Arguments Based on Domain Rules in Prediction Justifications

Joeri Peters[1,2,*], Floris Bex[1,3] and Henry Prakken[1]

[1]Utrecht University, Utrecht, The Netherlands
[2]Netherlands National Police, Driebergen, The Netherlands
[3]Tilburg University, Tilburg, The Netherlands

Abstract

Ensuring the interpretability of trained machine learning models is often paramount, particularly in high-stakes domains such as counter-terrorism and other forms of law enforcement. Post hoc techniques have emerged as a promising avenue for justifying the predictions of complex models. However, while these approaches provide valuable insights, they often lack the ability to directly reference familiar domain rules and make use of these rules to guide explanations. This paper introduces a method for incorporating arguments about the applicability of domain rules in justifying classifier predictions.

Keywords

Case-Based Argumentation, Precedential Constraint, Explainable AI, Domain Knowledge

1. Introduction

This paper is concerned with explainability in machine learning (ML). Specifically, we focus on enhancing the explainable artificial intelligence (XAI [12]) approach known as '*a fortiori* case-based argumentation' (AF-CBA [17]). AF-CBA justifies binary classification predictions using the theory of precedential constraint [10], that is, referencing precedential cases from a case base constructed from training (or historical [15]) data. Our goal is to extend this framework by incorporating domain rules, recognising that domain-specific knowledge plays a pivotal role in decision-making processes.

ML models are often regarded as 'black boxes' when their opacity is high, whether due to relative complexity or proprietary protection [11, 8]. Neural networks serve as a typical example of intricate models that have revolutionised predictive accuracy at the cost of increased opacity. Transparency and explainability concerns become particularly critical in high-stakes domains, such as law enforcement, where decisions may carry significant consequences for individuals or court cases. Predictions have to be highly accurate—possibly necessitating opaque models—yet explainable. Post hoc approaches like AF-CBA are aimed at solving this problem by justifying ML predictions 'after the fact', meaning that the approach does not access the ML model itself and is therefore model agnostic. In our experience, the need for such an approach arises relatively frequently in practice. ML models can be inaccessible at the moment an explanation is required or the type of explanation it can offer is too technical for the intended users, thereby rendering it a black box. Furthermore, there can be situations when the performance metrics of an interpretable alternative to black box approaches is deemed unsatisfactory, necessitating a post hoc solution. AF-CBA produces such justifications on the basis of earlier cases (*precedents*).

Applicable scenarios can be drawn from the domain of counter-terrorism, where ML classifiers can be used to quickly yet objectively distinguish between two outcomes. For example, there may be a need to decide whether a particular incident is the responsibility of a specific terrorist organisation, judging by the *modus operandi* and objectives of its members. Another binary categorisation could be between the incident forming a part of a large-scale coordinated attack and it being a 'lone-wolf' incident, the

CMNA'24: The 24th International Workshop on Computational Models of Natural Argument, September 17, 2024, Hagen, Germany
*Corresponding author.
j.g.t.peters@uu.nl (J. Peters); f.j.bex@uu.nl (F. Bex); h.prakken@uu.nl (H. Prakken)
0009-0009-1493-9872 (J. Peters); 0000-0002-5699-9656 (F. Bex); 0000-0002-3431-7757 (H. Prakken)

outcome of which warrants a different police response. As a running example, we adopt the scenario in which officials seek to determine whether a violent event should be classified as an act of terrorism. It is realistic that a classifier should be used in order to facilitate quick yet valid judgement in such a situation, to avoid responders acting on gut feeling alone. However, the number of applicable precedents is relatively low and much contextual knowledge is involved in making this decision. Hence, the system should be transparently constrained by experts' knowledge of this domain. Our approach is in line with a tradition of viewing rule- and case-based reasoning as complementary. For instance, the two were combined in an overall architecture by Golding & Rosenbloom [7] to allow the latter to produce analogies in order to handle exceptions to the (incomplete) rule set. A similar integration of rules and cases was used by Rissland & Skalak [18] in their CABARET system, aimed at an area of income tax law. Our goal, however, is to let AF-CBA make use of and refer to such domain knowledge in its justifications of the predicted outcomes of a black-box model.

The rest of this paper is structured as follows. We introduce our XAI approach in Section 2 before considering how to incorporate domain rules in Section 3. Finally, we discuss conclusions and future work directions in Section 4.

2. Preliminaries

In justifying binary class labels, the predictions of a classifier trained on labelled data during its training phase can be likened to court decisions based on judicial precedents. In this vein, Prakken & Ratsma [17] propose a top-level model, afterwards dubbed AF-CBA, drawing on AI & Law research and utilising case-based argumentation inspired by Horty's model of *a fortiori* reasoning [9]. AF-CBA is influenced by CATO [1] and work by Čyras et al. [4, 3, 5]. Contrary to [3], AF-CBA is not its own explainable classification approach, but a post hoc approach used to justify the classification predictions of another ML model.

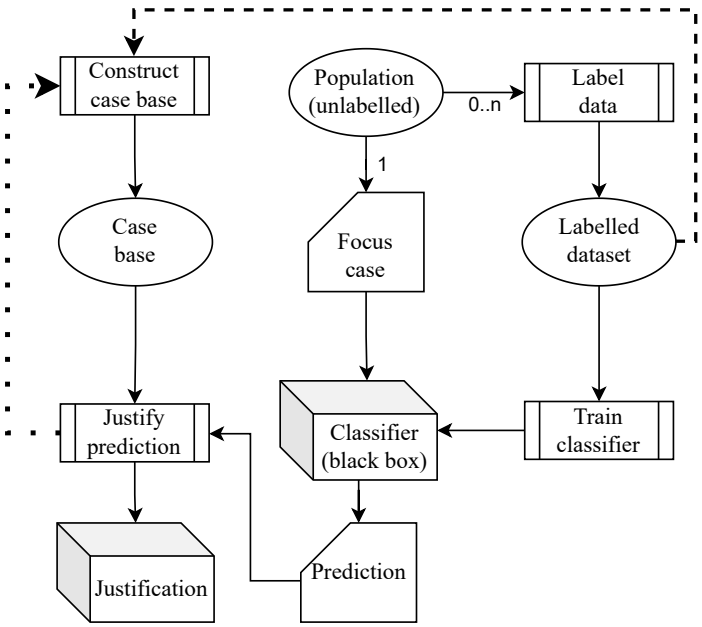

Figure 1: A schematic depiction of AF-CBA's workflow. The case base is constructed either instantly on the basis of the labelled data (dashed line) or stepwise on the basis of earlier predictions (dotted line, as in [15]).

The context of AF-CBA is illustrated in Figure 1. A labelled dataset constitutes a random sample from the overall population, to which annotators or decision-makers assign labels, and on which a classifier is trained (supervised ML). A *focus case* represents a single, random sample from the same population, and the classifier assigns a predicted outcome to it. Due to the black-box nature of the classifier, it lacks the capability to explain the rationale behind the prediction. AF-CBA addresses this limitation by utilising

either the labelled set or an archive of previous case predictions [15] as a case base, engaging in an argument game between a *proponent* and an *opponent* of the predicted outcome. In this argument game, cases which are similar to the focus case are cited in order to argue that the focus case ought to receive the same outcome. The decision is *forced* when no relevant differences exist between the focus case and the precedent. Moves in the argumentation game follow Dung's abstract argumentation framework [6], with the game modelled on grounded semantics [16]. The notion of precedential constraint is that a focus case ought to receive the same outcome as a precedential case if any differences between those cases only serve to strengthen the focus case for that particular outcome. A winning strategy for the proponent is then presented as a justification for the predicted outcome in the form of an argument graph.

An abstract argument framework (AF), introduced by Dung [6], consists of a pair $AF = \langle A, attack \rangle$, where A represents a set of arguments, and $attack$ is a binary relation on A. A subset B of A is termed *conflict-free* if no arguments in B attacks arguments in B and *admissible* if it is both conflict-free and capable of defending itself against attacks. In other words, if an argument A_1 is in B, and some argument A_2 in A attacks A_1, then some argument in B must attack A_2. There are different types of admissible sets, known as *extensions*. We focus on the *grounded extension*, which has the additional properties that it contains all arguments it defends and is subset-minimal for these conditions.

Formally, a *case* in the *case base* (CB) comprises an *outcome* and a *fact situation*. The case's outcome is a binary label, denoted as o or o'. Variables s and $\bar{s}$ represent the two sides, such that $s = o$ if $\bar{s} = o'$ and vice versa. The fact situation includes *dimensions* (features), where each dimension is a tuple $d = (V, \leq_o, \leq_{o'})$. The tuple consists of a value set V and two partial orderings on V, $\leq_o$ and $\leq_{o'}$, such that $v \leq_o v'$ if and only if $v' \leq_{o'} v$ for $v, v' \in V$. Each dimension has a *tendency*, with a positive tendency indicating a higher value is associated with one outcome (e.g., 1 or $true$), and vice versa for the other. The tendency is sometimes given explicitly, that is: d_i^+ or d_i^-. A value assignment, represented as (d, v), signifies the value x of dimension d in case $c \in CB$ as $v(d, c) = x$. The collective value assignments for all dimensions d in the non-empty set D form a fact situation denoted as F. We assume that two fact situations pertain to the same set D. Defining a case as $c = (F, outcome(c))$ where $outcome(c) \in \{o, o'\}$, the fact situation of case c can be expressed as $F(c)$.

When assessing two fact situations, one may find that one case is 'stronger' or 'better' for a specific outcome than the other. The outcome of a focus case is considered forced if there exists a precedent in the CB with the same outcome, and all differences between the focus case and that precedent serve to strengthen the focus case for that very outcome [10].

Definition 1 (Preference relation for fact situations)**.** *Given two fact situations F and F', $F \leq_s F'$ iff $v \leq_s v'$ for all $(d, v) \in F$ and $(d, v') \in F'$.*

Definition 2 (Precedential constraint)**.** *Given case base CB and fact situation F, deciding F for s is forced iff CB contains a case $c = (F', s)$ such that $F' \leq_s F$.*

A fact situation could be forced for both outcomes o and o' by different precedents, in which case we can speak of an inconsistent CB:

Definition 3 (Case base consistency)**.** *A case base CB is consistent iff it does not contain two cases $c = (F, s)$ and $c' = (F', \bar{s})$ such that $F \leq_s F'$. Otherwise it is inconsistent.*

A *best precedent* has the same outcome as the focus case and as few as possible *relevant differences*. Multiple cases can meet these criteria.

Definition 4 (Differences between cases)**.** *Let $c = (F(c), outcome(c))$ and $f = (F(f), outcome(f))$ be two cases. The set $D(c, f)$ of differences between c and f is $D(c, f) = \{(d, v) \in F(c) \mid v(d, c) \not\leq_{outcome(f)} v(d, f)\}$.*

Definition 5 (Best precedent)**.** *Let $c = (F(c), outcome(c))$ and $f = (F(f), outcome(f))$ be two cases, where $c \in CB$ and $f \notin CB$. c is a best precedent for f iff:*

- $outcome(c) = outcome(f)$ *and*

- *there is no $c' \in CB$ such that $outcome(c') = outcome(c)$ and $D(c', f) \subset D(c, f)$.*

The two players argue about differences between the focus case and precedents from the CB. The proponent does so in favour of the focus case's predicted outcome and the opponent to the contrary. The proponent starts by citing a best precedent. The opponent aims to respond to the proponent's initial citation by either presenting a counterexample or making a distinguishing move $Worse(c,x)$ (the focus case is inferior to precedent c in dimensions x). A distinguishing move can be countered with a compensation move $Compensates(c,y,x)$ (dimensions y make up for the shortcomings in dimensions x compared to precedent c). Finally, there is the transformation move $Transformed(c,c')$ (the citation can be transformed into a case where $D(c,f) = \emptyset$). The proponent can respond using these moves, then the opponent can do the same in turn, and this back-and-forth continues until the opponent cannot make any more moves. Note that y in $Compensates(c,y,x)$ can be the empty set. This is intended to guarantee the possibility of using a compensation move, ensuring the existence of a winning strategy for the proponent and thus that of a justification for the focus case's predicted outcome.

Definition 6 outlines the argumentation framework. The compensation move utilises the set *sc*, containing compensation definitions. The specifics and structure of *sc* were intentionally left open by Prakken & Ratsma [17]. In the most straightforward scenario, *sc* serves as a partial ordering on dimensions, indicating, for example, when a high value for d_i compensates for a low value for d_j. Essentially, *sc* imparts specific domain knowledge. In this paper, we employ the set *sc* to explicitly introduce domain rules into the framework for use by the compensation move.

Definition 6 (Case-based argumentation framework)**.** *Given a case base CB, a focus case $f \notin CB$, and definitions of compensation sc, an abstract argumentation framework AF is a pair $< \mathscr{A}, attack >$, where:*

- $\mathscr{A} = CB \cup M$,
  with $M = \{Worse(c,x) \mid c \in CB,\ x \neq \emptyset$ *and* $x = \{(d,v) \in F(f) \mid v(d,f) <_{outcome(f)} v(d,c)\}\} \cup$
  $\{Compensates(c,y,x) \mid c \in CB,\ y \subseteq \{(d,v) \in F(f) \mid v(d,c) <_{outcome(f)} v(d,f)\},$
  $x = \{(d,v) \in F(f) \mid v(d,f) <_{outcome(f)} v(d,c)\}$ *and y compensates x according to sc*$\} \cup$
  $\{Transformed(c,c') \mid c \in CB$ *and c can be transformed into* c' *and* $D(c',f) = \emptyset\}$
- *A attacks B iff:*
  - $A, B \in CB$ *and* $outcome(A) \neq outcome(B)$ *and* $D(B,f) \not\subset D(A,f)$*;*
  - $B \in CB$ *with* $outcome(B) = outcome(f)$ *and A is of the form* $Worse(B,x)$*;*
  - *B is of the form* $Worse(c,x)$ *and A is of the form* $Compensates(c,y,x)$*;*
  - $B \in CB$ *and* $outcome(B) \neq outcome(f)$ *and A is of the form* $Transformed(c,c')$*.*

3. Argument-Based Compensation Moves

The set *sc* is used in Definition 6 as a placeholder for any construct that incorporates domain knowledge of some type—including CATO-like hierarchical relations [1, 19] and perhaps complex ontologies [14]—depending on how the phrase "...*y* compensates *x* according to *sc*..." is interpreted. In this paper, we restrict ourselves to domain rules about fact situations to underpin compensation moves. To this end, we want to construct arguments with conclusions of the form $compensates(c,y,x)$ on the basis of such domain rules. We formalise *sc* as a reasoning system, which takes the form of an argumentation framework $AF_{sc} = \langle A, attack \rangle$ containing arguments constructed by instantiating argumentation schemes based on domain rules [20]. Using AF_{sc}, we want to evaluate which compensation moves are available—that is, which conclusions are part of the grounded extension.

We assume a body of domain rules exists. It may be provided by a single domain expert or multiple, or through some unspecified statistical mechanism like a rule discovery technique. We allow for the possibility that not every single situation relevant to the domain is explicitly covered by these domain rules. Instead, we assume there may be exceptions to these rules which may themselves be informative to the user who interprets the justification of the compensation move.

3.1. Compensation as an Argument Scheme

The application of a domain rule for compensation can be described by Scheme 1, which states that the compensation is the conclusion from the premises that the fact situation is worse for f in dimensions D_w (premise w) while it is better for f in dimensions D_b (premise b) and that D_b is *preferred* over D_w (premise p) according to the *preference relation* $D_w \prec D_b$. The fact that the focus case f has worse values for dimensions in D_w than the precedent c can then be downplayed, because f has better values for dimensions in D_b and those dimensions are deemed more relevant for the outcome of f. Note that 'worse' and 'better' are relative to the tendency of the dimension in question and therefore do not necessarily correspond to 'lower' and 'higher' values, respsectively.

Argumentation Scheme 1 (Compensation)**.** *Let $c \in CB$ be a precedent, f be a focus case, and $D_b, D_w \subseteq D$ two sets of dimensions where $D_b \cap D_w = \emptyset$, then the compensation scheme $COMP(f, c, D_b, D_w)$ is defined as the following reasoning pattern:*

$$\begin{array}{rl} w: & D_w = \{d \in D \mid d(f) <_{outcome(f)} d(c)\} \\ b: & D_b = \{d \in D \mid d(f) >_{outcome(f)} d(c)\} \\ p: & D_w \prec D_b \\ \hline Conc: & compensates(c, D_b, D_w) \end{array}$$

Table 1
Precedent case c and focus case f.

Case	$d^+_{casualties}$	d^-_{weapon}	...	Outcome
c	5	low	...	True
f	10	high	...	True

In Table 1, f has a higher weapon sophistication (d_{weapon}) than c, which is associated with non-terrorist incidents, making f worse than c on this dimension. However, f also has a higher number of casualties ($d_{casualties}$), which is a strong indicator of a terrorist incident. Using the domain rule that a higher number of casualties compensates for a higher weapon sophistication, we instantiate the following argument:

$$\begin{array}{rl} \textbf{COMP}(f, c, D_b, D_w)\textbf{:} & \\ \text{w:} & D_w = \{d_{weapon}\} \\ \text{b:} & D_b = \{d_{casualties}\} \\ \text{p:} & \{d_{weapon}\} \prec \{d_{casualties}\} \\ \hline \text{Conc:} & \text{compensates}(c, \{d_{casualties}\}, \{d_{weapon}\}) \end{array}$$

In this example, the argument states that although f has a higher ('worse') weapon sophistication than c, the higher ('better') number of casualties of f compensates for this, justifying the predicted outcome of *true* for f on the basis of this precedent.

We assume that the fact situations of both the precedents from the CB and the focus case are known. Therefore, we cannot argue against the first two premises of this scheme. Furthermore, the scheme is strict in that the conclusion of this scheme cannot be negated if all its premises are true. But first, we must consider how we know that premise p (the preference relation underpinning the compensation move) is true. In practice, conditions may apply for a preference relation and we will now consider the various forms these conditions may take.

3.2. Conditional Preference Relations

There may be situations where a specific threshold value must be met for a preference relation to be considered to hold through application of by Scheme 1. For instance, the aforementioned relation $\{d_{weapon}\} \prec \{d_{casualties}\}$ may only hold for high numbers of casualties, say at least 4. Fewer casualties may not be considered a good enough reason to compensate for the fact that the highly sophisticated

weapon used in this incident is so irregular. In other words, the premise p of this instance of Scheme 1 depends on the condition that $d_{casualties} \geq 4$. We consider additional examples of conditions below, but for now we summarise the sets of conditions for a preference relation $D_w \prec D_b$ with an abstract premise Ψ.

Argumentation Scheme 2 (Preference)**.** *Let f be a focus case, $s \in \{o, o'\}$, $D_b, D_w \subseteq D$ be two sets of dimensions where $D_b \cap D_w = \emptyset$, Ψ be an abstract placeholder whose thruth value represents whether the preference conditions are fulfilled. Then the preference scheme $PREF(f, D_b, D_w, D)$ is defined as the following reasoning pattern:*

ψ: Ψ *(preference conditions fulfilled)*
==========
Conc: $D_w \prec D_b$

Scheme 2 evaluates whether Ψ holds in a particular instance. If so, the relevant preference relation can be concluded and subsequently used as a premise p in the instantiation of Scheme 1. In Table 1, the focus case f has a 'worse' level of sophistication in the weapon used (d_{weapon}) and a 'better' number of casualties ($d_{casualties}$), with respect to the outcome true. Instantiating Schemes 2 ($PREF(f, D_b, D_w, D)$) and 1 ($COMP(f, c, D_b, D_w)$) lets us construct the following argument:

PREF(f, D_b, D_w, D):
ψ: $d_{casualties}(f) \geq 4$
==========
Conc: $\{d_{weapon}\} \prec \{d_{casualties}\}$

COMP(f, c, D_b, D_w):
w: $D_w = \{d_{weapon}\}$
b: $D_b = \{d_{casualties}\}$
p: $\{d_{casualties}\} \prec \{d_{weapon}\}$

Conc: compensates$(c, \{d_{casualties}\}, \{d_{weapon}\})$

Furthermore, there can be more than one threshold as part of Ψ. We could have a preference relation that states that $d_{casualties}$ in combination with d_{fear} (a numerical expression of public fear) is more relevant than (i.e. is preferred over) d_{weapon} if both dimensions exceed their respective thresholds. We would then instantiate Scheme 2 as:

PREF(f, D_b, D_w, D):
ψ: $d_{casualties}(f) \geq 4 \wedge d_{fear}(f) \geq 10$
==========
Conc: $\{d_{weapon}\} \prec \{d_{casualties}, d_{fear}\}$

Whether certain dimensions surpass certain thresholds is a type of condition that presumes that each dimension must *independently* meet a sub-condition. Alternatively, the condition for a preference relation might hinge on a combination of dimensions, in the form of some evaluation function surpassing a single threshold. For an example of such a rule, we can imagine a preference relation $\{d_{weapon}\} \prec \{d_{casualties}, d_{wounded}\}$ and its condition that $d_{casualties}(f) + d_{wounded}(f) \geq 10$. Here, the evaluation function is the sum of the number of fatal casualties and non-fatally wounded that compensates for a high level of weapon sophistication. In other words, the distinction between fatally and non-fatally harmed victims is of no consequence in this domain rule; what matters is the number of victims.

PREF(f, D_b, D_w, D):
ψ: $d_{casualties}(f) + d_{wounded}(f) \geq 10$
==========
Conc: $\{d_{weapon}\} \prec \{d_{casualties}, d_{wounded}\}$

Alternatively, one can imagine rules in which the difference between the number of perpetrators and victims plays a role in distinguishing terrorist incidents from, say, assassinations. Or perhaps the ratio between wounded and deceased victims modulates the impact of weapon sophistication in some hypothetical rule. A weighted mean of several dimensions may have to surpass a certain value. And so

on, domain experts may have dozens of rules for a domain that is particularly well understood and rich in descriptive dimensions, possibly assisted by some statistical analysis or rule discovery approach. More complex functions are also possible. One could argue that at least some of such evaluations ought to be captured in the feature engineering phase before training a model, rather than in post-hoc justifications; we remind the reader that our approach is model- and data-agnostic, so we should generally support relevant evaluations.

Consider the following scenario: an attack involving a sophisticated bomb (d_{weapon}) that does not result in a high number of casualties ($d_{casualties}$). Under normal circumstances, the sophistication of the weapon might suggest a targeted assassination rather than a terrorist attack. However, if the event generates an exceptionally high level of public fear (d_{fear}), this could compensate for the lower casualty count, as the primary goal of terrorism is often to instil fear and disrupt societal normalcy. In this case, the evaluation function might give significant weight to d_{fear}, such that a weighted sum of d_{fear} and $d_{casualties}$ is compared to a threshold value.

PREF(f, D_b, D_w, D):

ψ: $0.3 \cdot d_{casualties}(f) + 0.7 \cdot d_{fear}(f) \geq 10$

==========

Conc: $\{d_{weapon}\} \prec \{d_{casualties}, d_{fear}\}$

Aforementioned thresholds form conditions on the very dimensions within the preference relation, D_w and D_b. However, there may be situations where contextual factors influence the applicability of the preference relation. For example, the additional dimension $d_{measures}$ (number of security measures in place) might in certain cases modulate the impact of a the number of casualties in compensating for weapon sophistication. In this condition, the threshold value pertains to a dimension that is itself not involved in the preference relation. The conditions for a preference relation can also involve spatiotemporal factors. For instance, the same set of dimensions might have different thresholds or weights depending on whether the event occurs in a region currently experiencing political instability. This adaptability is crucial in counter-terrorism, where the nature of threats and societal impact can change rapidly. When trying to attribute historical incidents to terrorist organisations, one would have to take into account that an organisation was founded at a certain moment in time, or was only active within a particular part of the world. For example, IS (ISIS/ISIL) did not rise to prominence until 2014 in areas of Syria and Iraq. Any domain rule that is concerned with characteristics of IS incidents or public claims by this organisation is likely specific to the appropriate time and place. The same type of corncern applies to the Taliban in Afghanistan before the American invasion in 2001 or their departure in 2021, or the Troubles in Ireland and Great Britain between 1966 and 1998. For a (simplified) example, consider the following, where the fact that an incident takes place during the Troubles in Belfast means that $\{d_{wounded}\}$ compensates for $\{d_{casualties}, d_{weapon}\}$:

PREF(f, D_b, D_w, D):

ψ: $d_{year}(f) = 1969 \wedge d_{location}(f) = \text{Belfast}$

==========

Conc: $\{d_{casualties}, d_{weapon}\} \prec \{d_{wounded}\}$

Alternatively, this particular insight from the domain expert could be used to construct an empty compensation move on the basis of domain knowledge. Note that if $D_b = \emptyset$, Scheme 1 describes the special case of empty compensation. We may want to overlook poor values for $\{d_{casualties}, d_{claims}\}$ out of hand. Normally in AF-CBA, we allow for compensation moves with $D_b = \emptyset$ in order to guarantee a winning strategy (Section 2), as somewhat of an unsatisfactory but necessary default substituting for a more informative justification. With an argument such as the following, we can actually provide expert-informed justifications why values in D_b are not relevant to the outcome of the focus case despite not having any compensating dimensions, making an empty compensation move more valuable than it would otherwise be:

PREF(f, D_b, D_w, D)**:**

ψ: $d_{year}(f) = 1969 \wedge d_{location}(f) = \text{Belfast}$

==========

Conc: $\{d_{casualties}, d_{weapon}\} \prec \emptyset$

COMP(f, c, D_b, D_w)**:**

w: $D_w = \{d_{casualties}, d_{weapon}\}$

b: $D_b = \emptyset$

p: $\{d_{casualties}, d_{weapon}\} \prec \emptyset$

Conc: $\text{compensates}(c, \emptyset, \{d_{casualties}, d_{weapon}\})$

Transitivity (where $\{d_1\} \prec \{d_2\}$ and $\{d_2\} \prec \{d_3\}$ implies $\{d_1\} \prec \{d_3\}$) and antisymmetry (where $\{d_1\} \prec \{d_2\}$ implies $\{d_2\} \not\prec \{d_1\}$) are not generally assumed and depend on the domain. Symmetric preference relations, such as $\{d_{casualties}\} \prec \{d_{weapon}\}$ and $\{d_{weapon}\} \prec \{d_{casualties}\}$, can coexist for the same focus case, indicating that a better value in one dimension can compensate for a worse value in another. For instance, a high number of casualties ($d_{casualties}$) may compensate for high weapon sophistication (d_{weapon}) and vice versa. This symmetry may imply that dimensions are equivalent in their influence on an outcome, acting as proxies for a more abstract notion. For example, d_{alert} (security alerted) and $d_{measures}$ (number of security measures) could be subcategories of a dimension $d_{security}$ (overall security preparedness). This implies a certain kind of equivalence. Thus our approach implicitly allows for the drawing of *abstract parallels* similar to the factor hierarchies in CATO [1].

3.3. Arguing About Preference Relations

As mentioned, we do not assume the body of domain knowledge to be uncontested. While Schemes 1 and 2 provide an approach to assess whether the conditions of a compensation move have been met, exceptions may be possible and premises can be contested. Exactly what kinds of attacks are possible may depend on the domain, but in general we can state that attacks between arguments can be modelled in a structured argumentation framework like ASPIC+ [13] or ABA [2].

For example, a domain expert may consider there to be another caveat for the preference relation $\{d_{casualties}\} \prec \{d_{weapon}\}$ besides $d_{casualties}(f) \geq 4$, namely that it does not hold if the weapon sophistication is extremely high. The abstract placeholder Ψ could then simply refer to two separate thresholds for this instance of $PREF(f, D_b, D_w, D)$:

PREF(f, D_b, D_w, D)**:**

ψ: $d_{casualties}(f) > 4 \wedge d_{weapon}(f) <$ 'Extremely high'

==========

Conc: $\{d_{casualties}\} \prec \{d_{weapon}\}$

However, one might argue that it is more informative if exceptions are modelled explicitly as separate arguments. The preference relation $\{d_{casualties}\} \prec \{d_{weapon}\}$ would then be attacked by an exception argument detailing how it can be concluded from the fact that $d_{weapon}(f) \not<$ 'Extremely high' that $\{d_{casualties}\} \prec \{d_{weapon}\}$ does not hold for f. This exception argument would have to be successfully attacked in order for $\{d_{casualties}\} \prec \{d_{weapon}\}$ to be usable in Scheme 1, perhaps by an exception to the exception. For instance, the exception $d_{weapon}(f) \not<$ 'Extremely high' might be considered irrelevant if the number of casualties is sufficiently high, e.g. $d_{casualties} \geq 30$. This second exception would attack the first exception, thereby defending the preference relation from Scheme 2 and thus reinstating the compensation move from Scheme 1. And so on for additional exceptions. This notion is illustrated in the argument graph Figure 2.

We allow for chains of arguments about preference relations. Whether long, complex arguments are always desirable is up to the domain experts themselves, based on what they deem appropriate for the intended user. Our approach allows them to decide on how elaborately to justify the domain knowledge used to justify ML predictions as they see fit. The goal is always to justify compensation moves in the eyes of the user, who may or may not be a domain expert, in order to provide an appropriate level of

justification for ML predictions.

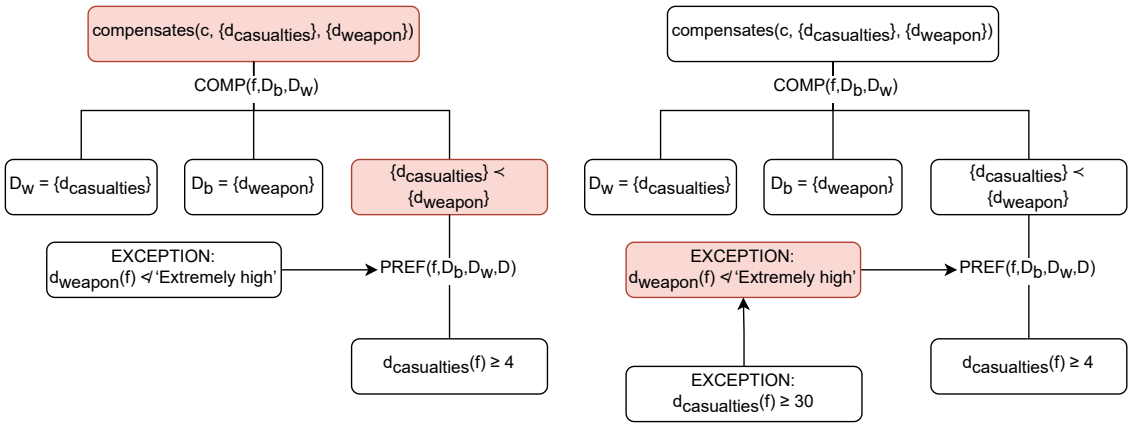

Figure 2: An illustration of a how an exception to an exception can defend a preference relation and thus a compensation move. Shaded boxes are not in the grounded extension, attacks are indicated by arrows.

Other types of argument could be valueable too, such as expert opinions (after Walton et al. [20]) on the basis of additional domain knowledge—based on professional experience, domain literature, or statistical/data analysis (e.g. rule discovery). Of course, there are domain-specific reasons why a preference relation is included in the first place, which suggests that conflicting opinions are possible. Similarly to how we could allow chains of exceptions to argue about preference relations, experts may decide that it is equally informative to explicitly model dissent between experts and perhaps to show how the latest analysis, literature research or most senior expert wins the debate.

For example, the preference relation $\{d_{casualties}\} \prec \{d_{weapon}\}$ could be attacked by an opinion stating that it does not hold, based on the experience of the expert. This could itself be attacked by an opinion on the basis of statistical analysis suggesting that even in scenarios where weapon sophistication was extremely high, the number of casualties had a more significant impact on outcomes. Thus, the argument from statistical analysis would then successfully defend the original preference relation.This example only describes the approach generally and more detailed decisions regarding possible arguments and attack types have to be made when it is implemented using a structured argumentation framework.

4. Conclusion

We have extended the XAI approach AF-CBA by adding a mechanism by which the justifications' compensation moves are determined using arguments based on domain knowledge provided by domain experts. This not only allows compensation moves to be informed by established domain rules, but also communicates reasons for those compensation moves in terms that are likely to be familiar to domain experts. We have implemented this mechanism as a secondary argument graph, which likewise can be shown to the user as a justification. This secondary argument graph provides an avenue for future extensions aimed at more in-depth justifications and possibly disputes about the domain knowledge itself.

Our extension of AF-CBA relies on argumentation schemes to capture defeasible reasoning patterns, providing a foundation for persuasive justifications. Further extending these patterns within a structured argumentation framework would enhance the sophistication of arguments, allowing arguments about premisses or about the implied entailment of preference relations. This could, for instance, take the form of arguments stemming from different sources of domain knowledge. The current paper is not on rule discovery or information extraction, but those techniques could be integrated in a larger framework in which any disagreements between sources of domain knowledge have to be resolved. Refining rules (e.g. thresholds) is another aspect that deserves attention in future work. Another possible future work direction is to take an experimental approach in the form of usability studies, which would allow us to evaluate various design choices for AF-CBA from a user's perspective.

References

[1] Vincent Aleven. Using background knowledge in case-based legal reasoning: A computational model and an intelligent learning environment. *Artificial Intelligence*, 150(1-2):183–237, 2003.

[2] Andrei Bondarenko, P. M. Dung, R. A. Kowalski, and F. Toni. An abstract, argumentation-theoretic approach to default reasoning. *Artificial Intelligence*, 93(1):63–101, 1997.

[3] Kristijonas Čyras, Ken Satoh, and Francesca Toni. Abstract Argumentation for Case-Based Reasoning. In *Proceedings of the Fifteenth International Conference on Principles of Knowledge Representation*, 2016.

[4] Kristijonas Čyras, Ken Satoh, and Francesca Toni. Explanation for Case-Based Reasoning via Abstract Argumentation. In *Proceedings of COMMA 2016*, pages 243–254. IOS Press, 2016.

[5] Kristijonas Čyras, David Birch, Yike Guo, Francesca Toni, Rajvinder Dulay, Sally Turvey, Daniel Greenberg, and Tharindi Hapuarachchi. Explanations by arbitrated argumentative dispute. *Expert Systems with Applications*, 127:141–156, 2019.

[6] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 2(77):321–357, 1995.

[7] Andrew R. Golding and Paul S. Rosenbloom. Improving accuracy by combining rule-based and case-based reasoning. *Artificial Intelligence*, 87(1-2):215–254, 1996.

[8] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5): 93:1–93:42, 2018.

[9] John Horty. Reasoning with dimensions and magnitudes. *Artificial Intelligence and Law*, 27(3): 309–345, 2019.

[10] John F. Horty. Rules and reasons in the theory of precedent. *Legal Theory*, 17(1):1–33, 2011.

[11] Zachary Lipton. The mythos of model interpretability. *Communications of the ACM*, 61:96–100, 2016.

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

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

[14] Joeri G.T. Peters and Floris J. Bex. Towards a Story Scheme Ontology of Terrorist MOs. In *2020 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 1–6, 2020.

[15] Joeri G.T. Peters, Floris J. Bex, and Henry Prakken. Model- and data-agnostic justifications with A Fortiori Case-Based Argumentation. In *Proceedings of the 19th International Conference on Artificial Intelligence and Law*, pages 207–216, Braga, Portugal, 2023. Association for Computing Machinery. ISBN 9798400701979.

[16] Henry Prakken. Dialectical proof theory for defeasible argumentation with defeasible priorities (preliminary report). In John-Jules Ch. Meyer and Pierre-Yves Schobbens, editors, *Formal Models of Agents*, Lecture Notes in Computer Science, pages 202–215, Berlin, Heidelberg, 1999. Springer. ISBN 978-3-540-46581-2.

[17] Henry Prakken and Rosa Ratsma. A top-level model of case-based argumentation for explanation: Formalisation and experiments. *Argument & Computation*, 13(2):159–194, 2022.

[18] Edwina L. Rissland and David B. Skalak. CABARET: Rule interpretation in a hybrid architecture. *International Journal of Man-Machine Studies*, 34(6):839–887, 1991.

[19] Wijnand van Woerkom, Davide Grossi, Henry Prakken, and Bart Verheij. Hierarchical Precedential Constraint. In *Proceedings of the 19th International Conference on Artificial Intelligence and Law*, pages 333–342, Braga, Portugal, 2023. Association for Computing Machinery.

[20] Douglas Walton, Christopher Reed, and Fabrizio Macagno. *Argumentation Schemes*. Cambridge University Press, 2008. ISBN 978-1-316-58313-5.