Justification in Case-Based Reasoning

Wijnand van Woerkom[1], Davide Grossi[2,3,4], Henry Prakken[1,5] and Bart Verheij[2]

[1]Department of Information and Computing Sciences, Utrecht University, The Netherlands
[2]Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, The Netherlands
[3]Amsterdam Center for Law and Economics, University of Amsterdam, The Netherlands
[4]Institute for Logic, Language and Computation, University of Amsterdam, The Netherlands
[5]Faculty of Law, University of Groningen, The Netherlands

Abstract

The explanation and justification of decisions is an important subject in contemporary data-driven automated methods. Case-based argumentation has been proposed as the formal background for the explanation of data-driven automated decision making. In particular, a method was developed in recent work based on the theory of precedential constraint which reasons from a case base, given by the training data of the machine learning system, to produce a justification for the outcome of a focus case. An important role is played in this method by the notions of citability and compensation, and in the present work we develop these in more detail. Special attention is paid to the notion of compensation; we formally specify the notion and identify several of its desirable properties. These considerations reveal a refined formal perspective on the explanation method as an extension of the theory of precedential constraint with a formal notion of justification.

Keywords

Precedential constraint, Interpretability, Law

1. Introduction

In [1] a case-based reasoning method is proposed to explain data-driven automated decisions for binary classification, based on the theory of precedential constraint introduced in [2, 3]. This method is motivated by an analogy between the way in which a machine learning system draws on training data to assign a label to a new data point and the way in which a court of law draws on previously decided cases to make a decision about a new fact situation, because in both of these situations the precedent that has been set must be adhered to as closely as possible. The theory of precedential constraint, which has been developed to describe the type of a fortiori reasoning used for legal decision making on the basis of case law, can therefore be applied to analyze machine-learned decisions that are made on the basis of training data.

More specifically, the method of [1] formally models the kind of dialogue in which lawyers cite precedents to argue in favor of their preferred outcome of the new fact situation. These citations, and the way in which they attack the opponent's citation, are formalized using an

1st International Workshop on Argumentation for eXplainable AI (ArgXAI, co-located with COMMA '22), September 12, 2022, Cardiff, UK

w.k.vanwoerkom@uu.nl (W. van Woerkom)
https://webspace.science.uu.nl/~woerk003/ (W. van Woerkom)

abstract argumentation framework as in [4]. A winning strategy in the grounded argument game on this framework, starting with an initial citation of a suitable precedent case, is taken as the explanation of the decision of the new fact situation.

In the present work, we examine the explanation model of [1] in detail and make various suggestions and modifications for improvement. Particularly close attention is paid to the subject of *compensation*; the way in which important differences between a new fact situation and a precedent case can be compensated for by features of the focus case. We make the formal nature of this subject more explicit, and specify various desirable properties it may have. Subsequently, we show that the model can be equivalently viewed as extending the theory of precedential constraint with notions of *justification* and *citability*, the combination of which constitutes the explanations produced by the model. This equivalent formulation only uses the simple notion of relation, thus greatly simplifying the specification of the model. The resulting view may be more broadly applied to the type of downplaying attacks seen in similar systems such as CATO [5].

We begin this work by summarizing the relevant aspects of the theory of precedential constraint in Section 2. In Section 3 we give a description of the explanation method of [1]. In Section 4 we revisit the definition of best citability, suggest some improvements, and demonstrate their potential experimentally. Then in Section 5 we reconsider the compensation relation and formulate desirable properties. These considerations lead us to give an equivalent formulation of the model just in terms of relations, which we do in Section 6. We conclude in Section 7 with some final thoughts and remarks.

2. Precedential Constraint

The theory of precedential constraint was developed in [2, 3] to describe the a fortiori reasoning involved with case law. It is taken as the point of departure of the explanation method in [1] and so we begin by recalling those aspects of it that are necessary for the rest of this work. The contents of this section are largely similar to [6, Section 2].

In order to describe the *fact situation* of a case we use what are called *dimensions* in the AI & law literature, which are formally just partially ordered sets.

Definition 2.1. A *dimension* is a partially ordered set $(d, \preceq)$.

We will frequently omit explicit reference to the dimension order $\preceq$ and instead refer to just the set d when we speak of a dimension. A model of precedential constraint of a specific domain assumes there is a set of these dimensions D, relative to which the rest of the definitions are specified.

Definition 2.2. A *fact situation* P is a choice function on the set of dimensions D, i.e. for each dimension $d \in D$ an element $P(d) \in d$ of that dimension is chosen by P. A *case* p is a fact situation P paired with an outcome $s \in \{0, 1\}$, written $p = P : s$. A set *CB* of cases is called a *case base*. If $p = P : s$ we may write $p(d)$ instead of $P(d)$.

In the context of a case $p, q, r, \ldots$ we will refer to its fact situation by the corresponding upper case letters $P, Q, R, \ldots$ without further explicit mention.

The order $\preceq$ of a dimension d specifies the relative preference the elements of d have towards either of two outcomes 0 and 1. More specifically, if $v \prec w$ for $v, w \in d$ this means w prefers outcome 1 relative to v, and conversely v prefers outcome 0 relative to w. Usually we want to compare preference towards an arbitrary outcome s, so to do this we define for any dimension $(d, \preceq)$ the notation $\preceq_s := \preceq$ if $s = 1$ and $\preceq_s := \succeq$ if $s = 0$.

Definition 2.3. Given fact situations P and Q we say Q is *at least as good* as P for an outcome s, denoted $P \preceq_s Q$, if it is at least as good for s on every dimension d:

$$P \preceq_s Q \quad \text{if and only if} \quad P(d) \preceq_s Q(d) \text{ for all } d \in D.$$

If moreover $p = P : s$ is a previously decided case we say that p *forces* the decision of Q for s. A case base *CB* forces the decision of Q for s if it contains a case that does so.

Definition 2.4. Given two cases $p = P : s$ and $q = Q : s$ such that $P \preceq_s Q$ we say that the outcome of q for s was *forced* by the case p, and write $p \preceq q$.

To give some intuition for these definitions we consider a running example of risk of recidivism, as in [6, Example 2.1].

Example 2.1. Convicts are described along three dimensions: age $(\texttt{Age}, \preceq_{\texttt{Age}})$, the number of prior offenses $(\texttt{Priors}, \preceq_{\texttt{Priors}})$, and sex $(\texttt{Sex}, \preceq_{\texttt{Sex}})$. Age and number of priors have the natural numbers as possible values, so $\texttt{Age} := \mathbb{N}$ and $\texttt{Priors} := \mathbb{N}$. The values for sex are $\texttt{Sex} := \{\text{M}, \text{F}\}$. The outcome for this domain is a judgement of whether the person is at high (1) or low (0) risk of recidivism. The associated orders are as follows:

$$\begin{aligned}(\texttt{Age}, \preceq_{\texttt{Age}}) &:= (\mathbb{N}, \geq),\\ (\texttt{Priors}, \preceq_{\texttt{Priors}}) &:= (\mathbb{N}, \leq),\\ (\texttt{Sex}, \preceq_{\texttt{Sex}}) &:= (\{\text{M}, \text{F}\}, \{(\text{F}, \text{F}), (\text{M}, \text{M}), (\text{F}, \text{M})\}).\end{aligned}$$

If a relation R is defined on all dimension we can, for fact situations P and Q, refer to the set of dimensions on which R holds with $[R(P, Q)] := \{d \in D \mid R(P(d), Q(d))\}$. For instance, instantiating $R := \not\preceq_s$ we have $[P \not\preceq_s Q] = \{d \in D \mid P(d) \not\preceq_s Q(d)\}$; the dimensions on which Q is not at least as good for s as P. Besides fact situations we will also consider *partial* fact situations, i.e. fact situations defined only on a particular subset of the dimensions. We can do so conveniently using the well established notation for function restriction. Let $f : X \to Y$ and $Z \subseteq X$, we obtain a function $f \restriction Z : Z \to Y$ by restriction: $f \restriction Z := \{(x, y) \in f \mid x \in Z\}$. For cases p and q with the same outcome s we write $W(p, q) := Q \restriction [P \not\preceq_s Q]$, the values of q on which q is *worse* than p for s, and $B(p, q) := Q \restriction [P \preceq_s Q]$, the values of q on which q is *better* than p for s.

Example 2.2. Suppose we have a case base of recidivism risk judgements, and two cases p, q with outcome 1 (i.e. judged high risk of recidivism) such that:

$$\begin{aligned} p(\texttt{Age}) &= 45, & p(\texttt{Priors}) &= 4, & p(\texttt{Sex}) &= \text{M},\\ q(\texttt{Age}) &= 50, & q(\texttt{Priors}) &= 5, & q(\texttt{Sex}) &= \text{M}. \end{aligned}$$

Now we can compute that $W(p, q) = \{(\texttt{Age}, 50)\}$ and $B(p, q) = \{(\texttt{Priors}, 5), (\texttt{Sex}, \text{M})\}$.

3. A Case-Based Reasoning Explanation Method

In this section we detail the workings of the dimension-based model of explanation of [1], which was inspired by the work of [7]. A more detailed comparison between [1], [7], and other related works, can be found in [1, Section 8]. The method is built upon the theory of precedential constraint of [2, 3] and conceptually tries to mimic the arguments relating to precedent used by lawyers with respect to case law. In such discussions, precedent cases are cited by both sides as a means of arguing that the present (focus) case should be decided similarly as the precedent. Both sides may attack the other's citations, by pointing to important differences between the citation and the focus case; and they may defend themselves against such attacks, by pointing to aspects of the focus case which compensates for these differences. Each of the elements of such a discussion – case citations, pointing to differences, and compensating for differences – has its counterpart in the formal model of explanation.

A key idea underlying the approach is that a tabular dataset for binary classification can be interpreted as a case base *CB* in the sense of Definition 2.2. The method assumes access to the training data used by the system, and interprets each of the features in the data as a dimension in the sense of Definition 2.1. The corresponding dimension orders may be determined by knowledge engineering, statistical methods, or a combination thereof. This gives us a body of precedent *CB* upon which the machine learning system bases its decisions.

Under this interpretation the machine learning system can be seen as deciding new fact situations for sides. The goal is to explain a particular decision of a fact situation F for a side s, called the *focus case* $f = F : s$. This explanation is provided in the form of a *best citable precedent* $p \in CB$ together with an *explanation dialogue* in which the choice for this p is justified. This dialogue is formalized as a winning strategy in the grounded argument game of a particular abstract argumentation framework.

Before we can apply the theory of precedential constraint, we should specify the dimensions as in Definition 2.1, and we begin in Section 3.1 by mentioning the method used for doing so in [1, 6]. Any explanation dialogue should start with the citation of a best citable case. A suggestion for the definition of this notion is given in [1] and we continue by recalling it in Section 3.2, after which we explain and motivate the presence of the arguments occurring in the argumentation framework in Sections 3.3 and 3.4. We are then ready to give the formal definition of the framework in Section 3.5, explain what it means to have a winning strategy in the argument game it induces, and as such what constitutes an explanation according to the model.

3.1. Determining the Dimension Orders

In order to instantiate the explanation method for a particular dataset, we should specify the dimension orders as in Definition 2.1. As just noted, this may be done on the basis of knowledge engineering and/or statistical methods. In [1] a general method for determining the orders corresponding to the dimensions was proposed, using a function c that associates each ordinal feature x in the data with a coefficient expressing the degree to which the values in the range of x prefer outcome 1. See [6, Section 4.2] for a more detailed explanation.

3.2. Citability of Cases

An important aspect of the explanations produced by the method of [1] is the selection of the precedent case p with which it initiates its explanation of the outcome of the focus case f. We will now describe how this selection procedure works; later in Section 4 we return to this topic to suggest improvements. We begin with the notion of citability.

Definition 3.1. A case p is *citable* for a case f if

(a) both cases have the same outcome s; and

(b) there is a dimension d such that $p(d) \preceq_s f(d)$.

Since this is a quite weak requirement there may in general be very many citable cases p for any given f. For this reason the notion is strengthened by requiring that p should have a *minimal* number of relevant differences with f, according to some suitable notion of minimality. To make this formal we should first define what a relevant differences is. This is accomplished by [1, Definition 11], which we repeat here.

Definition 3.2. The set $D(p, f)$ of *relevant differences* between $p = P : s$ and $f = F : t$ is

$$D(p, f) := P \restriction [P \not\preceq_s F] = \{(d, P(d)) \mid d \in D,\ P(d) \not\preceq_s F(d)\}.$$

In other words, the relevant differences are given by the values of the precedent p on the dimensions on which f is not better than p for s. Now a *best* citable precedent should minimize this set of differences, in the following sense.

Definition 3.3. A case p is a *best citable* case for a case f if

(a) p is citable for f; and

(b) there is no other q satisfying (a) for which $D(q, f) \subset D(p, f)$.

3.3. Compensation of Relevant Differences

An idea central to the explanation dialogues is that when a precedent p does not force a focus case f, the values $W(p, f)$ on which f is worse than p for their outcome can be *compensated* for by the values $B(p, f)$ on which f is better than p. This idea is often encountered in the literature on case-based reasoning, see e.g. [8], where certain compensations are described as "*showing that at a more abstract level, a parallel exists between the cases, arguing in effect that the apparent distinction is merely a mismatch of details.*"

In our context we assume the existence of a relation *SC* on partial fact situations x, y, where *SC*(y, x) says that y compensates for x. This is used in practise as follows. Consider a precedent p and a focus case f, both with outcome s. If p forces the decision of f then f is at least as good as p for s on all dimensions, so $\emptyset = W(p, f)$ or equivalently $B(p, f) = f$. If this is not the case, then $\emptyset \subset W(p, f)$ or equivalently $B(p, f) \subset f$, and for p to justify the outcome of f we should have that $B(p, f)$ compensates for $W(p, f)$ as determined by whether *SC*$(B(p, f), W(p, f))$ holds.

3.4. Opposing Citations and Case Transformations

The last component of the dialogue is opposing citations, to which a response is possible through the use of case transformations. The idea is that the proponent of the decision of f for its outcome s can have their citation countered by the citation of a case q with outcome $\bar{s}$, as a means of saying that q should be a more appropriate precedent to draw on. This is analogous to the argument between lawyers in a legal case.

Definition 3.4. We define a semantics function $[\![\cdot]\!]$ on the compensation arguments by:

$$[\![\mathrm{Compensates}_p(y,x)]\!] := (P \setminus P \restriction \mathrm{dom}(x)) \cup x : s.$$

A case p *can be transformed into* q iff $p = q$ or there exists $X \in \mathcal{A}_p$ such that $[\![X]\!] = q$.

The goal of the semantics function is to change p into a case q that forces the outcome of f. It does so by replacing the values of the precedent case with those of the focus case, on those dimensions on which the focus case is not at least as good as the precedent.

3.5. An Abstract Argumentation Framework for Explanation

We are now ready to describe the formal account of the explanation dialogues in [1] through the use of an *abstract argumentation framework*, a concept introduced in [4]. An abstract argumentation framework $AF = (\mathrm{Arg}, \mathrm{Attack})$ is a directed graph, in which the nodes are interpreted as arguments and the edges as an attack relation between them.

An argumentation framework $(\mathrm{Arg}, \mathrm{Attack})$ is defined in [1] that combines the types of arguments defined in the preceding Sections 3.2, 3.3, and 3.4, relative to a *focus case* $f = F : s$. To do so we first define, for a particular precedent $p = P : s$ that may be cited in defense of the decision of F for s, a subset $\mathcal{A}_p \subseteq \mathrm{Arg}$ as follows:

$$\mathcal{A}_p := \bigcup \Big\{ \begin{array}{l} \{\mathrm{Worse}_p(x) \mid x = W(p,f) \neq \emptyset\}, \\ \{\mathrm{Compensates}_p(y,x) \mid \mathrm{Worse}_p(x) \in \mathcal{A}_p,\ y \subseteq B(p,f),\ SC(y,x)\}, \\ \{\mathrm{Transformed}_p(q) \mid p \text{ can be transformed into a case } q \text{ with } q \preceq f\} \end{array} \Big\}. \tag{1}$$

Definition 3.5. Given a finite case base *CB*, a focus case $f = F : s$, and a compensation relation *SC*, an *abstract argumentation framework for explanation with dimensions* is a pair $AF = (\mathrm{Arg}, \mathrm{Attack})$ where the arguments Arg are given by

$$\mathrm{Arg} := CB \cup \bigcup \{\mathcal{A}_p \mid p \in CB \text{ if } p \text{ has the same outcome as } f\},$$

and for arguments $X, Y \in \mathrm{Arg}$ we have $\mathrm{Attack}(X,Y)$ if and only if either:

- $X, Y \in CB$ have different outcomes and $[X \not\preceq f] \not\subset [Y \not\preceq f]$;
- $Y \in CB$ and X is of the form $\mathrm{Worse}_Y(x)$;
- Y is of the form $\mathrm{Worse}_p(x)$ and X is of the form $\mathrm{Compensates}_p(y,x)$; or
- $Y \in CB$ has outcome $\bar{s}$ and X is of the form $\mathrm{Transformed}_p(q)$.

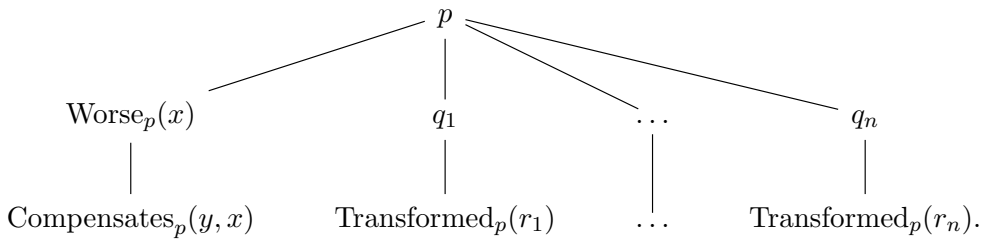
A dialogue now takes the form of a *grounded argument game* played on $(\mathrm{Arg}, \mathrm{Attack})$. For the sake of brevity we only give an intuitive explanation of how this works, the reader is referred to [1] for a detailed treatment of the subject.

An argument game on an AF (A, R) is a two-player game, in which the players take turns playing arguments from A which must attack the previously played argument according to the attack relation R. A player can win the game by moving an argument to which the other player cannot reply, and a *winning strategy* for a player is a method of playing that ensures a win regardless of how the opponent plays.

We now have the formal machinery in place to define explanations as in [1].

Definition 3.6. An *explanation* of a focus case f is a winning strategy in the grounded argument game starting with the citation of a best citable precedent $p \in CB$ for f, played on the abstract argumentation framework for explanation with dimensions $(\mathrm{Arg}, \mathrm{Attack})$.

The winning strategies may be viewed as trees and have the following general shape:



4. On the Citability of Cases

Let us now consider some possible modifications of Definition 3.3 to better formalize the intuitive notion of a most closely related case p of our focus case f.

Firstly, since Definition 3.2 does not gather just the dimensions on which f is worse than p but also the value of p at that dimension, a situation can arise where there is some case q with $[Q \not\preceq_s F] \subset [P \not\preceq_s F]$ but $Q \restriction [Q \not\preceq_s F] \not\subset P \restriction [P \not\preceq_s F]$, just because there is some dimension $d \in [Q \not\preceq_s F]$ with $Q(d) \neq P(d)$. It does not seem correct to dismiss q as a good citation simply because it disagrees with p on a single dimension, especially when $[Q \not\preceq_s F]$ is only a very small subset of $[P \not\preceq_s F]$. Let us look at an example to illustrate this point.

Example 4.1. We consider three cases p, q, f with outcome 1 (meaning they were judged high risk of recidivism) in the recidivism scenario of Example 2.1:

$$\begin{array}{lll} p(\texttt{Age}) = 20, & p(\texttt{Sex}) = \mathrm{M}, & p(\texttt{Priors}) = 3, \\ q(\texttt{Age}) = 50, & q(\texttt{Sex}) = \mathrm{M}, & q(\texttt{Priors}) = 1, \\ f(\texttt{Age}) = 40, & f(\texttt{Sex}) = \mathrm{M}, & f(\texttt{Priors}) = 2. \end{array}$$

We have that $D(p, f) = \{(\texttt{Age}, 20), (\texttt{Priors}, 3)\}$ and $D(q, f) = \{(\texttt{Priors}, 1)\}$. Therefore, even though there are fewer dimensions on which q has relevant differences with f – as $\{\texttt{Priors}\} \subset \{\texttt{Age}, \texttt{Priors}\}$ – this does not prevent p from being considered a best citable precedent for f – as $\{(\texttt{Priors}, 1)\} \not\subset \{(\texttt{Age}, 20), (\texttt{Priors}, 3)\}$.

This consideration suggests the definition should require minimality of $[P \not\preceq_s F]$ instead of $P \restriction [P \not\preceq_s F]$. However, this modification leaves room for a second type of scenario where there is some precedent q which is intuitively much closer to the focus case relatively to some other p, without hindering p from being considered best citable. To see why we consider a set of $n+1$ dimensions $\{d_0, \ldots, d_n\}$. Now we may have that $[Q \not\preceq_s F] = \{d_0\}$ and $[P \not\preceq_s F] = \{d_1, \ldots, d_n\}$. This means that the presence of q does not hinder p's being considered a best citable precedent for f, even though f is worse than p on n times as many dimensions as it is worse on than q. To remedy this, we could require minimality of the number of dimensions rather than the set of dimensions itself, i.e. of $|[P \not\preceq_s F]|$.

In addition to looking just at differences between the precedent and focus case it may be beneficial to also consider the similarities since after all, the *stare decisis* doctrine states that similar cases must be decided similarly. To achieve this we can require the best citable precedent to subsequently maximize $|[P = F]|$, so that it both minimizes differences and maximizes similarities. In all, this leads us to the following definition.

Definition 4.1. A case p is a *best citable* case for a case f if it satisfies the conditions (a) p is citable for f; (b) there is no other q satisfying (a) with $|[Q \not\preceq_s F]| < |[P \not\preceq_s F]|$; (c) there is no other q satisfying (a) and (b) with $|[Q = F]| > |[P = F]|$.

Experimental results in [1] showed that there are in general many cases satisfying Definition 3.3 for any f. Measured on three datasets, the mean and standard deviation of the number of best citable cases were respectively 82 ± 123.6, 76 ± 134, and 106 ± 116.5 [1, Table 5]. Recalculating these statistics for the same datasets with Definition 4.1 instead results in respectively 5.6 ± 2.0, 2.1 ± 2.6, and 2.6 ± 2.5 average number of best citable cases; a substantial decrease. Still, the definition remains somewhat ad-hoc, and more research is needed to assess its adequacy.

5. Specifying the Compensation Relation

In [1] no further explicit assumptions are made of the compensation relation *SC*. However in order for this relation to function according to our intuitions it may be necessary to do so, and we now consider a few such requirements. Let us first illustrate *SC* through a continuation of Example 2.2.

Example 5.1. We saw two example cases p, q where q was worse than p on the dimensions `Age` and `Sex`, but better on `Priors`. Suppose that for a number of priors higher than 4, we no longer care about values besides the number of priors. Then we may define

$$SC(y, x) \quad \text{if and only if} \quad y(\texttt{Priors}) \geq 4.$$

In this case the worse values $W(p, q)$ would indeed be compensated for by the better values $B(p, q)$, since $q(\texttt{Priors}) = 5$.

A point to consider is whether the compensation relation should itself adhere to an a fortiori principle. That is to say, if a set y is capable of compensating for a set x, should a superset $z \supseteq y$ be capable of compensating for x as well?

Definition 5.1. A compensation relation *SC* is *monotone* if for any partial fact situations x, y, z it holds that *SC*(y, x) implies *SC*$(y \cup z, x)$.

The same goes for values that are being compensated for; if a set y can compensate for a set x then we might require of it to compensate any subset $z \subseteq x$ as well.

Definition 5.2. A compensation relation *SC* is *antitone* if for any partial fact situations x, y, z it holds that *SC*(y, x) implies *SC*$(y, x \cap z)$.

In the factor based model of explanation in [1], i.e. the special case where the dimensions are all two element sets with a linear order, it is possible to compensate for a set of worse values in parts through the use of a pSubstitutes(y, x, c)&cCancels(y', x', c) move [1, Definition 5]. We can translate this to the dimensional setting as follows.

Definition 5.3. A compensation relation *SC* is *linear* if for any partial fact situations w, x, y, z it holds that *SC*(w, x) and *SC*(y, z) imply *SC*$(w \cup y, x \cup z)$.

A more fundamental question regarding the compensation relation is that of *context dependence*; should the compensation of two sets be allowed to depend on the context in which it takes place? This question and its consequences are the subject of Section 6.

6. Justification as an Extension of Forcing

An interesting way to think of the compensation relation is as an extension of the notion of forcing between cases. In essence a compensation says that while a precedent p might not force the decision of some other case q, the obstructing relevant differences can be compensated, and so the precedent p may still be said to *justify* the outcome of q.

6.1. Context-Dependent Compensations

A downside of the formal specification of this compensation relation is that it is defined on partial fact situations, rather than just fact situations. This makes it impossible for compensations to take the values of the precedent into account when allowing compensations to be made.

Example 6.1. In Example 2.2 the difference in age between p and q is only 5, and we may want to say that $B(p, q)$ compensates for $W(p, q)$ in this case if we find this difference small enough to be insignificant. To make this compensation possible formally we would need to postulate *SC*$(\{(\texttt{Age}, 50)\}, \{(\texttt{Priors}, 5), (\texttt{Sex}, \text{M})\}$ but this would inadvertently sanction compensations where the age of the precedent case is, say, 20, in which case we may find the difference in age large enough to be significant.

Modifying *SC* so that it takes the precedents' values into account yields a relation on full fact situations. A natural requirement of any such relation is that it *extends* the forcing relation $\preceq$ of Definition 2.4. This is akin to saying that any set can compensate for the empty set. This leads us to the following definition.

Definition 6.1. A relation $\sqsubseteq$ on cases is called a *justification* relation if it extends the forcing relation $\preceq$, i.e. if $\preceq \subseteq \sqsubseteq$.

Note that any compensation relation *SC* gives rise to a justification relation $\sqsubseteq_{SC}$:

$$p \sqsubseteq_{SC} q \quad \text{if and only if} \quad p \preceq q \text{ or } SC(B(p,q), W(p,q)). \tag{2}$$

The converse does not hold, precisely because a justification relation takes into account the *context* of the compensation. To see this, consider the naïve approach of obtaining a compensation relation $SC_{\sqsubseteq}$ from a justification relation $\sqsubseteq$:

$$SC_{\sqsubseteq}(y, x) \quad \text{if and only if} \quad p \sqsubseteq q \text{ for some } p, q \text{ with } x = W(p,q), y = B(p,q). \tag{3}$$

The problem is that this definition is not necessarily *well defined*, meaning that the truth value of $SC_{\sqsubseteq}(y, x)$ may depend on the particular representatives p and q that are used for its evaluation. This leads us to define the notion of a context-independent $\sqsubseteq$, requiring exactly that the relation $SC_{\sqsubseteq}$ above is well defined.

Definition 6.2. A justification relation $\sqsubseteq$ is *context-independent* if for any four cases p, q, r, s with $W(p,q) = W(r,s)$ and $B(p,q) = B(r,s)$ it holds that $p \sqsubseteq q$ iff $r \sqsubseteq s$.

6.2. Winning Strategies and Justification

The terminology of Definition 6.1 is inspired by [1], where an argument is said to be justified if and only if the proponent has a winning strategy in the grounded argument game about the argument. We will now formally justify this comparison by showing that for any compensation relation *SC* the proponent of an initial citation p has a winning strategy in the game on the argumentation framework if and only if $p \sqsubseteq_{SC} f$ (of Eq. (2)).

Let us fix a precedent case p and a focus case f, and introduce some shorthand terminology to ease our work. We will say a case p has a winning strategy if the proponent has a winning strategy in the grounded argument game on the explanation AF $(\mathrm{Arg}, \mathrm{Attack})$ of Definition 3.5, starting with a citation of p. Following [1] we distinguish between *nontrivial* winning strategies for p, in which p can be attacked by a $\mathrm{Worse}_p(x)$ move, and *trivial* winning strategies for p, in which there is no $\mathrm{Worse}_p(x)$ attack possible. In other words, a winning strategy for p is nontrivial if $\mathrm{Worse}_p(x) \in \mathcal{A}_p$ and trivial if $\mathrm{Worse}_p(x) \notin \mathcal{A}_p$, with $\mathcal{A}_p$ as defined in Eq. (1).

Proposition 6.1. There is a trivial winning strategy for p if and only if $p \preceq f$.

Proof. Note that $\mathrm{Worse}_p(x) \notin \mathcal{A}_p$ iff $W(p,f) = \emptyset$ iff $p \preceq f$. Hence left to right is immediate. For right to left we note in addition that any citation made by the opponent can be attacked with a $\mathrm{Transformed}_p(p)$ move, and so since there is no reply possible to a Transformed move the proponent has a (trivial) winning strategy for p. □

Proposition 6.2. There is a nontrivial winning strategy for p if and only if $W(p,f) \neq \emptyset$ and $SC(B(p,f), W(p,f))$.

Proof. Suppose the proponent has a winning strategy. Since $\mathrm{Worse}_p(x) \notin \mathcal{A}_p$ attacks the initial citation of p there should be a $\mathrm{Compensates}_p(y,x)$ response to the $\mathrm{Worse}_p(x)$ move available to the proponent, with $y = B(p,q)$. This implies that $SC(B(p,f), W(p,f))$.

For the other direction we begin by noting that because $W(p,q) \neq \emptyset$ there is $\mathrm{Worse}_p(x) \in \mathcal{A}_p$, and so the assumption $SC(B(p,f),W(p,f))$ guarantees that there is $C = \mathrm{Compensates}_p(y,x) \in \mathcal{A}_p$. Now, there are two types of moves available to the opponent to which we need a reply.

1. The first is $\mathrm{Worse}_p(x) \in \mathcal{A}_p$. As mentioned we have a reply C available, and since a compensation move cannot be replied to the game is won by the proponent.

2. The second is the citation of a case $q \in CB$ with outcome $\bar{s}$ for which it holds that $[q \not\preceq f] \not\subset [p \not\preceq f]$. By Definition 3.4 we have that p can be transformed into $p' = [\![C]\!]$, and so we can reply to the citation with $\mathrm{Transformed}_p(q) \in \mathcal{A}_p$. There are no more moves available to the opponent and so the proponent wins the game. □

Corollary 6.2.1. There is a winning strategy for p if and only if $p \sqsubseteq_{SC} f$.

Proof. Applying Eq. (2) and then Propositions 6.1 and 6.2 we get

$$\begin{aligned} p \sqsubseteq_{SC} f \text{ iff } & p \preceq q \text{ or } SC(B(p,q),W(p,q)) \\ & \text{iff } p \text{ has a trivial winning strategy or } p \text{ has a nontrivial winning strategy} \\ & \text{iff } p \text{ has a winning strategy.} \end{aligned}$$ □

Under this view of the winning strategies, and employing a fully general definition of compensation through a justification relation $\sqsubseteq$, we can now rephrase Definition 3.6 of explanations in the following way.

Definition 6.3. An *explanation* of a case f is a best citable precedent $p \in CB$ with $p \sqsubseteq f$.

The theory of precedential constraint describes how the outcome of a fact situation can be forced by precedent. However the collection of precedents may not be sufficient to force the outcome of all possible new fact situations. If such an undecided fact situation presents itself there may still be a precedent which, on the basis of additional reasoning, can be argued to *justify* an outcome for the fact situation. This is the view suggested by Corollary 6.2.1; a justification relation goes beyond the forcing relation by sanctioning citations of precedents that do not strictly force the outcome of the focus case.

6.3. A Relational Description of the Explanation Model

Having shown that a justification relation in some sense corresponds to the winning strategies underlying the explanations of [1], we can give a succinct description of the explanation method just through the use of relations on cases. Let us think of citability as a relation $\trianglelefteq$, then those $p \in CB$ related to the focus case through the intersection $\sqsubseteq \cap \trianglelefteq$ with f are said to explain the focus case f, i.e. those p with $p \sqsubseteq f$ and $p \trianglelefteq f$.

The model in [1] is a top-level model as it does not give explicit definitions of these notions, apart from suggesting a definition for the citability relation $\trianglelefteq$ as in Definition 3.3, and a method for determining $\preceq$ on the basis of Pearson correlation coefficients. In its running example and the experiments in [1, Section 6] all compensations are allowed, so that $\sqsubseteq \cap \trianglelefteq = \trianglelefteq$. Through

the relational view we summarize these inputs as follows: 1. The forcing relation $\preceq$, determined by specifying the dimensions and their orders. 2. The justification relation $\sqsubset$, determined by specifying the compensations. 3. The citability relation $\trianglelefteq$, determined by the definition of a best citable precedent. This view considerable simplifies the presentation of the model as it does not rely on the concepts of argumentation frameworks and winning strategies.

7. Discussion and Conclusion

We have described the explanation model of [1] in Section 3, which provides explanations as winning strategies on the grounded argument game of an abstract argumentation theory. In Section 6 we showed that this model admits an equivalent rephrasing in terms of relations, in which explanations are provided as cases related to the focus case through justification and citation relations. Most notably this shows that the explanation model can in some sense be seen as adding a notion of justification to the theory of precedential constraint as a relation $\sqsubset$ extending the forcing relation $\preceq$.

We conclude by noting an important consideration for future work on the topic. In order to apply this notion of justification to the explanation of machine-learned decisions, it is imperative that the input parameters – that being the forcing, justification, and citation relations – are constructed in such a way that they are faithful to the rationale of the black-box under consideration, because otherwise such a justification runs a high risk of becoming a rationalization if it does not reflect the real reasons behind the decision.

Acknowledgments

This research was (partially) funded by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, grant number 024.004.022. We also thank the referees for very useful commentary and suggestions.

References

[1] H. Prakken, R. Ratsma, A top-level model of case-based argumentation for explanation: Formalisation and experiments, Argument & Computation 13 (2022) 159–194.
[2] J. F. Horty, Rules and reasons in the theory of precedent, Legal Theory 17 (2011) 1–33.
[3] J. Horty, Reasoning with dimensions and magnitudes, Artificial Intelligence and Law 27 (2019) 309–345.
[4] P. Dung, On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games, Artificial Intelligence 77 (1995) 321–357.
[5] V. Aleven, K. D. Ashley, Evaluating a learning environment for case-based argumentation skills, in: Proceedings of the 6th international conference on Artificial intelligence and law, 1997, pp. 170–179.

[6] W. van Woerkom, D. Grossi, H. Prakken, B. Verheij, Landmarks in case-based reasoning: From theory to data, in: Proceedings of the First International Conference on Hybrid Human-Machine Intelligence, Frontiers of AI, IOS Press, 2022, p. tbd.
[7] K. Čyras, D. Birch, Y. Guo, F. Toni, R. Dulay, S. Turvey, D. Greenberg, T. Hapuarachchi, Explanations by arbitrated argumentative dispute, Expert Systems with Applications 127 (2019) 141–156.
[8] V. Aleven, Using background knowledge in case-based legal reasoning: A computational model and an intelligent learning environment, Artificial Intelligence 150 (2003) 183–237.