A top-level model of case-based argumentation for explanation

Prakken, Hendrik

A Top-level Model of Case-based Argumentation for Explanation

Henry Prakken[1]

Abstract. This paper proposes a formal top-level model of explaining the outputs of machine-learning-based decision-making applications. The model draws on AI & law research on argumentation with cases, which models how lawyers draw analogies to past cases and discuss their relevant similarities and differences. The model is top-level in that it can be extended with more refined accounts of similarities and differences between cases.

1 INTRODUCTION

There is currently an explosion of interest in automated explanation of machine-learning applications [1, 16, 20]. Some methods assume access to the learned model (model-aware explanation) while other methods assume no such access (model-agnostic explanation). This paper presents a model-agnostic method for explaining learned classification models, motivated by the fact that access to a learned model often is impossible (since the application is proprietary) or uninformative (since the learned model is not transparent). We only assume access to the training data and the possibility to observe a learned model's output given input data. We take an example-based approach, in which case outcomes are explained by comparing the case to similar cases in the training set. We in particular draw on AI & law research on argumentation with cases, which models how lawyers draw analogies to past cases and discuss their relevant similarities and differences. A case-based approach is natural since the training data of machine-learning algorithms can be seen as collections of cases. Our explanation model is top-level in that it can be extended with more refined accounts of case similarity.

There is so far little work on argumentation for model-agnostic explanation of machine-learning algorithms but recent research suggests the feasibility of an argumentation approach. We are inspired by the work of Čyras et al. [30, 29], also applied by [14]. They define cases as sets of binary features plus a binary outcome. Then they explain the outcome of a 'focus case' in terms of a graph structure that essentially utilises an argument game for grounded semantics of abstract argumentation semantics [15, 21]. We want to use the latter idea while overcoming some limitations of Čyras et al.'s approach. First, they do not consider the tendency of features to favour one side or another, while in many applications information on these tendencies will be available. Second, their features are binary, while many realistic applications will have multi-valued features. Finally, they leave the precise nature in which their graph structures explain an outcome somewhat implicit. We want to address all three limitations in terms of recent AI & law work on case-based reasoning.

[1] Department of Information and Computing Sciences, Utrecht University, and Faculty of Law, University of Groningen, The Netherlands, email: h.prakken@uu.nl

This paper is organised as follows. We present preliminaries in Section 2 and outline our general approach in Section 3. We then present a boolean-factor-based definition of case-based explanation dialogues in Section 4 and extend it to multi-valued factors or 'dimensions' in Section 5. We then briefly discuss how our top-level model can be extended with more refined accounts of similarities and differences between cases in Section 6, after which we conclude.

2 PRELIMINARIES

Many **AI & law accounts of argumentation with cases** (for an excellent overview see [8]) are applied to problems that are not decided by a clear rule but by weighing sets of relevant factors pro and con a decision. Legal data-driven algorithms are often applied to such factor-based problem domains [6, 13]. The seminal work is Rissland & Ashley's [28, 4, 5] work on the HYPO system for US trade secrets law. HYPO generates argument moves for analogizing or distinguishing precedents and hypothetical cases. Precedents can be cited to argue for the same outcome in the current case. Citations can then be distinguished by pointing at relevant differences between the precedent and the current case, and counterexamples, i.e., precedents with the opposite outcome, can be cited.

In AI & Law research, factors are legally relevant fact patterns assumed to favour one side or the other. Factors can be boolean (e.g. 'the secret was obtained by deceiving the plaintiff', 'a non-disclosure agreement was signed' or 'the product was reverse-engineerable') or multi-valued (e.g. the number of people to whom the plaintiff had disclosed the secret or the severity of security measures taken by the plaintiff). Multi-valued factors are often called dimensions; henceforth the term 'factor' will be reserved for boolean factors. In a factor-based approach, cases are defined as two sets of factors pro and con a decision (for example, there was misuse of trade secrets) plus (in case of precedents) the decision. Dimensions are not simply pro or con an outcome but are stronger or weaker for a side depending on their value in a case. Accordingly, in dimension-based approaches cases are defined as collections of value assignments to dimensions plus (for precedents) the decision. While HYPO-style work mainly focuses on rhetoric (generating persuasive debates), other work addresses the logical question how precedents constrain decisions in new cases. An important idea here is that precedents are sources of preferences between factor or dimension-value sets [25, 17, 9, 27, 18] and that these preferences are often justified by balancing underlying legal or societal values [12, 11].

An **abstract argument framework**, as introduced by Dung [15] is a pair $AF = \langle \mathcal{A}, attack \rangle$, where $\mathcal{A}$ is a set of arguments and *attack* a binary relation on $\mathcal{A}$. A subset $\mathcal{B}$ of $\mathcal{A}$ is *conflict-free* if no argument in $\mathcal{B}$ attacks an argument in $\mathcal{B}$ and it is *admissible* if it is

conflict-free and *defends* itself against any attack, i.e., if an argument A_1 is in $\mathcal{B}$ and some argument A_2 in $\mathcal{A}$ but not in $\mathcal{B}$ attacks A_1, then some argument in $\mathcal{B}$ attacks A_2. The theory of AFs identifies sets of arguments (called *extensions*) which are all admissible but may differ on other properties. In this paper we focus on the *grounded extension*, which is always unique. In particular, our explanations will take the form of an *argument game* between a proponent and opponent of an argument (in our approach a case citation for an outcome to be explained) that can be used to verify whether an individual argument is in the grounded extension. The game is sound and complete with respect to grounded semantics [23, 21]. The game starts with an argument by the proponent and then the players take turns after each argument: the opponent must attack the proponent's last argument while the proponent must one-way attack the opponent's last argument (i.e., the attacked argument does not in turn attack the attacker). A player *wins an argument game* iff the other player cannot move. An argument is *justified* (i.e., in the grounded extension) iff the proponent has a winning strategy in a game about the argument, i.e., if the proponent can make the opponent run out of moves in whatever way the opponent plays. A strategy for the proponent can be displayed as a tree of games which only branches after the proponent's moves and which contains all attackers of this move. A strategy for a player is *winning* if all games in the tree end with a move by that player.

For describing **factor-based models of precedential constraint** we first recall some notions concerning factors and cases often used in AI & law (e.g. in [17, 27, 18]), although sometimes with some notational differences. Let o and o' be two outcomes and Pro and Con two disjoint sets of atomic propositions favouring, respectively, outcome o and o'. The variable s (for 'side') ranges over $\{o, o'\}$ and $\overline{s}$ denotes o' if $s = o$ while it denotes o if $s = o'$. A set $F \subseteq Pro \cup Con$ *favours* side s (or F is pro s) if $s = o$ and $F \subseteq Pro$ or $s = o'$ and $F \subseteq Con$. For any set F of factors the set $F^s \subseteq F$ consists of all factors in F that favour side s. A *fact situation* is any subset of $Pro \cup Con$. A *case* is then a triple $(pro(c), con(c), outcome(c))$ where $outcome(c) \in \{o, o'\}$. Moreover, $pro(c) \subseteq Pro$ if $outcome(c) = o$ and $pro(c) \subseteq Con$ if $outcome(c) = o'$. Likewise, $con(c) \subseteq Con$ if $outcome(c) = o$ and $con(c) \subseteq Pro$ if $outcome(c) = o'$. Finally, a *case base* CB is a set of cases.

We next summarise Horty's [17] factor-based 'result' model of precedential constraint (the differences with his 'reason model' are for present purposes irrelevant, which we therefore do not discuss).

Definition 1 [Preference relation on fact situations [17].] Let X and Y be two fact situations. Then $X \leq_s Y$ iff $X^s \subseteq Y^s$ and $Y^{\overline{s}} \subseteq X^{\overline{s}}$.

$X <_s Y$ is defined as usual as $X \leq Y$ and $Y \not\leq X$. This definition says that Y is at least as good for s as X iff Y contains at least all pro-s factors that X contains and Y contains no pro-$\overline{s}$ factors that are not in X.

Definition 2 [Precedential constraint with factors [17].] Let CB be a case base and F a fact situation. Then, given CB, deciding F for s is *forced* iff there exists a case $c = (X, Y, s)$ in CB such that $X \cup Y \leq_s F$.

Horty thus models *a fortiori reasoning* in that an outcome in a focus case is forced if a precedent with the same outcome exists such that all their differences make the focus case even stronger for their outcome than the precedent. As for terminology, a case base CB is *inconsistent* if and only if there exists a fact situation F such that, given CB, both deciding F for s and deciding F for $\overline{s}$ is forced.

As our running example we use a small part of the US trade secrets domain of the HYPO and CATO systems. We assume the following six factors along with whether they favour the outcome 'misuse of trade secrets' (π for 'plaintiff') or 'no misuse of trade secrets' (δ for 'defendant'): the defendant had obtained the secret by deceiving the plaintiff (π_1) or by bribing an employee of the plaintiff (π_2), the plaintiff had taken security measures to keep the secret (π_3), the product is not unique (δ_1), the product is reverse-engineerable (δ_2) and the plaintiff had voluntarily disclosed the secret to outsiders (δ_3). We assume the following precedents:

$c_1(\pi)$: *deceived*$_{\pi 1}$, *measures*$_{\pi 3}$, *not-unique*$_{\delta 1}$, *disclosed*$_{\delta 3}$
$c_2(\delta)$: *bribed*$_{\pi 2}$, *not-unique*$_{\delta 1}$, *disclosed*$_{\delta 3}$

Clearly, deciding a fact situation F for π is forced iff it has at least the π-factors $\{\pi_1, \pi_3\}$ and at most the δ-factors $\{\delta_1, \delta_3\}$ (by precedent c_1), since then we have $\{\pi_1, \pi_3\} \subseteq F^\pi$ and $F^\delta \subseteq \{\delta_1, \delta_3\}$. Likewise, deciding a fact situation for δ is forced iff it has at least the δ-factors $\{\delta_1, \delta_3\}$ and at most the π-factor $\{\pi_2\}$ (by precedent c_2).

Consider next the following fact situation:

F_1: *bribed*$_{\pi 2}$, *measures*$_{\pi 3}$, *reverse-eng*$_{\delta 2}$, *disclosed*$_{\delta 3}$

Comparing F_1 with c_1 we must check whether $\{\pi_1, \pi_3, \delta_1, \delta_3\} \leq_\pi \{\pi_2, \pi_3, \delta_2, \delta_3\}$. This is not the case, for two reasons. We have $\{\pi_1, \pi_3\} \not\subseteq F_1^\pi = \{\pi_2, \pi_3\}$ and we have $F_1^\delta = \{\delta_2, \delta_3\} \not\subseteq \{\delta_1, \delta_3\}$. Next, comparing with precedent c_2 we must check whether $\{\pi_2, \delta_1, \delta_3\} \leq_\delta \{\pi_2, \pi_3, \delta_2, \delta_3\}$. This is also not the case for two reasons. We have $\{\delta_1, \delta_3\} \not\subseteq F_1^\delta = \{\delta_2, \delta_3\}$ and we have $F_1^\pi = \{\pi_2, \pi_3\} \not\subseteq \{\pi_2\}$. So neither deciding F_1 for π nor deciding F_1 for δ is forced. Henceforth we will assume it was decided for π.

We finally recall some ideas and results of [24] and add a new result to them. In [24] a similarity relation is defined on a case base given a focus case and a correspondence is proven with Horty's factor-based model of precedential constraint. The similarity relation is defined in terms of the relevant differences between a precedent and the focus case. These differences are the situations in which a precedent can be distinguished in a HYPO/CATO-style approach with factors [4, 2], namely, when the new case lacks some factors pro its outcome that are in the precedent or has new factors con its outcome that are not in the precedent. To define the similarity relation, it is relevant whether the two cases have the same outcome or different outcomes.

Definition 3 [Differences between cases with factors [24].] Let c and f be two cases. The set $D(c, f)$ of differences between c and f is defined as follows.

1. If $outcome(c) = outcome(f)$ then $D(c, f) = pro(c) \setminus pro(f) \cup con(f) \setminus con(c)$.
2. If $outcome(c) \neq outcome(f)$ then $D(c, f) = pro(f) \setminus con(c) \cup pro(c) \setminus con(f)$.

Consider again our running example and consider first any focus case f with outcome π and with a fact situation that has at least the π-factors $\{\pi_1, \pi_3\}$ and at most the δ-factors $\{\delta_1, \delta_3\}$. Then $D(c, f) = \emptyset$. Likewise with any focus case f with outcome δ and with a fact situation that has at least the δ-factors $\{\delta_1, \delta_3\}$ and at most the π-factor $\{\pi_2\}$. Next, let f be a focus case with fact situation F_1 and outcome π. We have

$D(c_1, f) = \{$*deceived*$_{\pi 1}$, *reverse-eng*$_{\delta 2}\}$
$D(c_2, f) = \{$*measures*$_{\pi 3}$, *not-unique*$_{\delta 1}\}$

The following result, which yields a simple syntactic criterion for determining whether a decision is forced, is proven in [24].

Proposition 1 Let CB be a case base CB and f a focus case with fact situation F. Then deciding F for s is forced given CB iff there exists a case c with outcome s in CB such that $D(c, f) = \emptyset$.

We call a case *citable* given f iff it shares at least one factor pro its outcome with f and they have the same outcome [4]. Then clearly every case c such that $D(c, f) = \emptyset$ is citable. A new result is that for any two cases with opposite outcomes that both have differences with the focus case, their sets of differences with the focus case are mutually incomparable (as with c_1 and c_2 in our running example).

Proposition 2 Let CB be a case base, f a focus case and c and c' two cases with opposite outcomes and with non-empty sets of differences with f. Then $D(c, f) \not\subseteq D(c', f)$ and $D(c', f) \not\subseteq D(c, f)$.

3 Approach and assumptions

We next sketch our general approach and its underlying assumptions. For a given classification model resulting from supervised learning we assume knowledge of the set of the model's input features, i.e., factors or dimensions and a binary outcome, plus the ability to observe the output of the learned model for given input. We also assume knowledge about the tendency of the input factors or dimensions towards a specific outcome, plus access to the training set from which the classification model was learned (data plus label). We then want to generate an explanation for a specific input-output pair of the classification model (the focus case) in terms of similar cases in the training set. Later we will briefly discuss a more general task where further domain specific information may be used to generate the explanation.

Since we have no access to the classification model, we do not know how the decision makers reasoned when deciding the cases in the training set. All we can do is generate the explanations in terms of a reasoning model that is arguably close to the domain, such as the above-described AI & law models of case-based argumentation. Accordingly, our aim is to investigate to what extent an explanation can be given in terms of these argumentation models.

It may happen that the outcomes of the classification and argumentation models disagree for a given input. Such a discrepancy does not imply that the argumentation model is wrong. It may also be that the learned classification model is wrong, since such models are rarely 100% accurate. If the two models disagree, it may be informative to show the user under which assumptions the outcome of the learned model is forced according to the argumentation model. The user can then decide whether to accept these assumptions. Accordingly, the information our explanations should provide is twofold: whether the focus case is forced, and if not, then what it takes to make it forced. Our explanation model can thus not only provide understanding of the learned model but also grounds to critique it.

4 EXPLANATION WITH FACTORS

We now present our top-level model for case-based explanation dialogues with factors, formalised as an application of the grounded argument game to a case-based abstract argumentation framework. The idea is that the proponent starts a dialogue for the explanation of a given focus case f by citing a most similar precedent in the case base CB with the same outcome as the focus case. Then the opponent can cite counterexamples and can distinguish the initial precedent on its differences with the focus case. The proponent then replies to the distinguishing moves with arguments why these differences are irrelevant and to the counterexamples in a way explained below.

Definition 4 formalises these ideas. We first informally introduce it. The set $\mathcal{A}$ of arguments consists of a case base of precedents assumed to be citable given a focus case, plus a set $\mathcal{M}$ of arguments about precedents. Conflicts between precedents are resolved by using the similarity relation as a preference ordering on $\mathcal{A}$. The attack relations from members of $\mathcal{M}$ on members of $\mathcal{A}$ or $\mathcal{M}$ implicitly define the flow of the dialogue. The first two moves in $\mathcal{M}$ are meant as 'distinguishing' attacks on an initial citation of a precedent c. $MissingPro(c, x)$ says that the focus case f lacks pro-s factors x of precedent c, while $NewCon(c, x)$ says that the focus case f contains new con-s factors x that are not in precedent c. These moves correspond to the two ways of distinguishing a case in [4, 2]. In our running example a citation of c_1 can be a attacked by $MissingPro(c_1, \{deceived_{\pi 1}\})$ and by $NewCon(c_1, \{$*reverse-eng*$_{\delta 2}\})$ (all moves in our running example are shown in Figure 1).

The next six moves are meant as replies to such distinguishing moves. They are inspired by the 'downplaying a distinction' moves from [2] (although that work does not contain counterparts of our $cSubstitutes$ and $cCancels$ moves). The first two downplay a *MissingPro* move. First, a $pSubstitutes(y, x, c)$ move says that the missing pro-s factors x are in a sense still in f, since they can be substituted with the new, similar pro-s factors y, so that the old preference in c for $pro(c)$ over $con(c)$ also holds for $pro(f)$ over $con(c)$. For example, in the US trade secrets domain both bribing an employee of the plaintiff and deceiving the plaintiff are questionable means to obtain the trade secret [2]. So in our running example the proponent can reply with $pSubstitutes(\{bribed_{\pi 2}\}, \{deceived_{\pi 1}\}, c_1)$. Second, a $cCancels(y, x, c)$ reply says that the negative effect of the missing pro-s factors x in f is cancelled by the positive effect of the missing con-s factors y in f, so that the old preference in c for $pro(c)$ over $con(c)$ still holds for $pro(f)$ over $con(f)$. For example, the $MissingPro(c_1, \{deceived_{\pi 1}\})$ attack can be counterattacked with $cCancels(c_1, \{$*not-unique*$_{\delta 1}\}, \{deceived_{\pi 1}\})$.

There are also two ways to downplay a *NewCon* distinction. The $cSubstitutes(y, x, c)$ move says that the new con-s factors y in f are in a sense already in the old case since they are similar to the old con-s factors x in c, so that the old preference in c for $pro(c)$ over $con(c)$ also holds for $pro(c)$ over $con(f)$. This move mirrors a *p-substitutes* move. In the US trade secrets domain, the two pro-δ factors that the product was not unique and that it was reverse-engineerable can both be seen as cases where the piece of trade information was known or elsewhere available [2]. So in our running example the proponent can reply with $cSubstitutes(\{$*reverse-eng*$_{\delta 2}\}, \{$*not-unique*$_{\delta 1}\}, c_1)$. Second, $pCancels(y, x, c)$ says that the negative effect of the new con-s factors x in f is cancelled by the positive effect of the new pro-s factors y in f, so that the old preference in c for $pro(c)$ over $con(c)$ also holds for $pro(f)$ over $con(f)$. This moves mirrors a *c-cancels* move. For example, the $NewCon(c_1, \{$*reverse-eng*$_{\delta 2}\})$ attack can be counterattacked with $pCancels(c_1, \{bribed_{\pi 2}\}, \{$*reverse-eng*$_{\delta 2}\})$.

For now all these moves will simply be formalised as statements. Later, in Section 6, we briefly discuss how full-blown arguments can be constructed with premises supporting these statements. To this end, our formal definition of the set of arguments assumes an unspecified set sc of definitions of p- and c-substitution and p- and c-cancellation relations, as placeholders for explicit accounts of these notions. Note that all downplaying moves allow the factor sets used to downplay a distinction to be empty, as ways of saying that the differences between the precedent and the focus case do not matter.

A complication is that a *MissingPro* or *NewCon* argument can be attacked in different ways on different subsets of the missing pro-s or new con-s factors. For instance, two different missing pro factors

may be p-substituted with two different new pro factors, or one subset of the missing pro-factors can be p-substituted by new pro-factors while another subset can be c-cancelled by missing con-s factors. The first situation can be accounted for in definitions in the set sc and will therefore be left implicit below. To deal with the second situation, the downplaying attacks will be formalised as combinations of an elementary *p(c)-substitutes* and/or *c(p)-cancels* move.

The last move is meant as a reply to a counterexample. For now its underlying idea can only be outlined. It is meant to say that an initial citation of a most similar case for the outcome of f can be transformed by the downplaying moves into a case with no relevant differences with f and which can therefore attack the counterexample. A more formal explanation can only be given after Definition 5.

Definition 4 [Case-based argumentation frameworks for explanation with factors.] Given a finite case base CB, a focus case $f \notin CB$ such that all cases in CB are citable given f, and definitions sc of substitution and cancellation, an *abstract argumentation framework for explanation with factors* $eAF_{CB,f,sc}$ is a pair $\langle \mathcal{A}, attack \rangle$ where:

- $\mathcal{A} = CB \cup \mathcal{M}$ where $\mathcal{M} =$
  $\{MissingPro(c,x) \mid x \neq \emptyset$ and $x = D(c,f) \cap pro(c)\} \cup$
  $\{NewCon(c,x) \mid x \neq \emptyset$ and $x = D(c,f) \cap con(f)\} \cup$
  $\{pSubstitutes(y,x,c) \mid x = D(c,f) \cap pro(c)$ and $y \subseteq pro(f) \setminus pro(c)$ and y p-substitutes x according to $sc\} \cup$
  $\{cSubstitutes(y,x,c) \mid x = con(c) \setminus con(f)$ and $y \subseteq D(c,f) \cap con(f)$ and y c-substitutes x according to $sc\} \cup$
  $\{pCancels(y,x,c) \mid x = D(c,f) \cap con(f)$ and $y \subseteq pro(f) \setminus pro(c)$ and y p-cancels x according to $sc\} \cup$
  $\{cCancels(y,x,c) \mid x = D(c,f) \cap pro(c)$ and $y \subseteq con(c) \setminus con(f)$ and y c-cancels x according to $sc\} \cup$
  $\{pSubstitutes(y,x,c) \& \{cCancels(y',x',c) \mid$
  $pSubstitutes(y,x,c) \in \mathcal{A}$ and $cCancels(y',x',c) \in \mathcal{A}\} \cup$
  $\{cSubstitutes(y,x,c) \& \{pCancels(y',x',c) \mid$
  $cSubstitutes(y,x,c) \in \mathcal{A}$ and $pCancels(y',x',c) \in \mathcal{A}\} \cup$
  $\{Transformed(c,c') \mid c \in CB$ and c can be transformed into $c'\}$
- A attacks B iff:
  - $A, B \in CB$ and $outcome(A) \neq outcome(B)$ and $D(B,f) \not\subset D(A,f)$;
  - $B \in CB$ and $outcome(B) = outcome(f)$ and A is of the form $MissingPro(B,x)$ or $NewCon(B,x)$;
  - B is of the form $MissingPro(c,x)$ and:
    - A is of the form $pSubstitutes(y,x,c)$ or $cCancels(y,x,c)$ and in both cases $x = D(c,f) \cap pro(c)$; or
    - A is of the form $pSubstitutes(y,x,c) \& cCancels(y',x',c)$ and $x \cup x' = D(c,f) \cap pro(c)$;
  - B is of the form $NewCon(c,x)$ and
    - A is of the form $cSubstitutes(y,x,c)$ or $pCancels(y,x,c)$ and in both cases $x = D(c,f) \cap con(f)$; or
    - A is of the form $cSubstitutes(y,x,c) \& pCancels(y',x',c)$ and $y \cup x' = D(c,f) \cap con(f)$;
  - $B \in CB$ and $outcome(B) \neq outcome(f)$ and A is of the form $Transformed(c,c')$ and $c \in CB$ is a case with $outcome(c) \neq outcome(f)$ and a subset-minimal $D(c,f)$ among the cases with the same outcome.

Henceforth the arguments that attack a $MissingPro$ or $NewCon$ move are sometimes called *downplaying moves*.

The grounded argument game now directly applies. The idea (inspired by [30, 29]) now is to explain the focus case f by showing a winning strategy for the proponent in the grounded game, which guarantees that the citation of the focus case is in the grounded extension of the argumentation framework defined in Definition 4. In our approach, the game should start with a 'best' precedent c in CB with the same outcome s as the focus case f (best in that there is no $c' \in CB$ with the same outcome as f and such that $D(c',f) \subset D(c,f)$). Moreover, any $Transformed(c,c')$ move must have as c the dialogue's initial move and as c' the transformation of c into c' during the dialogue according to Definition 5 below. Any strategy for the proponent that satisfies these constraints is called an *explanation* for f. As will become clear below, these further constraints do not affect the existence of a winning strategy.

For a focus case with outcome s, three situations are relevant.

(1) s is forced and $\overline{s}$ is not forced. Then the proponent has a trivial winning strategy, namely, to move a precedent with no relevant differences with the focus case, after which the opponent has no reply.

(2) Neither s nor $\overline{s}$ is forced. Then the proponent has a winning strategy if sc is *explanation complete* in that it always contains at at least one legal reply in the grounded game to a $MissingPro(c,x)$ or $NewCon(c,x)$ move. Then in a winning strategy T all branches are three moves deep: either citation - distinction - downplaying the distinction or citation - counterexample - attacking the counterexample. Moreover, the root of T has at most one $MissingPro$ reply and at most one $NewCon$ reply and at least one such reply, plus zero or more counterexample replies.

(3) $\overline{s}$ is forced. Then the proponent also has a winning strategy if if there is a citable precedent with outcome s and if sc is *explanation complete*, since any counterexample with no differences with the focus case that the opponent can move can be attacked with a *Transformed* move. This follows from Proposition 3 below since a substituting or cancelling set can, as explained above, be empty. Admittedly, such a justification of the outcome of the focus case is weak, but at least it informs a user that justifying the outcome of the focus case requires making the case base inconsistent.

One idea of our approach is that all moves in an explanation dialogue receive their meaning from (or are thus justified by) the formal theory of precedential constraint. To make this formal, we now specify the following operational semantics of the downplaying arguments in $\mathcal{M}$ as functions on the set of cases. The idea is that together these moves modify the root precedent of a strategy for the proponent into a case that makes f forced. Below $S^{y/x}$ stands for the set obtained by replacing subset x of S with y.

Definition 5 [Downplaying with factors: operational semantics] Given an $eAF_{CB,f,sc}$ and a case $c \in CB$ with outcome s:

- $pSubstitutes(y,x,c) = (pro(c)^{y/x}, con(c), s)$;
- $cSubstitutes(y,x,c) = (pro(c), con(c)^{y/x}, s)$;
- $pCancels(y,x,c) = (pro(c) \cup \{y\}, con(c) \cup \{x\}, s)$;
- $cCancels(y,x,c) = (pro(c) \setminus \{x\}, con(c) \setminus \{y\}, s)$;
- $pSubstitutes(y,x,c) \& cCancels(y',x',c) =$
  $pSubstitutes(y,x, cCancels(y',x',c))$;
- $cSubstitutes(y,x,c) \& pCancels(y',x',c) =$
  $pCancels(y,x, cSubstitutes(y',x',c))$.

A sequence $m_1(y_1,x_1,c_1), \ldots, m_n(y_n,x_n,c_n)$ of downplaying moves is an *explanation sequence* iff for every pair $m_i(y_i,x_i,c_i), m_{i+1}(y_{i+1},x_{i+1},c_{i+1})$ $(1 \leq i < n)$ it holds that $c_{i+1} = m_i(y_i,x_i,c_i)$.

In our running example we henceforth assume that O_{1a} is attacked

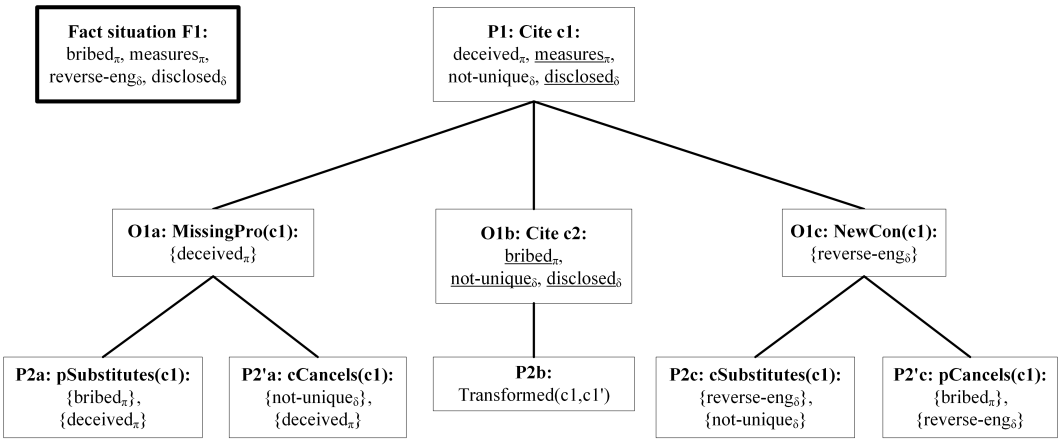

Figure 1. Example dialogue game tree.

with P_{2a} and O_{1c} with P_{2c}. Then c_1 is transformed into a case c_1' as follows. First, $pSubstitutes(\{bribed_{\pi 2}\}, \{deceived_{\pi 1}\}, c_1)$ yields

$c_1''(\pi)$: *bribed*$_{\pi 2}$, *measures*$_{\pi 3}$, *not-unique*$_{\delta 1}$, *disclosed*$_{\delta 3}$

Then $cSubstitutes(\{reverse\text{-}eng_{\delta 2}\}, \{not\text{-}unique_{\delta 1}\}, c_1)$ gives

$c_1'(\pi)$: *bribed*$_{\pi 2}$, *measures*$_{\pi 3}$, *reverse-eng*$_{\delta 2}$, *disclosed*$_{\delta 3}$

Note that $D(c', f) = \emptyset$, so adding c' to the case base would make deciding F_1 for π forced. The following result shows that this holds in general for when the proponent has a winning strategy.

Proposition 3 Let T be a winning strategy for P in an explanation dialogue and let $M = m_1, \ldots, m_n$ be any explanation sequence of all downplaying moves in T. Then the output of m_n is a case (X, Y, s) such that $pro(f) \cup con(f) \leq_s X \cup Y$.

PROOF. (Sketch) According to Definition 1 it must be shown that $X \subseteq pro(f)$ and $con(f) \subseteq Y$. Four cases must be considered. If T contains just one move, then f is forced and the result follows immediately by Definition 2. Otherwise, either T contains a $MissingPro$ reply but no $NewCon$ reply, or T contains a $NewCon$ reply but no $MissingPro$ reply, or T contains both a $MissingPro$ reply and $NewCon$ reply. In all three cases it is straightforward to verify that the initially cited case is gradually transformed into a case that makes the focus case forced, by successively applying the functions from Definition 5. QED

This proposition formally captures the sense in which the focus case is explained (for consistent case bases). If the focus case is forced, then any precedent with no relevant differences explains the focus case. Otherwise, an explanation sequence of downplaying moves derived from the winning strategy explains what has to be accepted to make the focus case forced; this information can be used to critique the explanation.

5 EXPLANATION WITH DIMENSIONS

We next adapt the above-defined factor-based explanation model to cases with dimensions. We first outline some formal preliminaries.

5.1 Dimension-based precedential constraint

We adopt from [18] the following technical ideas (again with some notational differences). A *dimension* is a tuple $d = (V, \leq_o, \leq_{o'})$ where V is a set (of values) and $\leq_o$ and $\leq_{o'}$ two partial orders on V such that $v \leq_o v'$ iff $v' \leq_{o'} v$. Given a dimension d, a *value assignment* is a pair (d, v), where $v \in V$. The functional notation $v(d) = x$ denotes the value x of dimension d. Then given a set D of dimensions, a *fact situation* is an assignment of values to all dimensions in D, and a *case* is a pair $c = (F, outcome(c))$ such that F is a fact situation and $outcome(c) \in \{o, o'\}$. Then a case base is as before a set of cases, but now explicitly assumed to be relative to a set D of dimensions in that all cases assign values to a dimension d iff $d \in D$. As for notation, $F(c)$ denotes the fact situation of case c and $v(d, c)$ denotes the value of dimension d in case or fact situation c. Finally, $v \geq_s v'$ is the same as $v' \leq_s v$.

Note that the set of value assignments of a case is unlike the set of factors of a case not partitioned into two subsets pro and con the case's outcome. The reason is that with value assignments it is often hard to say in advance whether they are pro or con the case's outcome. All that can often be said in advance is which side is favoured more and which side less if a value of a dimension changes, as captured by the two partial orders $\leq_s$ and $\leq_s'$ on a dimension's values.

In HYPO [28, 5], two of the factors from our running example are actually dimensions. *Security-Measures-Adopted* has a linearly ordered range, below listed in simplified form (where later items increasingly favour the plaintiff so decreasingly favour the defendant):

- *Minimal-Measures*, *Access-To-Premises-Controlled*, *Entry-By-Visitors-Restricted*, *Restrictions-On-Entry-By-Employees*

Moreover, *disclosed* has a range from 1 to some high number, where higher numbers increasingly favour the defendant so decreasingly favour the plaintiff. For the remaining four factors we assume that they have two values 0 and 1, where presence (absence) of a factor means that its value is 1 (0) and where for the pro-plaintiff factors we have $0 <_\pi 1$ (so $1 <_\delta 0$) and for the pro-defendant factors we have $0 <_\delta 1$ (so $1 <_\pi 0$).

Accordingly, we change our running example as follows.

$c_1(\pi)$: *deceived*$_{\pi 1}$, *measures* = *Entry-By-Visitors-Restricted*, *not-unique*$_{\delta 1}$, *disclosed* = 20

$c_2(\delta)$: *bribed*$_{\pi 2}$, *measures* = *Minimal*, *not-unique*$_{\delta 1}$, *disclosed* = 5

F_1: *bribed*$_{\pi 2}$, *measures* = *Access-To-Premises-Controlled*, *reverse-eng*$_{\delta 2}$, *disclosed* = 10

In Horty's [18] dimension-based result model of precedential con-

straint a decision in a fact situation is forced iff there exists a precedent c for that decision such that on each dimension the fact situation is at least as favourable for that decision as the precedent. He formalises this idea with the help of the following preference relation between sets of value assignments.

Definition 6 [Preference relation on dimensional fact situations [18].] Let F and F' be two fact situations with the same set of dimensions. Then $F \leq_s F'$ iff for all $(d, v) \in F$ and all $(d, v') \in F'$ it holds that $v \leq_s v'$.

In our running example we have for any fact situation F' that $F(c_1) \leq_\pi F'$ iff F' has π_1 but not δ_3 and $v(\textit{measures}, F') \geq_\pi$ *Entry-By-Visitors-Restricted* and $v(\textit{disclosed}, F') \geq_\pi 20$ (so ≤ 20). Likewise, $F(c_2) \leq_\delta F'$ iff F' has δ_1 but not π_1 and $v(\textit{measures}, F') =$ *Minimal* and $v(\textit{disclosed}, F') \geq_\delta 5$ (so ≥ 5).

Then adapting Definition 2 to dimensions is straightforward.

Definition 7 [Precedential constraint with dimensions [18].] Let CS be a case base and F a fact situation given a set D of dimensions. Then, given CB, deciding F for s is *forced* iff there exists a case $c = (F', s)$ in CB such that $F' \leq_s F$.

In our running example, deciding F_1 for π is not forced, for two reasons. First, $v(\textit{deceived}, c_1) = 1$ while $v(\textit{deceived}, F_1) = 0$ and for *deceived* we have that $0 <_\pi 1$. Second, $v(\textit{measures}, c_1) =$ *Entry-By-Visitors-Restricted* while $v(\textit{measures}, F_1) =$ *Access-To-Premises-Controlled* and *Access-To-Premises-Controlled* $<_\pi$ *Entry-By-Visitors-Restricted*. Deciding F_1 for δ is also not forced, since $v(\textit{measures}, c_2) =$ *Minimal* while $v(\textit{measures}, F_1) =$ *Access-To-Premises-Controlled* and *Minimal* $<_\delta$ *Access-To-Premises-Controlled*.

We next recall [24]'s adaptation of Definition 3 to dimensions. Unlike with factors, there is no need to indicate whether a value assignment favours a particular side, since we have the $\leq_s$ orderings.

Definition 8 [Differences between cases with dimensions [24].] Let $c = (F(c), \textit{outcome}(c))$ and $f = (F(f), \textit{outcome}(f))$ be two cases. The set $D(c, f)$ of differences between c and f is defined as follows.

1. If $\textit{outcome}(c) = \textit{outcome}(f) = s$ then $D(c, f) = \{(d, v) \in F(c) \mid v(d, c) \not\leq_s v(d, f)$.
2. If $\textit{outcome}(c) \neq \textit{outcome}(f)$ where $\textit{outcome}(c) = s$ then $D(c, f) = \{(d, v) \in F(c) \mid v(d, c) \not\geq_{\overline{s}} v(d, f)$.

Let c be a precedent and f a focus case. Then clause (1) says that if the outcomes of the precedent and the focus case are the same, then any value assignment in the focus case that is not at least as favourable for the outcome as in the precedent is a relevant difference. Clause (2) says that if the outcomes are different, then any value assignment in the focus case that is not at most as favourable for the outcome of the focus case as in the precedent is a relevant difference. In our running example, we have:

$D(c_1, f) =$ $\{(\textit{deceived}, 1)$, (*reverse-eng*,0), (*measures*, *Entry-By-Visitors-Restricted*)$\}$

$D(c_2, f) =$ $\{(\textit{measures}, \textit{Minimal})$,(*not-unique*,0)$\}$

The following counterpart of Proposition 1 is proven in [24].

Proposition 4 Let, given a set D of dimensions, CB be a case base and f a focus case with fact situation F. Then deciding F for s is forced given CB iff there exists a case in CB with outcome s such that $D(c, f) = \emptyset$.

The counterpart of Proposition 2 can be proven as a new result.

Proposition 5 Let, given a set D of dimensions, CB be a case base, f a focus case and c and c' be two cases with opposite outcomes and both with a non-empty set of differences with f. Then $D(c, f) \not\subseteq D(c', f)$ and $D(c', f) \not\subseteq D(c, f)$.

PROOF. Suppose first that c and f have the same outcome and suppose that $(d, v) \in D(c, f)$. Then $v(d, c) \not\leq_s v(d, f)$, so $v(d, c) \not\geq_{\overline{s}} v(d, f)$, so $(d, v) \notin D(c', f)$. Suppose next that c and f have different outcomes and suppose that $(d, v) \in D(c, f)$. Then $v(d, c) \not\geq_{\overline{s}} v(d, f)$, so $v(d, c) \not\leq_s v(d, f)$, so $(d, v) \notin D(c', f)$. QED

5.2 Adding dimensions to the top-level model of explanation

When extending our explanation model with dimensions, it would at first sight seem that factors are simply a special case of dimensions with just two values 0 and 1 where $0 <_s 1$ while $1 <_{\overline{s}} 0$. However, upon closer inspection this is not the case, since with factors there is more to say then just that the two sides have opposed preferences over the presence or absence of a factor. Consider, for example, in the trade-secrets domain the factor *bribed*. That the defendant bribed one of the plaintiff's employees surely is a factor pro misuse of trade secrets, but that the defendant did not bribe any of the plaintiff's employees does not have to be regarded as a factor con that outcome: it can also be regarded as neutral with respect to that outcome. Therefore, it makes sense to treat factors differently than dimensions.

Accordingly, we introduce some new terminology. Each two-valued dimension in D comes with a partial function $t_d : V \longrightarrow \{o, o'\}$ that assigns to zero, one or both values of the dimension an outcome subject to two constraints (henceforth if d is two-valued and v is one value of d then $\overline{v}$ denotes the other value of d):

1. if $t_d(v) = o$ then $t_d(\overline{v}) = o'$ or $t_d(\overline{v})$ is undefined.
2. if $t_d(v) = o$ then $\overline{v} <_o v$.

The t_d function captures which outcome is favoured by a value of d, if any. Any value assignment (d, v) to a two-valued dimension d such that $t_d(v) = o$ is called a pro-o factor. The terminology of Section 4 also applies to such factors. Then D^t is the subset of D of two-valued dimensions for which t_d is defined for at least one value, and $D^m = D \setminus D^t$. A dimensional fact set F^t (F^m) assigns values to all dimensions in D^t (D^m).

In our running example, we assume that $t_d(v) = \pi$ for *deceived* and *bribed* with value 1 and $t_d(v) = \delta$ for *not-unique* and *reverse-eng* with value 1. In all other cases $t_d(v)$ is undefined.

It can be proven that if D^m is empty, that is, we only have factors, then Definition 7 of dimension-based precedential constraint reduces to its factor-based counterpart.

Proposition 6 Let $AF_{CB,f} = \langle \mathcal{A}, \textit{attack} \rangle$ given D be such that $D^m = \emptyset$, let $\textit{Pro} = \{(d, v) \mid t_d(v) = o\}$, let $\textit{Con} = \{(d, v) \mid t_d(v) = o'\}$ and for any dimensional fact situation F, let F^s be $\{d^v \mid v(d) \in F$ and $t_d(v(d)) = s\}$. Then f is forced according to Definition 7 iff f is forced according to Definition 2.

Next, we adapt Definition 4 of case-based argumentation frameworks for explanation to dimensions as follows. The idea of extending the explanation model with dimensions is to treat relevant differences differently according to whether they concern 'factors', i.e. elements of D^t, or 'dimensions', i.e., elements of D^m. First, that a precedent is *citable* given a focus case f now means that they have the same

outcome s and at least one dimension has a value in f that is at is at least as favourable for s as in the precedent and if all such dimensions are two-valued, then at least one yields a pro-s factor in the precedent and f. Next, the set $\mathcal{A}$ of arguments still includes the arguments of Definition 4 for when the sets x and y are in D^t, while for sets of value assignments in D^m the following arguments are added:

Definition 9 [Case-based argumentation frameworks for explanation with dimensions.] Given a finite case base CB, a focus case $f \notin CB$ such that all cases in CB are citable given f and definitions sc of substitution and cancellation, an *abstract argumentation framework for explanation with dimensions* $eAF_{CB,f,sc}$ is a pair $\langle \mathcal{A}, attack \rangle$ where:

- $\mathcal{A} = CB \cup \mathcal{M}$ where $\mathcal{M} =$
  $\{m \in \mathcal{M}$ from Definition 4 $\mid x, y$ in m assign values to dimensions in $D^t\} \cup$
  $\{Worse(c, x) \mid x \neq \emptyset$ and $x = \{(d, v) \in F^m(f) \mid v(d, f) <_{outcome(f)} v(d, c)\}\} \cup$
  $\{Compensates(y, x, c) \mid y = \{(d, v) \in F^m(f) \mid v(d, c) <_{outcome(f)} v(d, f)\}\}$
- A attacks B iff:
  - A attacks B according to Definition 4; or
  - $B \in CB$ and $outcome(B) = outcome(f)$ and A is of the form $Worse(B, x)$;
  - B is of the form $Worse(c, x)$ and A is of the form $Compensates(y, x, c)$; or
  - $B \in CB$ and $outcome(B) \neq outcome(f)$ and A is of the form $Compensates(y, x, c)$.

The $Compensates$ move is an additional *downplaying move*. It says that the factors on which the focus case is not at least as good for its outcome than the precedent are compensated by the factors on which the focus case is better for its outcome than the precedent. Like with the factor-based downplaying moves, a compensating set can be empty, as a way of saying that the values in the *Worse* set are still not bad enough to change the outcome. In our running example, a citation of c_1 by the proponent can now additionally be attacked by $Worse(c_1, \{measures\})$, since *Access-To-Premises-Controlled* $<_\pi$ *Entry-By-Visitors-Restricted*. This attack can be downplayed by $Compensates(\{disclosed\}, \{measures\}, c_1)$, since $20 <_\pi 10$.

Definition 5 is now extended as follows.

Definition 10 [Downplaying with dimensions: operational semantics] Given an $eAF_{CB,f,sc}$ and a case $c \in CB$ with outcome s:

- The semantics of the moves from Definition 4 is as in Definition 5;
- $Compensates(y, x, c) = (F^t(c) \cup F^{cm}(c), s)$, where $(d, v) \in F^{cm}(c)$ iff $(d, v) \in F^m(c) \setminus x \cup y$ or else $(d, v) \in x \cup y$.

In other words, on the dimensions with relevant differences, the precedent's values are replaced with the focus case's values. This way of downplaying dimensional differences is admittedly somewhat crude but more refined ways can only be defined if additional information is available (cf. Section 6 below). With this semantics for $Compensates$ moves, the proof of Proposition 3 can easily be adapted to the explanation model with dimensions. We omit the proposition and its proof for reasons of space.

6 EXTENDING THE TOP-LEVEL MODEL

So far we have modelled explanation dialogues that only use information from the case base, that is, from the training set of the machine-learning application. However, more relevant information may be available, provided in advance by a knowledge engineer or during an explanation dialogue by a user. It is for this reason that our explanation model contains a thus far undefined set sc of definitions of why downplaying arguments can be played (hence the qualification 'top-level' model). We now briefly discuss how explicit definitions of this set can be given and how they can be used to provide the premises of downplaying arguments.

AI & law provides many insights here [8]. For example, the premises of *pSubstitutes* and *cSubstitutes* claims can be founded on a 'factor hierarchy' as defined for the CATO system [3, 2]. We gave examples of this above. Furthermore, the *pCancels* and *cCancels* arguments can be said to express a preference for a set of pro factors over a set of con factors. In AI & law accounts have been developed of basing such preferences on underlying legal, moral or societal values. Arguments according to these accounts can provide the premises for the *pCancels* and *cCancels* claims. For example, move P'_{2c} from our running example could be based on a preference for promoting honesty over stimulating economic competition.

Applying these ideas requires that arguments have a richer internal structure, where the various claims become conclusions of inferences from sets of premises. One way to achieve this is to formalise relevant argument schemes in a suitable structured formal account of argumentation [19]. This approach was followed in the context of the *ASPIC*$^+$ framework [22] in [26, 10, 7]. It can be straightforwardly adapted to the present context in any formalism suitable for modelling reasoning with argument schemes. In *ASPIC*$^+$ the attacks defined in the present paper would reduce to the three general *ASPIC*$^+$-types of rebutting, undercutting and undermining attacks.

7 CONCLUSION

In this paper we have presented an argumentation-based top-level model of explaining the outcomes of machine-learning applications where access to the learned model is impossible or uninformative. The argumentation model can be used to explain but also to critique the outcome of the learned model (the latter in cases where the outcomes of the learned model and the argumentation model disagree). The presented model is still theoretical groundwork. It is top-level in that it can be extended with more refined accounts of similarities and differences between cases. Its suitability as an explanation model must still be tested. We have built on earlier work of [30, 29] but extended it to multi-valued features and to (boolean or multi-valued) features with a tendency towards a particular outcome. We have also discussed links with more refined case-based argumentation models and added a formally precise account of the sense in which argumentation dialogues explain an outcome.

Research to test our approach faces several challenges. Our running example was kept small for ease of explanation but with many features our approach might become unmanageable or non-informative. Techniques can be studied for focusing on subsets of a feature set, and other distance measures or functions can be studied as alternatives to our rather coarse similarity relation between cases. It would also be interesting to investigate whether it is realistic to allow users to add relevant information during an explanation dialogue. Another topic is adapting the present approach to contrastive explanation styles [20], in which an outcome is explained by contrasting it with similar cases with the opposite outcome. Finally, the ultimate test is whether our method helps users to better understand or to better critique outcomes of machine-learning applications.

REFERENCES

[1] A. Addadi and M. Berrada, 'Peeking inside the black box: a survey on explainable artificial intelligence (XAI)', *IEEE Access*, (2018). doi: 10.1109/ACCESS.2018.2870052.

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

[3] V. Aleven and K.D. Ashley, 'Doing things with factors', in *Proceedings of the Fifth International Conference on Artificial Intelligence and Law*, pp. 31–41, New York, (1995). ACM Press.

[4] K.D. Ashley, 'Toward a computational theory of arguing with precedents: accomodating multiple interpretations of cases', in *Proceedings of the Second International Conference on Artificial Intelligence and Law*, pp. 39–102, New York, (1989). ACM Press.

[5] K.D. Ashley, *Modeling Legal Argument: Reasoning with Cases and Hypotheticals*, MIT Press, Cambridge, MA, 1990.

[6] K.D. Ashley, *Artificial Intelligence and Legal Analytics. New Tools for Law Practice in the Digital Age*, Cambridge University Press, Cambridge, 2017.

[7] K.D. Atkinson, T.J.M. Bench-Capon, H. Prakken, and A.Z. Wyner, 'Argumentation schemes for reasoning about factors with dimensions', in *Legal Knowledge and Information Systems. JURIX 2013: The Twenty-sixth Annual Conference*, ed., K.D. Ashley, 39–48, IOS Press, Amsterdam etc., (2013).

[8] T.J.M. Bench-Capon, 'HYPO's legacy: introduction to the virtual special issue', *Artificial Intelligence and Law*, **25**, 205–250, (2017).

[9] T.J.M. Bench-Capon and K.D. Atkinson, 'Dimensions and values for legal CBR', in *Legal Knowledge and Information Systems. JURIX 2017: The Thirtieth Annual Conference*, eds., A.Z. Wyner and G. Casini, 27–32, IOS Press, Amsterdam etc., (2017).

[10] T.J.M. Bench-Capon, H. Prakken, A.Z. Wyner, and K. Atkinson, 'Argument schemes for reasoning with legal cases using values', in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*, pp. 13–22, New York, (2013). ACM Press.

[11] T.J.M. Bench-Capon and G. Sartor, 'A model of legal reasoning with cases incorporating theories and values', *Artificial Intelligence*, **150**, 97–143, (2003).

[12] D.H. Berman and C.D. Hafner, 'Representing teleological structure in case-based legal reasoning: the missing link', in *Proceedings of the Fourth International Conference on Artificial Intelligence and Law*, pp. 50–59, New York, (1993). ACM Press.

[13] L.K. Branting, 'Data-centric and logic-based models for automated legal problem solving', *Artificial Intelligence and Law*, **25**, 5–27, (2017).

[14] O. Cocarascu, K. Čyras, and F. Toni, 'Explanatory predictions with artificial neural networks and argumentation', in *Proceedings of the IJCAI/ECAI-2018 Workshop on Explainable Artificial Intelligence*, pp. 26–32, (2018).

[15] P.M. Dung, 'On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming, and *n*–person games', *Artificial Intelligence*, **77**, 321–357, (1995).

[16] 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, (2019).

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

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

[19] *Argument and Computation*, ed., A.J. Hunter, volume 5, 2014. Special issue with Tutorials on Structured Argumentation.

[20] T. Miller, 'Explanation in artificial intelligence: Insights from the social sciences', *Artificial Intelligence*, **267**, 1–38, (2019).

[21] S. Modgil and M. Caminada, 'Proof theories and algorithms for abstract argumentation frameworks', in *Argumentation in Artificial Intelligence*, eds., I. Rahwan and G.R. Simari, 105–129, Springer, Berlin, (2009).

[22] S. Modgil and H. Prakken, 'The ASPIC+ framework for structured argumentation: a tutorial', *Argument and Computation*, **5**, 31–62, (2014).

[23] H. Prakken, 'Dialectical proof theory for defeasible argumentation with defeasible priorities (preliminary report)', in *Formal Models of Agents*, eds., J.-J.Ch. Meyer and P.-Y. Schobbens, number 1760 in Springer Lecture Notes in AI, pp. 202–215, Berlin, (1999). Springer Verlag.

[24] H. Prakken, 'Comparing alternative factor- and precedent-based accounts of precedential constraint', in *Legal Knowledge and Information Systems. JURIX 2019: The Thirty-Second Annual Conference*, eds., M. Araszkiewicz and V. Rodriguez-Doncel, 73–82, IOS Press, Amsterdam etc., (2019).

[25] H. Prakken and G. Sartor, 'Modelling reasoning with precedents in a formal dialogue game', *Artificial Intelligence and Law*, **6**, 231–287, (1998).

[26] H. Prakken, A.Z. Wyner, T.J.M. Bench-Capon, and K. Atkinson, 'A formalisation of argumentation schemes for legal case-based reasoning in ASPIC+', *Journal of Logic and Computation*, **25**, 1141–1166, (2015).

[27] A. Rigoni, 'Representing dimensions within the reason model of precedent', *Artificial Intelligence and Law*, **26**, 1–22, (2018).

[28] E.L. Rissland and K.D. Ashley, 'A case-based system for trade secrets law', in *Proceedings of the First International Conference on Artificial Intelligence and Law*, pp. 60–66, New York, (1987). ACM Press.

[29] 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, (2019).

[30] K. Čyras, K. Satoh, and F. Toni, 'Abstract argumentation for case-based reasoning', in *Principles of Knowledge Representation and Reasoning: Proceedings of the Fifteenth International Conference*, pp. 549–552. AAAI Press, (2016).