Modelling and Explaining Legal Case-Based Reasoners Through Classifiers

Xinghan Liu, Emiliano Lorini, Antonino Rotolo, Giovanni Sartor

Legal Knowledge and Information Systems
E. Francesconi et al. (Eds.)

doi:10.3233/FAIA220451

Modelling and Explaining Legal Case-Based Reasoners Through Classifiers

Xinghan LIU [a], Emiliano LORINI [a], Antonino ROTOLO [b] and Giovanni SARTOR [b]
[a] *IRIT-CNRS, University of Toulouse, France*
[b] *Alma Human AI, University of Bologna, Italy*

Abstract. This paper brings together factor-based models of case-based reasoning (CBR) and the logical specification of classifiers. Horty [8] has developed the factor-based models of precedent into a theory of precedential constraint. In this paper we combine binary-input classifier logic (BCL) to classifiers and their explanations given by Liu & Lorini [13,14] with Horty's account of factor-based CBR, since both a classifier and CBR map sets of features to decisions or classifications. We reformulate case bases in the language of BCL, and give several representation results. Furthermore, we show how notions of CBR can be analyzed by notions of classifier explanation.

Keywords. Case-based reasoning, Modal Logic for classifiers, Explainable AI

1. Introduction

This paper brings together two lines of research: factor-based models of case-based reasoning (CBR) and the logical specification of classifiers.

Logical approaches to classifiers capture the connection between features and outcomes in classifier systems. They are well-suited for modeling and computing a large variety of explanations of a classifier's decisions [18,5,12,11,4,13], e.g., prime implicants, abductive, contrastive and counterfactual explanations. Consequently, they enable detecting biases and discrimination in the classification process. They can thus contribute to provide controllability and explainability over automated decision-making (as required, e.g., by Art. 22 GDPR and by Art. 6 ECHR relative to judicial decisions).

Factor-based reasoning [2,1] is a popular approach to precedential reasoning in AI&law research. The key idea is that a case can be represented as a set of factors, where a factor is a legally relevant aspect. Factors are assumed to have a direction, i.e., to favor certain outcomes. Usually both factors and outcomes are assumed to be binary, so that each factor can be labelled with the outcome it favors (usually denoted as π, the outcome requested by the plaintiff, and δ, the outcome requested by the defendant). The party which is interested in a certain outcome in a new case can support her request by citing a past case that has the same outcome, and shares with the new case some factors supporting that outcome. The party that is interested in countering that outcome can respond with a distinction, i.e., can argue that some factors which supported that outcome in the precedent are missing in the new case or that some additional factors against that outcome are present in the new case. Horty [7,9] has developed the factor-based models

of precedent into a theory of precedential constraints, i.e., of how a new case must be decided, in order to preserve consistency in the case law. In [8,6], he takes into account the fact that judges may also provide explicit reasons for their choice of a certain outcome. This leads to the distinction between the result and the reason model of precedents. In the first model, the message conveyed by the case is only that all factors supporting the case-outcome (pro-factors) outweigh all factors against that outcome (con-factors). In the second, the message is that the factors for the case outcome indicated by the judge outweigh all factors against that outcome.

In this paper we shall combine Liu & Lorini's modal logic approach to classifiers and their explanations [13,14] with Horty's account of factor-based CBR. The combination is based on the fact that both a classifier and CBR map sets of features to decisions or classifications. In this way, our contribution is at least twofold.

First, we explore the relation between two apparently unrelated reasoning systems. While the connection between CBR and reasoning about classifier systems is of interest in itself, we believe that, through this relation, new research perspectives can be offered, since we could in the future investigate CBR by exploiting several techniques and results from modal logic. We will see that the challenge of this paper is to adapt the formal representation of a classifier to the bidirectionality of factors in the HYPO model. Once this is solved, we can provide a logical model and a semantics for factor-based CBR.

Second, we investigate the idea of normative explanation: While the literature on the concept of explanation is immense, the AI community is now paying attention to it due to the development of explainable AI (XAI) [15,3]. Our paper, by connecting CBR and reasoning about classifier systems, explores different notions of explanation in law, such as abductive and contrastive explanations for the outcome suggested by the case-based reasoner. Our model allows for building explainable case-based reasoners, which could also be deployed to reproduce and analyze the functioning of opaque predictors of the outcome of cases. We import notions such as prime implicant and contrastive explanation in the domain of XAI for classifiers to showcase how to analyze CBR in the field of XAI.

The the paper is organised as follows. Section 2 presents Horty's models of CBR. Section 3 introduces the notion of classifier model (CM) for the binary-input classifier logic BCL. Section 4 studies the connection between CBR and classifier models.Section 5 shows that notions for classifier explanation in XAI help study case base. Finally, Section 6 discusses related work and concludes. Proofs and the axiomatics are in the appendix.[1]

2. Horty's Two Models of Case-Based Reasoning

In this section we account for the two models of case-based reasoning / precedential constraint proposed by Horty. We simply say *result model* for "the factor-based result model of precedential constraint" and *reason model* for "the factor-based reason model of precedential constraint".

Let $Atm_0 = Plt \cup Dfd$, where Plt and Dfd are disjoint sets of factors favoring the plaintiff and defendant respectively. In addition, let $Val = \{1, 0, ?\}$ where elements stand for *plaintiff wins*, *defendant wins* and *indeterminacy* respectively. Let $Dec = \{\mathtt{t}(x) : x \in$

[1] The paper with appendix is available here: https://arxiv.org/abs/2210.11217.

Val} and read $\mathtt{t}(x)$ as "the actual decision/outcome (of the judge/classifier) takes value x". An outcome $\mathtt{t}(1)$ or $\mathtt{t}(0)$ means that, the judge is predicted to decide for the plaintiff or for the defendant (the classifies "forces" one of the two outcomes). The outcome $\mathtt{t}(?)$ means either outcome would be consistent: the judge may develop the law in one direction or the other. This reflects the incompleteness nature of CBR. We use *Atm* to denote $Atm_0 \cup Dec$.

We call $s \subseteq Atm_0$ a *fact situation*. A set of atoms X is called a *reason* for an outcome (decision) x if it a set of factors all favoring the same outcome: $X \subseteq Plt$ is a reason for 1 and $X \subseteq Dfd$ is a reason for 0. A (defeasible) *rule* consist of a reason and the corresponding outcome: $X \mapsto x$ is rule, if $X \subseteq Plt$ and $x = 1$, or $X \subseteq Dfd$ and $x = 0$. For readability, we make a convention that, for $x \in \{0,1\}$, let $\overline{x} = 1 - x$ and $\overline{\overline{x}} = x$. Moreover, let $Atm_0^x = Plt$ if $x = 1$, and $Atm_0^x = Dfd$ if $x = 0$.

In the reason model, a *precedent case* (precedent) is a triple $c = (s, X, x)$, where $s \subseteq Atm_0$, $X \subseteq Atm_0^x, x \in \{0,1\}$. In plain words, $s \cap Atm_0^x$ contains all *pro-factors* in s for x, while $s \cap Atm_0^{\overline{x}}$ all *con-factors* in s for x. X is the *reason of the case*, namely a subset of the pro-factors which the judge considers sufficient to support that outcome, relative to all con-factors in the case.

A *case base CB* (for reason model) is a set of precedential cases. When the reason contains all pro-factors within the situation (i.e., when $c = (s, s \cap Atm_0^x, x)$) all such factors are considered equally decisive. If a case base only contains cases of this type, we obtain what Horty calls "the result model", and note such a case base CB^{res}.[2] The class of all *CBs* and $CB^{\text{res}}s$ are noted **CB** and $\mathbf{CB}^{\text{res}}$ respectively.

Example 1 (Running example)**.** *In the paper we refer to the following running example taken from [16]. Let us assume the following six factors, each of which either favors the outcome 'misuse of trade secrets' ('the plaintiff wins') or rather favors the outcome no misuse of trade secrets ('the defendant wins'): 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)*. Hence in our running example* $Atm = \{\pi_1, \pi_2, \pi_3, \delta_1, \delta_2, \delta_3, \mathtt{t}(0), \mathtt{t}(1), \mathtt{t}(?)\}$ *Let us consider a case base* $CB^{ex} = \{c_1, c_2\}$ *where* $c_1 = (\{\pi_1, \pi_3, \delta_1, \delta_3\}, \{\pi_1\}, 1)$; $c_2 = (\{\pi_2, \delta_1, \delta_3\}, \{\delta_3\}, 0)$*, which means:*

- c_1 *has factors (fact situation)* $s_1 = \{\pi_1, \pi_3, \delta_1, \delta_3\}$*, reason* $\{\pi_1\}$ *and outcome* 1*;*
- c_2 *has outcome* δ*, factors* $s_2 = \{\pi_2, \delta_1, \delta_3\}$*, reason* $\{\delta_3\}$ *and outcome* 0

A case base can be inconsistent when two precedents map the same fact situation to different outcomes. Another scenario is that a consistent case base becomes inconsistent after *update*, namely after expanding it with some new case. Hence maintaining consistency is the crucial concern of case-based reasoning. But first of all, one need to define these notions. The following definitions, except symbolic difference, are based on [8,16].

Definition 1 (Preference relation derived from a case)**.** *Let* $c = (s, X, x)$ *be a case. Then the preference relation* $<_c$ *derived from c is s.t. for any two reasons* Y, Y' *favoring* x *and* $\overline{x}$ *respectively,* $Y' <_c Y$ *iff* $Y' \subseteq s \cap Atm_0^{\overline{x}}$ *and* $X \subseteq Y$.

[2] So we view result model as a special kind of reason model, as [8, p. 25] also mentioned.

Definition 2 (Preference relation derived from a case base)**.** *Let CB be a case base. Then the preference relation $<_{CB}$ derived from CB is s.t. for any two reasons Y, Y' favoring x and $\bar{x}$ respectively, $Y' <_{CB} Y$ iff $\exists c \in CB$ s.t. $Y' <_c Y$.*

Definition 3 ((In)consistency)**.** *A case base CB is inconsistent, if there are two reasons Y, Y' s.t. $Y' <_{CB} Y$ and $Y <_{CB} Y'$. CB is consistent if it is not inconsistent.*

Definition 4 (Precedential constraint)**.** *Let CB be a consistent case base, X is a reason for x in CB and applicable in a new fact situation s', i.e. $X \subseteq s'$. Updating CB with the new case (s', X, x) meets the precedential constraint, iff $CB \cup \{(s', X, x)\}$ is still consistent.*

There is more than one way to satisfy the precedential constraint, depending on how the precedents in CB interacts with the new case. The requirement of consistency dictates the outcome when the "a fortiori" constraint applies: if reason X for x outweighs (i.e., is stronger that) reason $s \cap Atm_0^{\bar{x}}$, a fortiori any superset of X outweighs any subset of $s \cap Atm_0^{\bar{x}}$, so that only by deciding for x rather than for $\bar{x}$ consistency is maintained.[3]

Example 2 (Running example)**.** *Let us consider two fact situations according to case base CB^{ex} running example.*

- *In $s_3 = \{\pi_1, \pi_3, \delta_1\}$, only a decision for 1 in s_3 is consistent with CB^{ex}, since a decision for 0 would entail that $\{\delta_1\} >_{CB^{ex}} \{\pi_1\}$, contrary to the preference $\{\pi_1\} >_{CB^{ex}} \{\delta_1\}$, which is derivable from c_1.*
- *In $s_4 = \{\pi_2, \delta_2\}$ both $(s_4, \{\pi_2\}, 1)$ and $(s_4, \{\delta_2\}, 0)$ are consistent with CB^{ex}, since neither $\{\pi_2\} >_{CB^{ex}} \{\delta_2\}$ nor $\{\delta_2\} >_{CB^{ex}} \{\pi_2\}$.*

3. Classifier Model of Binary-input Classifier Logic

In this section we introduce the language and semantics of binary-input classifier logic BCL first appeared in [13]. Recall that $Atm = Atm_0 \cup Dec$, where $Atm_0 = Dfd \cup Plt$, and $Dec = \{\mathsf{t}(x) : x \in Val = \{0, 1, ?\}\}$. The modal language $\mathscr{L}(Atm)$ of BCL is defined as:

$$\varphi \quad ::= \quad p \mid \mathsf{t}(x) \mid \neg\varphi \mid \varphi \wedge \varphi \mid [X]\varphi,$$

where p ranges over Atm_0, $\mathsf{t}(x)$ ranges over Dec, and X is a finite subset of Atm_0.[4] Operator $\langle X \rangle$ is the dual of $[X]$ and is defined as usual: $\langle X \rangle \varphi =_{def} \neg[X]\neg\varphi$. Finally, for any $X \subseteq Y \subseteq Atm_0$, the following definition syntactically expresses a valuation on Y s.t. all variables in X are assigned as true, while all the rest in Y are false.

$$\mathsf{cn}_{X,Y} =_{def} \bigwedge_{p \in X} p \wedge \bigwedge_{p \in Y \setminus X} \neg p.$$

The language $\mathscr{L}(Atm)$ is interpreted relative to classifier models defined as follows.

[3] We generalize a fortiori constraint from only acting on result models in [8] to also on reason models in the same manner as viewing a result models as a special reason model, whose reason contains all pro-factors.

[4] Atm is finite since the factors in case-based reasoning are supposed to be finite. Notice p ranging over $Dfd \cup Plt$, i.e. p can be some δ or some π. X can denote a reason (an exclusive set of plaintiff/defendant factors), or any subset of Atm_0, which is clear from the context. Last but not least, p and $\mathsf{t}(x)$ have different statuses regarding negation: $\neg p$ means that the input variable p takes value 0, but $\neg\mathsf{t}(x)$ merely means the output does not take value x: we do not know which value it takes, since the output is trinary.

Definition 5 (Classifier model)**.** *A classifier model (CM) is a pair $C = (S, f)$ where:*

- $S \subseteq 2^{Atm_0}$ *is a set of states (or fact situations), and*
- $f : S \longrightarrow Val$ *is a decision (or classification) function.*

The class of classifier models is noted **CM**.

A pointed classifier model is a pair (C, s) with $C = (S, f)$ a classifier model and $s \in S$. Formulas in $\mathscr{L}(Atm)$ are interpreted relative to a pointed classifier model, as follows.

Definition 6 (Satisfaction relation)**.** *Let (C, s) be a pointed classifier model with $C = (S, f)$ and $s \in S$. Then:*

$$
\begin{aligned}
(C,s) &\models p \iff p \in s,\\
(C,s) &\models \mathtt{t}(x) \iff f(s) = x,\\
(C,s) &\models \neg\varphi \iff (C,s) \not\models \varphi,\\
(C,s) &\models \varphi \wedge \psi \iff (C,s) \models \varphi \textit{ and } (C,s) \models \psi,\\
(C,s) &\models [X]\varphi \iff \forall s' \in S : \textit{ if } (s \cap X) = (s' \cap X) \textit{ then } (C,s') \models \varphi.
\end{aligned}
$$

A formula φ of $\mathscr{L}(Atm)$ is said to be satisfiable relative to the class **CM** if there exists a pointed classifier model (C, s) with $C \in \mathbf{CM}$ such that $(C, s) \models \varphi$. It is said to be valid if $\neg\varphi$ is not satisfiable relative to **CM** and noted as $\models_{\mathbf{CM}} \varphi$.

We can think of a pointed model (C, s) as a pair (s, x) in f with $f(s) = x$. The formula $[X]\varphi$ is true at a state s if φ is true at all states that are modulo-X equivalent to state s. It has the *selectis paribus* (SP) (selected things being equal) interpretation "features in X being equal, necessarily φ holds (under possible perturbation on the other features)". $[Atm_0 \setminus X]\varphi$ has the standard *ceteris paribus* (CP) interpretation "features other than X being equal, necessarily φ holds (under possible perturbation of the features in X)". Notice when $X = \emptyset$, $[\emptyset]$ is the S5 universal modality since every state is modulo-$\emptyset$ equivalent to all states, viz. $(C, s) \models [\emptyset]\varphi \iff \forall s' \in S, (C, s') \models \varphi$.

4. Representation between Consistent Case Base and CM

In this section we shall show that the language of case bases can be translated into the language $\mathscr{L}(Atm)$; hence case bases can be studied by classifier models. More precisely, a case base is consistent iff its translation, together with the following two formulas that we abbreviate as `Compl` and `2Mon`, is satisfiable in the class **CM**:

$$
\mathtt{Compl} =_{def} \bigwedge_{X \subseteq Atm_0} \langle\emptyset\rangle \mathsf{cn}_{X, Atm_0}
$$

$$
\mathtt{2Mon} =_{def} \bigwedge_{x \in \{0,1\}, X \subseteq Atm_0^x, Y \subseteq Atm_0^{\overline{x}}} \Big(\langle\emptyset\rangle(\mathsf{cn}_{X \cup Y, Atm_0} \wedge \mathtt{t}(x)) \rightarrow \bigwedge_{Atm_0^x \supseteq X' \supseteq X, Y' \subseteq Y} [\emptyset](\mathsf{cn}_{X' \cup Y', Atm_0} \rightarrow \mathtt{t}(x)) \Big)
$$

According to `Compl`, *every possible* situation description must be satisfied by the classifier, where a situation description is a conjunction of factors (those being present X) and negations of factors (those being absent, $Atm_0 \setminus X$).

`2Mon` introduces a *two-way monotonicity*, which is meant to implement the *a fortiori* constraint: if the classifier associates a situation s to an outcome x, then it must assign the same outcome to every situation s' such that both (a) s' includes all factors for x that are in s and (b) s' does *not include* factors for $\bar{x}$ that are *outside of* s. This formula is meant to maintain consistency with respect to the preference relation, as Definition 1 indicates: if a case including reason X for x and factors Y for $\bar{x}$, has outcome x, it means that $X > Y$. Thus it cannot be that outcome $\bar{x}$ is assigned to a situation s' including both a superset $X' \supseteq X$ of factors for x and only a subset $Y' \subseteq Y$ of factors for $\bar{x}$. In fact, if $X > Y$, then is must be the case that also $X' > Y'$, while a decision for $\bar{x}$ would entail that $X' < Y'$.

Let $\mathbf{CM}^{prec} = \{C = (S, f) \in \mathbf{CM} : \forall s \in S, (C, s) \models \mathtt{Compl} \wedge \mathtt{2Mon}\}$, where $\mathbf{CM}^{prec}$ means the class of CMs for precedent theory. Satisfiability and validity relative to $\mathbf{CM}^{prec}$ are defined in an analogous way as $\mathbf{CM}$.

To translate a result-model case-base CB^{res} into a classifier model (C, f), we need to ensure that all precedents in the case-base are satisfied by the classifier, with regard to both their factors and their outcome.

Definition 7 (Translation of case base for result model)**.** *The translation function tr_1 maps each case from a case base CB^{res} to a corresponding formula in the language $\mathscr{L}(Atm)$. It is defined as follows:*

$$tr_1(s, s \cap Atm_0^x, x) =_{def} \langle \emptyset \rangle (\mathsf{cn}_{s, Atm_0} \wedge \mathsf{t}(x)).$$

We generalize it to the entire case base CB^{res} as follows:

$$tr_1(CB^{\text{res}}) =_{def} \bigwedge_{(s, s \cap Atm_0^x, x) \in CB} tr(s, s \cap Atm_0^x, x).$$

Therefore, in the result model a precedent $(s, s \cap Atm_0^x, x)$ is viewed as a situation s being classified by f as x.

Example 3 (Running example)**.** *The case $(\{\pi_1, \pi_2, \delta_1\}, \{\pi_1, \pi_2\}, 1)$ is translated as $\langle \emptyset \rangle (\pi_1 \wedge \pi_2 \wedge \delta_1 \wedge \neg \pi_3 \wedge \neg \delta_2 \wedge \neg \delta_3 \wedge \mathsf{t}(1))$, which means that $f(\pi_1, \pi_2, \delta_1) = 1$*

In translations for the reason model we need to capture the role of reasons. This is obtained by ensuring that for every case (s, X, x), not the fact situation s directly, but the one consisting only of reason X and all $\bar{x}$-factors in s (i.e. $s \cap Atm_0^{\bar{x}}$) is classified as x. It reflects that the precedent finds x-factors in s outside of X dispensable for the outcome.

Definition 8 (Translation of case base for reason model)**.** *The translation function tr_2 maps each case from a case base CB to a corresponding formula in the language $\mathscr{L}(Atm)$. It is defined as follows:*

$$tr_2(s, X, x) =_{def} \langle \emptyset \rangle (\mathsf{cn}_{X \cup (s \cap Atm_0^{\bar{x}}), Atm_0} \wedge \mathsf{t}(x)).$$

We generalize it to the entire case base CB as follows:

$$tr_2(CB) =_{def} \bigwedge_{(s,s\cap Atm_0^x,x)\in CB} tr_2(s,s\cap Atm_0^x,x).$$

Notice that the function tr_1 for the result model is a special case of the function tr_2 for the reason model, since $((s\cap Atm_0^x)\cup(s\cap Atm_0^{\overline{x}}) = s$.

Fact 1. $tr_1(s,s\cap Atm_0^x,x) = tr_2(s,s\cap Atm_0^x,x)$.

The formulas `2Mon` and `Compl` require that the outcome x supported by reason X in a precedent is assigned to all possible cases including X that do not contain additional factors against x. If both formulas are satisfiable then the case base is consistent, as stated by the following theorem.

Theorem 1. *Let $CB \in \mathbf{CB}$ be a case base. Then, CB is consistent iff $tr_2(CB)$ is satisfiable in $\mathbf{CM}^{prec}$.*

In light of the theorem and the fact above, the representation of a case base for result model turns to be a corollary.

Corollary 1. *Let $CB^{\text{res}} \in \mathbf{CB}^{\text{res}}$ be a case base for the result model. Then, CB^{res} is consistent iff $tr_1(CB^{\text{res}})$ is satisfiable in $\mathbf{CM}^{prec}$.*

Similarly, the precedential constraint can also be represented as a corollary.

Corollary 2. *Let $CB \in \mathbf{CB}$ be a consistent case base and (s',X,x) a case. Updating CB with (s',X,x) meets the precedential constraint, iff $tr_2(CB)\wedge tr_2(s',X,x)$ is satisfiable in $\mathbf{CM}^{prec}$.*

Example 4 (Running example)**.** *Case $c_3 = (\{\pi_1,\pi_2,\delta_2\},\{\delta_2\},0)$ is incompatible with the CB^{ex}. According to $tr_2(CB^{ex}\cup\{c_3\})$, `2Mon` and `Compl`, the fact situation $\{\pi_1,\pi_2,\delta_1\}$ should be classified both as 1, based on CB^{ex}, and 0, based on c_3.*

5. Explanations

The representation results above pave the way to providing explanations for the outcomes of cases. For this purpose it is necessary to introduce the following notations. Let λ denote a conjunction of finitely many literals, where a literal is an atom p (positive literal) or its negation $\neg p$ (negative literal). We write $\lambda\subseteq\lambda'$, call λ a part (subset) of λ', if all literals in λ also occur in λ'; and $\lambda\subset\lambda'$ if $\lambda\subseteq\lambda'$ but not $\lambda'\subseteq\lambda$. We write $Lit(\lambda), Lit^+(\lambda), Lit^-(\lambda)$ to mean all literals, all positive literals and all negative literals in λ respectively. By convention $\top$ is a term of zero conjuncts. In the glossary of Boolean classifier (function), λ is called a *term* or *property* (of the instance s). The set of terms is noted *Term*. A key role in our analysis is played by the notion of a (prime) implicant, i.e., a (subset-minimal) term which makes a classification necessarily true.

Definition 9 (Implicant (Imp) and prime implicant (PImp))**.** *We write $\mathtt{Imp}(\lambda,x)$ to mean that λ is an* implicant *for x and define it as $\mathtt{Imp}(\lambda,x) =_{def} [\emptyset](\lambda\rightarrow \mathtt{t}(x))$. We write $\mathtt{PImp}(\lambda,x)$ to mean that λ is a* prime implicant *for x and define it as*

$$\mathtt{PImp}(\lambda,x) =_{def} [\emptyset]\Big(\lambda\rightarrow\big(\mathtt{t}(x)\wedge\bigwedge_{p\in Atm(\lambda)}\langle Atm(\lambda)\setminus\{p\}\rangle\neg\mathtt{t}(x)\big)\Big).$$

According to the definition, λ being an implicant for x means that any state s verifying λ is necessarily classified as x (necessity); and λ being a prime implicant for x means that any proper subset of λ is not an implicant for x (minimality).[5] Implicants explain the classifier in the sense that to know an implicant satisfied at a state is to know the classification of the state.

Intuitively, for a case base containing precedent (s,X,x) to be consistent, s must be incompatible with every prime implicant λ for $\bar{x}$. To guarantee that, either λ must have some literal $\neg p$, where p is in X and hence is true at s; or λ must have some literal p, where $p \notin s \cap Atm_0^{\bar{x}}$ and hence is false at s.

Proposition 1. *Let CB be a consistent case base and* $(s,X,x) \in CB$, *and* $C \in \mathbf{CM}^{prec}$ *s.t.* $(C,s) \models tr_2(CB)$. *Then,* $\forall \lambda \in Term_{,}$ *if* $(C,s) \models \mathtt{PImp}(\lambda,\bar{x})$, *then either* $X \cap Atm(Lit^{-}(\lambda)) \neq \emptyset$ *or* $s \cap Atm_0^{\bar{x}} \not\supseteq Atm(Lit^{+}(\lambda))$.

Example 5. *Let* $C = (S,f) \in \mathbf{CM}^{prec}$ *and* $tr_2(CB^{ex})$ *is satisfiable in C. Obviously* π_1 *cannot be PImp for* 0, *otherwise* $f(s_1) = 0$, *contrary to* c_1. *Also* $\neg\delta_2 \wedge \pi_2$ *cannot be PImp for* 1, *otherwise* $f(\{\pi_2,\delta_1,\delta_3\}) = 1$, *contrary to* c_2.

In XAI, people [18,5,12] also focus on "local" (prime) implicants, namely (prime) implicants true at a given state. We adopt the definitions of abductive explanations in [12,10], and express these notions in $\mathscr{L}(Atm)$ as follows:

Definition 10 (Abductive explanation (AXp) and weak abductive explanation (wAXp))**.** *We write* $\mathtt{AXp}(\lambda,x)$ *to mean that* λ abductively explains *the decision x and define it as* $\mathtt{AXp}(\lambda,x) =_{def} \lambda \wedge \mathtt{PImp}(\lambda,x)$. *We write* $\mathtt{wAXp}\ (\lambda,x)$ *to mean that* λ weak-abductively explains *the decision x and define it as* $\mathtt{wAXp}\ (\lambda,x) =_{def} \lambda \wedge \mathtt{Imp}(\lambda,x)$.

The proposition below states that to be the reason (of a fact situation) is to be the positive part of some weak AXp of that situation. Notice a reason is not always the positive part of some AXp, since reasons in precedent do not in general respect minimality.

Proposition 2. *Let CB be a consistent case base,* $(s,X,x) \in CB$, *and* $C \in \mathbf{CM}^{prec}$ *s.t.* $(C,s) \models tr_2(CB)$. *Then* $\exists \lambda \in Term$ *s.t.* $Atm(Lit^{+}(\lambda)) = X$ *and* $(C,s) \models \mathtt{wAXp}\ (\lambda,x)$.

In fact, we always know one weak AXp for a precedent (s,X,x) in a consistent case base, i.e., the conjunction of all factors in X and negations of all $\bar{x}$-factors that are in s.

Proposition 3. *Let CB be a consistent case base,* $(s,X,x) \in CB$, *and* $C \in \mathbf{CM}^{prec}$ *s.t.* $(C,s) \models tr_2(CB)$. *Then we have* $(C,s) \models \mathtt{wAXp}\ (\mathtt{cn}_{X,X\cup(s\cap Atm_0^{\bar{x}})},x)$.

Example 6. *Let* $C \in \mathbf{CM}^{prec}$ *be a model of* $tr_2(CB^{ex})$. *Then we have* $(C,s_1) \models \mathtt{wAXp}\ (\pi_1 \wedge \neg\delta_2,1)$ *and* $(C,s_2) \models \mathtt{wAXp}\ (\delta_2 \wedge \neg\pi_1 \wedge \neg\pi_2,0)$. *Notice that* $(C,s_2) \models \neg\mathtt{wAXp}\ (\delta_2,0)$, *because e.g.* $(C,s_1) \models \delta_2 \wedge \neg\mathtt{t}(0)$.

The idea of contrastive explanation is dual with abductive explanation, since it points to a minimal part of a situation whose change would falsify the current decision, and

[5] Notice that we have not fully used the expressive power of $[X]\varphi$ and $\langle X\rangle\varphi$ until now for minimality. The intuitive meaning of $\langle Atm(\lambda)\setminus\{p\}\rangle\neg\mathtt{t}(x)$ in the formula is that even if we just perturb one variable p in λ from its actual value, the classification will possibly no longer be x.

the duality between their weak versions is similar [10]. A conjunction of literals λ is a contrastive explanation for outcome x in situation s, if the following conditions are satisfied: (a) λ is true at s, and s has outcome x, (b) if all literals in λ were false then the outcome would be different, (c) λ is the subset-minimal literals satisfiying (a) and (b). A weak contrastive explanation is only based on conditions (a) and (b).

Definition 11 (Contrastive explanation (CXp) and weak contrastive explanation (wCXp))**.** *We write* $\mathtt{CXp}(\lambda,x)$ *to mean that* λ constrastively explains *the decision* x *and define it as*

$$\mathtt{CXp}(\lambda,x) =_{def} \lambda \wedge \langle Atm_0 \setminus Atm(\lambda)\rangle \neg \mathtt{t}(x) \wedge \bigwedge_{p \in Atm(\lambda)} [(Atm_0 \setminus Atm(\lambda)) \cup \{p\}]\mathtt{t}(x).$$

We write $\mathtt{wCXp}\ (\lambda,x)$ *to mean that* λ weak-contrastively explains *the decision* x *and define it as* $\mathtt{wCXp}\ (\lambda,x) =_{def} \lambda \wedge \mathtt{t}(x) \wedge \langle Atm_0 \setminus Atm(\lambda)\rangle \neg \mathtt{t}(x)$.

Intuitively speaking, we can test whether λ is a wCXp of situation s having outcome x by "flipping" its positive literals to negative, and negative to positive, and observe if the resulting state is classified differently from x. CXp is the subset-minimal wCXp.

Weak CXps can be used to study the preferences between reasons in a case base. The next proposition indicates that given a precedent (s,X,x), if the absence of Y at s, by itself *alone* can weakly contrastively explain x, then Y is "no weaker than" X in CB.

Proposition 4. *Let CB be a consistent case base and* $(s,X,x) \in CB$*, and* $C \in \mathbf{CM}^{prec}$ *s.t.* $(C,s) \models tr_2(CB)$*. If* $(C,s) \models \mathtt{wCXp}\ (\mathtt{cn}_{\emptyset,Y},x)$*, then it is not the case that* $Y <_{CB} X$.

Example 7. *Let* $C \in \mathbf{CM}^{prec}$ *be a model of* $tr_2(CB^{ex})$*. Since* $\{\delta_3\} <_{CB^{ex}} \{\pi_1\}$*, we have* $(C,s_2) \models \mathtt{wCXp}\ (\mathtt{cn}_{\emptyset,\{\pi_1\}},0)$*. Indeed* $f(\pi_1,\delta_1,\delta_3) = 0$ *by* `2Mon` *according to* s_1.

6. Related Work and Conclusion

In this paper, we have shown that through the concept of classifier a novel logical model of factor-based case-based reasoning can be provided, which allows for a rigorous analysis of case bases and of the inferences they support.

As noted in the introduction, our work is based upon the case-based reasoning models of HYPO and CATO [2,1] and upon the analysis of precedential constraint by Jeff Horty [8,9]. Further approaches exist that make use of logic in reasoning with cases. For instance, [17] provided a factor-based model based on formal defeasible argumentation. More recently [19,20] represent precedents as propositional formulas and compare precedents by (propositional) logical entailment.

However, this propositional representation does not fully use the power of logic, in the sense that it does not provide a proof theory (axiomatics) for reasoning with precedents. By contrast, besides the semantic framework presented here, we can make syntactic derivations of properties of CBR using the axiomatics of BCL (see in Appendix).

Moreover, our representation results allow for exploring different notions of explanation, such as abductive and contrastive explanations. We can accordingly explain why a case-based reasoning suggests a particular outcome (rather then a different one) in a new case. Thus, out model could be used to build explainable case-based reasoners,

which could also be deployed to reproduce and analyze the functioning of opaque predictors of the outcome of cases. Thus, by bringing CBR into the broader context of classifier systems, we connect three lines of research: legal case-based reasoning, AI&Law approaches on to explanation [3], techniques and results developed in the context of XAI.

In future work we will examine more deeply the relation between classifiers, explanations, and reasoning with legal precedents. Interesting developments pertain to addressing analogical reasoning beyond the a fortiori constraint considered here and to deploying ideas of explanation to extract knowledge out of cases (e.g., to determine the direction of factors and the way in which they interact).

References

[1] Vincent Aleven. Using background knowledge in case-based legal reasoning: a computational model and an intelligent learning environment. *Artificial Intelligence*, 150(1-2):183–237, 2003.
[2] Kevin D. Ashley. *Modeling Legal Argument: Reasoning with Cases and Hypotheticals*. MIT, 1990.
[3] Katie Atkinson, Trevor Bench-Capon, and Danushka Bollegala. Explanation in ai and law: Past, present and future. *Artificial Intelligence*, 289:103387, 2020.
[4] Gilles Audemard, Steve Bellart, Louenas Bounia, Frédéric Koriche, Jean-Marie Lagniez, and Pierre Marquis. On the computational intelligibility of boolean classifiers. In *Proceedings of KR 2021*, number 1, pages 74–86, 2021.
[5] Adnan Darwiche and Auguste Hirth. On the reasons behind decisions. In *Proceedings of ECAI 2020*, pages 712–720. IOS Press, 2020.
[6] John Horty. Reasoning with dimensions and magnitudes. In *International Conference on Artificial Intelligence and Law, ICAIL2017*. ACM, 2017.
[7] John F. Horty. The result model of precedent. *Legal Theory*, 10:19–31, 2004.
[8] John F. Horty. Rules and reasons in the theory of precedent. *Legal theory*, 17:1–33, 2011.
[9] John F. Horty and Trevor J. M. Bench-Capon. A factor-based definition of precedential constraint. *Artificial intelligence and Law*, 20:181–214, 2012.
[10] Xuanxiang Huang, Yacine Izza, Alexey Ignatiev, Martin Cooper, Nicholas Asher, and Joao Marques-Silva. Tractable explanations for d-dnnf classifiers. In *Proceedings of AAAI 2022*, pages 5719–5728, 2022.
[11] Alexey Ignatiev, Nina Narodytska, Nicholas Asher, and Joao Marques-Silva. From contrastive to abductive explanations and back again. In *International Conference of the Italian Association for Artificial Intelligence*, pages 335–355. Springer, 2020.
[12] Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. Abduction-based explanations for machine learning models. In *Proceedings of AAAI 2019*, pages 1511–1519, 2019.
[13] Xinghan Liu and Emiliano Lorini. A logic for binary classifiers and their explanation. In *Proceedings of the 4th International Conference on Logic and Argumentation (CLAR 2021)*, pages 302–321. Springer-Verlag, 2021.
[14] Xinghan Liu and Emiliano Lorini. A unified logical framework for explanations in classifier systems. *Journal of Logic and Computation*, forthcoming.
[15] Tim Miller, Robert Hoffman, Ofra Amir, and Andreas Holzinger, editors. *Artificial Intelligence journal: Special issue on Explainable Artificial Intelligence (XAI)*, volume 307, 2022.
[16] Henry Prakken. A formal analysis of some factor and precedentbased accounts of precedential constraint. *Artificial Intelligence and Law*, 2021.
[17] Henry Prakken and Giovanni Sartor. Modelling reasoning with precedents in a formal dialogue game. *Artificial Intelligence and Law*, 6:231–87, 1998.
[18] Andy Shih, Arthur Choi, and Adnan Darwiche. Formal verification of bayesian network classifiers. In *International Conference on Probabilistic Graphical Models*, pages 427–438. PMLR, 2018.
[19] Heng Zheng, Davide Grossi, and Bart Verheij. Case-based reasoning with precedent models: Preliminary report. In *Computational Models of Argument*, pages 443–450. IOS Press, 2020.
[20] Heng Zheng, Davide Grossi, and Bart Verheij. Precedent comparison in the precedent model formalism: theory and application to legal cases. In *Proceedings of the EXplainable and Responsible AI in Law (XAILA) Workshop at JURIX*, 2020.