Justifying Black-Box Predictions with Domain Knowledge

Joeri G.T. Peters
j.g.t.peters@uu.nl
Utrecht University,
Netherlands National Police
The Netherlands

Floris J. Bex
f.j.bex@uu.nl
Utrecht University,
Tilburg University
The Netherlands

Henry Prakken
h.prakken@uu.nl
Utrecht University
The Netherlands

Abstract

AF-CBA uses case-based argumentation to justify classifier predictions by arguing about differences between cases. We extend the mechanism by modelling which differences can compensate for each other by constructing arguments using domain knowledge. This involves a secondary argumentation framework. To assist experts in defining the appropriate domain knowledge, we use a rule-based classifier for semi-automated knowledge induction. We use the resulting rule set to derive arguments and demonstrate this with an evaluation procedure.

CCS Concepts

• **Computing methodologies** → *Knowledge representation and reasoning*; *Machine learning*; • **Applied computing** → Law.

Keywords

XAI, Justification, Argumentation, Domain Knowledge

ACM Reference Format:
Joeri G.T. Peters, Floris J. Bex, and Henry Prakken. 2025. Justifying Black-Box Predictions with Domain Knowledge. In *Proceedings of 20th International Conference on Artificial Intelligence and Law (ICAIL 2025).* ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The ability to justify machine learning (ML) models has become a critical area of interest with the rise of black-box models. Understanding the rationale behind predictions is essential for establishing trust, especially in high-stakes domains such as finance, medicine, and security. This also fosters better domain understanding, which can be a goal unto itself.

This paper is focussed on enhancing the explainable artificial intelligence (XAI [19]) approach called '*a fortiori* case-based argumentation' (AF-CBA [24]). AF-CBA justifies binary classification predictions using the theory of precedential constraint [12], referencing cases from a case base built from training (or historical [22]) data. This reasoning model thus follows and is inspired by legal patterns of reasoning, where established principles are paramount. Our aim is to extend this approach by incorporating domain knowledge, acknowledging its importance in decision-making and using established knowledge to make justifications more understandable to domain experts. The first part of this paper is based on preliminary work[20], which we elaborate and extend in the current paper.

A ML classifier can be considered a black box due to its technical complexity or propietary nature [10, 15]. Neural networks, for example, typically have improved predictive accuracy compared to decision trees, at the expense of increased opacity. With an opaque model, biases can remain hidden. Transparency is often a legal requirement for decision-making processes [2], both within and outside the legal domain. This is especially concerning in high-stakes domains like law enforcement, where decisions have significant consequences. One might argue that such tasks preclude black-box models. However, if the alternative model does not perform well enough, a trade-off becomes unavoidable. Post hoc approaches like AF-CBA address this by justifying ML predictions without accessing the model itself, making it model-agnostic.

Our research is concerned with a high-stakes and sensitive domain, counter-terrorism, in which classifiers might predict whether an incident is linked to a specific terrorist organisation on the basis of its *modus operandi*, or whether law enforcement should respond to it as a coordinated attack versus a 'lone-wolf' incident. In this paper, we use a scenario where officials use a black-box classifier to predict whether an incident is an act of terrorism or not. Such a prediction is then justified using AF-CBA. We want AF-CBA to refer to expert domain knowledge to justify its decisions. Our approach follows a tradition of combining rule- and case-based reasoning, as demonstrated by Golding & Rosenbloom [8] and Rissland & Skalak [25], who integrated these methods to handle rule exceptions.

An important caveat is that the domain knowledge itself is not necessarily available in a suitable form. Therefore, we address the challenge of semi-automatically discovering domain knowledge. Specifically, we focus on identifying *preference relations* between sets of features (dimensions) in a dataset, where one set is more crucial than another for predicting an outcome. These relations are discovered within the dataset and can be debated or modified by domain experts.

The rest of this paper is structured as follows. We describe the preliminaries of AF-CBA in Section 2 and our extension

ICAIL 2025, Chicago, IL, USA

ACM ISBN XXXXXXX.XXXXXXX
https://doi.org/XXXXXXX.XXXXXXX

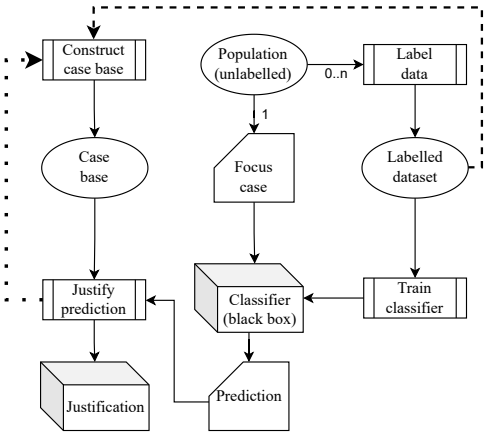

Figure 1: A schematic depiction of our XAI approach. The dashed line represents the scenario in which the case base is created from training data, the dotted line in which it is created from historical predictions.

using a secondary argument framework in Section 3. We then look at possible knowledge induction approaches to inform this extended approach in Section 4 with domain knowledge derived from data and evaluate this approach in Section 5 and some related literature in Section 6. Finally, we discuss conclusions and future work in Section 7.

2 AF-CBA

Justifying binary class predictions made by a classifier is analogous to court rulings based on legal precedents. Prakken and Ratsma [24] introduced a top-level model, later named AF-CBA, which draws from AI & Law research and employs case-based argumentation influenced by Horty's theory of *precedential constraint* reasoning [11]. AF-CBA builds on CATO [1] and the work of Čyras et al. [5, 6]. AF-CBA is, however, not a standalone classifier, but a post hoc approach designed to justify the predictions of an existing ML model.

AF-CBA's procedure is graphically depicted in Figure 1. A random sample from the broader population is assigned labels by annotators or decision-makers, and a classifier is trained on this data (supervised ML). A *focus case* is a new, individual random sample, with the classifier predicting an outcome for it. However, this classifier is a black box and cannot justify its decision. AF-CBA resolves this by using either the labelled data or previous case predictions [22] as a case base, and conducts an argument game between a *proponent* and an *opponent* of the predicted outcome. Cases similar to the focus case are invoked to argue that the focus case should receive the same outcome, provided that any differences between the two only serve to reinforce the outcome for the focus case (precedential constraint). The game is based on Dung's abstract argumentation framework [7], using grounded semantics [23]. The proponent's winning strategy is presented as a justification for the predicted outcome as an argument graph.

An abstract argumentation framework (AF), as introduced by Dung [7], is defined as a pair $AF = \langle A, attack \rangle$, where A is a set of arguments, and *attack* is a binary relation on A. A subset B of A is considered *conflict-free* if no argument in B attacks any other argument in B, and *admissible* if it is both conflict-free and able to defend itself against attacks. More specifically, if an argument A_1 belongs to B, and some argument A_2 in A attacks A_1, then there must be an argument in B that counters A_2. There are various types of admissible sets, referred to as *extensions*. This work focuses on the *grounded extension*, which has the additional properties of containing all the arguments it defends and being the minimal subset that satisfies these conditions.

Formally, a *case* within the *case base* (CB) consists of an *outcome* and a *fact situation*. The outcome is a binary label, denoted by either o or o'. The variables s and $\bar{s}$ represent the two sides, with $s = o$ if $\bar{s} = o'$, and vice versa. The fact situation contains *dimensions*, where each dimension is a tuple $d = (V, \leq_o, \leq_{o'})$. This tuple includes 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 $v' \leq_{o'} v$, where $v, v' \in V$. Each dimension has an associated *tendency*, where a positive tendency indicates that higher values are linked with one particular outcome (e.g., 1 or *true*), and the reverse tendency applies for the opposing outcome. The tendency can sometimes be explicitly given, for example as d_i^+ or d_i^-. In other words, a dimension is a feature with a tendency. A value assignment, written as (d, v), specifies the value x of dimension d in case $c \in CB$, represented as $v(d, c) = x$. The collective set of value assignments for all dimensions d within the non-empty set D constitutes a fact situation denoted by F. It is assumed that two fact situations refer to the same set D. A case is defined as $c = (F, outcome(c))$, where $outcome(c) \in \{o, o'\}$, and the fact situation for case c is expressed as $F(c)$. When comparing two fact situations, it is possible to determine that one case is 'stronger' or 'better' for a particular outcome than the other. As an example, Table 1 shows a precedent c with a better value for dimension d^-_{weapon} than the focus case f, but a worse value for $d^+_{casualties}$. The outcome of a focus case is considered *forced* if there is a precedent within the case base (CB) with the same outcome, and all the differences between the focus case and the precedent only reinforce the focus case's suitability for that outcome [12].

Table 1: Precedent case c and focus case f.

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

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 $F' \leq_s F$.*

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 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). A *best precedent* (Definition 5) has the same outcome as the focus case and as few as possible *relevant differences* (Definition 4). 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 debate differences between the focus case and cited precedents. The proponent argues in favour of the predicted outcome; the opponent challenges it. The proponent initiates the argument by citing a best precedent. The opponent's objective is to respond either by presenting a counterexample or by making a distinguishing move, $Worse(c, x)$, indicating that the focus case is weaker than precedent c in dimensions x. A distinguishing move can in turn be countered with a compensation move, $Compensates(c, y, x)$, where dimensions y offset the deficiencies in dimensions x relative to precedent c. Additionally, the transformation move, $Transformed(c, c')$, indicates that the cited case can be transformed into one where $D(c, f) = \emptyset$. The proponent may use any of these moves to respond, after which the opponent can reply, and this back-and-forth continues until the opponent is unable to make further moves. Note that y in $Compensates(c, y, x)$ may be the empty set, ensuring the possibility of a compensation move. This guarantees the existence of a winning strategy for the proponent, making AF-CBA 'explanation complete' [24] and thus ensuring that the predicted class can always be justified.

Definition 6 presents the argumentation framework. Compensation relies on *sc*, which was deliberately left unspecified by Prakken and Ratsma [24]. In its simplest form, it functions as a partial ordering on dimensions, indicating when a strong value for d_i compensates for a poor value for d_j. This essentially conveys domain knowledge. In this paper, we utilise *sc* to incorporate domain knowledge into the framework.

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 Compensation Moves and Preference Relations

In Definition 6, the set *sc* represents any construct that encompasses some form of domain knowledge. This could include hierarchical relationships akin to CATO [1, 26] or ontologies [21]. The interpretation of the phrase "...y compensates x according to *sc*..." varies based on context. We aim to develop arguments with conclusions of the type $compensates(c, y, x)$ derived from domain knowledge. The relevant domain knowledge does not have to be uncontested—a specific insight might have exceptions, for instance. We conceptualise *sc* as an argumentation framework denoted by $AF_{sc} = \langle A, attack \rangle$, which consists of arguments formed by instantiating argumentation schemes based on domain knowledge [27]. Utilising AF_{sc}, we ascertain the available compensation moves by determining which conclusions are in the grounded extension.

3.1 Compensation as an Argument Scheme

Domain knowledge can be used as in Scheme 1: a conclusion drawn from premises indicating that the fact situation is less favourable for f in dimensions D_w (premise w), while being more favourable in dimensions D_b (premise b), coupled with the condition that D_b is *preferred* over D_w (premise p), according to the *preference relation* $D_w \prec D_b$. The worse values of f in relation to c regarding D_w can be compensated by the better values of f in D_b, and these dimensions are regarded as more significant for the outcome. Note that the terms 'worse' and 'better' are contextual and do not necessarily equate to lower and higher values, respectively.

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:*

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: $D_w \prec D_b$

Conc: $compensates(c, \{(d,v) \in F(f) \mid d \in D_b\}, \{(d,v) \in F(f) \mid d \in D_w\})$

In Table 1, f has a higher weapon sophistication (d_{weapon}) than c, which is associated with non-terrorist incidents, rendering f less favourable on this dimension. However, f also records a higher number of casualties ($d_{casualties}$), an indicator of a terrorist incident. By employing the domain knowledge that a greater number of casualties offsets a higher weapon sophistication, we derive the following argument:

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

Conc: $compensates(c, \{(d_{casualties}, 10)\}, \{(d_{weapon}, high)\})$

In this scenario, the argument asserts that while f possesses a higher ('worse') level of weapon sophistication than c, the higher ('better') casualties figure associated with f counterbalances this, justifying the predicted outcome of *true* for f. We presume that the fact situations of both the precedents from the CB and the focus case are known, preventing any challenge to the first two premises of this scheme. Moreover, this scheme is strict in that the conclusion cannot be attacked if all its premises are satisfied. However, we must ascertain the truth of premise p (the preference relation that supports the compensation move). In practice, various conditions may influence a preference relation, which we will now examine in detail.

3.2 Conditional Preference Relations

A threshold must be satisfied for a preference relation to hold, as per Scheme 2. For example, the aforementioned relation $\{d_{weapon}\} \prec \{d_{casualties}\}$ might only be applicable when the number of casualties is sufficiently high, say at least 4. A lower casualty count may not be regarded as sufficient justification to offset the fact that the sophisticated weapon involved in this incident is atypical. Thus, the validity of premise p for this particular instance of Scheme 2 relies on the condition that $d_{casualties} \geq 4$. While additional conditions will be considered later, we will first summarise the sets of conditions for a preference relation $D_w \prec D_b$ under 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 truth 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 assesses whether Ψ holds in a given instance. If it does, the corresponding preference relation can be inferred and subsequently employed as premise p in the instantiation of Scheme 1. In Table 1, the focus case f exhibits a 'worse' level of weapon sophistication (d_{weapon}) alongside a 'better' count of casualties ($d_{casualties}$). By instantiating Schemes 2 ($PREF(f, D_b, D_w, D)$) and 1 ($COMP(f, c, D_b, D_w)$), we can formulate 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$(c, \{(d_{casualties}, 10)\}, \{(d_{weapon}, high)\})$

Multiple thresholds can exist within Ψ. For instance, a preference relation might assert that $d_{casualties}$, combined with d_{fear} (a numerical measure of public fear), takes precedence over d_{weapon} when both dimensions exceed their respective thresholds. In such a case, Scheme 2 would be instantiated 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 exceed particular thresholds represents a condition where each dimension must *independently* satisfy a sub-condition. Alternatively, a preference relation might depend on a combination of dimensions, where an evaluation function surpasses a single threshold. For instance, consider a preference relation $\{d_{weapon}\} \prec \{d_{casualties}, d_{wounded}\}$ with the condition that $d_{casualties}(f) + d_{wounded}(f) \geq 10$. In this case, the evaluation function is the sum of fatal and non-fatal casualties, which compensates for a high level of weapon sophistication. In other words, the distinction between fatally and non-fatally harmed victims is irrelevant within this domain.

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 could envision expert knowledge where the difference between the number of perpetrators and victims helps to differentiate terrorist incidents from assassinations, for example. Likewise, the ratio of wounded to deceased victims might influence the impact of weapon sophistication in some hypothetical domain-specific insight. A weighted mean of several dimensions may need to exceed a certain value. Domain experts might develop dozens of such expressions of very specific domain knowledge in areas that are well understood and rich in descriptive dimensions, potentially aided by statistical analysis or knowledge induction methods. More complex functions are also conceivable. One might argue that at least some of these evaluations should be incorporated during the feature engineering phase, before model training, rather than

in post-hoc justifications. It is important to note, however, that our approach remains model- and data-agnostic, so we should generally support such evaluations.

Consider the following scenario: an attack involving a sophisticated bomb (d_{weapon}) that does not result in a large number of casualties ($d_{casualties}$). Under typical circumstances, the sophistication of the weapon might suggest a targeted assassination rather than a terrorist act. However, if the event generates an exceptionally high level of public fear (d_{fear}), this could compensate for the casualty count, as terrorists aim to instil fear and disrupt society. In this case, the evaluation function might assign significant weight to d_{fear}, such that a weighted sum of d_{fear} and $d_{casualties}$ is compared against 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 establish conditions on the dimensions within the preference relation, D_w and D_b. However, contextual factors may also influence the applicability of a preference relation. For instance, an additional dimension $d_{measures}$ (the number of security measures in place) might modulate the impact of the casualty count in compensating for weapon sophistication. In this scenario, the threshold relates to a dimension that is not part of the preference relation itself.

Moreover, conditions for a preference relation can encompass spatiotemporal factors. For example, 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 domains like counter-terrorism, where the nature of threats and their societal impact can change rapidly. When attributing historical incidents to terrorist organisations, it is essential to consider that an organisation was founded at a specific moment in time or was only active in a particular geographical area. For instance, ISIS (ISIL) did not rise to prominence until 2014 in regions of Syria and Iraq. Any piece of domain knowledge concerning characteristics of ISIS incidents or public claims made by this organisation is likely to be specific to the relevant time and location. The same concerns apply to the Taliban in Afghanistan, both before the American invasion in 2001 and after America's departure in 2021, or to the Troubles in Ireland and Great Britain from 1966 to 1998.

For a simplified example, consider the following: the occurrence of an incident during the Troubles in Belfast implies 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 employed to construct an empty compensation move based on domain knowledge. It is important to note that if $D_b = \emptyset$, Scheme 1 represents the special case of empty compensation. In AF-CBA, we permit compensation moves with $D_b = \emptyset$ to ensure a winning strategy (see Section 2), serving as a somewhat unsatisfactory but necessary default that substitutes for a more informative justification. With an argument such as the following, we can provide expert-informed justifications for why the values in D_b are not relevant to the outcome of the focus case, despite the absence of any compensating dimensions. This makes an empty compensation move more informative 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: compensates$(c, \emptyset, \{(d_{casualties}, 10), (d_{weapon}, high)\})$

Transitivity (where $\{d_1\} \prec \{d_2\}$ and $\{d_2\} \prec \{d_3\}$ imply $\{d_1\} \prec \{d_3\}$) and antisymmetry (where $\{d_1\} \prec \{d_2\}$ implies $\{d_2\} \not\prec \{d_1\}$) cannot be universally presumed, but 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, suggesting that a superior value in one dimension can compensate for an inferior value in another. For example, a high number of casualties ($d_{casualties}$) may offset high weapon sophistication (d_{weapon}) and vice versa. This symmetry may indicate that the dimensions are equivalent in their influence on an outcome, functioning as proxies for a more abstract concept. For instance, d_{alert} (security alert status) and $d_{measures}$ (number of security measures) could be subcategories of a broader dimension $d_{security}$ (overall security preparedness), suggesting a certain equivalence. Consequently, our approach implicitly permits the drawing of *abstract parallels* similar to the factor hierarchies in CATO [1].

3.3 Arguing About Preference Relations

As previously noted, we do not assume the body of domain knowledge to be uncontested. While Schemes 1 and 2 offer a framework for evaluating whether the conditions of a compensation move have been satisfied, exceptions may exist, and the premises can be challenged. The specific types of attacks that may arise depend on the domain; however, in general, attacks between arguments could be modelled within a structured argumentation framework, such as ASPIC+ [17] or ABA [3].

For instance, a domain expert may identify an additional caveat for the preference relation $\{d_{casualties}\} \prec \{d_{weapon}\}$ beyond the condition $d_{casualties}(f) \geq 4$; specifically, this preference may not hold if the weapon sophistication is exceedingly high. The abstract placeholder Ψ could then refer to two distinct thresholds for this instance of $PREF(f, D_b, D_w, D)$:

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

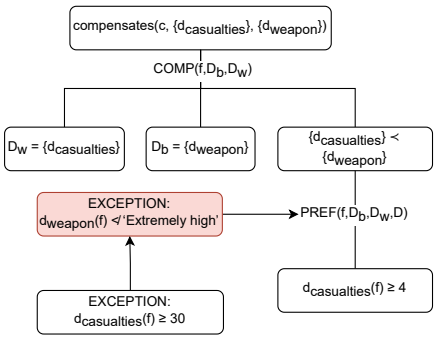

Figure 2: An illustration an exception to an exception defending a preference relation. The shaded box is not in the grounded extension, attacks are indicated by arrows.

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

==========

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

However, one might contend that it is more informative to model exceptions explicitly as separate arguments. The preference relation $\{d_{casualties}\} \prec \{d_{weapon}\}$ would then be challenged by an exception argument that asserts $\{d_{casualties}\} \prec \{d_{weapon}\}$ does not hold for f due to the condition $d_{weapon}(f) \not< $ 'Extremely high'. For $\{d_{casualties}\} \prec \{d_{weapon}\}$ to be applicable in Scheme 1, this exception argument must be successfully countered, possibly by introducing an exception to the exception.

For instance, the exception $d_{weapon}(f) \not< $ 'Extremely high' might be deemed irrelevant if the number of casualties is sufficiently high, such as $d_{casualties} \geq 30$. This second exception would attack the first exception, thereby defending the preference relation from Scheme 2 and reinstating the compensation move from Scheme 1. This process can continue for any additional exceptions. This concept is illustrated in the argument graph shown in Figure 2.

We permit chains of arguments regarding preference relations. Whether lengthy, complex arguments are always beneficial depends on the domain experts, who can determine what is suitable for the intended user. Our approach enables them to decide how thoroughly to justify the domain knowledge employed to support ML predictions, according to their preferences. The objective is always to justify compensation moves from the perspective of the user, who may or may not possess domain expertise, ensuring an appropriate level of justification.

Other types of arguments could be valuable, such as expert opinions (following Walton et al. [27]) based on professional experience, domain literature, or statistics. Naturally, there are domain-specific reasons for the inclusion of a preference relation in the first place, which implies that conflicting opinions may arise. Just as we can permit chains of exceptions to preference relations, experts might find it equally informative to explicitly model dissent. This approach could illustrate how the most recent analysis, literature review, or the input of the most senior expert resolves the debate.

For example, the preference relation $\{d_{casualties}\} \prec \{d_{weapon}\}$ could be challenged by an opinion asserting that it does not hold, drawing on the expert's experience. This assertion could itself be countered by another opinion based on statistical analysis, which indicates that even in cases of extremely high weapon sophistication, the number of casualties had a more significant impact on outcomes. In this scenario, the argument stemming from statistical analysis would successfully defend the original preference relation.

4 Knowledge Induction

Preference relations and their conditions (Scheme 2) may be provided by experts, but could also be discovered from data by some means. In this section, we look into ways of discovering them through knowledge induction techniques, that is, using rule-based classifiers. Rule-based classifiers do not typically yield domain knowledge directly in the form required for generating such conditional preference relations. We therefore combine a rule-based classifier with feature importance scoring, in order to arrive at possible instances of Scheme 2. We use terms like "feature importance" alongside "dimension" in this section, since "feature" is the preferred term in ML literature.

Importance scores lack explicit conditions that specify when one dimension should be prioritised over another. Rule-based classifiers, conversely, offer explicit, interpretable *decision rules*, which link dimensions to outcomes. The decision rules in such a *decision rule set* can be viewed as domain knowledge for how to classify cases; they are not the same as our preference relations on dimension sets. Decision rules do not directly rank feature importance across the dataset, as they are typically just a set of conditions and an outcome. For example, a rule-based classifier might generate decision rules like r_1 : `IF` $d_{location} = urban$ `AND` $d_{weapon} = Bombing$ `THEN` Outcome = true. When the fact situation (dimension-value pairs) of a case meets all the conditions of a decision rule, we say that that case *satisfies* those conditions. For instance, given rule r_1, if we find that $d_{location}$ has consistently higher importance scores than some other dimension d_x for cases satisfying Ψ_1 (r_1's conditions), we can derive a preference relation $\{d_x\} \prec \{d_{location}\}$ when urban bombings are involved.

Using a transparent knowledge induction technique is crucial in XAI, as the goal is to make model decisions understandable to users. Rule-based classifiers provide clear, human-readable decision rules that can be scrutinised and validated. However, excessively complex decision rules risk losing interpretability. We have also argued that the notion of black-box models can be a function of someone's expertise, rather than simply an inherent function of the model itself. As such, we should strike a balance when selecting an appropriate rule-based classifier with which to induce a decision rule set (that is, domain knowledge used to classify data).

Several approaches exist for generating decision sets with rule-based classifiers. RIPPER (Repeated Incremental Pruning to Produce Error Reduction)[4] appears well-suited for our purposes. It generates decision rules through a sequential covering process, where decision rules are iteratively learned and refined to cover remaining positive examples while excluding negative ones. Like decision trees, RIPPER employs pruning techniques to handle noisy data and prevent overfitting. However, RIPPER typically produces more compact and generalisable decision rule sets compared to decision trees. RIPPER's focus on producing concise, accurate decision rules while maintaining interpretability makes it an good choice for deriving conditional preference relations. It is, however, worth noting that rule-based classifiers like RIPPER are typically limited to fairly straightforward decision rules. More sophisticated conditions like the function-based thresholds we described in Section 3.2 wil not result from such a classifier.

4.1 Determining Feature Importance Scores

Feature importance scores obtained through methods such as *SHAP* [16] offer a means of assessing the influence of individual features on model outputs. SHAP calculates the average contribution of each feature across all possible subsets of features, providing a globally consistent measure of feature importance. We can interpret this as a preference relation. For instance, if dimension d_1 has a higher importance score than dimension d_2, this could imply a preference relation $\{d_2\} \prec \{d_1\}$, suggesting that d_1 is more influential in driving classification decisions. Given a decision rule k from a decision rule set with a condition set Ψ_k, let $CB_k \subseteq CB$ be the subset of cases that satisfy that condition. For any set of dimensions $D' \subseteq D$, let ϕ_d^k represent the squared mean SHAP value of d given CB_k and Ψ_k. This provides a single score representing the collective importance of the dimensions in D' under Ψ_k. We can then define a preference relation between two disjoint sets of dimensions D'_i and D'_j under condition Ψ_k if:

$$\frac{\bar{\phi}^k_{D'_i} - \bar{\phi}^k_{D'_j}}{\bar{\phi}^k_{D'_j}} > \delta \tag{1}$$

where δ is a relative significance threshold, expressing the difference between the squared importances of the two dimension sets relative to one another. If this condition holds, then Scheme 2 can be instantiated with $D_b = D'_i$, $D_w = D'_j$ given a focus case f that meets the condtions in Ψ_k.

It is not trivial to derive feature importance scores from a decision rule set, but notions like *support* and *coverage* could possibly be used in this regard. We instead opt for using the black-box classifier (whose predictions AF-CBA is meant to justify) to calculate them. Not only is this less ambiguous, it means we are tying our justification approach more closely to the classifier—using somewhat dubious feature importance scores from our rule-based classifier would widen the gap between the black-box classifier and our justification. In this combined approach, we can infer conditional preference relations that are both interpretable and relevant to specific decision-making contexts. The process involves three main steps:

DEFINITION 7 (PREFERENCE RELATION INDUCTION). *Let CB be a case base, D its set of dimensions, and $Clas_B$ and $Clas_R$ two classifiers that make predictions on cases from CB, where $Clas_B$ is a statistical classifier (black box) and $Clas_R$ a rule-based classifier that produces a decision rule set. For any case $c \in CB$ and dimension $d \in D$, let $\phi_d(c)$ denote the squared feature importance score of dimension d for case c as computed by an XAI method (e.g., SHAP, LIME) applied to $Clas_B$. The process of deriving conditional preference relations proceeds as follows:*

- ***Induction of Classification Rules**:*
  Using $Clas_R$, identify a set of decision rules $R = \{r_1, ..., r_K\}$, where each rule $r_k \in R$ is of the form:

  $$r_k : \texttt{IF } \Psi_k \texttt{ THEN } outcome = s_k,$$

  where Ψ_k is the context of the decision rule, defined as a conjunction of conditions on dimension-value pairs (e.g., $d_1 < t \wedge d_2 > t'$), and $s_k \in \{0,1\}$ is the binary outcome predicted by the decision rule.
- ***Aggregation of Importance Scores**:*
  For each decision rule r_k, let $CB_k \subseteq CB$ be the subset of cases satisfying Ψ_k. For any set of dimensions $D' \subseteq D$, compute the conditional mean feature importance from the individual importance scores derived from $Clas_B$:

  $$\bar{\phi}^k_{D'} = \frac{1}{|CB_k|} \sum_{d \in D'} \sum_{c \in CB_k} \phi_d(c).$$

  This computes the importance of the entire set of dimensions within the context Ψ_k.
- ***Recognition of Preference Relations**:*
  For any two sets of dimensions $D'_i, D'_j \subseteq D$, where $D'_i \cap D'_j = \emptyset$, define a conditional preference relation under Ψ_k if:

  $$\frac{\bar{\phi}^k_{D'_i} - \bar{\phi}^k_{D'_j}}{\bar{\phi}^k_{D'_j}} > \delta,$$

  where δ is a relative significance threshold to determine of Scheme 2 can be instantiated with $D_b = D'_i$, $D_w = D'_j$ given a focus case f that meets the condtions in Ψ_k. We write the set of instantiations made possible in this way as the set $\mathscr{P}$, where:

  $$\mathscr{P} = \bigcup_{k=1}^{K} \{\ (\Psi_k,\ D'_j \prec D'_i)\ \}.$$

For each decision rule r_k, the importance of each possible set of dimensions D' is aggregated as the mean importance score across cases satisfying the decision rule. Preference relations between non-overlapping dimension sets are then recognised by comparing these aggregated scores. For example, if $(\bar{\phi}^{r_1}_{d_{location}} - \bar{\phi}^{r_1}_{d_{casualties}})/\bar{\phi}^{r_1}_{d_{casualties}} > \delta$ under the condition that $d_{casualties} < 10$, a preference relation $\{d_{location}\} \prec \{d_{casualties}\}$ is established for a focus case for which $d_{casualties}(f) < 10$ is true.

Originally ([24]), AF-CBA was made 'explanation complete' by guaranteeing a winning strategy through empty compensation moves. Now that we have made the derivation of preference relations explicit, we can likewise guarantee that AF-CBA can justify any outcome by explicitly adding preference relations with $D_b = \emptyset$ without any conditions in Ψ_k to the set $\mathscr{P}$.

When we determine feature importance scores for dimension sets by summing (and normalising) individual feature importance scores, we are assuming these dimensions are not highly correlated. This assumption is often unrealistic, as many real-world datasets contain dimensions that are either redundant or highly correlated. Such correlations can lead to an inflation of feature importance for dimension subsets that include correlated dimensions, as their individual SHAP values may reflect overlapping contributions.

One could argue that highly correlated dimensions should already be minimised within the context of our XAI framework, especially if we assume we are justifying a well-performing ML model. For a model to perform well, ideally, it should avoid reliance on redundant, highly correlated dimensions. This assumption would simplify our task, allowing us to compute aggregated feature importance scores without adjusting for high dimension correlations.

We can compromise by introducing a final argument scheme, which attacks $PREF(f, D_b, D_w, D)$ (Scheme 2) if D_b contains highly correlated dimensions. Scheme 3 describes how, when the correlation between two dimensions within either dimension set exceeds a given value, the preference relation cannot be said to hold. This could potentially be elaborated for mixed dimension types, but we keep this scheme simple for now. An instance of this scheme attacks the corresponding instance of the preference scheme, thereby preventing that preference relation from being used in a compensation move and informing the user of the reason why.

ARGUMENTATION SCHEME 3 (CORRELATION). *Let $D_b, D_w \subseteq D$ be two sets of dimensions where $D_b \cap D_w = \emptyset$, the Pearson correlation coefficient $\rho(d_i, d_j)$, and a corresponding threshold value ε. Then the correlation scheme $CORR(D_b, D_w)$ is defined as the following reasoning pattern:*

$$\max(\rho(d_i, d_j)) > \varepsilon$$
$$==========$$
Conc: $\neg(D_w \prec D_b)$,

where $d_i, d_j \in D_b$.

5 Example & Evaluation

In this section, we demonstrate our approach using a real-world dataset: the Global Terrorism Database [13] (GTD). This dataset contains data of a very serious nature and serves as a proxy for similar such datasets within the law-enforcement domain, for which the unfortunate trade-off between performance and justifiability warrants the use of a post-hoc approach in our estimation. We use parameters δ (the minimum required relative difference in aggregated feature importance scores) and ε (the maximum allowed correlation coefficient). This experiment showcases how our approach can be used in practice to derive preference relations from data for AF-CBA's justifications of classification predictions.[1]

Applying our approach within AF-CBA is not computationally expensive, but evaluating all preference relations given a large number of features can be. As higher-level parameters, we have two integers to restrict the procedure's search space: the maximum set size and the n most important dimensions (3 and 8, resp.). The intuition is that preference relations with overly large dimension sets are both unrealistic and uninterpretable for experts, and that preference relations with important features are more likely to be relevant. These parameters can be set to higher values if sufficient computational resources are available to handle the resulting combinatorial explosion.

The GTD is a comprehensive dataset documenting terrorist incidents globally, spanning the period from 1970 to 2017 [13]. In this work, we restrict our analysis to the post-1997 segment of the dataset. As our outcome label (the target feature), we take *suicide*—whether or not the terrorist incident constitutes a suicide attack. This is relevant in practice in the direct aftermath of an incident, because there can be doubts as to whether a second attack is likely to take place and this affects the appropriate response from the authorities. For the sake of realism, we use only those features which are likely to be known shortly after an incident, namely numerical data and the attack type (see appendix for other combinations).

Our pipeline includes a data preprocessing class that performs data filtering, missing value handling (whereby we map various missing value codes to a single value for all features), caching and optional encoding of categorical values as defined per feature in a single configuration file (see our code appendix). After data preprocessing, we train ($F_1 = 0.80$, Acc$= 0.98$) a classifier as a representative stand-in for black-box classifiers in general, including neural network approaches. Because of its ease of use, we choose HistGradientBoosting from the Scikit-learn package. We also train our rule-based classifier (RIPPER from the Wittgenstein package using default parameters) and use 5-fold cross-validation to derive a final decision rule set, where we dismiss decision rules that only appear in one fold. We then determine and aggregate feature importance scores as described, given each rule in the decision rule set. This gives us the desired conditional preference relations. We repeat this process while varying the parameters δ and ε in order to study the effects on the generated rule set, as a demonstration of our evaluation procedure. The effect of these two parameters is expressed using the following metrics:

- **The total number of preference relations (total and unique)**: We want a sufficient number of of preference relations to be available in order to construct compensation moves, but an excessively large set is unlikely to be very applicable, maintainable and interpretable. Because different decision rules can result in the same preference relations (albeit with different conditions), we also report the deduplicated number.

[1] The code repository is available at https://github.com/JGTP/preference-mining/tree/6f5903354ee1f1fa200b5366b304776775c9552d.

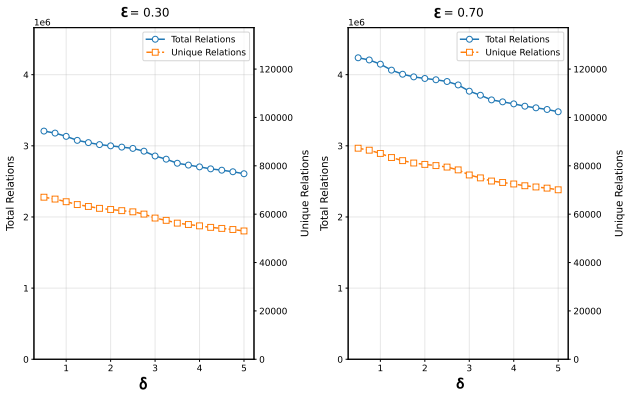

Figure 3: Total (blue circles) and unique (orange squares) preference relations generated as δ increases, shown for two values of ε (0.30 and 0.70). The analysis produced 56 stable rules.

- **The mean number of dimensions in D_w and D_b per preference relation**: Preference relations containing fewer dimensions are generally more widely applicable and thus more useful. They are also more interpretable.

The complexity and applicability of the decision rules themselves (including the complexity of their conditions) are relevant factors. However, those notions are a function of the decision rule set generated by the rule-based classifier, not of our subsequent approach of generating preference relations. Therefore, the usual evaluation procedure of decision rule sets applies there. What we aim to show here, is how our generation of preference relations can be evaluated so that parameters can be set and our approach used.

Consider the following decision rule obtained from RIPPER, which predicts incidents with fewer than two perpetrators, of attack type 3 (bombing/explosion), occurring between 2013 and 2014, to be suicide attacks.

Listing 1: Example decision rule from RIPPER

```
"conditions": [
  {"feature": "nperps", "operator": "==", "value":
      "<2.0"},
  {"feature": "attacktype1_3", "operator": "==", "
      value": 1.0},
  {"feature": "iyear", "operator": "==", "value": "
      2013.0 - 2014.0"}],
"outcome": "suicide"
```

Listing 2: Example preference relation

```
"set1": ["nkillter", "iyear"],
"set2": ["nwound", "ransomamtus"],
"set1_importance": 5.85,
"set2_importance": 0.34,
```

Above is a resulting preference relation indicating that the combination of terrorists killed and year has substantially higher importance (5.85 vs. 0.34) than the combination of wounded casualties and ransom amounts in US dollars.

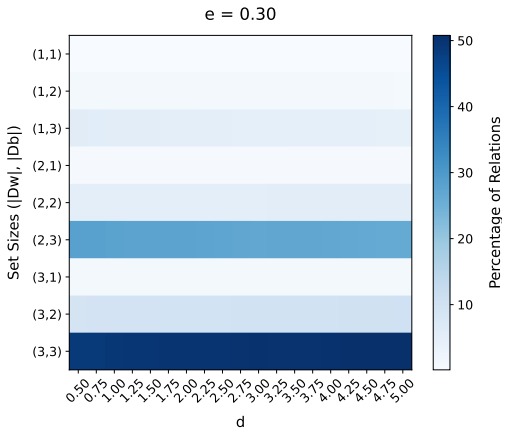

Figure 4: Distribution of set sizes $(|D_w|, |D_b|)$ across different values of δ for $\varepsilon = 0.30$, with $\varepsilon = 0.70$ having similar results. The distribution seems to vary little for this configuration. Regardless, this plot should be of use to domain experts evaluating the set of preference relations.

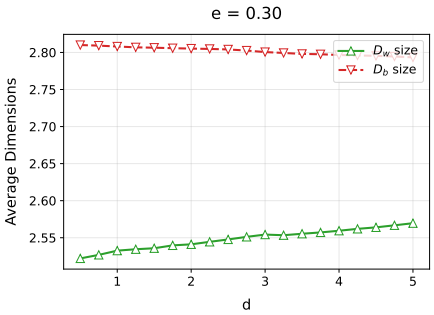

Figure 5: Mean size of D_w (lower) and D_b for $\varepsilon = 0.30$, with $\varepsilon = 0.70$ having similar results. These show little variation for the current configuration of parameters and features.

Figure 3 illustrates how increasing δ reduces the number of relations generated. Figure 4 shows that most relations involve feature sets of size 2, particularly when both D_w and D_b are of size 2. Figure 5 reveals that the average size of both feature sets remains relatively stable across different thresholds.

This shows that our approach succesfully generates preference relations from decision rules. How many decision rules and preference relations to allow is something for domain experts to decide. Inspecting each decision rule is quite realistic, inspecting the strongest preference relations is realistic and creating sufficient preference relations is realistic; inspecting every single preference relation generated in this way is clearly not realistic. This would call for an early pruning solution, alongside additional optimisation we reserve for future work.

6 Related Literature

Argument-Based Machine Learning [18] has been realised as an extension of the CN2 rule induction algorithm to ensure

induced rules align with expert-provided arguments. The approach identifies critical examples through an interactive process, improving both classification accuracy and model transparency.

PADUA [28] is a protocol to enable agents to engage in classification dialogues using arguments from association rules. This addresses classification errors by allowing agents to cite evidence rather than relying on predefined rules.

Value Judgment-based Argumentative Prediction (VJAP) predicts outcomes of trade secret misappropriation cases [9] using argument schemes and value-based legal reasoning, drawing on the Value Judgment Formalism to generate argument graphs. These incorporate the effects of specific facts on abstract legal values such as property interest, confidentiality and competition. The approach quantifies these effects using learned weights derived from past cases. Through analogy and distinction, VJAP grounds its predictions in legal norms.

7 Conclusion

We have extended the AF-CBA framework with a mechanism to determine the justifications for compensation moves based on domain knowledge. This presents the justifications in a manner that aligns with concepts familiar to domain experts. Our extension builds on argumentation schemes to capture defeasible reasoning patterns, offering a foundation for more persuasive justifications. Expanding these patterns within a formal argumentation framework could further enhance the complexity of the reasoning process, enabling arguments to address underlying premises or the inferred implications of preference relations.

The knowledge induction approach introduced in this paper extends the capability of the framework to infer preference relations between dimensions in binary classification tasks. By combining rule-based learning with the quantifiable impact of feature importance, we can derive preference relations. We demonstrate this as a possible evaluation procedure.

Finally, if we abandon the possibility of empty compensation moves, precedential constraint can be used as a reasoning model for classification unto itself, as opposed an XAI mechanism. Such as an approach is conceptually comparable to retrieval-augmented generation (RAG [14]) in its emphasis on domain knowledge. A comparison between RAG and precedential constraint is therefore a possible future work direction.

References

[1] V. 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] A. Bibal, M. Lognoul, A. de Streel, and B. Frénay. Legal requirements on explainability in machine learning. *Artificial Intelligence and Law*, 29(2): 149–169, June 2021.

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

[4] W. W. Cohen. Fast Effective Rule Induction. In A. Prieditis and S. Russell, editors, *Machine Learning Proceedings 1995*, pages 115–123. Morgan Kaufmann, San Francisco (CA), Jan. 1995.

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

[6] K. Čyras, D. Birch, Y. Guo, F. Toni, R. Dulay, S. Turvey, D. Greenberg, and T. Hapuarachchi. Explanations by arbitrated argumentative dispute. *Expert Systems with Applications*, 127:141–156, Aug. 2019.

[7] P. M. 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.

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

[9] M. Grabmair. Predicting trade secret case outcomes using argument schemes and learned quantitative value effect tradeoffs. In *Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law*, ICAIL '17, pages 89–98, New York, NY, USA, June 2017. Association for Computing Machinery.

[10] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5):93:1–93:42, 2018.

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

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

[13] G. LaFree and L. Dugan. Introducing the Global Terrorism Database. *Terrorism and Political Violence*, 19(2):181–204, Apr. 2007.

[14] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020.

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

[16] S. Lundberg and S.-I. Lee. A Unified Approach to Interpreting Model Predictions, Nov. 2017.

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

[18] M. Možina, J. Žabkar, and I. Bratko. Argument based machine learning. *Artificial Intelligence*, 171(10-15):922–937, July 2007.

[19] E. S. Ortigossa, T. Gonçalves, and L. G. Nonato. EXplainable Artificial Intelligence (XAI)—From Theory to Methods and Applications. *IEEE Access*, 12:80799–80846, 2024.

[20] J. Peters, F. Bex, and H. Prakken. Arguments based on domain rules in prediction justifications. *Proceedings of the 24th Workshop on Computational Models of Natural Argument co-located with 10th International Conference on Computational Models of Argument (COMMA 2024)*, 3769: 90–99, 2024.

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

[22] J. G. Peters, F. J. Bex, and H. 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.

[23] H. Prakken. Dialectical proof theory for defeasible argumentation with defeasible priorities (preliminary report). In J.-J. C. Meyer and P.-Y. Schobbens, editors, *Formal Models of Agents*, Lecture Notes in Computer Science, pages 202–215, Berlin, Heidelberg, 1999. Springer.

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

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

[26] W. van Woerkom, D. Grossi, H. Prakken, and B. 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.

[27] D. Walton, C. Reed, and F. Macagno. *Argumentation Schemes*. Cambridge University Press, Aug. 2008.

[28] M. Wardeh, T. Bench-Capon, and F. Coenen. PADUA: A protocol for argumentation dialogue using association rules. *Artificial Intelligence and Law*, 17(3):183–215, Sept. 2009.