A top-level model of case-based argumentation for explanation: formalisation and experiments

Henry Prakken [a,*], Rosa Ratsma [b]
[a] *Department of Information and Computing Sciences, Utrecht University, The Netherlands*
E-mail: h.prakken@uu.nl
[b] *Department of Information and Computing Sciences, Utrecht University, The Netherlands*
E-mail: rosaratsma@gmail.com

Abstract. This paper proposes a formal top-level model of explaining the outputs of machine-learning-based decision-making applications and evaluates it experimentally with three data sets. 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 in terms of relevant factors and dimensions in the problem domain. A case-based approach is natural since the input data of machine-learning applications can be seen as cases. While the approach is motivated by legal decision making, it also applies to other kinds of decision making, such as commercial decisions about loan applications or employee hiring, as long as the outcome is binary and the input conforms to this paper's factor- or dimension format. The model is top-level in that it can be extended with more refined accounts of similarities and differences between cases. It is shown to overcome several limitations of similar argumentation-based explanation models, which only have binary features and do not represent the tendency of features towards particular outcomes. The results of the experimental evaluation studies indicate that the model may be feasible in practice, but that further development and experimentation is needed to confirm its usefulness as an explanation model. Main challenges here are selecting from a large number of possible explanations, reducing the number of features in the explanations and adding more meaningful information to them. It also remains to be investigated how suitable our approach is for explaining non-linear models.

Keywords: explaining machine learning, argumentation, case-based reasoning

1. Introduction

There is currently an explosion of interest in automated explanation of machine-learning applications [1–3]. Some explanation methods explain the entire learned model (global explanation) while other methods explain an output in a specific case (local explanation). Also, some methods assume access to the model (model-aware explanation) while other methods assume no such access (model-agnostic explanation). This paper presents a model-agnostic method for (locally) explaining outcomes of learned classification models and evaluates it experimentally with three data sets. Our model-agnostic approach is 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 will only assume access to the training data and the possibility to observe a learned model's output given input data. There

[*] Corresponding author. E-mail: h.prakken@uu.nl.

1946-2166/0-1900/$35.00 © 0 – IOS Press and the authors. All rights reserved

are several model-agnostic explanation methods, including example-based methods, which explain a decision for a case by comparing it to similar cases in the training set [4]. We take a similar approach, drawing 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. The resulting explanation model does not explain how a learned model *reached* a decision (it cannot do so since it has no access to the learned model) but under which assumptions the decision can be *justified* in terms of an argumentation model. Our explanation model thus not only provides better understanding of the decision but also grounds to critique it.

Our explanation model is top-level in that it itself gives no reasons for why differences between cases matter or can be downplayed but can be extended with more refined accounts of similarities or differences between cases in which such reasons are given. The model is formally defined as an instance of Dung's well-known theory of abstract argumentation frameworks [5]. The results of our experimental evaluation studies indicate that our model is to some extent feasible in practice, but further development and more experimental studies, especially with human users and with extensions of our top-level model, are needed to confirm its usefulness as an explanation model. While our approach is motivated by legal decision making, it also applies to other kinds of decision making, such as commercial decisions about loan applications or employee hiring, as long as the outcome is binary and the input conforms to this paper's factor- or dimension format.

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. [6, 7], also applied by [8, 9]. They define cases as sets of binary features plus a binary outcome (in [7] also a case's 'stages' are considered, but for present purposes these can be ignored). 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 [5, 10]. 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. A more detailed comparison with [7] will be given in Section 8.

As suggested by [3], good explanations are selective, contrastive and social. That an explanation is *selective* means that only the most salient points relevant to an outcome are presented. That an explanation is *contrastive* means that it not just explains why the given outcome was reached but also why another outcome was not reached. Finally, that an explanation is *social* means that in the transfer of knowledge, assumptions about the user's prior knowledge and views influence what constitutes a proper explanation. In Section 10 we will argue that our model satisfies these criteria in several respects.

This paper is organised as follows. We present preliminaries in Section 2 and outline our general approach and its underlying assumptions 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. In Section 6 we report on experiments with datasets to evaluate our approach and in Section 7 we briefly discuss how our top-level model can be extended with more refined accounts of similarities and differences between cases. We then discuss related research in Section 8 after which in Section 9 we discuss in a more general sense whether our method is an explanation method or something else.

We conclude in Section 10. Sections 2–5 are an extended and somewhat revised version of [11] while Section 6 is adapted from [12].

2. Preliminaries

In this section we describe some preliminaries. After a brief overview of AI & law research on case-based reasoning, we briefly summarise the theory of abstract argumentation frameworks and outline the factor-based theory of precedential constraint. The dimension-based version of the latter theory will be discussed later in Section 5.

2.1. AI & law accounts of case-based reasoning

Many **AI & law accounts of argumentation with cases** (for an excellent overview see [13]) 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 [14, 15]. The seminal work on legal argumentation with factors is Rissland & Ashley's [16–18] 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 which tend to favour one side or the other. Factors have to be weighed or balanced in each case, unlike legal rules, of which the conditions are ordinarily sufficient for accepting their conclusion. 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. In his CATO system, Aleven [19, 20] also considers support and attack relations between less and more abstract factors (for instance, that the products are identical is a reason to believe that the information was used) and uses them to define argument moves for emphasizing or downplaying distinctions. 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 sets [21, 22] and that these preferences are often justified by balancing underlying legal or societal values [23, 24]. This work has recently been extended to dimensions [25–27]. Since new cases are rarely identical to precedents, these preferences and values often do not uniquely determine a decision in a new case. Therefore the systems based on these ideas suggest alternative decisions with their arguments pro and con.

2.2. Some theory of abstract argumentation frameworks

An **abstract argument framework**, as introduced by Dung [5] is a pair $AF = \langle \mathcal{A}, \mathit{attack} \rangle$, 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 $\mathcal{B}$. 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 and which is defined as the smallest conflict-free set $S \subseteq \mathcal{A}$ such that $A \in S$ if and only if S defends A against any attack. 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 [10, 28]. 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. As is well-known, 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.

2.3. Factor-based precedential constraint

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 [22, 26, 27]), 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 $s = o'$. We say that 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 $Pro \neq \emptyset$ and $outcome(c) \in \{o, o'\}$ (where we call $pro(c) \cup con(c)$ the fact situation of c). 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.

Note that given the way the tendency of factors towards a particular outcome is defined, we cannot model dependency relations between factors. For example, we cannot model that one factor is pro an outcome only if another factor is present. In AI & law research on case-based reasoning with factors or dimensions there is a general assumption that factors or dimensions are independent from each other but this assumption may not always be warranted. The assumption also means that our model may be less suitable for explaining non-linear decision-making models.

We next summarise Horty's [22] factor-based 'result' model of precedential constraint (the differences with his 'reason model' are irrelevant for present purposes, and hence not discussed here).

Definition 1. [Preference relation on fact situations [22].] Let X and Y be two fact situations. Then $X \leqslant_s Y$ iff $X^s \subseteq Y^s$ and $Y^{\overline{s}} \subseteq X^{\overline{s}}$.

$X <_s Y$ is defined as usual as $X \leqslant_s Y$ and $Y \nleqslant_s X$. 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 [22].] 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 $X \cup Y \leqslant_s F$.

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 as least as strong for their outcome as the precedent.

Definition 3. A case base CB is *inconsistent* if and only if it contains two cases c and c' with respectively, fact situations X and Y and outcomes s and $\overline{s}$ such that $X \leqslant_s Y$ (In that case deciding Y for both s and $\overline{s}$ is forced). And CB is *consistent* if and only if it is not inconsistent.

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 information is obtainable elsewhere (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:

$c_1(\pi)$: $\textit{deceived}_{\pi 1}$, $\textit{measures}_{\pi 3}$, $\textit{obtainable-elsewhere}_{\delta 1}$, $\textit{disclosed}_{\delta 3}$
$c_2(\delta)$: $\textit{bribed}_{\pi 2}$, $\textit{obtainable-elsewhere}_{\delta 1}$, $\textit{disclosed}_{\delta 3}$

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

Comparing F_1 with c_1 we must check whether $\{\pi_1, \pi_3, \delta_1, \delta_3\} \leqslant_\pi \{\pi_2, \pi_3, \delta_2, \delta_3\}$. This is not the case, for two reasons. We have $\{\pi_1, \pi_3\} \nsubseteq F_1^\pi = \{\pi_2, \pi_3\}$ and we have $F_1^\delta = \{\delta_2, \delta_3\} \nsubseteq \{\delta_1, \delta_3\}$. Next, comparing with precedent c_2 we must check whether $\{\pi_2, \delta_1, \delta_3\} \leqslant_\delta \{\pi_2, \pi_3, \delta_2, \delta_3\}$. This is also not the case for two reasons. We have $\{\delta_1, \delta_3\} \nsubseteq F_1^\delta = \{\delta_2, \delta_3\}$ and we have $F_1^\pi = \{\pi_2, \pi_3\} \nsubseteq \{\pi_2\}$. 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 [29] and add a new result to them. In [29] 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 [17, 20], 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.

If they have the same outcome, then a factor in the precedent that lacks in the focus case is only relevant if it is pro that outcome; otherwise its absence in the focus case only makes the focus case stronger than the precedent. Likewise, an additional factor in the focus case is only relevant if it is con the outcome. If the focus case and the precedent have opposite outcomes, then this becomes different. If the precedent has an additional factor that is pro the outcome in the precedent (so con the outcome in the focus case), then its missing in the focus case makes the focus case stronger than the other case, so it is a relevant difference. However, if the additional factor in the precedent is con the outcome in that case, then its missing in the focus case as (as a factor pro its outcome) makes the focus case weaker compared to the precedent. So this is not a relevant difference. On the other hand, if the focus case has an additional factor pro its outcome, then its addition to the precedent (as a factor con its outcome) might have changed its outcome, so this is a relevant difference between the cases. Finally, if the additional factor in the focus case is con its outcome, then adding it to the precedent (as a factor pro its outcome) strengthens the precedent, so this is not a relevant difference between the cases.

Definition 4. [Differences between cases with factors [29].] 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 $D(c, f) = pro(f) \setminus con(c) \cup pro(c) \setminus con(f)$.

Figure 1 illustrates this definition.

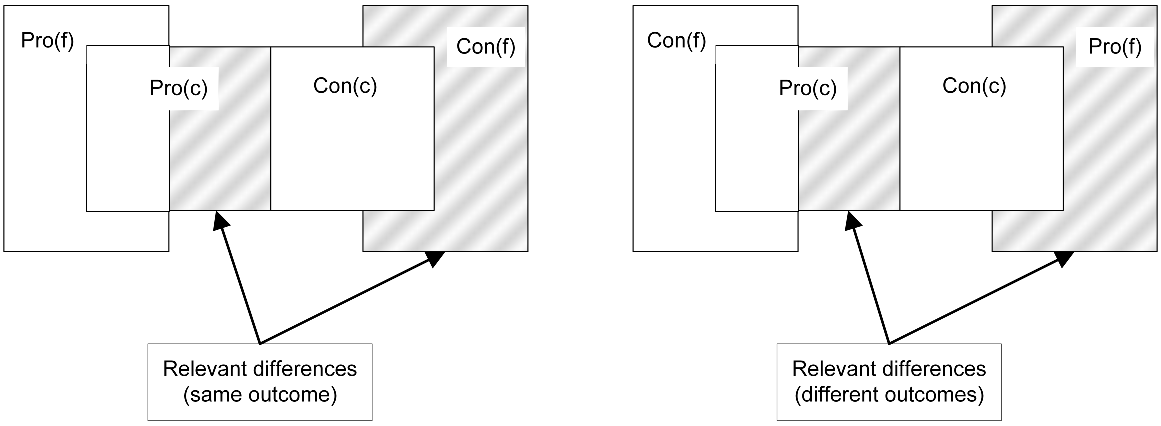
Pro(f)
Pro(c)
Con(c)
Con(f)
Relevant differences (same outcome)
Con(f)
Pro(c)
Con(c)
Pro(f)
Relevant differences (different outcomes)

Fig. 1. Relevant differences $D(c, f)$

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 $\{deceived_{\pi1}, measures_{\pi3}\}$ and at most the δ-factors $\{obtainable\text{-}elsewhere_{\delta1}, disclosed_{\delta3}\}$. 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 $\{obtainable\text{-}elsewhere_{\delta1}, disclosed_{\delta3}\}$ and at most the π-factor $\{bribed_{\pi2}\}$. Next, let f be a focus case with fact situation F_1 and outcome π. We have

$$D(c_1, f) = \{deceived_{\pi1},\ reverse\text{-}eng_{\delta2}\}$$
$$D(c_2, f) = \{measures_{\pi3},\ obtainable\text{-}elsewhere_{\delta1}\}$$

The following result, which yields a simple syntactic criterion for determining whether a decision is forced, is proven in [29].

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

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 [17]. Then clearly every case c such that $D(c, f) = \emptyset$ is citable (recall that every case contains at least one factor pro its outcome).

As a new result, it can be proven 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). The proofs of all results in this paper are given in Appendix A.1.

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, in Section 7, we will briefly discuss a more general task where further domain specific information may be used to generate the explanation. It should be noted that there is one important difference between our approach and the 'traditional' case-based argumentation models described above in Section 2.1: in our approach the outcome of the case to be explained is given as input while in traditional approaches the outcome of the new case is an output of the model.

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.

There are a few important differences between our approach and more familiar model-agnostic local explanation techniques. Feature summary approaches, such as LIME [30], explain a prediction in terms of the contribution of features to the model's decision. Our approach offers more context by explaining the decision relative to other instances (cases) in the data set. Providing this context can help answer contrastive questions, such as: 'Why did c have outcome x rather than y?'.

A second important difference is that our model does not explain how the learned model *reached* its decision but how this decision can be *justified* in terms of another model. It may happen that the outcomes of a black-box classification model and the argumentation-based model of precedential constraint disagree for a given input in that the output given by the classification model is not forced by the argumentation model. 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 better understanding of the decision but also grounds to critique it. In Section 9 we will discuss these points in more detail.

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 model is top-level in that it gives no reasons for why differences between cases matter or can be downplayed. In Section 7 we briefly discuss how our top-level model can be extended with more refined accounts of similarities and differences between cases in which such reasons can be given. The idea of our model 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. An explanation then amounts to a winning strategy for the proponent in the argument game that defines explanation dialogues; such a winning strategy shows how all possible attacks of the opponent on the initial citation can be counterattacked.

Definition 5 formalises these ideas in the form of a specification of a set of arguments plus an attack relation. We first informally introduce the definition. Figure 2 informally displays the two distinguishing moves and all ways to downplay them, while Figure 3 shows how these moves can be used in our running example. Further example explanations (with both factors and dimensions) are given in Appendix A.2.

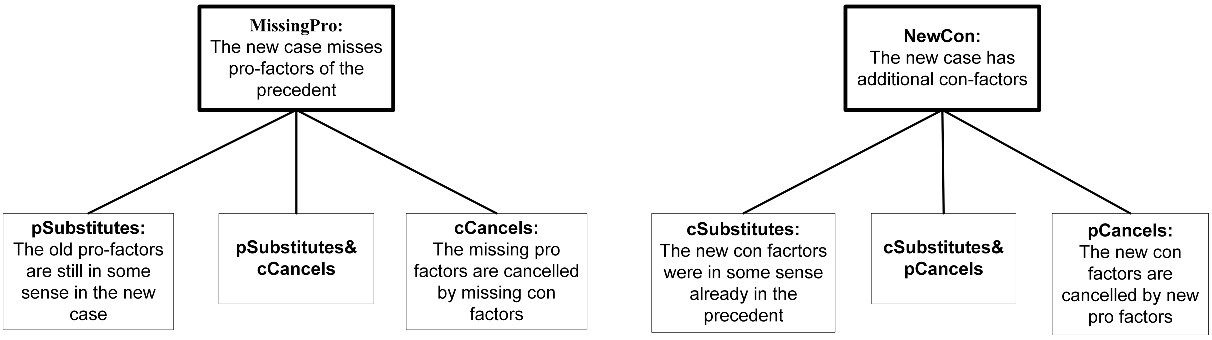
MissingPro:
The new case misses pro-factors of the precedent
NewCon:
The new case has additional con-factors
pSubstitutes:
The old pro-factors are still in some sense in the new case
pSubstitutes&
cCancels
cCancels:
The missing pro factors are cancelled by missing con factors
cSubstitutes:
The new con facrtors were in some sense already in the precedent
cSubstitutes&
pCancels
pCancels:
The new con factors are cancelled by new pro factors

Fig. 2. Distinguishing and downplaying distinctions.

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 with opposite outcomes are resolved by using the similarity relation between precedents as a preference ordering between the precedents in $\mathcal{A}$. 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. A $\mathit{MissingPro}(c, x)$ move says that the focus case f lacks pro-s factors x of precedent c, while $\mathit{NewCon}(c, x)$ 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 [17, 20]. In our running example a citation of c_1 can be a attacked by $\mathit{MissingPro}(c_1, \{\mathit{deceived}_{\pi 1}\})$ and by $\mathit{NewCon}(c_1, \{\mathit{reverse\text{-}eng}_{\delta 2}\})$.

The next six moves are meant as replies to such distinguishing moves. They are inspired by the 'downplaying a distinction' moves from [20] (although that work does not contain counterparts of our *cSubstitutes* and *cCancels* moves). The first two downplay a *MissingPro* move. First, a $\mathit{pSubstitutes}(y, x, c)$ 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 $\mathit{pro}(c)$ over $\mathit{con}(c)$ also holds for $\mathit{pro}(f)$ over $\mathit{con}(c)$. 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 [20]. So in our running example the proponent can reply with $\mathit{pSubstitutes}(\{\mathit{bribed}_{\pi 2}\}, \{\mathit{deceived}_{\pi 1}\}, c_1)$. Second, a $\mathit{cCancels}(y, x, c)$ 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 $\mathit{pro}(c)$ over $\mathit{con}(c)$ still holds for $\mathit{pro}(f)$ over $\mathit{con}(f)$. For example, the $\mathit{MissingPro}(c_1, \{\mathit{deceived}_{\pi 1}\})$ attack can be counterattacked with $\mathit{cCancels}(c_1, \{\mathit{obtainable\text{-}elsewhere}_{\delta 1}\}, \{\mathit{deceived}_{\pi 1}\})$.

There are also two ways to downplay a *NewCon* distinction. The $\mathit{cSubstitutes}(y, x, c)$ 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 $\mathit{pro}(c)$ over $\mathit{con}(c)$ also holds for $\mathit{pro}(c)$ over $\mathit{con}(f)$. This move mirrors a *p-substitutes* move. In the US trade secrets domain, the two pro-δ factors that the information is obtainable elsewhere and that it was reverse-engineerable can both be seen as cases where the piece of trade information was known or elsewhere available [20]. So in our running example the proponent can reply with $\mathit{cSubstitutes}(\{\mathit{reverse\text{-}eng}_{\delta 2}\}, \{\mathit{obtainable\text{-}elsewhere}_{\delta 1}\}, c_1)$. Second, $\mathit{pCancels}(y, x, c)$ 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 $\mathit{pro}(c)$ over $\mathit{con}(c)$ also holds for $\mathit{pro}(f)$ over $\mathit{con}(f)$. This moves mirrors a *c-cancels* move. For example, the $\mathit{NewCon}(c_1, \{\mathit{reverse\text{-}eng}_{\delta 2}\})$ attack can be counterattacked with $\mathit{pCancels}(c_1, \{\mathit{bribed}_{\pi 2}\}, \{\mathit{reverse\text{-}eng}_{\delta 2}\})$.

For now all these moves will simply be formalised as statements. Later, in Section 7, 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 how to handle 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 8.

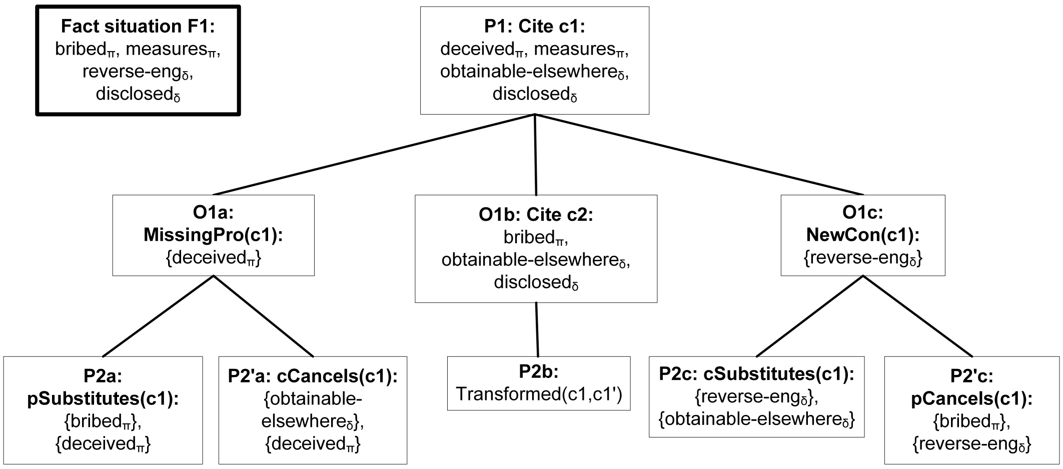
Fact situation F1:
bribedπ, measuresπ, reverse-engδ, disclosedδ
P1: Cite c1:
deceivedπ, measuresπ, obtainable-elsewhereδ, disclosedδ
O1a: MissingPro(c1):
{deceivedπ}
O1b: Cite c2:
bribedπ, obtainable-elsewhereδ, disclosedδ
O1c: NewCon(c1):
{reverse-engδ}
P2a: pSubstitutes(c1):
{bribedπ}, {deceivedπ}
P2'a: cCancels(c1):
{obtainable-elsewhereδ}, {deceivedπ}
P2b:
Transformed(c1,c1')
P2c: cSubstitutes(c1):
{reverse-engδ}, {obtainable-elsewhereδ}
P2'c: pCancels(c1):
{bribedπ}, {reverse-engδ}

Fig. 3. Example dialogue game tree.

Definition 5. [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 $\langle \mathcal{A}, \mathit{attack} \rangle$ 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 $sc\} \cup$
  $\{cSubstitutes(y,x,c) \mid x = D(c,f) \cap con(f)$ and $y \subseteq con(c) \setminus con(f)$ 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 $cCancels(y',x',c) \in \mathcal{A}$ and $x \cup x' = D(c,f)\} \cup$
  $\{cSubstitutes(y,x,c)\&\{pCancels(y',x',c) \mid$
  $cSubstitutes(y,x,c) \in \mathcal{A}$ and $pCancels(y',x',c) \in \mathcal{A}$ and $x \cup x' = D(c,f)\} \cup$
  $\{Transformed(c,c') \mid c \in CB$ and c can be transformed into c' and $D(c',f) = \emptyset\}$
- A attacks B iff:
  - $A, B \in CB$ and $outcome(A) \neq outcome(B)$ and $D(B,f) \not\subseteq D(A,f)$;
  - $B \in CB$ 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 $x \cup x' = D(c,f) \cap con(f)$;
- $*$ $B \in CB$ and $outcome(B) \neq outcome(f)$ and A is of the form $Transformed(c,c')$.

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

Definition 6. [Explanation-complete case-based AFs] An abstract argumentation framework for explanation with factors $\langle \mathcal{A}, attack \rangle$ is *explanation complete* iff every $MissingPro(c,x)$ or $NewCon(c,x)$ move has an attacker in *attack*.

The grounded argument game now directly applies. The idea (inspired by [6, 7]) 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 5. Such a winning strategy shows how all possible attacks of the opponent on the initial citation can be counterattacked. In our approach, an explanation dialogue should start with a 'best' citable 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' a transformation of c into c' during the dialogue according to Definition 8 below. Any strategy for the proponent that satisfies these constraints is an explanation for f. As will become clear below, these further constraints do not affect the existence of a winning strategy for at least one citation for the outcome of f.

Definition 7. Given an abstract argumentation framework for explanation with factors $eAF_{CB,f,sc} = (\mathcal{A}, F)$, a nonempty sequence of $Arg_1, \ldots, Arg_n, \ldots$, i.e., an *explanation dialogue*, satisfies the rules of the *explanation game* for $eAF_{CB,f,sc}$ iff $Arg_i \in \mathcal{A}$ for all i and:

(1) Arg_1 is a case from CB with the same outcome as f for which there is no $c' \in CB$ with the same outcome as f such that $D(c',f) \subset D(c,f)$ (i.e., c is a 'best' citable case for the outcome of f);
(2) if i and j are odd and $i \neq j$, then
  (a) $Arg_i \neq Arg_j$; and
  (b) Arg_i asymmetrically attacks Arg_{i-1};
(3) if i is even, then Arg_i attacks Arg_{i-1};
(4) if Arg_i is of the form $Transformed(c,c')$, then $Arg_1 = c$.

If i is odd (even), we say that the player of Arg_i is the *proponent* (opponent). A player *wins* an explanation dialogue iff the other player cannot move. An *explanation* for f on the basis of $eAF_{CB,f,sc}$ is any winning strategy (in the game-theoretic sense) for the proponent for the explanation game for $eAF_{CB,f,sc}$.

Recall that a strategy for a player can be displayed as a tree of dialogues which only branches after that player's moves and which contains all attackers of this move. A strategy for a player is then winning if all dialogues in the tree end with a move by that player; cf. [10].

One idea of our approach is that all moves in an explanation 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 8. [Downplaying with factors: operational semantics] Given an $eAF_{CB,f,sc}$ and a case $c \in CB$ with outcome s:

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

Then given $eAF_{CB,f,sc}$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 $m_i(y_i, x_i, c_i), m_{i+1}(y_{i+1}, x_{i+1}, c_{i+1})$ $(1 \leqslant i < n)$ it holds that $c_{i+1} = m_i(y_i, x_i, c_i)$. We say that the explanation sequence *transforms* c_1 into c_n, and we say that a case $c \in CB$ *can be transformed into* c' iff there exists an explanation sequence with all moves from $\mathcal{A}$ that transforms c into c'.

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

$c_1''(\pi)$: $bribed_{\pi 2}$, $measures_{\pi 3}$, $obtainable\text{-}elsewhere_{\delta 1}$, $disclosed_{\delta 3}$.

Then $cSubstitutes(\{reverse\text{-}eng_{\delta 2}\}, \{obtainable\text{-}elsewhere_{\delta 1}\}, c_1)$ gives

$c_1'(\pi)$: $bribed_{\pi 2}$, $measures_{\pi 3}$, $reverse\text{-}eng_{\delta 2}$, $disclosed_{\delta 3}$.

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.

Lemma 3. Given an $eAF_{CB,f,sc} = AF$ with explanation-complete sc, it holds for every $c \in CB$ with relevant differences with f but the same outcome as f that there exists an explanation sequence given AF that transforms c into a case c' such that $D(c', f) = \emptyset$.

Proposition 4. For any abstract argumentation framework for explanation with factors $eAF_{CB,f,sc}$ with explanation-complete sc and such that there are citable cases for the proponent, the proponent has a winning strategy in the explanation game for $eAF_{CB,f,sc}$ for any best citable case for the outcome of f.

Corollary 5. A case $c \in CB$ is in the grounded extension of $eAF_{CB,f,sc}$ if and only if the proponent has a winning strategy starting with c in the explanation game for $eAF_{CB,f,sc}$.

Note that Lemma 3 justifies that a *Transformed* move attacks a citation of a counterexample, since it says that the moves in an explanation dialogue that downplay a distinction transform the initially cited case in the dialogue to a case with no relevant differences with the focus case. This means that the counterexample to the initial citation moved by the opponent loses its force, since it is not a counterexample to the transformed case. This is one illustration of the idea that an explanation dialogue identifies the assumptions that have to be accepted to justify a predicted outcome of the focus case.

One underlying reason that Proposition 4 holds is that a substituting or cancelling set can be empty, which is necessary to make the proponent win if the opposite outcome of the focus case is forced. 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. In all other cases where the outcome of f is not forced, the following proposition tells us that a winning strategy in fact transforms the initial precedent into a precedent that does control the focus case.

Proposition 6. Let T be a winning strategy for the proponent in an explanation dialogue with an initial move c such that $D(c, f) \neq \emptyset$ and the opposite outcome as in f is not forced, 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 $X \cup Y \leqslant_s pro(f) \cup con(f)$.

Together, our results formally capture the sense in which an explanation according to Definition 7 explains the focus case. If the focus case is forced, then any precedent with no relevant differences explains the focus case; if neither outcome is forced, then an explanation sequence of downplaying moves derived from the winning strategy explains what has to be accepted to make the focus case forced; and if the opposite outcome is forced, then explaining the outcome boils down to explaining that justifying the outcome of f requires making the case base inconsistent.

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 [27] the following technical ideas (again with some notational differences). A *dimension* is a tuple $d = (V, \leqslant_o, \leqslant_{o'})$ where V is a set (of values) and $\leqslant_o$ and $\leqslant_{o'}$ two partial orders on V such that $v \leqslant_o v'$ iff $v' \leqslant_{o'} v$. 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 nonempty 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, $v \geqslant_s v'$ is the same as $v' \leqslant_s v$.

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 $\leqslant_s$ and $\leqslant_s'$ on a dimension's values.

Note also that, since the partial orders $\leqslant_o$ and $\leqslant_{o'}$ are defined on individual dimensions, we cannot model dependency relations between values of different dimensions. As noted in Section 2.3 about factor-based models, in AI & law research on case-based reasoning there is a general assumption that factors or dimensions are independent from each other but this assumption may not always be warranted. Finally, note that, given the way the two partial orders on a dimension are defined, we cannot model dimension intervals (for instance, in a medical example, low blood pressure is pro-disease, high blood pressure is pro-disease, normal blood pressure is con-disease). Future research should reveal how serious this limitation is in practical applications.

In HYPO [16, 18], 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*

(For simplicity we will below assume that each case contains exactly one security measure; generalisation to multiple measures is straightforward by defining the orderings on sets of measures.). 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.

$c_1(\pi)$: $\quad$ *deceived*$_{\pi 1}$, *measures* = *Entry-By-Visitors-Restricted*,
obtainable-elsewhere$_{\delta 1}$, *disclosed* $= 20$
$c_2(\delta)$: $\quad$ *bribed*$_{\pi 2}$, *measures* = *Minimal*, *obtainable-elsewhere*$_{\delta 1}$,
disclosed $= 5$
F_1: $\quad$ *bribed*$_{\pi 2}$, *measures* = *Access-To-Premises-Controlled*,
reverse-eng$_{\delta 2}$, *disclosed* $= 10$

In Horty's [27] dimension-based result model of precedential constraint 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 9. [Preference relation on dimensional fact situations [27].] Let F and F' be two fact situations with the same set of dimensions. Then $F \leqslant_s F'$ iff for all $(d, v) \in F$ and all $(d, v') \in F'$ it holds that $v \leqslant_s v'$.

Definition 3 of (in)consistent case bases now directly applies to case bases with dimensions.

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

Then adapting Definition 2 to dimensions is straightforward.

Definition 10. [Precedential constraint with dimensions [27].] Let CB 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 $F' \leqslant_s F$.

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

We next recall [29]'s adaptation of Definition 4 to dimensions. Unlike with factors, there is no need to indicate whether a value assignment favours a particular side, since we have the $\leqslant_s$ orderings.

Definition 11. [Differences between cases with dimensions [29].] Let $c = (F(c), \mathit{outcome}(c))$ and $f = (F(f), \mathit{outcome}(f))$ be two cases. The set $D(c, f)$ of differences between c and f is defined as follows.

(1) If $\mathit{outcome}(c) = \mathit{outcome}(f) = s$ then $D(c, f) = \{(d, v) \in F(c) \mid v(d, c) \not\leqslant_s v(d, f)\}$.
(2) If $\mathit{outcome}(c) \neq \mathit{outcome}(f)$ where $\mathit{outcome}(c) = s$ then $D(c, f) = \{(d, v) \in F(c) \mid v(d, c) \not\geqslant_{\bar{s}} v(d, f)\}$.

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:

$D(c_1, f) = \{(\mathit{deceived}, 1), (\mathit{reverse\text{-}eng}, 0), (\mathit{measures}, \mathit{Entry\text{-}By\text{-}Visitors\text{-}Restricted})\}$
$D(c_2, f) = \{(\mathit{measures}, \mathit{Minimal}), (\mathit{obtainable\text{-}elsewhere}, 0)\}$

The following counterpart of Proposition 1 is proven in [29].

Proposition 7. Let, given a set D of dimensions, CB be a case base and F a fact situation. Then deciding F for s is forced given CB iff there exists a case in CB with outcome s such that $D(c, f) = \emptyset$, where $f = (F, s)$.

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

Proposition 8. 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 $D(c', f) \not\subseteq D(c, f)$.

5.2. Adding dimensions to the top-level model of explanation

When extending our explanation model with dimensions, we see factors simply as a special case of dimensions with just two values 0 and 1 where $0 <_s 1$ while $1 <_{\bar{s}} 0$.[1] We then adapt Definition 5 of case-based argumentation frameworks for explanation to dimensions as follows. 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

[1] This simplification overcomes a limitation of [11], in which worse values of dimensions cannot be compensated with factors.

has a value in f that is at is at least as favourable for s as in the precedent. Next, instead of the factor-based distinguishing and downplaying moves of Definition 5 we now have one simple distinguishing *Worse* move, which says that the new case is on some dimensions worse than the precedent for the precedent's outcome. Moreover, we now have one simple downplaying *Compensates* move, which says that the factors on which the focus case is not at least as good for its outcome as 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.

Definition 12. [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 a definition sc of compensation, 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 $\mathcal{M} =$ $\{Worse(c,x) \mid x \neq \emptyset$ and $x = \{(d,v) \in F(f) \mid v(d,f) <_{outcome(f)} v(d,c)\}\} \cup$ $\{Compensates(y,x,c) \mid x = \{(d,v) \in F(f) \mid v(d,f) <_{outcome(f)} v(d,c)\}$ and $y \subseteq \{(d,v) \in (f) \mid v(d,c) <_{outcome(f)} v(d,f)\}$ and y compensates x according to $sc\} \cup$ $\{Transformed(c,c') \mid c \in CB$ and c can be transformed into c' and $D(c',f) = \emptyset\}$
- A attacks B iff:
  - $*$ $A, B \in CB$ and $outcome(A) \neq outcome(B)$ and $D(B,f) \not\subset D(A,f)$;
  - $*$ $B \in CB$ 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 $Compensates(y,x,c)$ and y compensates x according to sc; or
  - $*$ $B \in CB$ and $outcome(B) \neq outcome(f)$ and A is of the form $Transformed(c,c')$.

In our running example, a citation of c_1 by the proponent can now be attacked by

$$Worse(c_1, \{(deceived, 0), (measures, \textit{Access-To-Premises-Controlled}),(\textit{reverse-eng},1)\}$$

since $0 <_\pi 1$ for *deceived* and $1 <_\pi 0$ for *reverse-eng* and *Access-To-Premises-Controlled* $<_\pi$ *Entry-By-Visitors-Restricted* for *measures*. This attack can be downplayed by

$$Compensates(S, \{(deceived, 0), (measures, \textit{Access-To-Premises-Controlled}),(\textit{reverse-eng},1)\}, c_1)$$

for any subset of $S = \{(bribed, 1), (disclosed, 10)\}$ since $0 <_\pi 1$ for *bribed* and $20 <_\pi 10$ for *disclosed.*

At first sight, it would seem strange that the dimension-based set of moves is simpler than the factor-based set. However, on closer inspection we can see that the dimension-based downplaying move abstracts from things made explicit in the factor-based downplaying moves, namely, the reasons why the differences do not matter. For example, suppose that in a domain with two factors f_1 and f_2 and one dimension d a precedent c can be distinguished with $Worse(c, \{(f_1, 0), (d, 10)\})$, which can be downplayed with $Compensates(\{(f_2, 1)\}, \{(f_1, 0), (d, 10)\})$. Then a definition in sc could say that $\{(f_2, 1)\}$ compensates $\{(f_1, 0), (d, 10)\}$ since f_2 substitutes f_1. Thus the moves of Definition 5 can still be made available in the case of dimensions, but at a lower level of abstraction.

Definition 7 now directly applies to explanation with dimensions. The operational semantics for the new downplaying move is defined as follows.

Definition 13. [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 5 is as in Definition 8;
- $Compensates(y, x, c) = (F^t(c), s)$, where $(d, v) \in F^t(c)$ iff $(d, v) \in F(c) \setminus x \cup y$ or else $(d, v) \in x \cup y$.

When $Compensates(y, x, c) = c'$, we say that the *Compensates* move *transforms* c into c_n. And we say that a case $c \in CB$ *can be transformed into* c' iff there exists a move in $\mathcal{A}$ that transforms c into c'.

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 rather crude but more refined ways can only be defined if additional information is available (cf. Section 7 below). With this semantics for *Compensates* moves, the results for the explanation model with factores can easily be adapted to the version with dimensions.

Lemma 9. Given an $eAF_{CB,f,sc} = AF$ for explanation with dimensions with explanation-complete sc, it holds for every $c \in CB$ with relevant differences with f but the same outcome as f that there exists a move in AF that transforms c into a case c' such that $D(c', f) = \emptyset$.

Proposition 10. For any abstract argumentation framework for explanation with dimensions $eAF_{CB,f,sc}$ with explanation-complete sc and such that there are citable cases for the proponent, the proponent has a winning strategy in the explanation game for $eAF_{CB,f,sc}$ for any best citable case for the outcome of f.

Corollary 11. A case $c \in CB$ is in the grounded extension of $eAF_{CB,f,sc}$ if and only if the proponent has a winning strategy starting with c in the explanation game for $eAF_{CB,f,sc}$.

Proposition 12. Let T be a winning strategy for the proponent in an explanation dialogue with an initial move c such that $D(c, f) \neq \emptyset$ and the opposite outcome as in $f = (F, s)$ is not forced, and let $M = m_1, \dots, m_n$ be any explanation sequence of all downplaying moves in T. Then the output of m_n is a case (F', s) such that $F' \leqslant_s F$.

6. Evaluation

In this section we report on experiments with three data sets to evaluate our approach. We in particular want to gain some initial insights into the circumstances under which our approach is feasible, and on which aspects it should be further developed or tested. We will evaluate the outcomes in terms of the nature of the generated explanations (size, comprehensibility, need for trivial explanations) and on how inconsistencies are dealt with. We first introduce the three data sets that were used for the experiments. After that, we discuss how the data instances were transformed into cases, making up the case bases. We then report on the experiments and evaluate the method based on the results. Recall that some example explanations for the three data sets are displayed in Appendix A.2.

6.1. About the data sets

For the experiments, we used three publicly available data sets: Churn [31], Mushroom [32] and Graduate Admission [33]. All data sets are tabular, consisting of several features, together with an outcome variable. The Churn data set contains information about customers of a telecom service. The outcome variable *Churn* represents whether a customer continued using a company's telecom services or churned, that is, cancelled the subscription. The Mushroom set consists of descriptions of hypothetical samples corresponding to species of mushrooms in the Agaricus and Lepiota Family. All features of this set are categorical and the outcome variable represents whether the mushroom is either definitely edible or (possibly) poisonous. The Mushroom data set was also used in [8, 9] to test their ANNA method. The Admission set consists of features - such as exam scores - with which a prediction can be made about whether an applicant will be admitted to a master program.

We made a heterogeneous selection of data sets in terms of consistency (recall that according to Definition 3 a case base is inconsistent if a fact situation exists for which two opposite decisions are forced). The Churn data set is of a highly statistical nature; on average, we can tell which profiles are likely to stay or churn based on the features, but there are many exceptions. In the Mushroom data set, on the other hand, the features do seem to possess enough information to consider the outcome variables some 'truth' in that cases with the same features must have the same outcome. The Admission data set forms a middle ground between the statistical Churn and the more consistent Mushroom data set. Further details about the data sets can be found in Table 1.

	Mushroom	Churn	Admission
number of instances	8124	7032	500
distribution outcome	48% poisonous	27% churned	7.8% refused
number of features	22	21	7
number of categorical features	22	18	1
number of continuous features	0	3	6

Table 1

Data set statistics

We did our experiments in two variants: one with the full data sets and one with the full Admission data set plus random selections of 500 cases from the Mushroom and Churn data sets. This allows us to observe effects of size differences on the results.

6.1.1. Transforming the data instances into cases

To prepare the data, we made several modifications. The outcome values of the Graduate Admission data set were transformed into binary values by replacing every value below 0.5 with 0, and other values with 1. We removed one feature from the Mushroom data set - 'veil-type' - for which all instances have the same value assignment.

We used an automatic approach to establish the tendencies of features to promote outcomes, using Pearson correlation coefficients. When a continuous feature positively (negatively) correlates with outcome s, we concluded that $v' \leqslant_s v$ whenever $v' \leqslant v$ ($v \leqslant v'$). This also works for binary categorical features, i.e., factors, since presence, respectively, absence of a factor is modelled with, respectively, numerical values 1 and 0. For non-binary categorical features we measured the correlation of every category with the outcome variable. We then created an ordering based on the correlation values. When v has a more positive correlation with s than v', we concluded that $v' \leqslant_s v$ and $v \leqslant_{s'} v'$.

Every row of data was transformed into a case. The feature values of an instance were represented as a *fact situation*: a list of pairs, (F, v), where F is the feature and v the value for that instance. A case is then a pair $Case(F, o)$, where F is the fact situation, and o the outcome value for that instance. Per data set, we created one case base - a list of cases - consisting of all instances.

6.1.2. Consistency

Before applying our method, we measured the consistency of the three case bases. With the 3 x 500 data sets, the Mushroom case base was nearly consistent (98.60%). This was different for the other case bases; for about 16% of the Churn cases and 20% of the Admission cases, a case that is more favorable for that outcome but received the opposite outcome could be found. This inconsistency seems caused by a small number of 'exceptional' cases. As Table 2 shows, removing respectively 5.6% and 3.2% of the most inconsistent cases results in consistent Churn and Admission case bases. With the full-size data sets (see Table 3), the degree of consistency of the Churn case base was reduced to less than 50%. Removing 11.4% of the most inconsistent cases resulted in a consistent case base. The consistency of the Mushroom case base barely changed when using the full-size data set.

	Mushroom	Churn	Admission
percentage consistent cases	98.60%	84.20%	79.80%
number of removals for consistent CB	1 (0.20%)	28 (5.60%)	16 (3.20%)

Table 2

Consistency statistics (3 x 500)

	Mushroom	Churn	Admission
percentage consistent cases	98.82%	48.38%	79.80%
number of removals for consistent CB	26 (0.32%)	798 (11.35%)	16 (3.20%)

Table 3

Consistency statistics (full size)

In realistic applications data sets will often be inconsistent to some degree, for instance, due to imperfect assignment of features to cases, missed dependencies among features or incomplete information. For that reason, any explanation method has to be able to deal with inconsistency. In Section 6.3 below we will discuss how our method fares in the face of inconsistency.

6.2. Experiments

In this section we present the results of the conducted computer experiments. All instances of the data sets are transformed into cases and included in the case base. For a single experiment, every case in the case base is used once as the focus case, while all remaining cases are used as candidate precedents.

6.2.1. Selection of precedents

In our method, the first step in explaining a focus case is to cite a *best precedent*. In Section 4 we defined a best precedent as a case in the case base that:

(1) received the same outcome as the focus case
(2) has a minimal set of relevant differences with the focus case compared to other cases in the case base.

The following abstract example with factors illustrates this notion.

Example 13. Consider the following case base and a focus case with predicted outcome o, where p_1 and p_2 are pro-o factors and d_1 and d_2 are con-o factors.

$c_1(o)$: p_1, p_2, d_1
$c_2(o)$: p_1, p_2
$c_3(o')$: p_1, p_2, d_2
$f(o)$: p_1, d_1

Then we have $D(c_1, f) = \{p_2\}; D(c_2, f) = \{p_2, d_1\}; D(c_3, f) = \{d_2\}$. We have that $D(c_1, f) \subset D(c_2, f)$ and $D(c_3, f) \not\subset D(c_1, f)$, so c_1 but not c_2 is a best precedent to cite for outcome o, while c_3 is a best precedent to cite for o'.

In this experiment, we checked the average number of best precedents for focus cases. Arguably, the lower this number, the better, since with a high number a single explanation can be said to be somewhat arbitrary. Multiple cases can meet both criteria for best precedents. For the 3 x 500 case bases, Table 4 shows the average and standard deviation of the number of best precedents that can be found.

	Mushroom	Churn	Admission
all cases	26.27 (29.20)	9.16 (9.07)	105.92 (116.38))
non-trivial cases	39.55 (28.53)	14.05 (9.44)	6.22 (5.18)

Table 4

Average and standard deviation of the number of best precedents (3 x 500)

Table 5 shows the same numbers for the full-size case bases. These numbers indicate that the average number of best precedents is rather high for all three databases, especially if they are full-size. This points at the need for future research on meaningful additional selection criteria for best precedents.

	Mushroom	Churn	Admission
all cases	82.36 (123.59)	76.65 (133.96)	105.92 (116.49)
non-trivial cases	9.14 (3.44)	18.86 (12.99)	6.22 (5.25)

Table 5

Average and standard deviation of the number of best precedents (full size)

6.2.2. Trivial winning strategies

We next checked how many focus cases have a trivial winning strategy. When there are no relevant differences between the focus case and a precedent c_1 cited at the start of an explanation, the opponent cannot distinguish the focus case from the precedent. In that scenario, the explanation dialogue either immediately terminates or - if the case base is inconsistent for the focus case - the opponent can cite a counterexample c_2 that also has no relevant differences with the focus case. This citation can then be trivially downplayed with a *Transformed*(c_1, c_1) move licensed by a *Compensates*$(\emptyset, \emptyset, c_1)$ move in $\mathcal{A}$ (cf. Definition 13). In both cases we say that the focus case has a trivial winning strategy. Arguably, the lower the percentage of 'trivial' cases, the better, since otherwise a substantial part of our model (distinguishing and downplaying) often remains unused.

As shown by Table 6, for all three databases a majority of the focus cases has a trivial winning strategy, and this majority increases to over 90% for the full-size case bases. This increase is to be expected, since the larger the case base, the easier it is to find precedents with opposite outcomes for a focus case which both have no relevant differences with it.

We can conclude that the larger or the more inconsistent the data sets, the less frequently the distinguishing and downplaying features of our explanation model can be used. The same may hold the fewer

	Mushroom	Churn	Admission
3 x 500	77.00%	62.20%	92.00%
full size	99.83%	91.52%	92.00%

Table 6

Percentages trivial winning strategies

features a data set has (note that Admission has substantially fewer features than Mushroom and Churn). A possible remedy is that in trivial cases also some cases with the opposite outcome are shown with their relevant differences, even though this is not required by our explanation game. Also, other distance measures can be studied as alternatives to our rather coarse similarity relation between cases.

Finally, until now we have considered explanations of correct predictions, but it is also interesting to see what the explanation system tells us in case of an incorrect prediction. If for each focus case we switch its outcome to its opposite, then it follows from our definitions that the percentage of cases with a trivial winning strategy equals the percentage of cases for which the case base is inconsistent. Then we see that this percentage is for all case bases considerably lower, in most cases even far lower. This is as should be expected from an explanation model.

6.2.3. *Trivial downplaying*

The structure of the explanation game can differ between best precedents that do have relevant differences with the focus case. A simple measure with which best precedents can be compared is whether any empty downplaying moves are needed to defend against attacks on the citation. An empty downplaying move is a way of saying that the differences between the focus case and precedent cannot be downplayed by other features, but still do not matter. This can be seen as the weakest form of attack, so the lower the percentage of cases that require empty downplaying moves, the better.

For instance, in Example 13, after the proponent begins by citing c_1, the opponent can distinguish with $\mathit{MissingPro}(c_1, \{p_2\})$. Since f does not contain new pro-o factors, the proponent has to downplay with a trivial $\mathit{pSubstitutes}(\emptyset, \{p_2\}, c_1)$ move.

To obtain an idea of whether there exist relevant differences between the best precedents that are selected for a focus case, we divided selections of best precedents into four outcome classes: selections of only trivial winning strategies and selections in which none, a part of or all of the precedents need empty downplaying moves. We show the results for the 3 x 500 case bases, as the appearance of non-trivial winning strategies is rare with full-size case bases (see Table 6). In Table 7 the distribution over these four groups is shown.

	Mushroom	Churn	Admission
trivial winning strategy	385	311	460
non-trivial and no precedents need empty downplay	54	174	39
non-trivial and some precedents need empty downplay	60	15	1
non-trivial and all precedents need empty downplay	1	0	0

Table 7

Number of occurrences of each outcome class for the 500 focus cases. Non-trivial winning strategies are divided in three classes: focus cases for which none, some or all of the best precedents need empty downplay (3 x 500)

In only one case of the Churn case base was it necessary to use empty downplaying moves to defend any of the best precedents. In other cases, the system could at least defend part of the best precedents against the distinguishing moves by pointing to compensating features. We can conclude that the number of cases requiring empty downplay is rather low in both cases.

6.2.4. Using actual predictions

In the experiments so far, we considered the output of our method when presented with only correct or only incorrect predictions as input (since the focus cases are selected from the precedents in the case base, for which we know their outcome). By contrast, in an actual application the system would receive the predictions of another classification model as input. It is interesting to see whether the system responds differently to the specific instances that another model predicts incorrectly. In this final experiment, we will compare the percentage of focus cases for which a trivial winning strategy exists given correct and incorrect predictions of another classifier.

We first made a selection of classifiers to test which models perform best on our data sets. We selected three classifiers which are currently very popular in practical applications: DecisionTree, Support Vector Machine and Naive Bayes [34]. We also added two popular meta-algorithms used in combination with a DecisionTree, named RandomForest and AdaBoost. Finally, we added a simple white-box classifier in the form of a Logistic Regression. Decision Trees and coefficients of linear Support Vector Machines can be examined similar to Logistic Regressions and can therefore also be considered white boxes. We used built-in classifiers from the Python *sklearn* library. Details about the models can be found in Appendix A.3.

In order to select the best performing models for each data set, we measured the performance of the models using 5-fold cross validation. This entails that we shuffled the data sets and split them into 5 sets of equal size. Per set, a model was trained on the 80% of remaining data, after which it was tested on the set. Because of the uneven class distributions, we measured the performance of a model using the average weighted F1-score of both labels. The F1-score is calculated as:

$$F1 = 2x\frac{Precision * Recall}{Precision + Recall}$$

The label scores are weighted by support (the number of true instances for each label).

The analysis made clear that there exist multiple models that reach 100% prediction accuracy on the Mushroom data set. We therefore only continued this experiment with the Churn and Admission data sets, doing it for the full-size case bases only. For this experiment, we again used 5-fold cross-validation on all instances of the data sets. On the Churn data set, a Logistic Regression performed best with an F1-score of 0.797. On the Admission data set, the Random Forest was most accurate, reaching a F1-score of 0.964. The performance of all models is specified in Table 8.

	Mushroom	Churn	Admission
Decision Tree	**1.000**	0.724	0.947
Support Vector Machine	0.998	0.719	0.956
Naive Bayes	0.994	0.745	0.931
Random Forest	**1.000**	0.763	**0.964**
Adaboost	**1.000**	0.788	0.960
Logistic Regression	0.999	**0.797**	0.938

Table 8

Mean performance, measured as the weighted F1-score, of the 5-fold cross validation on the six classifiers.

Table 9 shows the percentages of focus cases with a trivial winning strategy for correct and incorrect predictions, using the best scoring models. To obtain the model's predictions, we split the data again into 5 sets of equal size. We then collected the predictions for each set by training it on the 4 other sets and testing it on the selected set. As we can see, on both data sets the percentage trivial winning strategies is substantially higher for correct predictions as desired, except for full-size Churn, when the difference is just over 3%. The high percentage of trivial winning strategies for incorrect outcomes using full-size Churn seems to be caused by a combination of high inconsistency and large case-base size.

	Churn 500	Churn full size	Admission
correct predictions	75.55%	96.88%	95.09%
incorrect predictions	59.34%	93.51%	68.75%

Table 9

Percentage of focus cases with a trivial winning strategy per data set for correct and incorrect predictions of the classifier. A logistic regression was used for the Churn predictions; a random forest for the Admission predictions.

Finally, we compared the performance of the best classifier on consistent and inconsistent cases. We did this for the full-size Churn base, omitting the Mushroom case base since the best performing models reach a 100% accuracy on all cases and omitting the Admission case base because of the small absolute number of inconsistent cases. Table 10 shows the results.

	Churn
consistent	0.891
inconsistent	0.722

Table 10

Performance (weighted F1-score) of the best performing classifier (logistic regression) on consistent and inconsistent cases.

We see that the difference in performance on consistent and inconsistent cases is quite large. We comment on the significance of these results in Section 6.3.

6.3. Discussion

The aim of our experiments was to gain some initial insights in the circumstances under which our approach is feasible, and on which aspects it should be further developed or tested. We now discuss our findings in light of these aims.

6.3.1. The nature of explanations

First of all, we found that in all three data sets the number of best precedents is rather high for all three databases, especially if they are full-size. This makes the selection of a best precedent with which to explain the focus case somewhat arbitrary if there are no additional meaningful selection criteria. We identified this as a first issue for further research. Second, we found for all data sets that for a substantial part of the focus cases a trivial winning strategy existed for the correct outcome, and that this is the more so the larger the data sets are. In such cases the full power of our explanation model is left unused. We discussed a possible remedy but more research is needed at this point. On the other hand, we also found that for cases with nontrivial explanations the percentage of cases that required non-trivial downplaying is rather low, which is encouraging.

We did not do user experiments on comprehensibility of our explanations but our three example explanations shown in Appendix A.2, especially the one for the Mushroom case base with its 22 features, suggest that research is needed to look for possibilities to decrease the number of features used in the method or presented to the user. In [8, 9] the use of an autoencoder neural network for this purpose is studied. As stated by Molnar [35], the interpretability of example-based methods crucially depends on the comprehensibility of a single instance in the data set.

We finally tested the intended use of our explanation model, by letting it generate explanations for predictions of several machine-learning classifiers applied to the same data sets. We found that for correct predictions of the classifiers our explanation model generated substantially more trivial explanations than for incorrect predictions. This is to be expected, since the incorrectly predicted outcome will often not be forced in terms of the theory of precedential constraint.

6.3.2. How to deal with inconsistencies

When transforming the data sets into case bases, we found that both the Churn- and the Admission data set were, to a considerable degree, inconsistent. This means that while using the case base, a non-negligible number of focus cases exist for which our method would find a trivial winning strategy for both outcomes. Is this a problem for our method? To start with, recall that above we observed that many realistic data sets will to some degree be inconsistent, so any explanation method has to be able to deal with inconsistency. In a data set like Churn, with data about customers of a company, a considerable degree of inconsistency is to be expected, since different customers can have widely varying preferences. In a data set like Admission a lower degree of inconsistency is more likely, since decision makers from the same or similar institutions can be expected to have more shared values and preferences. Finally, data sets like Mushroom can be expected to be highly consistent, since they reflect biological ground truths.

As shown by Table 10, the best-performing classifier performs better for consistent than for inconsistent cases. Therefore, informing a user that a focus case is (in)consistent can be useful. In addition, just as we proposed for trivial cases (of which inconsistent cases are a subclass), in case of inconsistency it can be useful to show cases with the opposite outcome that have relevant differences with the focus case, even though this is not required by our explanation game. In addition, it would be good to investigate what further useful information can be given in inconsistent cases.

A related topic for future research arises from the observation that the consistency of the databases partly depends on how we determined the tendency of features in our experiments. We did this by measuring the individual correlations with the outcome variable but this method is less suitable when there are too many dependencies between features. Using our current approach, we could, for example, find that both a large size and a purple colour of a mushroom positively correlate with being edible, while in reality a purple colour only promotes the chance of being edible when the mushroom is large. Recall that in AI & law research on case-based reasoning with factors or dimensions there is a general assumption that factors or dimensions are independent from each other but this assumption may not always be warranted. In any case, future research is needed on how to measure the tendency of features towards outcomes.

7. 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, depending on the nature of the application, 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 [13]. For example, the premises of *pSubstitutes* and *cSubstitutes* claims can be founded on a 'factor hierarchy' as defined for the CATO system [19, 20]. 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, for example, by saying that deciding for a side on the basis of the presence of particular factors promotes or demotes particular legal or societal values. Arguments according to these accounts can provide the premises of arguments 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 [36]. In [37–39] this approach was followed in the context of the *ASPIC*$^+$ framework [40]. To give an idea of how the approach of these papers can be applied in the present context, we now semiformally sketch a few relevant argument schemes. The first is a scheme for cancelling new con factors with new pro factors:

pCancels(y, x, c, f):

$$\begin{array}{l} x \text{ are the con factors in } f \text{ that are not in } c \\ y \text{ are pro factors in } f \text{ that are not in } c \\ y \text{ are preferred over } x \\ \hline\hline y \text{ pCancels } x \text{ in } c \end{array}$$

The conclusion of this scheme can be defined as an undercutter of a similar scheme with conclusion 'f has new con factors x that are not in c'. The pCancels scheme can be combined with a further argument scheme for arguing for preferences between factor sets on the basis of preferences between the sets of values promoted by these factor sets. The conclusion of this scheme in fact expresses that y pCancels x according to *sc* (cf. Definition 5).

Valuepref$(y, x, values_1, values_2)$:

$$\begin{array}{l} \text{deciding a case con since it contains factors } x \text{ promotes } values_1 \\ \text{deciding a case pro since it contains factors } y \text{ promotes } values_2 \\ values_2 \text{ are preferred over } values_1 \\ \hline\hline y \text{ are preferred over } x \end{array}$$

We intend to further elaborate on these ideas in future research.

8. Related research

In this section we discuss related research on using case-based argumentation for explanation. A recent overview of AI & law models of explanation is [41]. An early suggestion to use example-based approaches for explaining machine-learning outcomes was [4]. Recently, the idea of 'twin systems' was proposed to explain artificial neural networks with case-based reasoning [42]. In this work a neural-network 'black box' model is mapped onto a more interpretable case-based reasoning system that works with the same data set as the neural net and that can be used to justify outputs of the neural net. The main focus in [42] is on methods for learning feature weights from neural networks for use in a k-nearest neighbor classifier. Similar methods could be useful for extensions of our current model with factor or dimension weights. Recent related work from AI & law is of Grabmair [43], who, adapting ideas of [44], predicts outcomes of legal cases in terms of factors and value judgements and explains these predictions with argumentation. While inspiring for our proposed research, his method relies on the use of an explicit argumentation model in the formulation of the classification problem, which he then uses for explaining the outcomes of the classifier. So his explanation method is, unlike our approach, not model agnostic. Finally, a recent empirical study suggests that a case-based style of explanation may lead to a lower degree of perceived justice than other styles [45]. We note that these experiments do not test for understandability of an explanation; it may be that the degree of perceived justice for case-based explanations is lower since they are more understandable than explanations of different kinds.

The closest to our approach and a main source of inspiration for us is Čyras *et al.*'s approach in [6, 7]; it therefore deserves a detailed discussion and comparison. We here summarise the account of [7], where we will ignore their 'stages', which for present purposes are irrelevant. Cases in [7] are pairs of feature sets F and binary outcome o or o'. A case base (a set of cases) is assumed to be coherent in that two cases with the same sets of features cannot have opposite outcomes. On the basis of a case base an abstract argumentation framework is defined as follows. Arguments are cases. One case $c = (F, o)$ attacks another $c' = (F', o')$ iff $F' \subset F$ (specificity) and there is no third case $c'' = (F'', o)$ such that $F' \subset F'' \subset F$ (conciseness). Conciseness says that only a least-specific case that is more specific than its target can attack the target. With this definition of attack, the resulting attack graph is acyclic, so in all four classical argumentation semantics an AF has the same unique extension [5]. Explanations are given in terms of 'dispute trees', which essentially are winning strategies in the argument game for grounded semantics. Although Čyras *et al.* do not use game-theoretic terminology, we will below still do so to facilitate the comparison with our approach. Games are about a *focus case*, which is not part of the case base and the outcome of which is to be explained. Games start with a *default case* $(\emptyset, \delta)$, where $\delta \in \{o, o'\}$ is the default outcome. If the outcome of the focus case equals the default outcome, then the grounded game is played about the default case, otherwise it is played about any attacker of the default case. The idea is that the game identifies a dialogical proof of the default case if its outcome coincides with that of the focus case, or of at least one attacker of the default case if the outcome of the default case and the focus case differ.

Note that if F and F' are not subsets of each other, then cases with these factor sets do not attack each other even though they have opposite outcomes. At first sight, this would seem to yield a problem. Consider $\delta = o$ and we have two cases $c_1 = (\{f_1\}, o)$ and $c_2 = (\{f_2\}, o')$, of which c_2 is the case to be explained. The game about the default case (assuming it has outcome o) would then end after c_1 attacks the default case. Čyras *et al.* solve this problem by adding the focus case as a special argument to AF that attacks all arguments which contain features that are not in the focus case. Then the game in this example ends after c_2 is moved to attack c_1. The excess features of a strategy is the set of features of all cases in the strategy attacked by the focus case (here $\{f_1\}$). Čyras *et al.* then define an *explanation* of the focus case as (in our terms) a pair consisting of a winning strategy plus its set of excess features, where the winning strategy is for the default case if it has the same outcome as the focus case and for an attacker of the default case otherwise.

Now because of the chosen attack relations (and since we ignore Čyras *et al.*'s stages) there will, if adding the focus case to CB leaves CB coherent, always be a winning strategy for the outcome of the focus case, while, moreover, the union of the sets of factors of the pro arguments in the proof is a subset of those of the focus case. The focus case itself does not necessarily appear in the proof; it will only if cases exist with the opposite outcome and features that are not in the focus case. The only kind of case where the outcome of the focus case does not have to be in the grounded extension is when there is a case in CB with the same features as the focus case but the opposite outcome.

While this approach is very interesting, it also has some limitations. First of all, the features are binary, while many realistic applications will have multi-valued features. Furthermore, the model does not represent the tendency of features (from now on 'factors') to support a particular outcome. When factors can be pro or con an outcome, then cases with factors not shared by the focus case can be informative, so that the design decision to have cases attacked by the focus case can result in loss of information relevant for explaining an outcome. Suppose, for example, that a focus case has outcome o with pro-o factor p_1 and con-o factor c_1 while a case in the case base also has outcome o and all factors of the focus case plus con-o factor c_2. Then in our approach the outcome in the focus case can be explained by a fortiori reasoning from the second case. However, Čyras *et al.* generate a trivial dispute tree consisting just of the default case (if it has the same outcome as the focus case) or else of the default case followed by the focus case. The problem with this is that the case from the case base that justifies the decision in the focus case does not appear in the explanation.

Another limitation is that since Čyras *et al.* do not model the tendency of features, they are forced to have a broader notion of consistency (which they call 'coherency') for case bases, in which a case base is only inconsistent if it contains two cases with exactly the same features but opposite outcomes. This makes that situations where one case is even better for an outcome than another case but still has the opposite outcome, go unnoticed in their model, while in our approach they make the case base inconsistent. We regard this as an advantage of our approach, since many realistic applications will have case bases that are inconsistent in our sense and, as we argued above in Section 6.3, noting such inconsistencies may give meaningful information to a user.

We finally compare the explanations generated by Čyras *et al.* in our running example to the explanations generated by our method. If the default outcome is π then they generate the following winning strategy:

$$\begin{array}{ll} P_1: & (\emptyset, \pi) \\ O_1: & c_2 = (\{\mathit{bribed}_{\pi 2},\ \mathit{obtainable\text{-}elsewhere}_{\delta 1},\ \mathit{disclosed}_{\delta 3}\}, \delta) \\ P_2: & f = (\{\mathit{bribed}_{\pi 2},\ \mathit{measures}_{\pi 3},\ \mathit{reverse\text{-}eng}_{\delta 2},\ \mathit{disclosed}_{\delta 3}\}, \pi) \end{array}$$

with the excess features $\{\mathit{obtainable\text{-}elsewhere}_{\delta 1}\}$. If the default outcome is δ then they generate the following even simpler winning strategy:

$$P_1: \quad f = (\{\mathit{bribed}_{\pi 2},\ \mathit{measures}_{\pi 3},\ \mathit{reverse\text{-}eng}_{\delta 2},\ \mathit{disclosed}_{\delta 3}\}, \pi)$$

with an empty set of excess features. If we compare this to the explanation given by our method as displayed in Figure 3, we see that in Čyras *et al.*'s approach precedent c_2 does not appear in the explanation while in our approach it does, together with arguments why the similarities between c_2 and f are more important than the differences and why c_1 is preferred over c_2. So in terms of [3] our explanations are more *contrastive* in that they also explain why the opposite decision should not be taken. This arguably illustrates an advantage of our approach over the one of Čyras *et al.*, where it should be noted that this advantage is only available if the direction of features towards an outcome can be identified. It would be interesting to know the percentage of focus cases in applications that must defend itself (similar to the experiments we conducted on focus cases with trivial winning strategies), but Čyras *et al.* do not report on such experiments.

9. Explaining predictions of another classifier: explaining, justifying or comparing?

In the previous section we discussed alternative example- and argumentation-based methods for explaining black-box classifiers. Yet a more general discussion of how to deal with black-box models is in order. As we noted at the end of Section 3, it may happen that the outcomes of a black-box classification model and the argumentation-based model of precedential constraint disagree for a given input in that the output given by the classification model is not forced according to the argumentation model. We proposed that in such cases it may be informative to show the user under which assumptions the outcome of the learned model is forced according to the argumentation model. This is captured in our operational semantics of the moves in an explanation dialogue, which for a focus case with non-forced outcomes shows how a best precedent is transformed through the dialogue into a case that forces the outcome of the focus case. Our explanation model can thus provide grounds to critique a learned model. Yet this also means that our explanation model does not simply explain how the classification model made its classification for the focus case. Instead of explaining how a black-box model comes to a prediction, our method explains how a different, more interpretable model can *justify* the prediction made. As a result, the explanation can fully deviate from the way the prediction model works.

At first sight, this would seem to make one of Rudin's [46] criticisms apply to our method, namely, the criticism that our explanation system must as a separate classifier be worse than the prediction model, since otherwise the explanation system would make the prediction model redundant. Rudin's worry is that given that the explanation system is worse, the system might distract the user from following correct predictions of the black-box. Yet this is not how our explanation system works; as Propositions 4 and 10 show, our explanation method is not designed to generate separate predictions that can be compared to those of the black-box classifier but to generate justified arguments for the black-box predictions, where the dialectical proofs of why these arguments are justified reveal the assumptions under which they are justified.

Because of her worries, Rudin proposes that at least for high-stake decisions, the aim should not be to explain black-box machine learning models but to design interpretable models instead. In fact, in [8, 9] the model of Čyras *et al.* was not applied to explain a separate prediction model but to generate predictions itself, which were then claimed to be explainable since they were in argumentation-based form (see also the approach of [43]). For example, in [8] the model was applied to the Mushroom data set after an autoencoder neural network was first used to trim down the feature space. The argumentation classifier was then shown to perform better than a neural-network and a decision-tree classifier (though as noted above in Section 6.2.4, several 100% accurate classifiers for this data set are available). Although Rudin offers very valuable insights into the pros and cons of the various approaches, in our opinion the only way to know what is the right approach is by developing, applying and testing each of them. And in this paper we have developed and tested one way of modelling the explanation approach.

10. 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. We investigated the model on its formal properties and we evaluated it experimentally with three data sets. The results of our experimental evaluation studies indicate that our model may be feasible in practice but that further development and experimentation, especially with human users and with extensions of our top-level model, are needed to confirm its usefulness and practical applicability as an explanation model. Main challenges here are selecting from a large number of possible explanations, reducing the amount of features in the explanations and adding more meaningful information to them. Also, it remains to be investigated how suitable our approach is for explaining non-linear models, given the independence assumptions on factors and dimensions that we inherit from the AI & law models on which we build.

The core idea of our approach was to see the training data of the machine-learning application as decided cases, to see input to the learned model as a new case and to explain the new case with techniques from case-based argumentation, embedded in the theory of abstract argumentation frameworks. Explanations of the outcome in a new case take the form of winning strategies in the argument game for grounded semantics, either showing that the outcome of the new case is justified in grounded semantics, or revealing assumptions under which it can be made justified. Thus our explanation model allows for mismatches between the machine-learning- and argumentation models due to the possibility that one or both of these models may not be fully correct. In the latter case, our explanation model may be used to critique instead of explain the outcome of the learned model. Our approach is inspired by earlier work of [6, 7] but extends it to multi-valued features and to (boolean or multi-valued) features with a tendency towards a particular outcome.

As suggested by [3], good explanations are selective, contrastive and social. That an explanation is *selective* means that only the most salient points relevant to an outcome are presented. Our method is, to some extent, selective since in explaining outcomes it only presents those differences between a precedent and a focus case that are relevant. However, we noted that further research is needed to improve our method in this respect. That an explanation is *contrastive* means that it not just explains why the given outcome was reached but also why another outcome was not reached. Our explanations are indeed contrastive in this sense, as reflected in our mechanism for emphasising and downplaying a distinction, although more research is needed to avoid too many ‘trivial’ explanations. Finally, that an explanation is *social* means that in the transfer of knowledge, assumptions about the user’s prior knowledge and views influence what constitutes a proper explanation. The dialogical and argumentation-based form of our explanations create the prospects for truly social explanations, since users may be enabled to provide their knowledge and views (values) to the system by using argument schemes of the kind discussed in Section 7. Of course, future research should investigate to what extent it is realistic to allow users to add such relevant information during an explanation dialogue.

While our approach was motivated by legal decision making, it may also apply to other kinds of decision making, such as commercial decisions about loan applications, employee hiring or customer treatment, as long as the outcome is binary and the input conforms to this paper’s factor- or dimension format. Nevertheless, as our experiments with the Churn and Admission data sets showed, the question remains whether our model only applies to a relatively small set of cases (as is often the case in the law; cf. [41]), or whether it can also be made practically feasible with large data sets.

Throughout the paper we gave a number of specific suggestions for further research. We conclude by observing that, on the one hand, our model has several attractive features and received several encouraging experimental outcomes but that, on the other hand, the added value of our explanation method is not self-evident and needs further development and further experimentation with human test subjects. A main aim of our paper has been to lay the foundations for such further development and experiments.

Appendix A. Appendices

A.1. Proofs

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 $D(c',f) \not\subseteq D(c,f)$.

Proof. Suppose c and f have the same outcome and suppose that $\varphi \in D(c,f)$. Assume first $\varphi \in \mathit{pro}(c) \setminus \mathit{pro}(f)$. Then $\varphi \notin \mathit{pro}(f)$, so $\varphi \notin \mathit{pro}(f) \setminus \mathit{con}(c')$, so $\varphi \notin D(c',f)$.

Assume next $\varphi \in \mathit{con}(f) \setminus \mathit{con}(c)$. Then $\varphi \in \mathit{con}(f)$ so $\varphi \notin \mathit{pro}(c') \setminus \mathit{con}(f)$, so $\varphi \notin D(c',f)$.

Suppose now that c and f have different outcomes and suppose that $\varphi \in D(c,f)$. Assume first $\varphi \in \mathit{pro}(f) \setminus \mathit{con}(c)$. Then $\varphi \in \mathit{pro}(f)$, so $\varphi \notin \mathit{pro}(c') \setminus \mathit{pro}(f)$, so $\varphi \notin D(c',f)$.

Assume next $\varphi \in \mathit{pro}(c) \setminus \mathit{con}(f)$. Then $\varphi \notin \mathit{con}(f)$, so $\varphi \notin \mathit{con}(f) \setminus \mathit{con}(c')$, so $\varphi \notin D(c',f)$. □ □

Lemma 3 Given an $eAF_{CB,f,sc} = AF$ with explanation-complete sc, it holds for every $c \in CB$ with relevant differences with f but the same outcome as f that there exists an explanation sequence given AF that transforms c into a case c' such that $D(c',f) = \emptyset$.

Proof. Recall that $D(c,f) = \mathit{pro}(c) \setminus \mathit{pro}(f) \cup \mathit{con}(f) \setminus \mathit{con}(c)$.

(1) Assume that $\mathit{pro}(c) \setminus \mathit{pro}(f) \neq \emptyset$. Then $\mathcal{A}$ contains a move $\mathit{MissingPro}(c,x)$ such that $x \neq \emptyset$ and $x = \mathit{pro}(c) \setminus \mathit{pro}(f)$. But then $\mathcal{A}$ also contains at least one of the following three moves.

(i) a move $\mathit{pSubstitutes}(y,x,c)$ such that $y \subseteq \mathit{pro}(f) \setminus \mathit{pro}(c)$. In this case, all pro-$s$ factors in c that are missing in f are replaced by zero or more new pro-s factors in f, so that $c' = \mathit{pSubstitutes}(y,x,c) = (X,Y,s)$ such that $X \subseteq \mathit{pro}(f)$, so $\mathit{pro}(c') \setminus \mathit{pro}(f) = \emptyset$.

(ii) a move $\mathit{cCancels}(y,x,c)$ such that $y \subseteq \mathit{con}(c) \setminus \mathit{con}(f)$. In this case all pro-$s$ and all pro-s factors in c that are missing in f are deleted from c so that $c' = \mathit{cCancels}(y,x,c) = (X,Y,s)$ such that $X \subseteq \mathit{pro}(f)$, so $\mathit{pro}(c') \setminus \mathit{pro}(f) = \emptyset$.

(iii) a move $\mathit{pSubstitutes}(y,x,c)\&\mathit{cCancels}(y',x',c)$ such that $x \cup x' = D(c,f)$ and $y \subseteq \mathit{pro}(f)$ and $y' \subseteq \mathit{con}(c) \setminus \mathit{con}(f)$. This move combines the effects of the two previous moves, resulting in a case c' such that $\mathit{pro}(c') \setminus \mathit{pro}(f) = \emptyset$.

(2) We can now assume that there exist explanation sequences that transform c in a version of c such that $\mathit{pro}(c) \setminus \mathit{pro}(f) = \emptyset$. Assume next that $\mathit{con}(f) \setminus \mathit{con}(c) \neq \emptyset$. Then $\mathcal{A}$ contains a move $\mathit{NewCon}(c,x)$ such that $x \neq \emptyset$ and $x = \mathit{con}(f) \setminus \mathit{con}(c)$. But $\mathcal{A}$ also contains at least one of the following three moves.

(i) $\mathit{cSubstitutes}(y,x,c)$ such that $y \subseteq \mathit{con}(c) \setminus \mathit{con}(f)$. In this case, all con-$s$ factors in f that are not in c replace zero or more old con-s factors in c, so that $c' = \mathit{cSubstitutes}(y,x,c) = (\mathit{pro}(c'),Y,s)$ such that $\mathit{con}(f) \subseteq Y$, so $\mathit{pro}(c') \setminus \mathit{pro}(f) = \emptyset$ and $\mathit{con}(f) \setminus \mathit{con}(c) = \emptyset$.

(ii) a move $\mathit{pCancels}(y,x,c)$ such that $y \subseteq \mathit{pro}(f) \setminus \mathit{pro}(c)$. In this case, $\mathit{pro}(c)$ is enlarged with all new pro-s factors from f while $\mathit{con}(c)$ is enlarged with all new con-s factors from f. So again $\mathit{pro}(c') \setminus \mathit{pro}(f) = \emptyset$ and $\mathit{con}(f) \setminus \mathit{con}(c) = \emptyset$.

(iii) a move $\mathit{cSubstitutes}(y,x,c)\&\mathit{pCancels}(y',x',c)$ such that $x \cup x' = D(c,f)$ and $y \subseteq \mathit{con}(c) \setminus \mathit{con}(f)$ and $y' \subseteq \mathit{pro}(f) \setminus \mathit{pro}(c)$. This move combines the effects of the two previous moves, resulting in a case c' such that $\mathit{pro}(c') \setminus \mathit{pro}(f) = \emptyset$ and $\mathit{con}(f) \setminus \mathit{con}(c) = \emptyset$.

In all these cases c is with one or two moves from $\mathcal{A}$ transformed into a case c' such that $D(c',f) = \emptyset$. □

Proposition 4 For any abstract argumentation framework for explanation with factors $eAF_{CB,f,sc}$ with explanation-complete sc and such that there are citable cases for the proponent, the proponent has a winning strategy in the explanation game for $eAF_{CB,f,sc}$ for any best citable case for the outcome of f.

Proof. Let the focus case have outcome s. Then three situations must be distinguished.

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

(2) s is forced and $\overline{s}$ is forced. Then O's only reply to P_1 is a move from CB, after which P can win with a $\mathit{Transformed}(c, c')$ move by Lemma 3 since sc is explanation complete.

(3) s is not forced. Then O can reply to P_1 with a *MissingPro* move or a *NewCon* move or a move from CB. Since sc is explanation complete, P can reply to a *MissingPro* move with either $\mathit{pSubstitutes}(y, x, c)$ or $\mathit{cCancels}(y, x, c)$ or $\mathit{pSubstitutes}(y, x, c)\&\mathit{cCancels}(y, x, c)$, which moves cannot be attacked so O has no reply. Moreover, P can reply to a *NewCon* move with either $\mathit{cSubstitutes}(y, x, c)$ or $\mathit{pCancels}(y, x, c)$ or $\mathit{cSubstitutes}(y, x, c)\&\mathit{pCancels}(y, x, c)$, which moves cannot be attacked so O has no reply. Finally, P can reply to a move from CB with $\mathit{Transformed}(c, c')$ as in case (2). □

Corollary 5 A case $c \in CB$ is in the grounded extension of $eAF_{CB,f,sc}$ if and only if the proponent has a winning strategy starting with c in the explanation game for $eAF_{CB,f,sc}$.

Proof. This follows from Lemma 3, Proposition 4 and soundness and completeness of the original argument for grounded semantics (cf. Section 2.2), since Lemma 3 implies that the extra condition (4) on the proponent's moves that definition 7 adds to the game for grounded semantics does not change the set of winning strategies for the proponent. □

Proposition 6 Let T be a winning strategy for P in an explanation dialogue with an initial move c such that $D(c, f) \neq \emptyset$ and the opposite outcome as in f is not forced, and let $M = m_1, \dots, m_n$ be any explanation sequence of all downplaying moves in T. Then the output of m_n is a case (X, Y, s) such that $X \cup Y \leqslant_s \mathit{pro}(f) \cup \mathit{con}(f)$.

Proof. Four cases must be considered:

T contains a *MissingPro* reply but no *NewCon* reply. Then T contains either a $\mathit{pSubstitutes}(y, x, c)$ or $\mathit{cCancels}(y, x, c)$ reply to this move or a combination of these replies. Then we have a special case of case (1) of the proof of Lemma 3 in which $\mathit{con}(f) \setminus \mathit{con}(c) = \emptyset$, so the result follows.

T contains a *NewCon* reply but no *MissingPro* reply. Then T contains either a $\mathit{cSubstitutes}(y, x, c)$ or $\mathit{pCancels}(y, x, c)$ reply to this move or a combination of these replies. Then we have case (2) of the proof of Lemma 3, so the result follows.

T contains both a *MissingPro* reply and *NewCon* reply. Consider the sequence from T with first a reply to the *MissingPro* move and then a reply to the *NewCon* move. The first reply brings us in case (1), so transforms c into a case c' such that $\mathit{pro}(c') \setminus \mathit{pro}(f) = \emptyset$. Then the second reply brings us in case (2), which transforms c' into a case c'' such that both $\mathit{pro}(c'') \setminus \mathit{pro}(f) = \emptyset$ and $\mathit{con}(f) \setminus \mathit{con}(c'') = \emptyset$. □

Proposition 8 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 $v(d,c) \not\leqslant_s v(d,f)$, so $v(d,c) \not\geqslant_{\overline{s}} v(d,f)$, 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 $v(d,c) \not\geqslant_{\overline{s}} v(d,f)$, so $v(d,c) \not\leqslant_s v(d,f)$, so $(d,v) \notin D(c',f)$. □

Lemma 9 Given an $eAF_{CB,f,sc} = AF$ for explanation with dimensions with explanation-complete sc, it holds for every $c \in CB$ with relevant differences with f but the same outcome as f that there exists a move in AF that transforms c into a case c' such that $D(c',f) = \emptyset$.

Proof. Recall that $D(c,f) = \{(d,v) \in F(c) \mid v(d,c) \not\leqslant_s v(d,f)\}$ and assume that $D(c,f) \neq \emptyset$. Then $\mathcal{A}$ contains a move $Compensates(y, D(c,f), c)$ in $\mathcal{A}$ such that $= \{v(d,f) \mid v(d,c) \in D(c,f)\}$, which thus transforms c into a case c' such that clearly $D(c',f) \neq \emptyset$. □

Proposition 10 For any abstract argumentation framework for explanation with dimensions $eAF_{CB,f,sc}$ with explanation-complete sc and such that there are citable cases for the proponent, the proponent has a winning strategy in the explanation game for $eAF_{CB,f,sc}$ for any best citable case for the outcome of f.

Proof. Let the focus case have outcome s. Then three situations must be distinguished.

(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) s is forced and $\overline{s}$ is forced. Then O's only reply to P_1 is a move from CB, after which P can win with a $Transformed(c,c')$ move by Lemma 3 since sc is explanation complete.

(3) s is not forced. Then O can reply to P_1 with a *Worse* move or a move from CB. Since sc is explanation complete, P can reply to a *Worse* move with $Compensates(y,x,c)$, which moves cannot be attacked so O has no reply. Moreover, P can reply to a move from CB with $Transformed(c,c')$ as in case (2). □

Corollary 11 A case $c \in CB$ is in the grounded extension of $eAF_{CB,f,sc}$ if and only if the proponent has a winning strategy starting with c in the explanation game for $eAF_{CB,f,sc}$.

Proof. This follows from Lemma 9, Proposition 10 and soundness and completeness of the original argument for grounded semantics, since Lemma 9 implies that the extra condition (4) on the proponent's moves that definition 7 adds to the game for grounded semantics does not change the set of winning strategies for the proponent. □

Proposition 12 Let T be a winning strategy for P in an explanation dialogue with an initial move c such that $D(c,f) \neq \emptyset$ and the opposite outcome as in $f = (F,s)$ is not forced, and let $M = m_1, \dots, m_n$ be any explanation sequence of all downplaying moves in T. Then the output of m_n is a case (F',s) such that $F' \leqslant_s F$.

Proof. Since a winning strategy in this case contains precisely one *Worse* move, the proposition immediately follows from Definition 13. □

Dimension	Focus case	Precedent
Cap-shape	Knobbed	Convex
Cap-surface	Smooth	Scaly
Cap-color	Red	Red
Bruises	No	No
Odor	Fishy	Fishy
Gill-attachment	Free	Free
Gill-spacing	Close	Close
Gill-size	Narrow	Narrow
Gill-color	Buff	Buff
Stalk-shape	Enlarging	Enlarging
Stalk-root	Missing	Missing
Stalk-surface-above-ring	Silky	Silky
Stalk-surface-below-ring	Smooth	Smooth
Stalk-color-above-ring	Pink	White
Stalk-color-below-ring	White	White
Veil-color	White	White
Ring-number	One	One
Ring-type	Evanescent	Evanescent
Spore-print-color	White	White
Population	Several	Several
Habitat	Leaves	Woods
Outcome	?	Poisonous

Table 11

Focus case and precedent for an example explanation of a trivial case in the mushroom data set

A.2. Example explanations

To illustrate the experiments of Section 6, we give a few example explanations for each of the data sets. We do not list them in the formal format of our model but in some possible more user-friendly ways in which they could be given in actual applications. In Example 1 a single outcome is forced, in Example 2 both outcomes are allowed but not forced, and in Example 3 both outcomes are forced since the case is inconsistent.

Example 1) Forced - Mushroom

- Prediction classifier: poisonous
- Explanation: Outcome *poisonous* is forced. The mushroom has no relevant differences with a poisonous mushroom, while it has relevant differences with all edible mushrooms in the database; see Table 11.

Example 2) Non-trivial - Churn

- Prediction classifier: stay
- Explanation: Outcome *stay* is forced if the differences that make the focus customer (F) less likely to stay than the precedent (P) can be compensated by the other differences. The comparison of F and P is presented in Table 12.

Differences that make F less likely to stay than P (Worse)
Customer F has been a member for 3 months less than P
Customer F total costs are $21 less than those of P
Differences that make F more likely to stay than P (Better)
Customer F is a male, while P is a female
Customer F shares the membership with a partner, while P does not
Customer F is charged $7,- per month less
Customer F pays with an automatic bank transfer, while P uses a mailed check

Table 12

The comparison between the focus case (F) and precedent (P) for a non-trivial case in the Churn data set

Example 3) Inconsistency - Admission

- Prediction classifier: admitted
- Explanation: The system reaches an inconsistent outcome for this student. Outcome *admitted* and *declined* are both forced. The relevant cases are displayed in Table 13.

	Focus case	Precedent	Counterexample
GRE Score	0.18	0.16	0.34
TOEFL Score	0.29	0.00	0.32
University Rating	0.25	0.00	0.50
SOP Score	0.25	0.25	0.75
LOR Score	0.25	0.25	0.5
CGPA Score	0.35	0.35	0.45
Research	No experience	No experience	No experience
Outcome	?	Admitted	Declined

Table 13

Focus case, precedent and counterexample for an example explanation of an inconsistent case in the admission data set

A.3. Sklearn Models

The following models from the Python *sklearn* library were used:

(1) DecisionTreeClassifier()
(2) SVC(kernel=‘linear?)
(3) GaussianNB()
(4) LogisticRegression(solver = ‘lbfgs?)
(5) AdaBoostClassifier()
(6) RandomForestClassifier()

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] 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) (2019), 93:1–93:42.

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

[4] C. Nugent and P. Cunningham, A case-based explanation system for black-box systems, *Artificial Intelligence Review* **24** (2005), 163–178.

[5] P.M. 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.

[6] K. Čyras, K. Satoh and F. Toni, Explanation for case-based reasoning via abstract argumentation, in: *Computational Models of Argument. Proceedings of COMMA 2016*, P. Baroni, T.F. Gordon, T. Scheffler and M. Stede, eds, IOS Press, Amsterdam etc, 2016, pp. 243–254.

[7] 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** (2019), 141–156.

[8] 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*, 2018, pp. 26–32.

[9] O. Cocarascu, A. Stylianou, K. Čyras and F. Toni, Data-empowered argumentation for dialectically explainable predictions, in: *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI 2020)*, 2020, pp. 2449–2456.

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

[11] H. Prakken, A top-level model of case-based argumentation for explanation, in: *Proceedings of the ECAI 2020 Workshop on Dialogue, Explanation and Argumentation for Human-Agent Interaction (DEXA HAI 2020)*, 2020.

[12] R. Ratsma, Unboxing the Black Box Using Case-Based Argumentation, Master's thesis, Artificial Intelligence Programme, Utrecht University, Utrecht, 2020.

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

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

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

[16] 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*, ACM Press, New York, 1987, pp. 60–66.

[17] 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*, ACM Press, New York, 1989, pp. 39–102.

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

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

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

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

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

[23] 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*, ACM Press, New York, 1993, pp. 50–59.

[24] T.J.M. Bench-Capon and G. Sartor, A model of legal reasoning with cases incorporating theories and values, *Artificial Intelligence* **150** (2003), 97–143.

[25] 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*, A.Z. Wyner and G. Casini, eds, IOS Press, Amsterdam etc., 2017, pp. 27–32.

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

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

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

[29] H. Prakken, A formal analysis of some factor- and precedent-based accounts of precedential constraint, *Artificial Intelligence and Law* (2021), Doi: 10.1007/s10506-021-09284-6.

[30] M.T. Ribeiro, S. Singh and C. Guestrin, Why should I trust you?: Explaining the predictions of any classifier, in: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[31] Telco Customer Churn, 2018, https://www.kaggle.com/blastchar/telco-customer-churn, version 1.

[32] D. Dua and C. Graff, UCI Machine Learning Repository, 2019. http://archive.ics.uci.edu/ml.

[33] M.S. Acharya, A. Armaan and A.S. Antony, A comparison of regression models for prediction of graduate admissions, *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)* (2019). doi:10.1109/iccids.2019.8862140.

[34] K. Das and R.N. Behera, A survey on machine learning: concept, algorithms and applications, *International Journal of Innovative Research in Computer and Communication Engineering* **5**(2) (2017), 1301–1309.

[35] C. Molnar, *Interpretable Machine Learning*, 2019, https://christophm.github.io/interpretable-ml-book/.

[36] A.J. Hunter (ed.), *Argument and Computation*, Vol. 5, 2014, Special issue with Tutorials on Structured Argumentation.

[37] 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** (2015), 1141–1166.

[38] 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*, ACM Press, New York, 2013, pp. 13–22.

[39] 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*, K.D. Ashley, ed., IOS Press, Amsterdam etc., 2013, pp. 39–48.

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

[41] K. Atkinson, T.J.M. Bench-Capon and D. Bollegala, Explanation in AI and Law: Past, present and future, *Artificial Intelligence* **289** (2020), article number 103387.

[42] E.M. Kenny and M.T. Keane, Twin-systems to explain artificial neural networks using case-based reasoning: comparative tests of feature-weighting methods in ANN-CBR twins for XAI, in: *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI-19)*, 2019, pp. 2708–2715.

[43] M. Grabmair, Predicting trade secret case outcomes using argument schemes and learned quantitative value effect tradeoffs, in: *Proceedings of the 16th International Conference on Artificial Intelligence and Law*, ACM Press, New York, 2017, pp. 89–98.

[44] S. Brueninghaus and K.D. Ashley, Generating legal arguments and predictions from case texts, in: *Proceedings of the Tenth International Conference on Artificial Intelligence and Law*, ACM Press, New York, 2005, pp. 65–74.

[45] R. Binns, M. Van Kleek, M. Veale, U. Lyngs, J. Zhao and N. Shadbolt, ‘It’s reducing a human being to a percentage’; perceptions of justice in algorithmic systems, in: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI 2018)*, ACM Press, New York, 2018, pp. 377:1–377:14.

[46] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, *Nature Machine Intelligence* **1** (2019), 206–215.