal model, in turn defined in terms of the parent relation underpinning the model.

Definition 1. *The* influence graph *corresponding to a causal model $\langle U, V, E\rangle$ is the pair $\langle \mathcal{V}, \mathcal{I}\rangle$, where $\mathcal{V} = U \cup V$ and $\mathcal{I} \subseteq \mathcal{V} \times \mathcal{V}$ such that $\mathcal{I} = \{(W_1, W_2) | W_1 \in PA(W_2)\}$ (referred to as the set of* influences*).*

Note that influence graphs are closely related to the notion of *causal diagrams* (Pearl 1995). While straightforward, they are useful as they underpin much of what follows.

Throughout, for illustration we will use a toy example with a simple (binary) causal model $\langle U, V, E\rangle$ comprising $U = \{U_1, U_2\}$, $V = \{V_1, V_2\}$ and $\forall W_i \in U \cup V$, $\mathcal{D}(W_i) = \{0, 1\}$. Figure 1i gives the combinations of values for the variables resulting from the structural equations E (amounting to $V_1 = U_1 \wedge \neg U_2$ and $V_2 = V_1$) and Figure 1ii visualises the influence graph $\langle \{U_1, U_2, V_1, V_2\}, \{(U_1, V_1), (U_2, V_1), (V_1, V_2)\}\rangle$ (we ignore Figure 1iii for the moment: this will be discussed later). This causal model may represent a group's decision on whether to enter a restaurant, with variables U_1: *"margherita" is spelt correctly on the menu, not like the drink*; U_2: *there is pineapple on the pizzas*; V_1: *the pizzeria seems to be legitimately Italian*; and V_2: *the group chooses to enter the pizzeria*.

Influence graphs synthetically express which variables affect which others but do not give an account of how the influences actually occur in the context the user may be interested in, as expressed by the given values to the exogenous variables. For example, the influence graph in Figure 1ii alone shows which variables affect other variables but provides little intuition on *how* they do so. Thus, our standpoint is that each influence can be assigned an explanatory role, indicating how that influence is actually working in that context. We assume that each explanatory role is specified by a *relation characterisation*, i.e. a Boolean logical requirement, that is used as a *mould to forge explanations* to be presented to users by indicating which relations play a role therein.

Definition 2. *Given a causal model* $\langle U, V, E\rangle$ *with corresponding influence graph* $\langle \mathcal{V}, \mathcal{I}\rangle$*, an* explanation mould *is a non-empty set* $\{c_1,\ldots,c_m\}$ *where,* $\forall i \in \{1,\ldots,m\}$*,* $c_i : \mathcal{U}\times\mathcal{I} \rightarrow \{true, false\}$ *is a* relation characterisation*, in the form of a Boolean condition in some formal language.*

Here, we do not prescribe any formal language for specifying relation characterisations: several may be suitable. The use of this definition requires an up-front choice of the number of relations and their characterisations. This choice then applies to all inputs in need of explaining.

Given an input $\mathbf{u}$, based on an explanation mould we can obtain an AF including, as dialectical relations, the influences satisfying the (different) relation characterisations for the given $\mathbf{u}$. Thus, the choice of relation characterisations is to a large extent dictated by the specific form of *argumentative explanation* the intended users expect. In general, argumentative explanations can be generated as follows.

Definition 3. *Given a causal model* $\langle U, V, E\rangle$*, its corresponding influence graph* $\langle \mathcal{V}, \mathcal{I}\rangle$ *and an explanation mould* $\{c_1, \ldots, c_m\}$*, an* argumentative explanation for $\langle U, V, E\rangle$ and $\mathbf{u} \in \mathcal{U}$ *is an AF* $\langle \mathcal{A}, \mathcal{R}_1, \ldots \mathcal{R}_m\rangle$*, where* $\mathcal{A} \subseteq \mathcal{V}$*, and* $\mathcal{R}_1, \ldots, \mathcal{R}_m \subseteq \mathcal{I} \cap (\mathcal{A} \times \mathcal{A})$ *such that, for any* $i = 1 \ldots m$*,* $\mathcal{R}_i = \{(W_1, W_2) \in \mathcal{I} \cap (\mathcal{A} \times \mathcal{A}) | c_i(\mathbf{u}, (W_1, W_2)) = true\}$*.*

Note that these argumentative explanations are *local*, namely they focus on the causal model's behaviour for (any) input $\mathbf{u}$. Thus, different argumentative explanations may be obtained for different inputs. Note also that we have left open the choice of $\mathcal{A}$ (as a generic, possibly non-strict subset of $\mathcal{V}$). In practice, $\mathcal{A}$ may be the full $\mathcal{V}$, but we envisage that users may prefer to restrict attention to some variables of interest (for example, excluding variables not "involved" in any influence satisfying the relation characterisations). For example, an argumentative explanation of a counterfactual nature for the causal model in Figure 1i and the input in the first row may choose to neglect U_1 since changing its value in this case does not affect the other variables' values.

The choice of (number and form of) relation characterisations in explanation moulds is crucial for the generation of argumentative explanations. Here we demonstrate a novel concept for utilising properties of gradual semantics for AFs for this choice, based on "property inversion". The idea is to interpret the variable values in the causal model as generated by a "hypothetical" gradual semantics embedded in the model itself. This is similar, in spirit, to recent work to extract (weighted) BFs from multi-layer perceptrons (MLPs) (Potyka 2021), using the underlying computation of the MLPs as a gradual semantics, and to proposals to explain recommender systems via tripolar AFs (Rago, Cocarascu, and Toni 2018) or BFs (Rago et al. 2020), using the underlying predicted ratings as a gradual semantics. A natural semantic choice for causal models, given that we are trying to explain why endogenous variables are assigned specific values, given assignments to the exogenous variables, is to use the assignments as a gradual semantics.

Then, the idea of inverting properties of semantics to obtain dialectical relations in AFs can be recast to obtain relation characterisations in explanation moulds as follows: *given an influence graph and a selected value assignment to exogenous variables, if an influence satisfies a given, desirable semantics property, then the influence can be cast as part of a dialectical relation with explanatory purposes in the resulting AF*. We will illustrate this concept with the property of *bi-variate reinforcement* for BFs (Amgoud and Ben-Naim 2018), which we posit is intuitive in the realm of explanations. In our formulation of this property, we require that increasing the value of variables which are attackers (supporters) can only decrease (increase, respectively) the values of variables they attack (support, respectively).

Definition 4. *Given a gradual causal model* $\langle U, V, E\rangle$ *and influence graph* $\langle \mathcal{V}, \mathcal{I}\rangle$*, a* reinforcement explanation mould *is an explanation mould* $\{c^r_-, c^r_+\}$ *such that, given some* $\mathbf{u} \in \mathcal{U}$ *and* $(W_1, W_2) \in \mathcal{I}$*:*

- $c^r_-(\mathbf{u}, (W_1, W_2)) = true$ *iff:*
  1. $\forall w_+ \in \mathcal{D}(W_1)$ *such that* $w_+ > f_{W_1}[\mathbf{u}]$*, it holds that* $f_{W_2}[\mathbf{u}, set(W_1 = w_+)] <_* f_{W_2}[\mathbf{u}]$ *with* $<_* = \leq$*;*
  2. $\forall w_- \in \mathcal{D}(W_1)$ *such that* $w_- < f_{W_1}[\mathbf{u}]$*, it holds that* $f_{W_2}[\mathbf{u}, set(W_1 = w_-)] >_* f_{W_2}[\mathbf{u}]$ *with* $>_* = \geq$*;*
  3. $\exists_{\geq 1} w_+ \in \mathcal{D}(W_1)$ *or* $\exists_{\geq 1} w_- \in \mathcal{D}(W_1)$ *satisfying the conditions at points 1 and 2 with* $<_* = <$ *and* $>_* = >$*.*
- $c^r_+(\mathbf{u}, (W_1, W_2)) = true$ *iff:*
  1. $\forall w_+ \in \mathcal{D}(W_1)$ *such that* $w_+ > f_{W_1}[\mathbf{u}]$*, it holds that* $f_{W_2}[\mathbf{u}, set(W_1 = w_+)] >_* f_{W_2}[\mathbf{u}]$ *with* $>_* = \geq$*;*
  2. $\forall w_- \in \mathcal{D}(W_1)$ *such that* $w_- < f_{W_1}[\mathbf{u}]$*, it holds that* $f_{W_2}[\mathbf{u}, set(W_1 = w_-)] <_* f_{W_2}[\mathbf{u}]$ *with* $<_* = \leq$*;*
  3. $\exists_{\geq 1} w_+ \in \mathcal{D}(W_1)$ *or* $\exists_{\geq 1} w_- \in \mathcal{D}(W_1)$ *satisfying the conditions at points 1 and 2 with* $<_* = <$ *and* $>_* = >$*.*

We call any argumentative explanation for $\langle U,V,E\rangle$ *and* $\mathbf{u}$*, given a reinforcement explanation mould* $\{c^r_-,c^r_+\}$*, a* reinforcement explanation *(RX) (for* $\langle U, V, E\rangle$ *and* $\mathbf{u}$*).*

For illustration, Figure 1iii shows the RX for the causal model in Figure 1i and $\mathbf{u}$ as in the caption. Note that the causal model can only be understood by inspection of the structural equations; instead, the argumentative explanations provide a qualitative characterisation of influences, without requiring an understanding of the structural equations. Note also that conditions 1 and 2 for the attack and support relations in RXs correspond to a weak form of local monotonicity of the model. For instance, since U_2 attacks V_1, the user knows that, all else remaining the same, any increase in the value of U_2 cannot give rise to an increase of the value of V_1, while a decrease of U_2 will not decrease the value of V_1. Condition 3 adds a guarantee of effectiveness: there is at least one variation of U_2, which, all else remaining the same, enforces a variation of V_1. Thus RXs have a counterfactual nature, as they suggest to the user the kind of local changes with respect to the current situation that could give rise to a desired change of outcome. In this respect, note that the role assigned to variables refers to the selected value assignment to exogenous variables. For example, in the RX for the input in the first line of the table in Figure 1i, the fact that "margherita" is spelt correctly on the menu does not play a role in determining that the pizzeria is not legitimately Italian (indeed this is determined solely by pineapple being on the pizza), thus the support (U_1, V_1) is not present in the RX for this input. Such differences reflect the fact that only some (or possibly none) of the individual changes of variables U_1 and U_2 are guaranteed to produce a change in V_1's value, depending on the initial context. The local nature of RXs, corresponding to the local nature of bi-variate reinforcement, ensures simplicity and a rather intuitive interpretation but at the same time clearly limits expressiveness, in particular RXs are not meant to cover cases where multiple variable changes are needed to produce an effect. Other, explanation moulds may be needed to satisfy differing users' explanatory requirements. Note that some explanation moulds may be unsuitable to some causal models, e.g. our reinforcement explanation mould is not directly applicable to causal models with variables whose domains lack a partial order.

4 Properties

We perform a theoretical evaluation of RXs with regards to their satisfaction of various properties (besides a variant of bi-variate reinforcement, satisfied by design and easy to see from Definition 4). Unless specified otherwise, we assume some RX $\langle \mathcal{A}, \mathcal{R}_-, \mathcal{R}_+ \rangle$ for gradual causal model $\langle U, V, E \rangle$ and $\mathbf{u} \in \mathcal{U}$ with influence graph $\langle \mathcal{V}, \mathcal{I} \rangle$.

The first result shows the deterministic nature of RXs.

Proposition 1 (Uniqueness). *There is no* $\langle \mathcal{A}, \mathcal{R}'_-, \mathcal{R}'_+ \rangle$ *for* $\langle U, V, E \rangle$ *and* $\mathbf{u}$, *such that* $\langle \mathcal{A}, \mathcal{R}'_-, \mathcal{R}'_+ \rangle$ *is different from* $\langle \mathcal{A}, \mathcal{R}_-, \mathcal{R}_+ \rangle$ *(i.e. such that* $\mathcal{R}'_- \neq \mathcal{R}_-$ *or* $\mathcal{R}'_+ \neq \mathcal{R}_+$*).*

This guarantees *stability* (e.g. as discussed in (Sokol and Flach 2020)), i.e. a user would never be shown two different explanations for the same causal model and input given the choice of $\mathcal{A}$. We posit that this is important to avoid possible user uncertainty and confusion.

Further, RXs are acyclic when seen as graphs.

Proposition 2 (Acylicity). *Let* (n, e) *be the graph with* $n = \mathcal{A}$ *and* $e = \mathcal{R}_- \cup \mathcal{R}_+$. *Then,* (n, e) *is acyclic.*

This follows directly from acyclicity of causal models. It prevents potentially undesirable behaviour such as a self-attacking or self-supporting variable assignments in RXs.

An argument in RXs may not both attack and support another argument therein:

Proposition 3 (Unambiguity). $\forall W_1 \in \mathcal{A}$, $\mathcal{R}_-(W_1) \cap \mathcal{R}_+(W_1) = \emptyset$, *or, equivalently,* $\mathcal{R}_- \cap \mathcal{R}_+ = \emptyset$.

Violation of this property would clearly provide contradictory indications to the user.

The following proposition states that argumentative relations in RXs are derived from causal relationships.

Proposition 4 (Relevance). $\mathcal{R}_- \cup \mathcal{R}_+ \subseteq \mathcal{I}$.

Note that, while straightforward for RXs, this property may be violated by (model-agnostic) explanation methods which do not leverage upon the underlying causal model. This property is in the same spirit as other properties in the XAI literature, e.g. *Dummy* (Sundararajan and Najmi 2020), which states that a feature which does not affect a classification is given a zero attribution value. This may be particularly important in some cases, e.g. in the running example, it may not be enough to use the absence of pineapple on pizza ($U_2 = 0$) as a reason for entering a restaurant ($V_2 = 1$), and $V_1 = 1$ provides a useful intermediate justification that the restaurant seems to be legitimately Italian.

The following requires that changing attackers or supporters in binary causal models necessitates a change in the value of the argument they attack or support, respectively.

Proposition 5 (Bipolar Counterfactuality). *If* $\langle U, V, E \rangle$ *is binary, then* $\forall (W_1, W_2) \in \mathcal{I}$ *where* $(W_1, W_2) \in \mathcal{R}_- \cup \mathcal{R}_+$, *for every* $w \neq f_{W_1}[\mathbf{u}]$: $f_{W_2}[\mathbf{u}, set(W_1 = w)] \neq f_{W_2}[\mathbf{u}]$.

This is a powerful explanatory characteristic of RXs since attacks and supports indicate counterfactuals (in the binary case). For example, given the RX in Figure 1iii, a user can immediately see that changing the value of V_2 can be achieved by changing the value of V_1, which itself can be achieved by changing the value of U_1 or U_2.

The next property shows that behaviour similar to attacks and supports with discrete semantics (see (Cayrol and Lagasquie-Schiex 2005)) arises in RXs for binary models.

Proposition 6 ((Dis)agreement). *If* $\langle U, V, E \rangle$ *is binary, then,* $\forall W_1 \in \mathcal{A}$*: if* $\exists W_2 \in \mathcal{R}_-(W_1)$ *then* $f_{W_2}[\mathbf{u}] \neq f_{W_1}[\mathbf{u}]$*; if* $\exists W_3 \in \mathcal{R}_+(W_1)$ *then* $f_{W_3}[\mathbf{u}] = f_{W_1}[\mathbf{u}]$.

We thus observe that attacks indicate a contradiction between two arguments while supports indicate harmony between them. Clearly this is the case in Figure 1, where assigning U_2 value 1 will reduce the value of V_1 to 0: a contradiction. Meanwhile, any input which changes V_1's value to 0 will necessitate the same result in V_2: a harmony.

The set of arguments assigned value 1 in an RX for a binary causal model satisfies coherence (see §2).

Proposition 7 ((Internal and External) Coherence). *If* $\langle U, V, E \rangle$ *is binary, then the set of* accepted arguments $\mathcal{A}_a = \{W_1 \in \mathcal{A} | f_{W_1}[\mathbf{u}] = 1\}$ *is internally and externally coherent.*

This result indicates that the basic principles of argumentation are upheld in RXs, which hence can support some genuine forms of argumentative reasoning on the model by the user. For example, if the RX in Figure 1iii were given for an input $\mathbf{u}'$ with $f_{U_2}[\mathbf{u}'] = 1$ and $f_{V_1}[\mathbf{u}'] = 1$, the set of accepted arguments would contain a contradiction, which is not intuitive since the accepted attacker has no effect.

5 Future Work

We believe that our approach provides the groundwork for many future directions. The computational complexity of RXs deserves attention. Moulds inspired by other properties, and resulting in other forms of AFs, could be devised. A full empirical analysis of RXs, including user studies, also seems worthwhile. Links between our work and existing XAI methods, particularly those utilising argumentation, could be instructive, while counterfactuals and causality also warrant investigation in our approach.

Acknowledgements

Toni was partially funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101020934). Rago and Toni were partially funded by J.P. Morgan and by the Royal Academy of Engineering under the Research Chairs and Senior Research Fellowships scheme. Any views or opinions expressed herein are solely those of the authors listed. The authors would like to thank Fabrizio Russo and Emanuele Albini for their helpful contributions to discussions in the build up to this work.

References

Albini, E.; Baroni, P.; Rago, A.; and Toni, F. 2020a. PageRank as an argumentation semantics. In *COMMA*, 55–66.

Albini, E.; Rago, A.; Baroni, P.; and Toni, F. 2020b. Relation-based counterfactual explanations for Bayesian network classifiers. In *IJCAI*, 451–457.

Albini, E.; Rago, A.; Baroni, P.; and Toni, F. 2021. Influence-driven explanations for Bayesian network classifiers. In *PRICAI*, 88–100.

Amgoud, L., and Ben-Naim, J. 2018. Weighted bipolar argumentation graphs: Axioms and semantics. In *IJCAI*, 5194–5198.

Antaki, C., and Leudar, I. 1992. Explaining in conversation: Towards an argument model. *European Journal of Social Psychology* 22:181–194.

Atkinson, K.; Baroni, P.; Giacomin, M.; Hunter, A.; Prakken, H.; Reed, C.; Simari, G. R.; Thimm, M.; and Villata, S. 2017. Towards artificial argumentation. *AI Magazine* 38(3):25–36.

Baroni, P.; Comini, G.; Rago, A.; and Toni, F. 2017. Abstract games of argumentation strategy and game-theoretical argument strength. In *PRIMA*, 403–419.

Baroni, P.; Gabbay, D.; Giacomin, M.; and van der Torre, L., eds. 2018. *Handbook of Formal Argumentation*. College Publications.

Baroni, P.; Rago, A.; and Toni, F. 2018. How many properties do we need for gradual argumentation? In *AAAI*, 1736–1743.

Cayrol, C., and Lagasquie-Schiex, M. 2005. On the acceptability of arguments in bipolar argumentation frameworks. In *ECSQARU*, 378–389.

Cyras, K.; Rago, A.; Albini, E.; Baroni, P.; and Toni, F. 2021. Argumentative XAI: A survey. In *IJCAI*, 4392–4399.

de Graaf, M. M. A., and Malle, B. F. 2017. How people explain action (and autonomous intelligent systems should too). In *AAAI Fall Symposia*, 19–26.

Dejl, A.; He, P.; Mangal, P.; Mohsin, H.; Surdu, B.; Voinea, E.; Albini, E.; Lertvittayakumjorn, P.; Rago, A.; and Toni, F. 2021. Argflow: A toolkit for deep argumentative explanations for neural networks. In *AAMAS*, 1761–1763.

Dung, P. M. 1995. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. *Artificial Intelligence* 77(2):321–358.

Gabbay, D. M. 2016. Logical foundations for bipolar and tripolar argumentation networks: preliminary results. *Journal of Logic and Computation* 26(1):247–292.

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; and Pedreschi, D. 2019. A survey of methods for explaining black box models. *ACM Computing Surveys* 51(5):93:1–93:42.

Halpern, J. Y., and Pearl, J. 2001. Causes and explanations: A structural-model approach: Part 1: Causes. In *UAI*, 194–202.

Lundberg, S. M., and Lee, S. 2017. A unified approach to interpreting model predictions. In *NeurIPS*, 4765–4774.

Madumal, P.; Miller, T.; Sonenberg, L.; and Vetere, F. 2019. A grounded interaction protocol for explainable artificial intelligence. In *AAMAS*, 1033–1041.

Madumal, P.; Miller, T.; Sonenberg, L.; and Vetere, F. 2020. Explainable reinforcement learning through a causal lens. In *AAAI*, 2493–2500.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267:1–38.

Pearl, J. 1995. Causal diagrams for empirical research. *Biometrika* 82(4):669–710.

Pearl, J. 1999. Reasoning with cause and effect. In *IJCAI*, 1437–1449.

Pearl, J. 2012. The do-calculus revisited. In *UAI*, 3–11.

Potyka, N. 2021. Interpreting neural networks as quantitative argumentation frameworks. In *AAAI*, 6463–6470.

Rago, A.; Cocarascu, O.; Bechlivanidis, C.; and Toni, F. 2020. Argumentation as a framework for interactive explanations for recommendations. In *KR*, 805–815.

Rago, A.; Cocarascu, O.; and Toni, F. 2018. Argumentation-based recommendations: Fantastic explanations and how to find them. In *IJCAI*, 1949–1955.

Schwab, P., and Karlen, W. 2019. CXPlain: Causal explanations for model interpretation under uncertainty. In *NeurIPS*, 10220–10230.

Sokol, K., and Flach, P. A. 2020. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *FAT**, 56–67.

Sundararajan, M., and Najmi, A. 2020. The many shapley values for model explanation. In *ICML*, 9269–9278.

Teze, J. C.; Godo, L.; and Simari, G. R. 2018. An argumentative recommendation approach based on contextual aspects. In *SUM*, 405–412.

Timmer, S. T.; Meyer, J. C.; Prakken, H.; Renooij, S.; and Verheij, B. 2015. Explaining bayesian networks using argumentation. In *ECSQARU*, 83–92.

Vassiliades, A.; Bassiliades, N.; and Patkos, T. 2021. Argumentation and Explainable Artificial Intelligence: A Survey. *The Knowledge Engineering Review* 36(2).

Woodward, J. 1997. Explanation, invariance, and intervention. *Philosophy of Science* 64:S26–S41.