Forging Argumentative Explanations from Causal Models

Antonio Rago[a], Fabrizio Russo[a], Emanuele Albini[a], Pietro Baroni[b] and Francesca Toni[a]

[a]Department of Computing, Imperial College London, UK
[b]Dipartimento di Ingegneria dell'Informazione, Università degli Studi di Brescia, Italy

Abstract
We introduce a conceptualisation for generating argumentation frameworks (AFs) from causal models for the purpose of forging explanations for models' outputs. The conceptualisation is based on reinterpreting properties of semantics of AFs as explanation moulds, which are means for characterising argumentative relations. We demonstrate our methodology by reinterpreting the property of bi-variate reinforcement in bipolar AFs, showing how the extracted bipolar AFs may be used as relation-based explanations for the outputs of causal models.

Keywords
Explainable AI, Argumentation frameworks, Causal models

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 [1] for an overview). More recently, incorporating a causal perspective into explanations has been explored by some, e.g. [2, 3, 4]. The link between causes and explanations has long been studied [5]; indeed, the two have even been equated (under a broad sense of the concept of "cause") [6]. Causal reasoning is, in fact, how humans explain to one another [7], and so mimicking such a trend lends credence to the hypothesis that machines should do likewise. Further, research from the social sciences [8] 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.

Despite these findings, many of the approaches for generating explanations for AI models have, nevertheless, neglected causality as a potential drive for explainability. Some of the most popular methods are heuristic and model-agnostic [9, 10], and, although

5th Workshop on Advances In Argumentation In Artificial Intelligence (AI^3 2021)
antonio@imperial.ac.uk (A. Rago); fabrizio@imperial.ac.uk (F. Russo); emanuele@imperial.ac.uk (E. Albini); pietro.baroni@unibs.it (P. Baroni); f.toni@imperial.ac.uk (F. Toni)
https://www.doc.ic.ac.uk/~afr114/ (A. Rago); https://briziorusso.github.io/ (F. Russo); https://www.emanuelealbini.com/ (E. Albini); https://pietro-baroni.unibs.it/ (P. Baroni); https://www.imperial.ac.uk/people/f.toni (F. Toni)
0000-0001-5323-7739 (A. Rago); 0000-0003-2964-4638 (E. Albini); 0000-0001-5439-9561 (P. Baroni); 0000-0001-8194-1459 (F. Toni)

they are useful, particularly with regards to their wide-ranging applicability, they neglect how models are determining their outputs and therefore the underlying causes therein. This has arguably left a chasm between how explanations are provided by models at the forefront of XAI technology and what users actually require from explanations [11].

Meanwhile, computational argumentation (see [12, 13] for recent overviews) has received increasing interest in recent years as a means for providing explanations of the outputs of a number of AI models, e.g. recommender systems [14], classifiers [15], Bayesian networks [16] and PageRank [17]. Argumentative explanations have also been advocated in the social sciences [18, 8], and several works focus on the power of argumentation to provide a bridge between explained models and users, validated by user studies [19, 20]. While argumentative explanations are wide-ranging in their application (see [21, 22] for recent surveys), the links between causal models and argumentative explanations have remained largely unexplored to date.

In this paper, we introduce a conceptualisation for generating argumentation frameworks (AFs) with any number of dialectical relations as envisaged in [23, 24], from causal models for the purpose of forging explanations for the models' outputs. Like [25], we focus not on explaining by features, but instead by relations, hence the use of argumentation as the underpinning explanatory mechanism. After giving the necessary background (§2), we show how properties of argumentation semantics from the literature can be reinterpreted to serve as explanation moulds, i.e. means for characterising argumentative relations (§3). In (§4) we propose a way to define explanation moulds based on inverting properties of argumentation semantics. Briefly, the idea is to detect, inside a causal model, the satisfaction of the conditions specified by some semantics property: if these conditions are satisfied by some influence in the causal model, then the influence can be assigned an explanatory role by casting it as a dialectical relation, whose type is in correspondence with the detected property. The identified dialectical relations compose, altogether, an argumentation framework. We demonstrate our methodology by reinterpreting the property of bi-variate reinforcement [26] from bipolar AFs [27] and then showing in (§5) how the extracted bipolar AFs may be used as counterfactual explanations for the outputs of causal models representing different classification methods. Finally, we discuss related work (§6) before concluding, indicating potentially fruitful future work (§7).

Overall, we make the following main contributions:

- We propose a novel concept for defining relation-based explanations for causal models by inverting properties of argumentation semantics.
- We use this concept to define a novel form of reinforcement explanation (RX) for causal models.
- We show deployability of RXs with two machine-learning models, from which causal models are drawn.

2. Background

Our method relies upon causal models and some notions from computational argumentation. We provide core background for both.

Causal models. A causal model [28] 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 values in its associated domain; we refer to the domain of $W_i \in U \cup V$ as $\mathscr{D}(W_i)$;
- 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' parents $PA(V_i) \subseteq U \cup V \setminus \{V_i\}$.

Example 1. Let us consider a simple causal model $\langle U, V, E\rangle$ comprising $U = \{U_1, U_2\}$, $V = \{V_1, V_2\}$ and for all $W_i \in U \cup V$, $\mathscr{D}(W_i) = \{\top, \bot\}$. Figure 1i (we ignore Figure 1ii for the moment: this will be discussed later in §4) visualises the variables' parents, and Table 1 gives the combinations of values for the variables resulting from the structural equations E. This may represent a group's decision on whether or not 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.

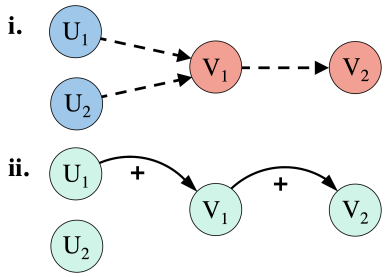

Figure 1: (i) Variables and parents for Example 1, with parents indicated by dashed arrows (for example $\{U_1, U_2\} = PA(V_1)$, i.e. U_1 and U_2 are the parents of V_1). (ii) SAF explanation (see §3) for the assignment to exogenous variables $\mathbf{u} \in \mathscr{U}$ such that $f_{U_1}[\mathbf{u}] = \top$ and $f_{U_2}[\mathbf{u}] = \top$.

Given a causal model $\langle U, V, E\rangle$ where $U = \{U_1, \ldots, U_i\}$, we denote with $\mathscr{U} = \mathscr{D}(U_1) \times \ldots \times \mathscr{D}(U_i)$ the a set of all possible combinations of values of the exogenous variables (realisations). With an abuse of notation, we refer to the value of any variable $W_i \in U \cup V$ given $\mathbf{u} \in \mathscr{U}$ as $f_{W_i}[\mathbf{u}]$: if W_i is an exogenous variable, $f_{W_i}[\mathbf{u}]$ will be its assigned value in $\mathbf{u}$; if W_i is

U_1	U_2	V_1	V_2
$\top$ margherita	$\top$ pineapple	$\bot$ ~Italian	$\bot$ ~enter
$\top$ margherita	$\bot$ ~pineapple	$\top$ Italian	$\top$ enter
$\bot$ margarita	$\top$ pineapple	$\bot$ ~Italian	$\bot$ ~enter
$\bot$ margarita	$\bot$ ~pineapple	$\bot$ ~Italian	$\bot$ ~enter

Table 1
Combinations of values ($\top$ or $\bot$) resulting from the structural equations for Example 1. Here we also indicate the intuitive reading of the assignment of values to variables according to the illustration in Example 1 (for example, the assignment of $\top$ to U_1 may be read as "margherita" is spelt correctly on the menu – simply given as 'margherita' in the table, and the assignment of U_2 to $\bot$ may be read as there is no pineapple on the pizzas – simply given as '~pineapple' in the table).

an endogenous variable, it will be the value dictated by the structural equations in the causal model.

We use the do operator [29] to indicate interventions, i.e., for any variable $V_i \in V$ and value thereof $v_i \in \mathcal{D}(V_i)$, $do(V = v_i)$ implies that the function f_{V_i} is replaced by the constant function v_i, and for any variable $U_i \in U$ and value thereof $u_i \in \mathcal{D}(U_i)$, $do(U_i = u_i)$ implies that U_i is assigned u_i.

Argumentation. In general, an argumentation framework (AF) is any tuple $\langle \mathscr{A}, \mathscr{R}_1, \ldots, \mathscr{R}_l \rangle$, with $\mathscr{A}$ a set (of arguments), $l > 0$ and $\mathscr{R}_i \subseteq \mathscr{A} \times \mathscr{A}$, for $i \in \{1, \ldots, l\}$, (binary and directed) dialectical relations between arguments [23, 24]. In the abstract argumentation [30] tradition, arguments in these AFs are unspecified abstract entities that can be instantiated differently to suit different settings of deployment. Several specific choices of dialectical relations can be made, giving rise to specific AFs instantiating the above general definition, including abstract AFs (AAFs) [30], with $l = 1$ (and $\mathscr{R}_1$ a dialectical relation of attack, referred to later as $\mathscr{R}_-$), support AFs (SAFs) [31], with $l = 1$ (and $\mathscr{R}_1$ a dialectical relation of support, referred to later as $\mathscr{R}_+$), and bipolar AFs (BAFs) [27], with $l = 2$ (and $\mathscr{R}_1$ and $\mathscr{R}_2$ dialectical relations of attack and support, respectively, referred to later as $\mathscr{R}_-$ and $\mathscr{R}_+$).

The meaning of AFs (including the intended dialectical role of the relations) may be given in terms of gradual semantics (e.g. see [24, 32] for BAFs), defined, for AFs with arguments $\mathscr{A}$, by means of mappings $\sigma : \mathscr{A} \to \mathbb{V}$, with $\mathbb{V}$ a given set of values of interest for evaluating arguments.

The choice of gradual semantics for AFs may be guided by properties that the mappings σ should satisfy (e.g. as in [26, 32]). We will utilise, in §4, a variant of the property of bi-variate reinforcement for BAFs from [26].

3. From Causal Models to Explanation Moulds and Argumentative Explanations

In this section 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, focusing on those from parents to children given by the causal structure underpinning the model, as follows.

Definition 1. Let $M = \langle U, V, E \rangle$ be a causal model. The influence graph corresponding to M is the pair $\langle \mathscr{V}, \mathscr{I} \rangle$ with:

- $\mathscr{V} = U \cup V$ is the set of all (exogenous and endogenous) variables;
- $\mathscr{I} \subseteq \mathscr{V} \times \mathscr{V}$ is defined as $\mathscr{I} = \{(W_1, W_2) | W_1 \in PA(W_2)\}$ (referred to as the set of influences).

Note that, while straightforward, the concept of influence graph (closely related to the notion of causal diagram [33]) is useful as it underpins much of what follows.

Next, the idea underlying explanation moulds is that, typically, inside the causal model, some variables affect others in a way that may not be directly understandable or even cognitively manageable by a user. The influence graph synthetically expresses which variables affect which others but does not give an account of how the influences actually occur in the context (namely, the values given to the exogenous variables) that a user may be interested in. Thus, the perspective we take is that each influence can be assigned an explanatory role, indicating how that influence is actually working in that context. The explanatory roles ascribable to influences can be regarded as a form of explanatory knowledge which is user specific: different users may be willing (and/or able) to accept explanations built using different sets of explanatory roles as they correspond to their understanding of how variables may affect each other. We assume that each explanatory role is specified by a relation characterisation, i.e. a Boolean logical requirement, which can be used to mould the explanations to be presented to the users by indicating which relations play a role in the explanations.

Definition 2. Given a causal model $\langle U, V, E \rangle$ and its corresponding influence graph $\langle \mathscr{V}, \mathscr{I} \rangle$, an explanation mould is a non-empty set:

$$\{c_1, \ldots, c_m\}$$

where for all $i \in \{1, \ldots, m\}$, $c_i : \mathscr{U} \times \mathscr{I} \rightarrow \{\top, \bot\}$ is a relation characterisation, in the form of a Boolean condition expressed in some formal language. Given some $\mathbf{u} \in \mathscr{U}$ and $(W_1, W_2) \in \mathscr{I}$, if $c_i(\mathbf{u}, (W_1, W_2)) = \top$ we say that the influence (W_1, W_2) satisfies c_i for $\mathbf{u}$.

Note that we are not prescribing any formal language for specifying relation characterisations, as several such languages may be suitable.

Given an assignment $\mathbf{u}$ to the exogenous variables, based on an explanation mould, we can obtain an AF including, as (different) 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 AF the intended users expect. Before defining argumentative explanations formally, we give an illustration.

Example 1 (Cont.). Let us imagine a situation where one would like to explain the behaviour of the causal model from Figure 1i and Table 1 with a SAF (see §2). We thus require one single form of relation (i.e. support) to be extracted from the corresponding influence graph $\langle\{U_1, U_2, V_1, V_2\}, \{(U_1, V_1), (U_2, V_1), (V_1, V_2)\}\rangle$. In order to define the explanation mould for such a situation, we note that the behaviour defining this relation could be characterised as changing the state of rejected arguments that it supports to accepted when the supporting argument's state is accepted. In our simple causal model, accepted arguments may amount to variables assigned to value $\top$ and rejected arguments may amount to variables assigned to value $\bot$. Thus, the intended behaviour can be captured by a relation characterisation c_s such that, given $\mathbf{u} \in \mathcal{U}$ and $(W_1, W_2) \in \mathcal{F}$:

$$c_s(\mathbf{u}, (W_1, W_2)) = \top \text{ iff}$$
$$(f_{W_1}[\mathbf{u}] = \top \wedge f_{W_2}[\mathbf{u}] = \top \wedge f_{W_2}[\mathbf{u}, do(W_1 = \bot)] = \bot) \vee$$
$$(f_{W_1}[\mathbf{u}] = \bot \wedge f_{W_2}[\mathbf{u}] = \bot \wedge f_{W_2}[\mathbf{u}, do(W_1 = \top)] = \top).$$

Then, for the assignment to exogenous variables $\mathbf{u} \in \mathcal{U}$ such that $f_{U_1}[\mathbf{u}] = \top$ and $f_{U_2}[\mathbf{u}] = \bot$, we may obtain the SAF in Figure 1ii (visualised as a graph with nodes as arguments and edges indicating elements of the support relation). For illustration, consider $(U_1, V_1) \in \mathcal{F}$ for this $\mathbf{u}$. We can see from Table 1 that $f_{V_1}[\mathbf{u}] = \top$ and also that $f_{V_1}[\mathbf{u}, do(U_1 = \bot)] = \bot$ and thus from the above it is clear that $c_s(\mathbf{u}, (U_1, V_1)) = \top$ and thus the influence is of the type of support that c_s characterises. Meanwhile, consider $(U_2, V_1) \in \mathcal{F}$ for the same $\mathbf{u}$: the fact that $f_{U_2}[\mathbf{u}] = \bot$ and $f_{V_1}[\mathbf{u}] = \top$ means that $c_s(\mathbf{u}, (U_2, V_1)) = \bot$ and thus the influence is not cast as a support. Indeed, if we consider the first and second rows of Table 1, we can see that U_2 being true actually causes V_1 to be false, thus it is no surprise that the influence is not cast as a support and plays no role in the resulting SAF. If we wanted for this influence to play a role, we could, for example, choose to incorporate an additional relation of attack into the explanation mould, to generate instead BAFs (see §2) as argumentative explanations. This example thus shows how explanation moulds must be designed to fit causal models depending on external explanatory requirements dictated by users. It should be noted also that some explanation moulds may be unsuitable to some causal models, e.g. the explanation mould with the earlier c_s would not be directly applicable to causal models with variables with non-binary or continuous domains.

In general, AFs serving as argumentative explanations can be generated as follows.

Definition 3. Given a causal model $\langle U, V, E\rangle$, its corresponding influence graph $\langle \mathcal{V}, \mathcal{F}\rangle$, some $\mathbf{u} \in \mathcal{U}$ and an explanation mould $\{c_1, \ldots, c_m\}$, an argumentative explanation is an AF $\langle \mathcal{A}, \mathcal{R}_1, \ldots \mathcal{R}_m\rangle$, where

- $\mathcal{A} \subseteq \mathcal{V}$, and

- $\mathscr{R}_1, \ldots, \mathscr{R}_m \subseteq \mathscr{I} \cap (\mathscr{A} \times \mathscr{A})$ such that, for any $i = 1 \ldots m$, $\mathscr{R}_i = \{(W_1, W_2) \in \mathscr{I} \cap (\mathscr{A} \times \mathscr{A}) | c_i(\mathbf{u}, (W_1, W_2)) = \top\}$.

Note that we have left open the choice of $\mathscr{A}$ (as a generic, possibly non-strict subset of $\mathscr{V}$). In practice, $\mathscr{A}$ may be the full $\mathscr{V}$, 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).

Example 1 (Cont.). The behaviour of the causal model from Figure 1i and Table 1 for $\mathbf{u}$ such that $f_{U_1}[\mathbf{u}] = \top$ and $f_{U_2}[\mathbf{u}] = \top$, using the explanation mould $\{c_s\}$ given earlier, can be captured by either of the two SAFs (argumentative explanations) below, depending on the choice of $\mathscr{A}$:

- the SAF in Figure 1ii, where every variable is an argument;
- the SAF with the same support relation but U_2 excluded from $\mathscr{A}$, as not "involved" and thus not contributing to the explanation.

Both SAFs explain that $f_{V_1}[\mathbf{u}] = \top$ is supported by $f_{U_1}[\mathbf{u}] = \top$, in turn supporting $f_{V_2}[\mathbf{u}] = \top$. Namely, the causal model recommends that the group should enter the pizzeria because the pizzeria seems legitimately Italian, given that "margherita" is spelt correctly on the menu. Note that the pineapple not being on the pizza could also be seen as a support towards the pizzeria being legitimately Italian, the inclusion of which could be achieved with a slightly more complex explanation mould.

4. Inverting Properties of Argumentation Semantics: Reinforcement Explanations

The choice (number and form) of relation characterisations in explanation moulds is crucial for the generation of explanations concerning the value assignments to endogenous variables in the causal models. Even after having decided which argumentative relations to include in the AF/argumentative explanation, the definition of the relation characterisations is non-trivial, in general. In this section we demonstrate a novel concept for utilising properties of gradual semantics for AFs for the definition of relation characterisations and the consequent extraction of argumentative explanations.

The common usage of these properties in computational argumentation can be roughly equated to: if a semantics, given an AF, satisfies some desirable properties, then the semantics is itself desirable (for the intended context, where those properties matter). We propose a form of inversion of this notion for use in our XAI setting, namely: if some desirable properties are identified for the gradual semantics of (still unspecified) AFs, then these properties can guide the definition of the dialectical relations underpinning the AFs. For this inversion to work, we need to identify first and foremost a suitable notion of gradual semantics for the AFs we extract from causal models. Given that, with our AFs, we are trying to explain the results obtained from underlying causal models, we cannot impose just any gradual semantics from the literature, but need to make sure that

we capture, with the chosen semantics, the behaviour of the causal model itself. This is similar, in spirit, to recent work to extract (weighted) BAFs from multi-layer perceptrons (MLPs) [34], using the underlying computation of the MLPs as a gradual semantics, and to the proposals to explain recommender systems (RSs) via tripolar AFs [35] or BAFs [20], using the underlying predicted ratings by the RSs as a gradual semantics.

A natural semantic choice for causal models, since we are trying to explain why endogenous variables are assigned specific values in their domains given assignments to the exogenous variables, is to use the assignments themselves 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 property, then the influence can be cast as part of a dialectical relation in the resulting AF.

Naturally, for this inversion to be useful, we need to identify useful properties from an explanatory viewpoint. We will illustrate this concept with the property of bi-variate reinforcement for BAFs [26], which we posit is generally intuitive in the realm of explanations. Bi-variate reinforcement is defined when the set of values $\mathbb{V}$ for evaluating arguments is equipped with a pre-order $<$. Intuitively, bi-variate reinforcement states that[1] strengthening an attacker (a supporter) cannot strengthen (cannot weaken, respectively) an argument it attacks (supports, respectively), where strengthening an argument amounts to increasing its value from $v_1 \in \mathbb{V}$ to $v_2 \in \mathbb{V}$ such that $v_2 > v_1$ (whereas weakening an argument amounts to decreasing its value from such v_2 to v_1). In our formulation of this property, we require that increasing the value of variables represented as attackers (supporters) can only decrease (increase, respectively) the values of variables they attack (support, respectively).

Property 1. Given a causal model $\langle U, V, E\rangle$ such that, for each $W_i \in U \cup V$, the domain $\mathscr{D}(W_i)$ is equipped with a pre-order $<$,[2] and given its corresponding influence graph $\langle \mathscr{V}, \mathscr{I}\rangle$, an argumentative explanation $\langle \mathscr{A}, \mathscr{R}_-, \mathscr{R}_+\rangle$ for $\mathbf{u} \in \mathscr{U}$ satisfies causal reinforcement iff for any $(W_1, W_2) \in \mathscr{I}$ where $w_1 = f_{W_1}[\mathbf{u}]$, for any $w_- \in \mathscr{D}(W_1)$ such that $w_- < w_1$, and for any $w_+ \in \mathscr{D}(W_1)$ such that $w_+ > w_1$:

- if $(W_1, W_2) \in \mathscr{R}_-$, then $f_{W_2}[\mathbf{u}, do(W_1 = w_+)] \leq f_{W_2}[\mathbf{u}]$ and $f_{W_2}[\mathbf{u}, do(W_1 = w_-)] \geq f_{W_2}[\mathbf{u}]$;
- if $(W_1, W_2) \in \mathscr{R}_+$, then $f_{W_2}[\mathbf{u}, do(W_1 = w_+)] \geq f_{W_2}[\mathbf{u}]$ and $f_{W_2}[\mathbf{u}, do(W_1 = w_-)] \leq f_{W_2}[\mathbf{u}]$.

We can then invert this property to obtain an explanation mould. In doing so, we introduce slightly stricter conditions to ensure that influencing variables that have no effect on influenced variables do not constitute both an attack and a support, a phenomenon which we believe would be counter-intuitive from an explanation viewpoint.

[1] Here, we ignore the intrinsic basic strength of arguments used in the formal definition in [26].

[2] With an abuse of notation we use the same symbol for all pre-orders.

Definition 4. Given a causal model $\langle U, V, E\rangle$ such that, for each $W_i \in U \cup V$, the domain $\mathscr{D}(W_i)$ is equipped with a pre-order $<$, and given its corresponding influence graph $\langle \mathscr{V}, \mathscr{I}\rangle$, a reinforcement explanation mould is an explanation mould $\{c_-, c_+\}$ such that, given some $\mathbf{u} \in \mathscr{U}$ and $(W_1, W_2) \in \mathscr{I}$, letting $w_1 = f_{W_1}[\mathbf{u}]$:

- $c_-(\mathbf{u}, (W_1, W_2)) = \top$ iff:
  1. $\forall w_+ \in \mathscr{D}(W_1)$ such that $w_+ > w_1$, it holds that $f_{W_2}[\mathbf{u}, do(W_1 = w_+)] \leq f_{W_2}[\mathbf{u}]$;
  2. $\forall w_- \in \mathscr{D}(W_1)$ such that $w_- < w_1$, it holds that $f_{W_2}[\mathbf{u}, do(W_1 = w_-)] \geq f_{W_2}[\mathbf{u}]$;
  3. $\exists_{\geq 1} w_+ \in \mathscr{D}(W_1)$ or $\exists_{\geq 1} w_- \in \mathscr{D}(W_1)$ satisfying strictly the inequality conditions in points 1 and 2 above.
- $c_+(\mathbf{u}, (W_1, W_2)) = \top$ iff:
  1. $\forall w_+ \in \mathscr{D}(W_1)$ such that $w_+ > w_1$, it holds that $f_{W_2}[\mathbf{u}, do(W_1 = w_+)] \geq f_{W_2}[\mathbf{u}]$;
  2. $\forall w_- \in \mathscr{D}(W_1)$ such that $w_- < w_1$, it holds that $f_{W_2}[\mathbf{u}, do(W_1 = w_-)] \leq f_{W_2}[\mathbf{u}]$;
  3. $\exists_{\geq 1} w_+ \in \mathscr{D}(W_1)$ or $\exists_{\geq 1} w_- \in \mathscr{D}(W_1)$ satisfying strictly the inequality conditions in points 1 and 2 above.

We call any argumentative explanation resulting from the explanation mould $\{c_-, c_+\}$ a reinforcement explanation (RX).

Note that, as for generic argumentative explanations, we do not commit in general to any choice of $\mathscr{A}$ in RXs.

Proposition 1. Any RX satisfies causal reinforcement.

Proof. Follows directly from the definition of Property 1 and Definition 4. □

The satisfaction of the property of causal reinforcement indicates how RXs could be used counterfactually, given that the results of changes to the variables' values on influenced variables are guaranteed. For example, if a user is looking to increase an influenced variable's value, supporters (attackers) indicate variables whose values should be increased (decreased, respectively). In the following sections, we will explore the potential of this capability when causal models provide abstractions of classifiers whose output needs explaining.

5. Reinforcement Explanations for Classification

In this section, we instantiate causal models for two different AI models commonly used for classification in the literature and preliminarily discuss a potential use of RXs in this context.

The two classification models that we use to instantiate causal models are Bayesian network classifiers (BCs) and classifiers built from feed-forward neural networks (NNs). Given some assignments to input variables I (from the variables' domains), these classifiers can be seen as determining the most likely value for classification variables, which, in

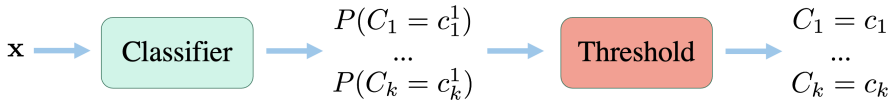

Figure 2: A schematic view of classification by BCs and NNs. We assume C = $\{C_1, \ldots, C_k\}$, for $k \geq 1$, with each C_i a binary classification variable, with values c_i^1 and c_i^0, such that $P(C_i = c_i^0) = 1 - P(C_i = c_i^1)$; c_i is the value for C_i whose probability P exceeds the threshold (θ). Assuming that the threshold is suitably chosen so that c_i is uniquely defined for each C_i, the classifier can be equated to the function $\mathscr{M}$ such that $\mathscr{M}(\mathbf{x}) = (c_1 \ldots, c_k)$.

this paper, we assume to be binary, in a given set C. Thus, the classification task may be seen as a mapping $\mathscr{M}(\mathbf{x})$ returning, for assignment $\mathbf{x}$ to input variables, either 1 or 0 (for the classification variables in C) depending on whether the probability exceeds a given threshold θ. We summarise the classification process in Figure 2. Note that the choice of threshold is crucial to guarantee that a single value c_i is determined by $\mathscr{M}$ for each classification variable C_i: if θ is too high, then no value may be computed, whereas if θ is too low, the probability of both values may exceed it. Note also that, in the case of NNs, the probabilities may result from using, e.g., a softmax activation for the output layer. Furthermore, note that for the purposes of this paper, the underpinning details of these classifiers and how they can be obtained are irrelevant and will be ignored. In other words, we treat the classifier as a black-box, as standard in much of the XAI literature, and explain its outputs in terms of its inputs.

We represent the classification task by a (naive) BC or by a NN with the following causal model:

Definition 5. A causal model for a naive BC or classifier built from a NN is a causal model $\langle U_C, V_C, E_C \rangle$, where:

- U_C consists of the input variables I of the classifier, with their respective domains;
- V_C = C such that, for each $C_i \in$ C, $\mathscr{D}(C_i) = \{c_i^1, c_i^0\}$;
- E_C corresponds to the computation of the probability values $P(C_i = c_i^1))$ by the classifier (see Figure 2).

$\mathscr{I}_C = U_C \times V_C$ represents the influences in the causal model for the classifier; these are such that the exogenous variables U_C are densely connected to the endogenous variables V_C.

In line with our assumptions for RXs, we assume that the variables' domains are equipped with a pre-order.

On this basis, we envisage the following use of RXs as actionable explanations in contexts where the user has a classification goal to reach and has control on (some of) the input variables.

Given a situation with an undesired classification outcome (e.g. a rejected loan application) and an explanation indicating the relevant attackers and supporters, if a user would like to decrease the probability of the current classification, s/he would look

to increase (decrease) the value of the corresponding variable's attackers (supporters, respectively), in line with Property 1.

A broader investigation of the possible uses of RXs is left to future work.

6. Related Work

The role of causality within explanations for AI models has received increasing attention of late. [2] define a framework for determining the causal effects between features and predictions using a variational autoencoder. The detection of causal relations and explanations between arguments within text has also proven effective within NLP [36]. [3] give causal explanations for NNs in that they train a separate NN by masking features to determine causal relations (in the original NN) from the features to the classifications. Generative causal explanations of black box classifiers [37] are built by learning the latent factors involved in a classification, which are then included in a causal model. [38] take a different approach, proposing a general framework for constructing structural causal models with deep learning components, allowing tractable counterfactual inference. Other approaches towards explaining NNs, e.g., [39, 40], take into account causal relations when calculating features' attribution values for explanation. Meanwhile, [4] introduce causal explanations for reinforcement learning models based on [5].

Computational argumentation has been widely used in the literature as a mechanism for explaining AI models, from data-driven explanations of classifiers' outputs [41], powered by AA-CBR [42], to the explanation of the PageRank algorithm [43] via bipolar AFs [17]. The outputs of Bayesian networks have been explained by SAFs [16], while decision-making [44] and scheduling [45] have also been targeted. Property-driven explanations based on bipolar [20] and tripolar [35] AFs have been extracted for recommendations, where the properties driving the extraction are defined in the orthodox manner (with respect to the resulting frameworks), rather than inversions thereof, as we propose. Other forms of argumentation have also proven effective in providing explanations for recommender systems [14], decision making [46] and planning [47].

Various works have explored the links between causality and argumentation. [48] shows that a propositional argumentation system in a full classical language is equivalent to a causal reasoning system, while [49] develops a formal theory combining "causal stories" and evidential arguments. Somewhat similarly to us, [50] present a method for extracting argumentative explanations for the outputs of causal models. However, their method requires more information than the causal model alone, namely, ontological links, and the argumentation supplements the rule-based explanations, rather than being the main constituent, as is the case in our approach.

7. Conclusions

We have introduced a novel approach for extracting AFs from causal models in order to explain the latter's outputs. We have shown how explanation moulds can be defined for particular explanatory requirements in order to generate argumentative explanations.

We focused, in particular, on inverting the existing property of argumentation semantics of bi-variate reinforcement to create an explanation mould, before demonstrating how the resulting reinforcement explanations (RXs) can be used to explain causal models representing different machine-learning-based classifiers.

One of the most promising aspects of this preliminary work is the vast array of directions for future investigation it suggests. First, an experimental validation of the proposed ideas and a comparison with other explanation approaches on a sufficiently large variety of case studies is needed.

Clearly, the wide-ranging applicability of causal models broadens the scope of explanation moulds and argumentative explanations well beyond machine learning models, and we plan to undertake an investigation into other contexts in which they may be useful, for example for decision support in healthcare.

We also plan to study inversions of different properties of argumentation semantics and different forms of AFs to understand their potential, e.g. counting for AAFs [51]. Within the context of explaining machine learning models, we plan to assess RXs' suitability for different data structures and different classifiers, considering in particular deeper explanations, e.g. including influences amongst input variables and/or intermediate, in addition to input and output, variables, in the spirit of [52, 25]. This may be aided by the deployment of methods for the extraction of more sophisticated causal models from classifiers, e.g., [53] for NNs.

Finally, while we posit that, when properly defined, the meaning and explanatory role of the dialectical relations can be rather intuitive at a general level, providing effective explanations to users through AFs will require the investigation of proper presentation and visualization methods, possibly tailored to users' competences and goals and to different application domains.

Acknowledgments

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). Further, Russo was supported by UK Research and Innovation [grant number EP/S023356/1], in the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence (www.safeandtrustedai.org). Finally, 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, and may differ, in particular, from the views and opinions expressed by J.P. Morgan or its affiliates. This material is not a product of the Research Department of J.P. Morgan Securities LLC. This material should not be construed as an individual recommendation for any particular client and is not intended as a recommendation of particular securities, financial instruments or strategies for a particular client. This material does not constitute a solicitation or offer in any jurisdiction.

References

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

[2] D. Alvarez-Melis, T. S. Jaakkola, A causal framework for explaining the predictions of black-box sequence-to-sequence models, in: Proc. EMNLP, 2017, pp. 412–421.

[3] P. Schwab, W. Karlen, CXPlain: Causal explanations for model interpretation under uncertainty, in: Proc. NeurIPS, 2019, pp. 10220–10230.

[4] P. Madumal, T. Miller, L. Sonenberg, F. Vetere, Explainable reinforcement learning through a causal lens, in: Proc. AAAI, 2020, pp. 2493–2500.

[5] J. Y. Halpern, J. Pearl, Causes and explanations: A structural-model approach: Part 1: Causes, in: UAI, 2001, pp. 194–202.

[6] J. Woodward, Explanation, invariance, and intervention, Philosophy of Science 64 (1997) S26–S41.

[7] M. M. A. de Graaf, B. F. Malle, How people explain action (and autonomous intelligent systems should too), in: Proc. AAAI Fall Symposia, 2017, pp. 19–26.

[8] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38.

[9] M. T. Ribeiro, S. Singh, C. Guestrin, "why should I trust you?": Explaining the predictions of any classifier, in: Proc. ACM SIGKDD, 2016, pp. 1135–1144.

[10] S. M. Lundberg, S. Lee, A unified approach to interpreting model predictions, in: Proc. NeurIPS, 2017, pp. 4765–4774.

[11] A. Ignatiev, Towards trustable explainable AI, in: Proc. IJCAI, 2020, pp. 5154–5158.

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

[13] P. Baroni, D. Gabbay, M. Giacomin, L. van der Torre (Eds.), Handbook of Formal Argumentation, College Publications, 2018.

[14] J. C. Teze, L. Godo, G. R. Simari, An argumentative recommendation approach based on contextual aspects, in: Proc. SUM, 2018, pp. 405–412.

[15] A. Dejl, P. He, P. Mangal, H. Mohsin, B. Surdu, E. Voinea, E. Albini, P. Lertvittayakumjorn, A. Rago, F. Toni, Argflow: A toolkit for deep argumentative explanations for neural networks, in: Proc. AAMAS, 2021.

[16] S. T. Timmer, J. C. Meyer, H. Prakken, S. Renooij, B. Verheij, Explaining Bayesian networks using argumentation, in: Proc. ECSQARU, 2015, pp. 83–92.

[17] E. Albini, P. Baroni, A. Rago, F. Toni, PageRank as an argumentation semantics, in: Proc. COMMA, 2020, pp. 55–66.

[18] C. Antaki, I. Leudar, Explaining in conversation: Towards an argument model, Europ. J. of Social Psychology 22 (1992) 181–194.

[19] P. Madumal, T. Miller, L. Sonenberg, F. Vetere, A grounded interaction protocol for explainable artificial intelligence, in: Proc. AAMAS, 2019, pp. 1033–1041.

[20] A. Rago, O. Cocarascu, C. Bechlivanidis, F. Toni, Argumentation as a framework for interactive explanations for recommendations, in: Proc. KR, 2020, pp. 805–815.

[21] K. Cyras, A. Rago, E. Albini, P. Baroni, F. Toni, Argumentative XAI: A survey, in: Proc. IJCAI, 2021, pp. 4392–4399.
[22] A. Vassiliades, N. Bassiliades, T. Patkos, Argumentation and Explainable Artificial Intelligence: A Survey, Knowledge Eng. Rev. 36 (2021).
[23] D. M. Gabbay, Logical foundations for bipolar and tripolar argumentation networks: preliminary results, J. Log. Comput. 26 (2016) 247–292.
[24] P. Baroni, G. Comini, A. Rago, F. Toni, Abstract games of argumentation strategy and game-theoretical argument strength, in: Proc. PRIMA, 2017, pp. 403–419.
[25] E. Albini, A. Rago, P. Baroni, F. Toni, Relation-based counterfactual explanations for bayesian network classifiers, in: Proc. IJCAI, 2020, pp. 451–457.
[26] L. Amgoud, J. Ben-Naim, Weighted bipolar argumentation graphs: Axioms and semantics, in: Proc. IJCAI, 2018, pp. 5194–5198.
[27] C. Cayrol, M.-C. Lagasquie-Schiex, On the acceptability of arguments in bipolar argumentation frameworks, in: Proc. ECSQARU, 2005, pp. 378–389.
[28] J. Pearl, Reasoning with cause and effect, in: Proc. IJCAI, 1999, pp. 1437–1449.
[29] J. Pearl, The do-calculus revisited, in: Proc. UAI, 2012, pp. 3–11.
[30] 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) 321–358.
[31] L. Amgoud, J. Ben-Naim, Evaluation of arguments from support relations: Axioms and semantics, in: Proc. IJCAI, 2016, pp. 900–906.
[32] P. Baroni, A. Rago, F. Toni, How many properties do we need for gradual argumentation?, in: Proc. AAAI, 2018, pp. 1736–1743.
[33] J. Pearl, Causal diagrams for empirical research, Biometrika 82 (1995) 669–710.
[34] N. Potyka, Interpreting neural networks as quantitative argumentation frameworks, in: Proc. AAAI, 2021, pp. 6463–6470.
[35] A. Rago, O. Cocarascu, F. Toni, Argumentation-based recommendations: Fantastic explanations and how to find them, ????
[36] Y. Son, N. Bayas, H. A. Schwartz, Causal explanation analysis on social media, in: Proc. EMNLP, 2018, pp. 3350–3359.
[37] M. R. O'Shaughnessy, G. Canal, M. Connor, C. Rozell, M. A. Davenport, Generative causal explanations of black-box classifiers, in: Proc. NeurIPS, 2020.
[38] N. Pawlowski, D. C. de Castro, B. Glocker, Deep structural causal models for tractable counterfactual inference, in: Proc. NeurIPS, 2020.
[39] A. Chattopadhyay, P. Manupriya, A. Sarkar, V. N. Balasubramanian, Neural network attributions: A causal perspective, in: Proc. ICML, 2019, pp. 981–990.
[40] T. Heskes, E. Sijben, I. G. Bucur, T. Claassen, Causal Shapley values: Exploiting causal knowledge to explain individual predictions of complex models, in: Proc. NeurIPS, 2020.
[41] O. Cocarascu, A. Stylianou, K. Cyras, F. Toni, Data-empowered argumentation for dialectically explainable predictions, in: Proc. ECAI, 2020, pp. 2449–2456.
[42] K. Cyras, K. Satoh, F. Toni, Abstract argumentation for case-based reasoning, in: Proc. KR, 2016, pp. 549–552.
[43] L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank Citation Ranking:

Bringing Order to the Web, WWW: Internet and Web Inf. Syst. 54 (1998) 1–17.
[44] L. Amgoud, H. Prade, Using arguments for making and explaining decisions, Artificial Intelligence 173 (2009) 413–436.
[45] K. Cyras, D. Letsios, R. Misener, F. Toni, Argumentation for explainable scheduling, in: Proc. AAAI, 2019, pp. 2752–2759.
[46] Q. Zhong, X. Fan, X. Luo, F. Toni, An explainable multi-attribute decision model based on argumentation, Exp. Syst. Appl. 117 (2019) 42–61.
[47] N. Oren, K. van Deemter, W. W. Vasconcelos, Argument-based plan explanation, in: Knowledge Engineering Tools and Techniques for AI Planning, Springer, 2020, pp. 173–188.
[48] A. Bochman, Propositional argumentation and causal reasoning, in: Proc. IJCAI, 2005, pp. 388–393.
[49] F. Bex, An integrated theory of causal stories and evidential arguments, in: Proc. ICAIL, 2015, pp. 13–22.
[50] P. Besnard, M. Cordier, Y. Moinard, Arguments using ontological and causal knowledge, in: Proc. FoIKS, 2014, pp. 79–96.
[51] L. Amgoud, J. Ben-Naim, Axiomatic foundations of acceptability semantics, in: Proc. KR, 2016, pp. 2–11.
[52] C. Olah, A. Satyanarayan, I. Johnson, S. Carter, L. Schubert, K. Ye, A. Mordvintsev, The building blocks of interpretability, Distill 3 (2018) e10.
[53] T. Kyono, Y. Zhang, M. van der Schaar, CASTLE: regularization via auxiliary causal graph discovery, in: Proc. NeurIPS, 2020.