Explaining Classifiers' Outputs with Causal Models and Argumentation

Antonio Rago, Fabrizio Russo, Emanuele Albini and Francesca Toni
Department of Computing, Imperial College London, UK
{antonio, fabrizio, emanuele, ft}@imperial.ac.uk

Pietro Baroni
Dipartimento di Ingegneria dell'Informazione, Università degli Studi di Brescia, Italy
pietro.baroni@unibs.it

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. We then evaluate our method empirically when the causal models represent (Bayesian and neural network) machine learning models for classification. The results show advantages over a popular approach from the literature, both in highlighting specific relationships between feature and classification variables and in generating counterfactual explanations with respect to a commonly used metric.

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] and causal models have been advocated as "explanations or understanding of how data are generated" [7]. Furthermore, some see causal reasoning as underpinning how humans explain to one another [8]. Also, research from the social sciences [9] 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. Given that "looking at how humans explain to each other can serve as a useful starting point for explanation in AI" [9], it does makes sense to draw explanations for AI models from causal models. However, it is also broadly understood that different users may need different forms of explanations [10], taking into account their cognitive abilities, their background and their specific goals when seeking explanations of AI systems, and work within the social sciences clearly points to humans favouring seemingly non-causal forms of explanations in some contexts, in particular: "the majority of what might look like causal attributions turn out to look like *argumentative* claim-backings"[11], and "people use *reasons* to explain or justify decisions already taken and beliefs already held" [12].

Meanwhile, *computational argumentation* (see [13, 14] 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 [15], classifiers [16], Bayesian networks [17] and *PageRank* [18]. Furthermore, 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 format and application (see [21, 22] for recent surveys), the links between causality and argumentative explanations have remained largely unexplored to date. In this paper, we aim to fill this gap and bring causality and argumentation together to support the XAI vision, focusing on the explanation of outputs of machine learning classifiers.

Specifically, 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 covering the most relevant work in the literature (Section 2) and giving the necessary background (Section 3), 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 (Section 4). Then (in Section 5) 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 (Section 6) how the extracted bipolar AFs may be used as counterfactual explanations for the outputs of causal models representing different classification methods. We then provide an empirical assessment of these explanations (Section 7), demonstrating how they can provide some important insights on the differences between different models' functionalities, while outperforming a popular approach from the literature along a counterfactual metric. Finally, we conclude, indicating potentially fruitful future work (Section 8).

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.
- We evaluate our proposal empirically: although preliminary, this evaluation shows promise and indicates directions for future work.

This work extends [28, 29] significantly, with Section 7 being completely new and the other sections being extended and improved.

2 Related Work

A dominant approach for model-agnostic explainability of AI models is the use of *feature attribution* methods, which assign a *signed* value to each feature (in input) to represent their importance towards the output of a classification model, for each of the inputs. LIME [30] and SHAP [31] are popular attribution methods, using different techniques to assess each feature's importance by measuring the outcome of changes to inputs. In a nutshell, LIME is based on sampling *perturbations* of the reference input, while SHAP is based on the notion of Shapley values from game theory, assessing the effect of the presence of a feature when added to all possible sets of other features (in practice a sampling over the possible *permutations* of features is used for an approximate evaluation since the exact calculation would be too costly for large sets of features). Alternatively, another model-agnostic approach is the use of *counterfactuals*, e.g. as in [32, 33, 34], in which a modified input which would result in the change in the classification is given. In the literature, feature attribution methods have been used to generate counterfactual explanations [35], and vice versa [36]. Various studies [37, 38, 39] have have highlighted how feature attribution methods (including SHAP) are often mis-interpreted and overly trusted. In line with [9], we regard counterfactual explanations as some of the most useful for understanding model behaviour. Hence, in this work, we analyse feature attribution explanations in a counterfactual manner as a baseline against which we assess our approach, demonstrating the advantages of incorporating causal information to explanations.

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 [40]. [3] give causal explanations for neural networks (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 [41] are built by learning the latent factors involved in a classification, which are then included in a causal model. [42] 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., [43, 44], take into account causal relations when calculating features' attribution values for explanation. Meanwhile, [4] introduce causal explanations for reinforcement learning models based on [5]. We take a different approach, drawing argumentative explanations from causal models.

Computational argumentation has been widely used in the literature as a mechanism for explaining AI models, from data-driven explanations of classifiers' outputs [45], powered by AA-CBR [46], to the explanation of the *PageRank* algorithm [47] via bipolar AFs [18]. The outputs of Bayesian networks have been explained by SAFs [17], while decision-making [48] and scheduling [49] have also been targeted. Property-driven explanations based on bipolar [20] and tripolar [50] 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 [15], decision making [51] and planning [52]. Our proposal in this paper adds to this line of work providing novel forms of argumentative explanations, but drawn from causal models.

Various works have explored the links between causality and argumentation. [53] shows that a propositional argumentation system in a full classical language is equivalent to a causal reasoning system, while [54] develops a formal theory combining "causal stories" and evidential arguments. Somewhat similarly to us, [55] 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.

Despite the clear potential of causality towards XAI, many of the approaches for generating explanations for AI models have neglected causality as a potential drive for explainability. Some of the most popular methods, as discussed earlier, are heuristic and model-agnostic [30, 31], and, although 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 [56]. On the other hand, while causal models provide the raw material for explanation, the latter is not limited to the selection of a set of appropriate causes [57]. We aim to address these problems by delivering explanations to users which are directly driven by, but not limited to, causal models themselves.

3 Background

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

Causal models. A *causal model* [58] 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 $\mathcal{D}(W_i)$;
- 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' *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* $W_i \in U \cup V$*,* $\mathcal{D}(W_i) = \{\top, \bot\}$*. Figure 1i (we ignore Figure 1ii for the moment: this will be discussed later in Section 5) visualises the variables' parents, and Table 1 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* $V_2 = V_1$*). 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*.*

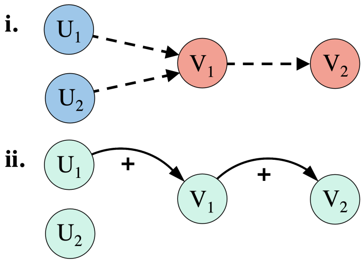

Figure 1: (i) Variables and parents for Example 1, with parents indicated by dashed arrows (for example $\{U_1, U_2\} = PA(V_1)$, i.e. U_1 and U_2 are the parents of V_1). (ii) SAF explanation (see Section 4) for the assignment to exogenous variables $\mathbf{u} \in \mathcal{U}$ such that $f_{U_1}[\mathbf{u}] = \top$ and $f_{U_2}[\mathbf{u}] = \top$.

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).

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 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 $\mathbf{u} \in \mathcal{U}$ 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 an endogenous variable, it will be the value dictated by the structural equations in the causal model.

We use the *do* operator [59] 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 $\langle \mathcal{A}, \mathcal{R}_1, \ldots, \mathcal{R}_l \rangle$, with $\mathcal{A}$ a set (of *arguments*), $l > 0$ and $\mathcal{R}_i \subseteq \mathcal{A} \times \mathcal{A}$, for $i \in \{1, \ldots, l\}$, (binary and directed) *dialectical relations* between arguments [23, 24]. In the abstract argumentation [60] 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) [60], with $l = 1$ (and $\mathcal{R}_1$ a dialectical relation of *attack*, referred to later as $\mathcal{R}_-$), *support AFs* (SAFs) [61], with $l = 1$ (and $\mathcal{R}_1$ a dialectical relation of *support*, referred to later as $\mathcal{R}_+$), and *bipolar AFs* (BAFs) [27], 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 $\mathcal{R}_+$).

The meaning of AFs (including the intended dialectical role of the relations) may be given in terms of *gradual semantics* (e.g. see [24, 62] for BAFs), defined, for AFs with arguments $\mathcal{A}$, 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 AFs may be guided by *properties* that the mappings σ should satisfy (e.g. as in [26, 62]). We will utilise, in Section 5, a variant of the property of *bi-variate reinforcement* for BAFs from [26].

4 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 $\langle \mathcal{V}, \mathcal{I}\rangle$ with:*

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

Note that, while straightforward, the concept of influence graph (closely related to the notion of causal diagram [63]) 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 $\langle \mathcal{V}, \mathcal{I}\rangle$, an* explanation mould *is a non-empty set:*

$$\{c_1, \dots, c_m\}$$

where for all $i \in \{1, \dots, m\}$, $c_i : \mathcal{U} \times \mathcal{I} \to \{\top, \bot\}$ is a relation characterisation*, in the form of a Boolean condition expressed in some formal language. Given some $\mathbf{u} \in \mathcal{U}$ and $(W_1, W_2) \in \mathcal{I}$, 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 Section 3). 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* $(W_1, W_2) \in \mathcal{I}$*:*

$$\begin{aligned}
&c_s(\mathbf{u}, (W_1, W_2)) = \top \textit{ iff}\\
&(f_{W_1}[\mathbf{u}] = \top \wedge f_{W_2}[\mathbf{u}] = \top \wedge f_{W_2}[\mathbf{u}, do(W_1 = \bot)] = \bot) \vee\\
&(f_{W_1}[\mathbf{u}] = \bot \wedge f_{W_2}[\mathbf{u}] = \bot \wedge f_{W_2}[\mathbf{u}, do(W_1 = \top)] = \top).
\end{aligned}$$

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* $(U_1, V_1) \in \mathcal{I}$ *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* $(U_2, V_1) \in \mathcal{I}$ *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 Section 3) 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 $\langle \mathcal{V}, \mathcal{I}\rangle$, 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*
- *$\mathcal{R}_1, \ldots, \mathcal{R}_m \subseteq \mathcal{I} \cap (\mathcal{A} \times \mathcal{A})$ such that, for any $i = 1 \ldots m$, $\mathcal{R}_i = \{(W_1, W_2) \in \mathcal{I} \cap (\mathcal{A} \times \mathcal{A}) | c_i(\mathbf{u}, (W_1, W_2)) = \top\}$.*

Note 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).

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 $f_{U_2}[\mathbf{u}] = \bot$, 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 $\mathcal{A}$:*

- *the SAF in Figure 1ii, where every variable is an argument;*
- *the SAF with the same support relation but U_2 excluded from $\mathcal{A}$, as it is not "involved" and thus does not contribute 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.*

5 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) [64], using the underlying computation of the MLPs as a gradual semantics, and to the proposals to explain recommender systems (RSs) via tripolar AFs [50] 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

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

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

- *if $(W_1, W_2) \in \mathcal{R}_-$, 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 $(W_1, W_2) \in \mathcal{R}_+$, 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.

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

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

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

3. $\exists_{\geq 1} w_+ \in \mathcal{D}(W_1)$ *or* $\exists_{\geq 1} w_- \in \mathcal{D}(W_1)$ *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 $\mathcal{A}$ 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.

6 Reinforcement Explanations for Classification

In this section, we first instantiate causal models for two families of classifiers commonly used in the literature. We then demonstrate how RXs can be used to explain these classifiers in a counterfactual manner, supplementing their structure with weights on the relations, which allows RXs to be compared with *feature attribution* methods (see Section 2).

The two families of classifiers that we use to instantiate causal models are Bayesian network classifiers (BCs) and classifiers built from feed-forward NNs. Given some assignments to *input variables* $\mathbf{I}$ (from the variables' domains), these classifiers can be seen as determining the most likely value for *classification variables*, which, in this paper, we assume to be binary, in a given set $\mathbf{C}$. Thus, the classification task may be seen as a mapping $\mathcal{M}(\mathbf{x})$ returning, for assignment $\mathbf{x}$ to input variables, either 1 or 0 (for the classification variables in $\mathbf{C}$) depending on whether the probability exceeds a given threshold θ. We summarise the classification process in Figure 2. Note 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

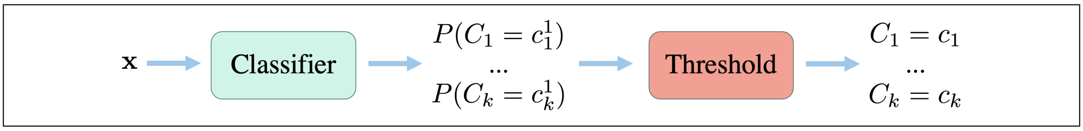

Figure 2: A schematic view of classification by BCs and NNs. We assume $\mathbf{C} = \{C_1, \ldots, C_k\}$, for $k \geq 1$, with each C_i a binary classification variable, with values c_i^1 and c_i^0, such that $P(C_i = c_i^0) = 1 - P(C_i = c_i^1)$; c_i is the value for C_i whose probability P exceeds 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 $\mathcal{M}$ such that $\mathcal{M}(\mathbf{x}) = (c_1 \ldots, c_k)$.

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

$\mathcal{I}_C = U_C \times V_C$ 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.

As discussed in Section 2, the purpose of feature attribution methods is to assign a signed importance value to each feature for a given input. Our motivation for this work is to explore an alternative direction, namely to interpret changes in outcomes with a *causal* lens and produce explanations that follow human intuition when presented to the users, while still maintaining feature attribution methods' goal of characterising the impact of each feature on a classification.

We aim to characterise (and rank) features based on their potential to change the outcome of the model. Our ingredients are: (i) The model outcome for the example to be explained in the form of a probability; (ii) A function to select the direction of change in the domain of the variable intervened; (iii) Interventions over the domain of the variables and the change in model probability resulting from them.

To arrive at the formulation for the counterfactual feature importance we propose, we introduce three functions that will help us scan features for their "counterfactual capabilities". They all refer to generic input variable U_j and classification variable C_i, for a given realisation $\mathbf{u}$ of the input variables. Note that for every variable $U_j \in U_C$, we assume that $\mathcal{D}(U_j)$ is finite and totally ordered and for each $u \in \mathcal{D}(U_j)$ we denote as $pos(u) \geq 1$ the natural number corresponding to its position in the ordering.

Potential Change in Outcome quantifies the change in probability of C_i given an intervention assigning the value u' to U_j:

$$\Delta f_{\mathbf{u},u'}^{(U_j,C_i)} = |f_{C_i}[\mathbf{u}|do(U_j = u')] - f_{C_i}[\mathbf{u}]|.$$

Relation Sign Function returns a positive or negative sign depending on the type of relation between U_j and C_i:

$$\delta(U_j, C_i, \mathbf{u}) = \begin{cases} 1 & \text{if } c_+(\mathbf{u}, (U_j, C_i)) = \top \\ -1 & \text{if } c_-(\mathbf{u}, (U_j, C_i)) = \top \\ 0 & \text{otherwise} \end{cases}$$

Domain Subset Function selects the subset of the domain of U_j to be considered to achieve a change in the classification outcome of C_i with respect to the one given by $\mathbf{u}$. The selection takes into account the threshold θ and the relation sign function δ:

$$\gamma(U_j,C_i,\mathbf{u}) = \begin{cases} \{u' \in \mathcal{D}(U_j) | u' > f_{U_j}[\mathbf{u}]\} & \text{if } (f_{C_i}[\mathbf{u}] - \theta) * \delta(U_j,C_i,\mathbf{u}) < 0 \\ \{u' \in \mathcal{D}(U_j) | u' < f_{U_j}[\mathbf{u}]\} & \text{if } (f_{C_i}[\mathbf{u}] - \theta) * \delta(U_j,C_i,\mathbf{u}) > 0 \\ \emptyset & \text{if } (f_{C_i}[\mathbf{u}] - \theta) * \delta(U_j,C_i,\mathbf{u}) = 0 \end{cases}$$

The idea is that the function γ selects the possible values of U_j which are greater than the current one in $\mathbf{u}$ in two cases: $f_{C_i}[\mathbf{u}]$ is above the threshold and U_j is an attacker; $f_{C_i}[\mathbf{u}]$ is below the threshold and U_j is a supporter. Analogously, γ selects the possible values of U_j which are lower than the current one in $\mathbf{u}$ in two cases: $f_{C_i}[\mathbf{u}]$ is above the threshold and U_j is a supporter; $f_{C_i}[\mathbf{u}]$ is below the threshold and U_j is an attacker.

On this basis, we formulate in the following our notion of counterfactual feature importance.

Counterfactual Importance ranks the input features based on the amount of change in probability that a value close to the current one can bring, provided that it produces a change in classification:

$$\omega(U_j, C_i, \mathbf{u}) = \sum_{u' \in \gamma(U_j, C_i, \mathbf{u})} \frac{\Delta f^{(U_j, C_i)}_{\mathbf{u}, u'} * \mathbb{1}((\theta - f_{C_i}[\mathbf{u}|do(U_j = u')]) \cdot (\theta - f_{C_i}[\mathbf{u}]) < 0)}{|pos(u') - pos(f_{U_j}[\mathbf{u}])|} \tag{1}$$

where $\mathbb{1}()$ is the indicator function taking value 1 if the expression in brackets is true and 0 otherwise. Note also that we assume by convention that $\omega(U_j, C_i, \mathbf{u}) = 0$ when $\gamma(U_j, C_i, \mathbf{u}) = \emptyset$.

The rationale behind the formulation is as follows: The sum includes a term for every possible value u' that can be used for an intervention on U_j coherently with the expected direction of change (these values are returned by $\gamma(U_j, C_i, \mathbf{u})$). Each of these values contributes to the sum proportionally to the potential change in probability of C_i (namely $\Delta f^{(U_j, C_i)}_{\mathbf{u}, u'}$) but only if it causes a change in the final classification, i.e. if the threshold is crossed in the desired direction (i.e. the difference between θ and f_{C_i} changes sign). Therefore, the indicator function filters the *"wanted"* changes and the interventions not producing a change are disregarded. Moreover, each of these terms is weighted according to the distance of u' from the current values of U_j: the greater the distance, the greater the denominator, the lower the contribution to the importance. This will improve the ranking of the variables that produce *actionable* changes, which are closest to the current input $\mathbf{u}$.

In representing classifiers as causal models and generating importance values for the relations of the resulting RXs, we are now able to directly compare RXs experimentally with feature attribution methods.

7 Experimental Evaluation

In this section we provide an empirical evaluation of our approach, focusing our evaluation on the property of causal reinforcement for RXs. The main research questions we aim to address are:

1. Can the attacks and supports in RXs be put in correspondence with positive and negative, respectively, *polarity* in feature attribution techniques?

2. Can relation importance in RXs be put in correspondence with the *magnitude* of the values associated to features in feature attribution techniques?

To answer both questions we compare RXs with a prominent feature attribution technique (i.e. SHAP [31], where, for the experiments, we use version 0.35.0 of the publicly available SHAP library). Concretely, we use SHAP in two ways: in Section 7.1 to extract *reasons for and against* classifications by classifiers (in comparison with supports and attacks in RXs); and in Section 7.2 as a way to determine reasons' importance as (the absolute values of) feature attribution values computed by SHAP (in comparison with our notion of relation importance). The sign of these feature attribution values is used to determine the sign of the reasons themselves.

In our experiments we use two publicly available datasets (FICO [65] and COMPAS [66]) and two different models, a naive BC and a NN, in line with Section 5. We implement the naive BCs using the scikit-learn implementation and the NNs using CASTLE [67]. For both datasets, there is a single, binary classification variable. We discretised continuous features using equally-sized bins limiting them to a maximum of 10 for FICO. For COMPAS, we used the existing variable domains, given that the variables are discrete (with a minimum of two values and a maximum of 17 values). Also, since Definition 4 and the definition of importance work under the assumption that variables' domains are ordered, a random ordering was generated for all variables with no inherent order. Some comments on the effect of this arbitrary ordering will be provided later. Additional details on the datasets are given in Table 2. Here, we can see how in the FICO dataset all features are continuous (and thus their domain is equipped with a natural total order) while in COMPAS 50% of the features lack an inherent order. We will show the consequences of this difference between the datasets in the results.

For each of the datasets, we trained a Naive BC and a NNs with 1 hidden layer and 32 hidden neurons. We trained the NN for a maximum of 200 epochs and with learning rate of 0.0005 and patience on the validation loss of 50 epochs. The naive BCs were fitted using Laplace estimation from the training set with $\alpha = 0.1$. Classification metrics for the two types of models on the two datasets (when trained on 75% of the samples and tested on the remaining 25%) are reported in Table 3. Note how the different models have similar performances on the same dataset. Note also that model performance optimisation was not the focus of this work and that we kept models as standard as possible.

7.1 Causal Reinforcement Analysis

In order to understand whether our RXs are able to handle different models while also unveiling differences in the way RXs operate when compared with SHAP, we measured: the prevalence of relations (i.e. the percentage of occurrences for each method) and agreement (between the two methods).

		FICO	COMPAS
	Number of samples	10,458	6,950
	Number of features	23	12
Size of Domain	Minimum	4	2
	Average	7	4.6
	Maximum	10	17
	Number of ordinal features	23	6
	% of ordinal features	100%	50%

Table 2: Dataset details. The number of samples for the dataset is the total. The number of features does not include the target classification variable. The size of the domains for the two datasets consist of deciles (where enough data were available) for continuous features and the original categories for categorical features. The number and % of ordinal features represents the features with a natural ordering, e.g. continuous, or with naturally ordered categories.

	FICO		COMPAS	
(*)	NN	NBC	NN	NBC
ROC-AUC	0.783	0.771	0.78	0.79
Accuracy	71.7%	71.9%	70.5%	71.6%
F1 Score	71.6%	71.9%	70.2%	71.5%
Precision	71.7%	71.9%	70.7%	72.6%
Recall	71.6%	71.8%	69.6%	70.5%

Table 3: Performances of the models. (*) NN (Neural Network) or NBC (Naive Bayesian network Classifier).

Prevalence of relations. We extracted RXs and SHAP explanations for all samples in the testing part of the two datasets and measured: for RXs, the percentage of influences in the causal models for the two models contributing attacks and supports, and, for SHAP, the percentage of negative and positive reasons.

The results are shown in Table 4. We note that there are large discrepancies across models and types of explanations for each of the two datasets, in contrast with similar performances by the classifiers (see Table 3). This is somewhat not surprising, as it could be a consequence of very different workings by the (very different) models to obtain classifications, and provides part of the motivation for the experiments in Section 7.2 to verify faithfulness of the explanations to the models, counterfactually. We also note that the total percentages of negative and positive attribution values established by SHAP are greater for FICO than for COMPAS, while the total percentages of influences that become part of the attack and support relations in RXs are considerably higher for NNs than NBCs independently of the dataset. This reflects the inner workings of the two models: NNs leverage the orderings over variables' domains since they assign weights that get multiplied with the value of the input variable, whose ordering (its value) has, therefore, a big influence on the final output. BCs on the other hand, mostly disregard these orderings since they calculate the probability of the classification variables for specific values of the input variables if categorical, or for a group/bucket of values, if numeric. BCs will therefore disregard ordering within the bucket, while across buckets the only link to the original ordering could come through the conditional probabilities, with a much less direct effect given that the value of the variable would be modified according to the class frequency in that band. In the case of the FICO dataset, whose continuous variables are all equipped with a natural ordering, RXs result in larger attack and support relations than for BCs, whereas in COMPAS, where some variables have been artificially and arbitrarily ordered to obtain RXs, the difference in relation size across the models is not so dramatic, somewhat confirming the expected dependence of RXs on the existence of natural orderings. Table 4 gives insight into the interactions between data, model, and RXs. For the FICO data, where all variables are numeric and hence have natural ordering, the difference between the amount of relations identified in NN and BC is much more significant than in the case of COMPAS (for FICO the difference is between 87.3% and 28.3%, while for COMPAS it is between 77.3% and 64.2%). This is to say that NN does a better job at leveraging numeric variables and shows an increased power to extract RXs that reflect the model behaviour for a given dataset, noting also that RXs need natural ordering to work at their best. NN does not support the extraction of many more relations than BCs for the COMPAS data instead, since there are not many natural orderings to leverage in the first place. Note that we do not assume that the larger the number of relations extracted the better. Instead, what we deem important is that the relationships that the model actually finds in the data are extracted for explanatory purposes. Investigation of this from different angles is provided in the following sections, highlighting how RXs are very effective at representing models that have extracted relationships from ordered variables in the data. Concerning the split between positive and negative reasons for SHAP, there seems to be a clear dominance of the former across datasets and models, but no clear pattern emerges for supports and attacks in RXs. We note though there are discrepancies in the +/- splits across the two different explanation methods, showing that they work differently and begging for further exploration of faithfulness in Section 7.2.

Agreement between RXs and SHAP. We also conducted a finer-grained analysis of the differences between the two forms of explanation, focusing on how many influences/reasons with opposite sign the two methods extract and on extracted influences/reasons versus ignored ones. Table 5 shows the results.

We note that SHAP and RXs agree less than 40% of the time for FICO and less than 20% for COMPAS. To understand this, we firstly looked at the cases where the models were establishing relations/reasons of opposite sign (Strong Disagree, i.e. + vs -) and noticed that this happened around 50% of the time for FICO NN and only 10% of the time for FICO NBC. Of course, this is a consequence of the number of extracted influences/reasons overall for this dataset, as seen in Table 4. Still, the amount of strong disagreement is quite high, but it does make intuitive sense when we think about the inner workings of the two explanation methods: for SHAP, a positive reason means that the current value of the corresponding variable is in favour of the current model output; according to Definition 4, instead, supporting variables are those whose values above the current one increase the probability of the value of the target classification variable (to be explained). In other words, our causal reinforcement definition focuses on the projection of possible changes to a variable that are guaranteed to have the expected behaviour on the target. At a general level this shows that apparently simple and superficially similar explanations elements may actually allow quite different interpretations. In our case, the generic idea of positive and negative influence can correspond to instances with significantly different meanings. Conveying the correct meaning to the users is obviously a crucial and nontrivial issue in this respect. Since we are assuming a context where users ascribe a counterfactual meaning to explanations, this observation brought us to the set of experiments in the next section, where we analyse the usefulness of Definition 4 for counterfactual purposes.

7.2 Causal Reinforcement for Counterfactuals

The second set of experiments assesses how we can apply Definition 4 to extract intuitive and actionable counterfactual behaviour from our models. One method for providing such an assessment is to compare with attribution methods functioning as counterfactual explanation methods, e.g. as in [68], a set-up which we use, along with the importance measure defined in Section 6. In doing so we evaluate the counterfactual nature of our explanations (see the relevant discussion in Section 2).

Again, we consider the same models and datasets, in comparison to SHAP, but this time we focus on applying interventions to input variables and observing the change in the models' outputs (or classification). To do this we couple Definition 4 with the counterfactual feature importance ω from (1) of how important the relations

		FICO		**COMPAS**	
		NN	NBC	NN	NBC
SHAP	−	30.3%	22.2%	21.9%	22.3%
	+	46.1%	66.1%	40.6%	40.3%
	Total	**76.4%**	**88.2%**	**62.5%**	**62.5%**
RXs	−	43.8%	10.9%	47.5%	40.6%
	+	43.5%	17.4%	29.7%	23.6%
	Total	**87.3%**	**28.3%**	**77.3%**	**64.2%**

Table 4: Prevalence of relations. Here + and − indicate, respectively, support and attack relations in RXs and positive and negative attribution values in SHAP. Totals do not sum up to 100% given that there can be influences/features that the methods do not extract.

	FICO		**COMPAS**	
RXs vs SHAP	NN	NBC	NN	NBC
Strong Disagree	52.9%	10.1%	30.2%	21.1%
Weak Disagree	47.1%	89.9%	69.8%	78.9%
Disagree	**60.2%**	**69.8%**	**84%**	**92.8%**
Agree	39.8%	30.2%	16%	7.2%

Table 5: Relation Agreement Summary. The 'Strong Disagree' row looks at influences/reasons that both RXs and SHAP extract, but with opposite signs (+ vs -, as per caption of Table 4). The 'Weak Disagree' row looks at the influences/reasons that one method extracts while the other does not (+ or - vs influences/reasons not extracted). 'Strong' and 'Weak Disagree' sum up to 100% and split the total of disagreements shown in the 'Disagree' row, while the 'Agree' row gives the remainders, i.e. the extracted influences/reasons with the same sign across explanation methods.

established by the models are. Concretely, we used the absolute value of $\omega(U_j,C_i,\mathbf{u})$ to select the input variables U_j to change in order to achieve a change in classification C_i (counterfactual output).

Definition 4 is useful in selecting the direction of change, given the current classification. Given that all input variables have categorical domains in this setting (after discretisation), we had to choose how many steps to move away from the current value u of U_j. We focused first on the most actionable change recommendations that the receiver of a model decision and explanation could want. Hence we analysed the change in classification for setting u' one step away from its current value $f_{U_j}[\mathbf{u}]$ (i.e. $|pos(u') - pos(f_{U_j}[\mathbf{u}])| = 1$). We did the same for SHAP. For both methods, we changed the sets of the most important features, increasing their size from 1 to 5 (Top $U_j = 1, \ldots, 5$, respectively) according to either SHAP or RXs.

The results are presented in Figure 3. In the FICO-NN setting, where the monotonic relationships are strong and well captured by the model, RXs perform well and outperform SHAP in all scenarios. In particular, a higher number of classification changes is achieved when allowing a greater number of variables to be changed while this does not happen in the case of SHAP. For FICO-NBC the situation is less clear-cut, though it can be observed that RXs do better than SHAP in the case where only one step away from the current value is allowed. It can be argued that this case is the most actionable and therefore relevant counterfactually. For COMPAS, RXs perform worse than SHAP in most cases. This again is expected given the mix of purely categorical and ordinal features in the data as well as the lower average number of categories. For the not naturally ordered variables we had to enforce a random ordering for the purposes of this tests, and this has evidently had an impact.

8 Conclusions & Future Work

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. We then performed an empirical evaluation of RXs, analysing the differences between the relations in RXs and the reasons for and against classification produced by the popular SHAP method [31]. We also introduced a preliminary measure of importance over the relations in RXs and used it to assess the counterfactuality of RXs. A deeper investigation on the notion of importance at a general level and the study of further, possibly more appropriate, definitions of this measure represent an important direction of future work.

Our preliminary empirical evaluation suggests that our approach outperforms SHAP in the cases where the conditions for its applicability are satisfied, and provides the basis for discussing the suitability of different approaches in different contexts. Our results also highlight the need for different explanation mechanisms depending on the users' needs. For instance, actionable explanations, concerning

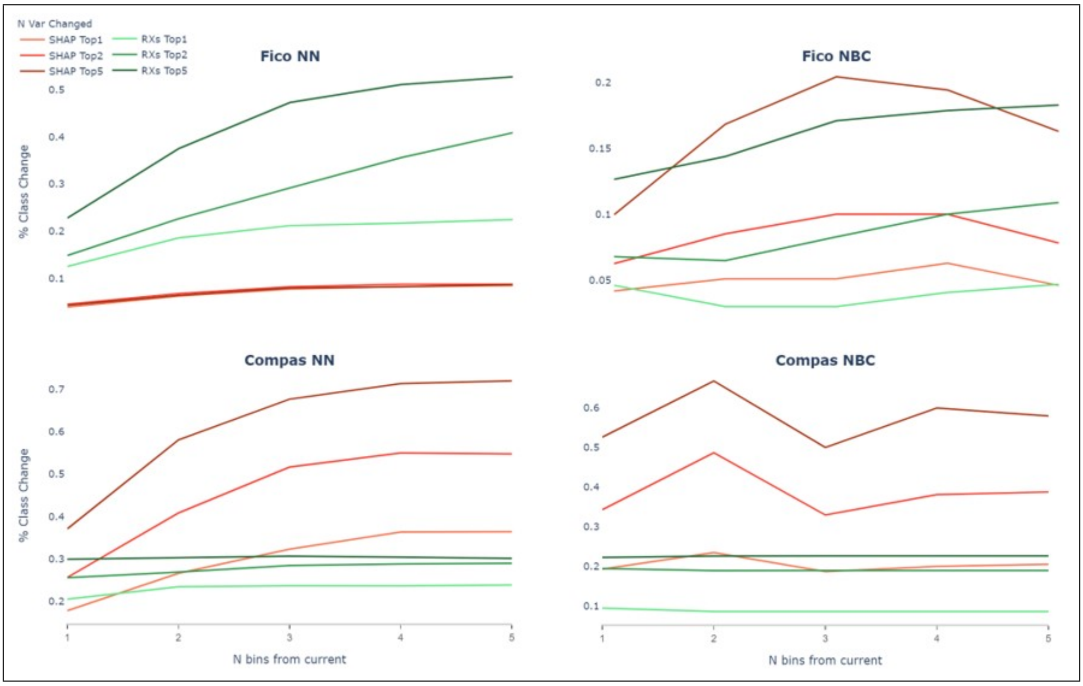

Figure 3: Proportion of successful counterfactual classification changes achieved by number of input variables changed (Top 1 to 5). The x axis represents the number of bins away from the current value i.e. distance from current position ($|pos(u') - pos(f_{U_j}[\mathbf{u}])|$) for each changed input variable U_j. The different shades of green are for changing the one, two and five most important variables for RXs, while the reds are for SHAP.

how to change the input of a model to get a different output, may not fit feature attribution techniques, and, in general, a one-size-fits-all approach to explanations cannot achieve this.

One of the most promising aspects of our work is the vast array of directions for future work it suggests. 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 [69]. 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 [70, 25]. This may be aided by the deployment of methods for the extraction of more sophisticated causal models from classifiers, e.g., [67] 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). Baroni was partially supported by the research project GNCS-INdAM CUP_E55F22000270001. 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 (5) (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] J. Pearl, Causality, Cambridge university press, 2009.

[8] 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.

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

[10] V. Arya, R. K. E. Bellamy, P. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilovic, S. Mourad, P. Pedemonte, R. Raghavendra, J. T. Richards, P. Sattigeri, K. Shanmugam, M. Singh, K. R. Varshney, D. Wei, Y. Zhang, AI explainability 360: Impact and design, in: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022, AAAI Press, 2022, pp. 12651–12657.
URL https://ojs.aaai.org/index.php/AAAI/article/view/21540

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

[12] H. Mercier, D. Sperber, The Enigma of Reason – A New Theory of Human Understanding, Penguin Books, 2018.

[13] 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 (3) (2017) 25–36.

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

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

[16] 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, pp. 1761–1763.

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

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

[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 (2) (2021).

[23] D. M. Gabbay, Logical foundations for bipolar and tripolar argumentation networks: preliminary results, J. Log. Comput. 26 (1) (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] A. Rago, F. Russo, E. Albini, P. Baroni, F. Toni, Forging argumentative explanations from causal models, in: Proceedings of the 5th Workshop on Advances in Argumentation in Artificial Intelligence 2021 co-located with the 20th International Conference of the Italian Association for Artificial Intelligence (AIxIA 2021), Milan, Italy, November 29th, 2021, 2021.
URL http://ceur-ws.org/Vol-3086/paper3.pdf

[29] A. Rago, P. Baroni, F. Toni, Explaining causal models with argumentation: the case of bi-variate reinforcement, in: Proceedings of the 19th International Conference on Principles of Knowledge Representation and Reasoning, KR 2022, Haifa, Israel. July 31 - August 5, 2022, 2022.
URL https://proceedings.kr.org/2022/52/

[30] 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.

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

[32] S. Wachter, B. D. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the GDPR, CoRR abs/1711.00399 (2017). `arXiv:1711.00399`.
URL http://arxiv.org/abs/1711.00399

[33] R. K. Mothilal, A. Sharma, C. Tan, Explaining machine learning classifiers through diverse counterfactual explanations, in: FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020, 2020, pp. 607–617. `doi:10.1145/3351095.3372850`.
URL https://doi.org/10.1145/3351095.3372850

[34] K. Kanamori, T. Takagi, K. Kobayashi, Y. Ike, K. Uemura, H. Arimura, Ordered counterfactual explanation by mixed-integer linear optimization, in: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on

Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, 2021, pp. 11564–11574.
URL https://ojs.aaai.org/index.php/AAAI/article/view/17376

[35] Y. Ramon, D. Martens, F. J. Provost, T. Evgeniou, A comparison of instance-level counterfactual explanation algorithms for behavioral and textual data: Sedc, LIME-C and SHAP-C, Adv. Data Anal. Classif. 14 (4) (2020) 801–819. doi:10.1007/s11634-020-00418-3.
URL https://doi.org/10.1007/s11634-020-00418-3

[36] E. Albini, J. Long, D. Dervovic, D. Magazzeni, Counterfactual Shapley additive explanations, in: FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022, 2022, pp. 1054–1070. doi:10.1145/3531146.3533168.
URL https://doi.org/10.1145/3531146.3533168

[37] I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, S. A. Friedler, Problems with Shapley-value-based explanations as feature importance measures, in: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, Vol. 119 of Proceedings of Machine Learning Research, PMLR, 2020, pp. 5491–5500.
URL http://proceedings.mlr.press/v119/kumar20e.html

[38] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. M. Wallach, J. W. Vaughan, Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning, in: R. Bernhaupt, F. F. Mueller, D. Verweij, J. Andres, J. McGrenere, A. Cockburn, I. Avellino, A. Goguey, P. Bjøn, S. Zhao, B. P. Samson, R. Kocielnik (Eds.), CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020, ACM, 2020, pp. 1–14. doi:10.1145/3313831.3376219.
URL https://doi.org/10.1145/3313831.3376219

[39] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. F. Moura, P. Eckersley, Explainable machine learning in deployment, in: M. Hildebrandt, C. Castillo, L. E. Celis, S. Ruggieri, L. Taylor, G. Zanfir-Fortuna (Eds.), FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020, ACM, 2020, pp. 648–657. doi:10.1145/3351095.3375624.
URL https://doi.org/10.1145/3351095.3375624

[40] Y. Son, N. Bayas, H. A. Schwartz, Causal explanation analysis on social media, in: Proc. EMNLP, 2018, pp. 3350–3359.

[41] M. R. O'Shaughnessy, G. Canal, M. Connor, C. Rozell, M. A. Davenport, Generative causal explanations of black-box classifiers, in: Proc. NeurIPS, 2020.

[42] N. Pawlowski, D. C. de Castro, B. Glocker, Deep structural causal models for tractable counterfactual inference, in: Proc. NeurIPS, 2020.

[43] A. Chattopadhyay, P. Manupriya, A. Sarkar, V. N. Balasubramanian, Neural network attributions: A causal perspective, in: Proc. ICML, 2019, pp. 981–990.

[44] 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.

[45] O. Cocarascu, A. Stylianou, K. Cyras, F. Toni, Data-empowered argumentation for dialectically explainable predictions, in: Proc. ECAI, 2020, pp. 2449–2456.

[46] K. Cyras, K. Satoh, F. Toni, Abstract argumentation for case-based reasoning, in: Proc. KR, 2016, pp. 549–552.

[47] L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web, WWW: Internet and Web Inf. Syst. 54 (1999-66) (1998) 1–17.

[48] L. Amgoud, H. Prade, Using arguments for making and explaining decisions, Artificial Intelligence 173 (3-4) (2009) 413–436.

[49] K. Cyras, D. Letsios, R. Misener, F. Toni, Argumentation for explainable scheduling, in: Proc. AAAI, 2019, pp. 2752–2759.

[50] A. Rago, O. Cocarascu, F. Toni, Argumentation-based recommendations: Fantastic explanations and how to find them, in: Proc. IJCAI, 2018, pp. 1949–1955.

[51] Q. Zhong, X. Fan, X. Luo, F. Toni, An explainable multi-attribute decision model based on argumentation, Exp. Syst. Appl. 117 (2019) 42–61.

[52] 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.

[53] A. Bochman, Propositional argumentation and causal reasoning, in: Proc. IJCAI, 2005, pp. 388–393.

[54] F. Bex, An integrated theory of causal stories and evidential arguments, in: Proc. ICAIL, 2015, pp. 13–22.

[55] P. Besnard, M. Cordier, Y. Moinard, Arguments using ontological and causal knowledge, in: Proc. FoIKS, 2014, pp. 79–96.

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

[57] M. J. Nathan, Causation vs. causal explanation: Which is more fundamental?, Foundations of Science (2020). doi:https://doi.org/10.1007/s10699-020-09672-2.

[58] J. Pearl, Reasoning with cause and effect, in: Proc. IJCAI, 1999, pp. 1437–1449.

[59] J. Pearl, The do-calculus revisited, in: Proc. UAI, 2012, pp. 3–11.

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

[61] L. Amgoud, J. Ben-Naim, Evaluation of arguments from support relations: Axioms and semantics, in: Proc. IJCAI, 2016, pp. 900–906.

[62] P. Baroni, A. Rago, F. Toni, How many properties do we need for gradual argumentation?, in: Proc. AAAI, 2018, pp. 1736–1743.

[63] J. Pearl, Causal diagrams for empirical research, Biometrika 82 (4) (1995) 669–710.

[64] N. Potyka, Interpreting neural networks as quantitative argumentation frameworks, in: Proc. AAAI, 2021, pp. 6463–6470.

[65] FICO, Fico xml challenge found at community.fico.com/s/xml (2017).

URL https://community.fico.com/s/explainable-machine-learning-challenge

[66] D. S. ProPublica, Compas recidivism risk score data and analysis (2016). URL https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analys

[67] T. Kyono, Y. Zhang, M. van der Schaar, CASTLE: regularization via auxiliary causal graph discovery, in: Proc. NeurIPS, 2020.

[68] A. White, A. d'Avila Garcez, Measurable counterfactual local explanations for any classifier, in: Proc. ECAI, 2020, pp. 2529–2535.

[69] L. Amgoud, J. Ben-Naim, Axiomatic foundations of acceptability semantics, in: Proc. KR, 2016, pp. 2–11.

[70] C. Olah, A. Satyanarayan, I. Johnson, S. Carter, L. Schubert, K. Ye, A. Mordvintsev, The building blocks of interpretability, Distill 3 (3) (2018) e10.

Received 26 May 2022