A Logic for Binary Classifiers and their Explanation

Xinghan Liu[1,2] and Emiliano Lorini[1,2]
[1]ANITI, Toulouse University, France
[2]IRIT-CNRS, Toulouse University, France

Abstract

Recent years have witnessed a renewed interest in Boolean function in explaining binary classifiers in the field of explainable AI (XAI). The standard approach of Boolean function is propositional logic. We present a modal language of a ceteris paribus nature which supports reasoning about binary classifiers and their properties. We study families of decision models for binary classifiers, axiomatize them and show completeness of our axiomatics. Moreover, we prove that the variant of our modal language with finite propositional atoms interpreted over these models is NP-complete. We leverage the language to formalize counterfactual conditional as well as a bunch of notions of explanation such as abductive, contrastive and counterfactual explanations, and biases. Finally, we present two extensions of our language: a dynamic extension by the notion of assignment enabling classifier change and an epistemic extension in which the classifier's uncertainty about the actual input can be represented.

1 Introduction

The notions of explanation and explainability have been extensively investigated by philosophers [11, 14] and are key aspects of AI-based systems given the importance of explaining the behavior and prediction of an artificial intelligent system. A variety of notions of explanations have been discussed in the area of explainable AI (XAI) including abductive, contrastive and counterfactual explanation [7, 1, 17, 13, 18, 16]. Recently, there has been a renewed interest for the notion of explanation in the context of classifier systems, i.e., explaining why a classifier has classified a given input instance in a certain way [6, 19]. Classifier systems can be seen as "black boxes" computing a given (Boolean) function in the context of a classification or prediction task. Artificial feedforward neural networks are special kinds of classifier systems aimed at learning or, at least approximating, the function mapping instances of the input data to their corresponding outputs. Explaining why the system has classified a given instance in a certain way and identifying the set of features that is necessary/(minimally) sufficient for the classification is crucial for making the system intelligible and for finding biases in the classification process.

In this paper we introduce a modal language for representing classifiers with binary input data. It includes modal operators to represent that a formula is necessarily true regardless of the truth or the falsity of certain atomic facts.

There are two reasons for a modal logic approach. First, a ceteris paribus aspect can be found in explaining the classifier. A sufficient reason (an abductive explanation) of the decision x for an instance α is such a minimal property λ, that the decision x necessarily takes place *regardless of other variables*. Correspondingly, λ' contrastively explains x for a given α, if changing the value of any variable in λ', *other variables being equal*, defects x.

Second, Boolean classifier is represented as Boolean function, and the latter is traditionally analyzed by, or even identical to propositional logic, since the work of Boole. Extending the logical framework to modal logic and more generally to non-classical logics, opens up new vistas including (i) defining counterfactual conditional and studying its relationship with the notions of abductive and contrastive explanation, (ii) representing the classifier's uncertainty about the actual instance to be classified through the use of epistemic logic [8], and (iii) modeling classifier dynamics through the use of of formal semantics for logics of communication and change [21, 24].

The paper is structured as follows. In Section 2 we introduce the ceteris paribus language which supports reasoning about binary classifiers and their properties. Moreover, we present the formal semantics based on the notion of decision model with respect to which the language is interpreted. Section 3 is devoted to the axiomatics and the complexity of satisfiability checking for our modal language. In Section 4, we define counterfactual conditional and highlights its relevance for understanding the behavior of a classifier. Then, in Section 5 we leverage the language to formalize the notions of abductive explanation (AXp), contrastive explanation (CXp) and bias and elucidating the connection between CXp and counterfactual conditional. Finally, in Section 6 we present two extensions of our language: (i) a dynamic extension by the notion of assignment enabling classifier change and (ii) an epistemic extension in which the classifier's uncertainty about the actual input can be represented. Except proof sketches of some theorems, all proofs are found in the appendix.

2 A Language for Binary Classifiers

In this section we introduce the language for modeling binary classifiers and present its semantics. The language has a *ceteris paribus* nature. It contains *ceteris paribus* operators of the form $[X]$ that allow us to express the fact that the classifier's actual decision (or classification) does not depend on the features of the input in the complementary set $Atm \setminus X$, with Atm the set of atomic propositions and X a finite subset of it. Such operators were introduced for the first time in [10].

2.1 Basic Language and Ceteris Paribus Model

Let Atm be a countable set of atomic propositions with elements noted $p, q, \ldots$ We define $AtmSet = 2^{Atm}$ and $AtmSet_0 = \{X \in AtmSet : X \text{ is finite}\}$.

Let us assume the set Atm includes special atomic formulas of type $\mathtt{t}(x)$, where $x \in Val$ and Val is a finite set of decision values with elements noted $x, y, \ldots$ We call them decision atoms and note $Dec = \{\mathtt{t}(x) : x \in Val\}$ the corresponding set. The decision atom $\mathtt{t}(x)$ has to be read "the actual decision takes value x".

The modal language $\mathcal{L}(Atm)$ is defined by the following grammar:

$$\varphi \quad ::= \quad p \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid [X]\varphi$$

where p ranges over Atm and X ranges over $AtmSet_0$.[1] The set of atomic propositions occurring in a formula φ is noted $Atm(\varphi)$.

The formula $[X]\varphi$ has a *ceteris paribus* (CP) reading: "φ is necessary all features in X being equal" or "φ is necessary regardless of the truth or falsity of the atoms in $Atm \setminus X$". Operator $\langle X\rangle$ is the dual of $[X]$ and is defined as usual: $\langle X\rangle\varphi =_{def} \neg[X]\neg\varphi$.

The natural interpretation of the modal operators $[X]$ is by means of *ceteris paribus* models of the following kind.

Definition 1 (Ceteris paribus model)**.** *A* ceteris paribus *model is a tuple* $M = (W, (\equiv_X)_{X\in AtmSet_0}, V)$ *where:*

- W *is a set of worlds,*
- *every* $\equiv_X$ *is a binary relation on* W*,*
- $V : W \longrightarrow 2^{Atm}$ *is a valuation function,*

and which satisfies the following constraints, for all $w, v \in W$*, and* $X \in AtmSet_0$*:*

(C1) $w \equiv_X v$ *if and only if* $V_X(w) = V_X(v)$*,*

where, for every $X \in AtmSet_0$*,* $V_X(w) = \big(V(w) \cap X\big)$*. The class of ceteris paribus models is noted* $\mathbf{M}$*.*

The relation $\equiv_X$ captures the concept of circumstantial indistinguishability (or equivalence modulo-X) which is needed for the intepretation of the modal operators $[X]$. In particular, $w \equiv_X v$ means that w and v are indistinguishable, with regard to the circumstances (the features) in X.

2.2 Decision Model and X-Properties

In order to be able to interpret the special decision atoms $\mathtt{t}(x)$ in the right way, we need a bit more structure. For this reason, we move from less structured *ceteris paribus* models to more structured decision models, which are defined as follows.

Definition 2 (Decision model)**.** *A decision model (DM) is a ceteris paribus model* $M = (W, (\equiv_X)_{X\in AtmSet_0}, V)$ *in the sense of Definition 1 which satisfies the following constraints, for all* $x, y \in Dec$ *and* $w, v \in W$*:*

(C2) $V_{Dec}(w) \neq \emptyset$*,*

[1] We do not include modal operators $[X]$ for X infinite, since we want our language to be finitary.

(C3) *if* $\mathtt{t}(x), \mathtt{t}(y) \in V(w)$ *then* $x = y$.

For every finite $X \subseteq \big(Atm \setminus Dec\big)$ *the decision model* M *is said to be* X*-definite, shortly* $def(X)$*, if it satisfies the following additional constraint, for all and* $w, v \in W$*:*

(C4) *if* $V_X(w) = V_X(v)$ *then* $V_{Dec}(w) = V_{Dec}(v)$.

It is said to be X*-complete, shortly* $comp(X)$*, if it satisfies the following additional constraint:*

(C5) $\forall X' \subseteq X, \exists w \in W$ *such that* $V_X(w) = X'$.

Finally, it is said to be X*-non trivial, shortly* $ntr(X)$*, if the following constraint is satisfied:*

(C6) $\exists w, v$ *s.t.* $V_X(w) \neq V_X(v)$ *and* $V_{Dec}(w) \neq V_{Dec}(v)$.

A pointed decision model is a pair (M, w) with $M = (W, (\equiv_X)_{X \in AtmSet_0}, V)$ a decision model and $w \in W$.

The class of decision models is noted $\mathbf{DM}$. Let $P \subseteq Prop$ with

$$Prop = \{def(X), comp(X), ntr(X) : X \subseteq \big(Atm \setminus Dec\big) \text{ and } X \text{ is finite}\}.$$

We denote by $\mathbf{DM}_P$ the class of decision models satisfying each property in P.

Definition 3 (Extension and Z-groundedness)**.** *Let* $M' = (W', (\equiv_X)_{X \in AtmSet_0}, V')$ *and* $M = (W, (\equiv_X)_{X \in AtmSet_0}, V)$ *be two decision models and* $Z \in AtmSet_0$.

We say that M' *is an extension of* M *(and* M *a submodel of* M'*), if* $W \subseteq W'$ *and* $V = V'|_W$.

We say M *is grounded on* Z*, if* M *is* Z*-definite, and* $\forall w \in W, \forall p \in Atm \setminus (Dec \cup Z), p \notin V(w)$.

The following proposition captures some interesting properties of X-definiteness, X-completeness and X-non triviality for decision models.

Proposition 1. *Let* $X, X' \subseteq \big(Atm \setminus Dec\big)$ *be finite and let* M *be a decision model. Then,*

1. *if* $M \in \mathbf{DM}_{\{def(X)\}}$ *and* $X \subseteq X'$ *then* $M \in \mathbf{DM}_{\{def(X')\}}$,
2. *if* $M \in \mathbf{DM}_{\{comp(X)\}}$ *and* $X' \subseteq X$ *then* $M \in \mathbf{DM}_{\{comp(X')\}}$,
3. *if* $M \in \mathbf{DM}_{\{def(X)\}}$*, then* M *can extend to an* $M' \in \mathbf{DM}_{\{def(X), comp(X)\}}$,
4. *if* $M \in \mathbf{DM}_{\{def(X), ntr(X), ntr(X'), comp(X \cup X')\}}$ *and* $X \cap X' = \emptyset$ *then* $M \notin \mathbf{DM}_{\{def(X')\}}$.

As the last item of the previous proposition highlights, there are cases in which the class of decision models collapses, i.e., for some set of semantic properties P, the model class $\mathbf{DM}_P$ is empty. However, there are also cases in which the model class surely does not collapse. For instance, as the following proposition indicates, the X-representative class of models $\mathbf{DM}_{\{def(X), comp(X), ntr(X)\}}$ is non-empty.

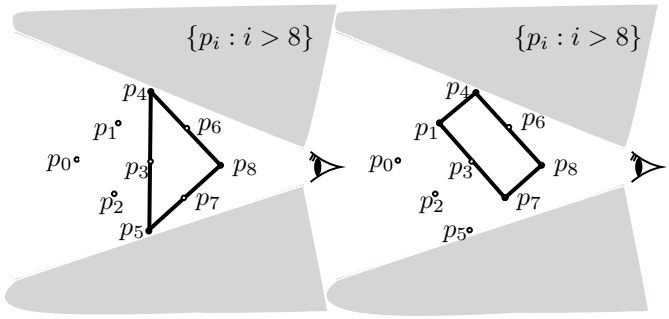

Figure 1: Example of classification of geometric figures

Proposition 2. *Let $P = \{def(X), comp(X), ntr(X)\}$. Then, $\mathbf{DM}_P$ is non-empty. Moreover, if Atm is infinite, then $\mathbf{DM}_P$ is infinite too.*

The intuitive idea behind the notion of Z-groundedness given in Definition 3 is that in practice, classifiers only take finitely many features in order to make decisions. The case is analogous to human's eyes, where though infinitely many objects are present in our eyes, perceivers can only be *aware of* finitely many pixels, due to limitation of resolution and perspective, and leave others in blind area. Figure 1 illustrates this idea. The eye represents a Z-definite DM M where $Z = \{p_0, \ldots, p_8\}$. Solid circle means $p \in V(w)$ and hollow circle means $p \notin V(w)$. All $Atm \setminus Z$ locates in the grey zone which is invisible to the eye, therefore for any $p_i \in Atm$ where $i > 8$, we have $p_i \notin V(w)$. Each configuration represents a world in the model. In the left world w_l we have $V_Z(w_l) = \{p_4, p_5, p_8\}$, in the right world w_r we have $V_Z(w_r) = \{p_1, p_4, p_7, p_8\}$. Figures result from linking all solid circles in the set Z stand for the classifications for the two figures, which are different. In particular, $V_{Dec}(w_l) = \{\text{“}\texttt{triangle}\text{”}\}$ and $V_{Dec}(w_r) = \{\text{“}\texttt{rectangle}\text{”}\}$.

Formulas in $\mathcal{L}(Atm)$ are interpreted relative to a pointed decision model, as follows.

Definition 4 (Satisfaction relation)**.** *Let (M, w) be a pointed decision model with $M = (W, (\equiv_X)_{X \in AtmSet_0}, V)$. Then:*

$$
\begin{array}{rcl}
(M, w) \models p & \iff & p \in V(w), \\
(M, w) \models \neg\varphi & \iff & (M, w) \not\models \varphi, \\
(M, w) \models \varphi \wedge \psi & \iff & (M, w) \models \varphi \textit{ and } (M, w) \models \psi, \\
(M, w) \models [X]\varphi & \iff & \forall v \in W, \textit{ if } w \equiv_X v \textit{ then } v \models \varphi.
\end{array}
$$

We abbreviate $(M, w) \models \varphi$ as $w \models \varphi$ when the context is clear.

A formula φ of $\mathcal{L}(Atm)$ is said to be satisfiable relative to the class $\mathbf{DM}_P$ if there exists a pointed decision model (M, w) with $M \in \mathbf{DM}_P$ such that $(M, w) \models \varphi$. It is said to be valid relative to $\mathbf{DM}_P$, noted $\models_{\mathbf{DM}_P} \varphi$, if $\neg\varphi$ is not satisfiable relative to $\mathbf{DM}_P$. Moreover, we say that that φ is valid in the model $M = (W, (\equiv_X)_{X \in AtmSet_0}, V)$, noted $M \models \varphi$, if $(M, w) \models \varphi$ for every $w \in W$.

2.3 Non-Redundant Model

The following definition introduces the concept of non-redundant decision model in which there are no multiple copies of the same valuation for the atomic formulas in $Atm \setminus Dec$.

Definition 5 (Non-redundant decision model)**.** *A non-redundant decision model (NDM) is a tuple $M^{nr} = (S, f)$ where:*

- $S \subseteq 2^{Atm \setminus Dec}$ *is a set of states, and*
- $f : S \longrightarrow Val$ *is a decision function.*

For every finite $X \subseteq \big(Atm \setminus Dec\big)$, the non-redundant decision model $M^{nr} = (S, f)$ is said to be X-definite (i.e., $def(X)$) if, for every $s, s' \in S$, if $s \cap X = s' \cap X$ then $f(s) = f(s')$.

It is is said to be X-complete (i.e., $comp(X)$) if $2^X \subseteq S$.

Finally, it is said to be X-non trivial (i.e., $ntr(X)$) if $\exists s, s'$ such that $(s \cap X) \neq (s' \cap X)$ and $f(s) \neq f(s')$.

The class of non-redundant decision models is noted $\mathbf{NDM}$. For every $P \subseteq Prop$, we denote by $\mathbf{NDM}_P$ the class of non-redundant decision models satisfying each property in P.

Notions of extension and Z-groundedness for decision models given in Definition 3 transfer to non-redundant decision models in the obvious way. Following the same proof strategy, all properties stated in Proposition 1 of also hold for appropriate NDMs.

The satisfaction relation for non-redundant decision models is defined by the following clauses for $M^{nr} = (S, f) \in \mathbf{NDM}$ and $s \in S$. (We omit boolean cases since they are defined in the usual way.)

$$\begin{array}{rcl} (M^{nr}, s) \models p & \iff & p \in s, \\ (M^{nr}, s) \models \mathsf{t}(x) & \iff & f(s) = x, \\ (M^{nr}, s) \models [X]\varphi & \iff & \forall s' \in S : \text{ if } (s \cap X) = (s' \cap X), \\ & & \text{then } (M^{nr}, s') \models \varphi. \end{array}$$

Notions of satisfiability and validity for formulas in $\mathcal{L}(Atm)$ relative to the class $\mathbf{NDM}_P$ as well as non-redundant decision model validity are defined in the usual way. As usual, we write $\models_{\mathbf{NDM}_P} \varphi$ to denote the fact that formula φ is validity relative to the class $\mathbf{NDM}_P$. As for DMs, we abbreviate $(M^{nr}, s) \models \varphi$ as $s \models \varphi$ when the context is clear.

The following theorem states the equivalence between models and non-redundant models with regard to the language $\mathcal{L}(Atm)$.

Theorem 1. *Let $P \subseteq Prop$ and $def(X) \in P$ for some $X \in AtmSet_0$, and $\varphi \in \mathcal{L}(Atm)$. Then, φ is satisfiable relative to the class $\mathbf{DM}_P$ iff it is satisfiable relative to the class $\mathbf{NDM}_P$.*

The reason of introducing NDMs is that they naturally encode classifiers in the real world. We view a binary classifier for the finite set of features $Z \in AtmSet_0$ as a

(possibly partial) Boolean function $g_Z : Input_Z \to Dec$, [2] where $Input_Z$ denotes the set of all possible inputs from Z, with an input from Z being a conjunction of elements from $Lit_Z = Z \cup \{\neg p : p \in Z\}$ where each atom in Z appears exactly once and, for each $p \in Z$, p is an element of the conjunction if and only if $\neg p$ is not. Members of $Input_Z$ (a.k.a. instances) are noted by $\alpha, \alpha', \beta \ldots$ Terms from Z, a.k.a. properties (of an instance), noted $\lambda, \lambda', \tau, \ldots$, are conjunctions of literals from Lit_Z where each atom in Z appears at most once and, for each $p \in Z$, p is an element of the conjunction if and only if $\neg p$ is not. In this sense, a term can be seen as a possibly incomplete input. By convention $\top$ is the conjunction of no literal. Let $\bar{\lambda}$ denote the term resulting from negating every literal occurring in λ. For sake of readability sometimes we abuse the notation $\lambda \subseteq \alpha$ to denote that λ is a "part" of α, in the sense that all literals occurring in λ also occurs in α.

Strictly speaking, our f takes a state s as an input. However, by virtue of the following facts, given any binary classifier, we can associate it to an NDM, where f mimics it.

Fact 1. *For every Z-definite non-redundant decision model $M^{nr} = (S, f)$ there exists a unique binary classifier $g_Z : Input_Z \longrightarrow Dec$ such that:*
(i) $\forall \alpha \in Dom(g), \exists s \in S$ *such that* $s \models \alpha$*, and*
(ii) $\forall \alpha \in Dom(g), \forall s \in S$*, if* $s \models \alpha$ *then* $f(s) = g(\alpha)$.

Fact 2. *For every binary classifier $g_Z : Input_Z \longrightarrow Dec$, there exists a Z-definite non-redundant decision model $M^{nr} = (S, f)$ such that:*
(i) $\forall \alpha \in Dom(g), \exists s \in S$ *such that* $s \models \alpha$*, and*
(ii) $\forall \alpha \in Dom(g), \forall s \in S$*, if* $s \models \alpha$ *then* $f(s) = g(\alpha)$.

Moreover, there is a unique Z-definite NDM grounded on Z satisfying (i) and (ii). We note $M^{nr}_{g_Z}$ this unique model.

In this sense, every classifier for the finite set of features Z induces some Z-definite NDM and every Z-definite NDM represents a classifier for Z. Moreover, a classifier for Z corresponds to a unique Z-definite NDM grounded on Z. Thus, we may bypass g_Z and study directly NDMs for it.

Let us end up this section with an illustrating toy example.

Example 1. *Given a finite language $\mathcal{L}(Atm)$ with $Atm = \{p_1, p_2, q_1, q_2, \mathtt{t}(x), \mathtt{t}(y)\}$. Consider a pointed NDM (M^{nr}, s) where $s = \{p_2, q_1\}, f(s) = y$. Name α the conjunction $\neg p_1 \wedge p_2 \wedge q_1 \wedge \neg q_2$. Hence $(M^{nr}, s) \models \alpha \wedge \mathtt{t}(y)$.*

We interpret p_1, p_2, q_1, q_2 in the previous example as gender (male or female), postcode (inner city or suburb), employment situation (employed or unemployed) and property ownership (possess or rent), respectively; f as a classifier of loan; x, y as acceptance and rejection by the bank, respectively; s as the state of an applicant, say, Alice. So this models a scenario that Alice is applying for a loan from her bank. She is female, employed, rents an apartment in the inner city. The bank decides to reject Alice's application.

[2] Strictly speaking it is a generalization of Boolean function, for Dec is not necessary binary. But it is also not semi-Boolean, for its outputs are not all real numbers.

Now Alice is asking for explanations of the decision, e.g., 1) which of her features (necessarily) leads to the current decision, 2) what would make a difference if she were another applicant, 3) perhaps most substantially, whether the decision for her is biased. In Section 5 we will show how to use the language $\mathcal{L}(Atm)$ and its semantics to answer these questions. But, before moving into the conceptual analysis of classifiers' explanations and biases, in the next section we focus on more technical issues about axiomatics and complexity for our language.

3 Axiomatization and Complexity

In this section we provide sound and complete axiomatics for the language $\mathcal{L}(Atm)$ relative to the formal semantics defined above. The following abbreviation is given for the sake of compactness, for every finite $X, Y \subseteq Atm$:

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

Definition 6 (Logic BCL)**.** *We define* BCL *(Binary Classifier Logic) to be the extension of classical propositional logic given by the following axioms and rules of inference:*

$$\big([\emptyset]\varphi \wedge [\emptyset](\varphi \to \psi)\big) \to [\emptyset]\psi \qquad (\mathbf{K}_{[\emptyset]})$$

$$[\emptyset]\varphi \to \varphi \qquad (\mathbf{T}_{[\emptyset]})$$

$$[\emptyset]\varphi \to [\emptyset][\emptyset]\varphi \qquad (\mathbf{4}_{[\emptyset]})$$

$$\varphi \to [\emptyset]\langle\emptyset\rangle\varphi \qquad (\mathbf{B}_{[\emptyset]})$$

$$[X]\varphi \leftrightarrow \bigwedge_{Y\subseteq X} \big(\mathsf{con}_{Y,X} \to [\emptyset](\mathsf{con}_{Y,X} \to \varphi)\big) \qquad (\mathbf{Red}_{[\emptyset]})$$

$$\bigvee_{x\in Val} \mathsf{t}(x) \qquad (\mathbf{AtLeast}_{\mathsf{t}(x)})$$

$$\mathsf{t}(x) \to \neg\mathsf{t}(y) \text{ if } x \neq y \qquad (\mathbf{AtMost}_{\mathsf{t}(x)})$$

$$\frac{\varphi \to \psi \quad \varphi}{\psi} \qquad (\mathbf{MP})$$

$$\frac{\varphi}{[\emptyset]\varphi} \qquad (\mathbf{Nec}_{[\emptyset]})$$

Moreover, for $\Xi \subseteq Ax$, we define BCL_Ξ *to be the extension of* BCL *by all axioms in Ξ, where*

$$Ax = \{\boldsymbol{Def}_X, \boldsymbol{Comp}_X, \boldsymbol{NTr}_X : X \subseteq \big(Atm \setminus Dec\big) \text{ and } X \text{ is finite}\},$$

and $\boldsymbol{Def}_X, \boldsymbol{Comp}_X$ *and* $\boldsymbol{NTr}_X$ *are the following axiom schemata:*

$$\bigwedge_{Y \subseteq X} \Big(\big(\mathsf{con}_{Y,X} \wedge \mathsf{t}(x)\big) \rightarrow [\emptyset]\big(\mathsf{con}_{Y,X} \rightarrow \mathsf{t}(x)\big)\Big) \qquad (\mathbf{Def}_X)$$

$$\bigwedge_{Y \subseteq X} \langle\emptyset\rangle \mathsf{con}_{Y,X} \qquad (\mathbf{Comp}_X)$$

$$\bigvee_{\substack{Y,Y' \subseteq X, x,y \in Val: \\ Y \neq Y' \text{and } x \neq y}} \Big(\langle\emptyset\rangle\big(\mathsf{con}_{Y,X} \wedge \mathsf{t}(x)\big) \wedge \langle\emptyset\rangle\big(\mathsf{con}_{Y',X} \wedge \mathsf{t}(y)\big)\Big) \qquad (\mathbf{NTr}_X)$$

Let us define the following bijective correspondence function cp mapping semantic properties in $Prop$ to axioms in Ax:

- $cp\big(def(X)\big) = \mathbf{Def}_X$,
- $cp\big(comp(X)\big) = \mathbf{Comp}_X$,
- $cp\big(ntr(X)\big) = \mathbf{NTr}_X$.

The following theorem highlights that the logics in the BCL family are all sound and complete relative to their corresponding semantics. The proof is entirely standard and based on a canonical model argument.

Theorem 2. *Let* $P \subseteq Prop$*. Then, the logic* $\mathsf{BCL}_{\{cp(pr):pr \in P\}}$ *is sound and complete relative to the class* $\mathbf{DM}_P$*.*

The following theorem provides a complexity result for checking satisfiability of formulas in $\mathcal{L}(Atm)$ relative to any class $\mathbf{DM}_P$ with $P \subseteq \{def(X), comp(X), ntr(X)\}$, under the assumption that the set of atomic propositions is finite. Note that when $X = Atm \setminus Dec$ and $P = \{def(X), comp(X), ntr(X)\}$, then $\mathbf{DM}_P$ is the class in which all features of the input are possibly relevant for the classifier's decision, the classifier can make a decision for any possible input and the classifier is not making trivial decisions (i.e., there are at least two different inputs with different decisions). The proof of the complexity result is based on three steps. First, we have that validity relative to the class $\mathbf{DM}_P$ is equivalent to validity relative to the models in the class $\mathbf{M}$ of Definition 1 that "globally" satisfy a finite set of formulas corresponding to the axioms in $\{cp(prop) : prop \in P\} \cup \{\mathbf{AtLeast}_{\mathsf{t}(x)}, \mathbf{AtMost}_{\mathsf{t}(x)}\}$. The size of such a set of formulas is constant and does not depend on the size of the formula to be checked. Therefore, we can polynomially reduce satisfiability checking relative to class $\mathbf{DM}_P$ to satisfiability checking relative to class $\mathbf{M}$. In [10] it is shown that, if the set of atomic propositions is finite, then the latter problem is in NP. This gives NP-membership for our logic. NP-hardness follows from NP-hardness of propositional logic.

Theorem 3. *Let* $\varphi \in \mathcal{L}(Atm)$ *and* Atm *be finite. Then, for every finite* $X \subseteq \big(Atm \setminus Dec\big)$ *and* $P \subseteq \{def(X), comp(X), ntr(X)\}$*, checking satisfiability of* φ *relative to* $\mathbf{DM}_P$ *is NP-complete.*

4 Counterfactual Conditional

In this section we investigate a simple notion of counterfactual conditional for binary classifiers, inspired from Lewis' notion [15]. In Section 5, we will elucidate its connection with the notion of explanation.

We start our analysis by defining the following notion of similarity between worlds in a model relative to a finite set of features X.

Definition 7. *Let $M = (W, (\equiv_X)_{X \in AtmSet_0}, V)$ be a decision model, $w, v \in W$ and $X \subseteq \big(Atm \setminus Dec\big)$ finite. The similarity between w and v in M relative to the set of features X, noted $sim_M(w,v,X)$, is defined as follows:*

$$sim_M(w,v,X) = |\{p \in X : (M,w) \models p \text{ iff } (M,v) \models p\}|.$$

A dual notion of distance between worlds can defined from the previous notion of similarity:

$$dist_M(w,v,X) = |X| - sim_M(w,v,X).$$

This notion of distance is in accordance of [5] in knowledge revision.[3] The following definition introduces the concept of conditional. Following Lewis' view, we evaluate a conditional at world of a decision model and state that the conditional holds if all closest worlds to the actual world in which the antecedent is true satisfy the consequent of the conditional.

Definition 8. *Let $M = (W, (\equiv_X)_{X \in AtmSet_0}, V)$ be a decision model, $w \in W$ and $X \subseteq \big(Atm \setminus Dec\big)$ finite. We say that "if φ was true then ψ would be true, relative to the set of features X" at w, noted $(M,w) \models \varphi \Rightarrow_X \psi$, if and only $closest_M(w,\varphi,X) \subseteq ||\psi||_M$ where*

$$closest_M(w,\varphi,X) = \arg\max_{v \in ||\varphi||_M} sim_M(w,v,X),$$

and for every $\varphi \in \mathcal{L}(Atm)$:

$$||\varphi||_M = \{v \in W : (M,v) \models \varphi\}.$$

As the following proposition highlights the previous notion of counterfactual conditional is expressible in the language $\mathcal{L}(Atm)$.

Proposition 3. *Let $M = (W, (\equiv_X)_{X \in AtmSet_0}, V)$ be a decision model, $w \in W$ and $X \subseteq \big(Atm \setminus Dec\big)$ finite. Then,*

$(M,w) \models \varphi \Rightarrow_X \psi$ *if and only if*

$$(M,w) \models \bigwedge_{0 \leq k \leq |X|} \big(\mathsf{maxSim}(\varphi,X,k) \rightarrow \bigwedge_{Y \subseteq X : |Y| = k} [Y](\varphi \rightarrow \psi)\big),$$

with

$$\mathsf{maxSim}(\varphi,X,k) =_{def} \bigvee_{Y \subseteq X : |Y| = k} \langle Y \rangle \varphi \wedge \bigwedge_{Y \subseteq X : k < |Y|} [Y] \neg \varphi.$$

[3] There are other options besides measuring distance by cardinality, e.g., distance in sense of subset relation as [2]. We will consider them in further research.

In light of the previous proposition, we can see $\varphi \Rightarrow_X \psi$ as an abbreviation of its corresponding $\mathcal{L}(Atm)$-formula. [4]

A remarkable fact is that, though our counterfactual conditional is Lewisian, it does not satisfy the *strong centering* in Lewis' VC, which says that the actual world is the only closest world when the antecedent is already true here. To see it, consider a toy model $M = (W, (\equiv_X)_{X \in AtmSet_0}, V)$ s.t. $W = \{w, v\}$ with $V(w) = \{p, q\}$ and $V(v) = \{p\}$. We have $closest_M(w,p,\emptyset) = W$, rather than $closest_M(w,p,\emptyset) = \{w\}$.

The interesting aspect of the previous notion of counterfactual conditional is that it can be used to represent a binary classifier's approximate decision for a given instance. Let us suppose the set of decision atoms Dec includes a special symbol ? meaning that the classifier has no sufficient information enabling it to classify an instance in a precise way. More compactly, ? means that the classifier abstains from making a precise decision. In this situation, the classifier can try to make an approximate decision: it considers the closest instances to the actual instance for which it has sufficient information to make a decision and checks whether the decision is uniform among all such instances. In other words, x is the classifier's approximate classification of (or decision for) the actual instance built from the set of relevant features X, noted $\mathsf{apprDec}(x,X)$, if and only if "if a precise decision was made for the input built from the set of features X then this decision would be x". Formally:

$$\mathsf{apprDec}(x,X) =_{def} \Big(\bigvee_{y \in Val : y \neq ?} \mathsf{t}(y) \Big) \Rightarrow_X \mathsf{t}(x).$$

The following proposition provides two interesting validities.

Proposition 4. *Let $x, y \in Val \setminus \{?\}$ and $P \subseteq Prop$. Then,*

$$\models_{\mathbf{DM}_P} \mathsf{apprDec}(x,X) \rightarrow \neg \mathsf{apprDec}(y,X) \text{ if } x \neq y,$$
$$\models_{\mathbf{DM}_P} \mathsf{t}(x) \rightarrow \mathsf{apprDec}(x,X) \text{ if } def(X) \in P.$$

According to the first validity, a classifier cannot make two different approximate decisions. According to the second validity, in a X-definite model, if the classifier is able to make a precise decision for a given instance, then its approximate decision coincides with it.

It is worth noting that the following formula is not valid relative to the class $\mathbf{DM}_{\{def(X)\}}$ for an arbitrary X:

$$\bigvee_{x \in Val \setminus \{?\}} \mathsf{apprDec}(x,X).$$

This means that, notwithstanding the fact that the classifier is X-definite, it may be unable to approximately classify the actual instance.

[4] A similar approach of ceteris paribus is [9]. They also refine Lewis' semantics for counterfactual by selecting the closest worlds according to not only the actual world and antecedent, but also a set of formulas where they note as Γ. The main technical difference is that they allow any counterfacual-free formula as a member of Γ, while in our setting X only contains atomic formulas.

5 Explanations and Biases

In this section we formalize some existing notions of explanation based on prime implicant in our setting, and deepen the current study. We focus our analysis in this section on the Z-representative class $\mathbf{NDM}_{\{def(Z),comp(Z),ntr(Z)\}}$ for an arbitrary finite $X \subseteq \big(Atm \setminus Dec\big)$. To simplify exposition, we abbreviate $\mathbf{NDM}_{\{def(Z),comp(Z),ntr(Z)\}}$ as $\mathbf{NDM}_{rep(Z)}$.

5.1 Prime Implicant in NDM

Now we are in position to formalize *prime implicant*, which plays a fundamental role in Boolean functions and in classifier explanations. We recall that a prime implicant is a *term λ from the finite set of atomic propositions Z*, as defined in Section 2.3.

Definition 9 (Prime Implicant)**.** *Let $M^{nr} = (S, f) \in \mathbf{NDM}_{rep(Z)}$ and $s \in S$. We say that property λ is an* implicant *of $\mathtt{t}(x)$ at (M^{nr}, s), if $(M^{nr}, s') \models \lambda \rightarrow \mathtt{t}(x)$ for all $s' \in S$.*

Moreover, we say that λ is a prime implicant *of $\mathtt{t}(x)$ at (M^{nr}, s), noted by $(M^{nr}, s) \models \mathsf{PImp}(\lambda, x)$, if λ is an implicant of $\mathtt{t}(x)$ at (M^{nr}, s) and there is no λ' s.t. $\lambda' \subseteq \lambda$, and λ' is an implicant of $\mathtt{t}(x)$ at (M^{nr}, s).*

Note that the notions of implicant and prime implicant are global properties of a model.

As the following proposition highlights, the notion of prime implicant is expressible in the language $\mathcal{L}(Atm)$. Thus, like counterfactual conditional, we will conceive $\mathsf{PImp}(\lambda, x)$ as an abbreviation of its corresponding $\mathcal{L}(Atm)$-formula. We will do the same for the notions of abductive explanation, contrastive explanation and bias which will also be seen as abbreviations of corresponding $\mathcal{L}(Atm)$-formulas.

Proposition 5. *Let $M^{nr} = (S, f) \in \mathbf{NDM}_{rep(Z)}$ and $s \in S$. We have $(M^{nr}, s) \models \mathsf{PImp}(\lambda, x)$ iff*

$$(M^{nr}, s) \models [\emptyset]\big(\lambda \rightarrow (\mathtt{t}(x) \wedge \bigwedge_{p \in Atm(\lambda)} \langle Atm(\lambda) \setminus \{p\}\rangle \neg \mathtt{t}(x))\big).$$

Theoretically, knowing all prime implicants makes the classifier transparent and understandable to humans. However, prime implicants are hard to enumerate. Therefore, many researchers focus on prime implicants of a given input. In such a way the explanation stays "local".

5.2 Abductive Explanation (AXp)

In the following parts we introduce some local explanations and their roles in detecting decision biases. Let us enrich the Alice's Example 1 for further illustration.

Example 2. *Suppose $Z = \{p_1, p_2, q_1, q_2\}$. Moreover, let $M^{nr} = (S, f) \in \mathbf{NDM}_{rep(Z)}$ such that $M^{nr} \models \mathtt{t}(x) \leftrightarrow ((q_1 \wedge q_2) \vee (p_1 \wedge q_1))$. So, the bank loan classifier only considers two cases as accepted: employed and owning property, or male and employed.*

Postcode plays no role. But f is a black box to Alice, and she asks for explanations regarding her own state.

Let us now formalize the concept of abductive explanation.

Definition 10 (Abductive Explanation (AXp))**.** *Let $M^{nr} = (S, f) \in \mathbf{NDM}_{rep(Z)}$ and $s \in S$. We say that λ abductively explains the decision x for the input α at (M^{nr}, s), noted $(M^{nr}, s) \models \mathsf{AXp}(\lambda, \alpha, x)$, if $(M^{nr}, s) \models \alpha$, $\lambda \subseteq \alpha$ and λ is a prime implicant of $\mathsf{t}(x)$ at (M^{nr}, s).*

The notion of abductive explanation is also expressible in the language $\mathcal{L}(Atm)$.

Proposition 6. *Let $M^{nr} = (S, f) \in \mathbf{NDM}_{rep(Z)}$ and $s \in S$. We have $(M^{nr}, s) \models \mathsf{AXp}(\lambda, \alpha, x)$ iff*

$$(M^{nr}, s) \models \alpha \wedge \lambda \wedge \mathsf{PImp}(\lambda, x).$$

In Alice's case, since $\neg p_1 \wedge \neg q_2 \subseteq \alpha$ (recall $\alpha = \neg p_1 \wedge p_2 \wedge q_1 \wedge \neg q_2$), we have $(M^{nr}, s) \models \mathsf{AXp}(\neg p_1 \wedge \neg q_2, \alpha, \mathsf{t}(y))$, namely that Alice's being female and not owning a property abductively explains her rejection.

Many names besides AXp are found in literature, e.g., PI-explanation and sufficient reason. [6] showed semantically that every decision (for an instance) has a sufficient reason. The interesting aspect of our approach is that we can prove them syntactically. Recall our M^{nr} belongs to the class $\mathbf{NDM}_{rep(Z)}$, $Atm(\alpha) = Z$ and α equals $\mathsf{con}_{Y,Z}$ for some $Y \subseteq Z$. In order to prove Darwiche & Hirth's result, we first prove the following validity:

$$\models_{\mathbf{NDM}_{rep(Z)}} (\alpha \wedge \mathsf{t}(x)) \rightarrow \bigvee_{\lambda \subseteq \alpha} \mathsf{PImp}(\lambda, x).$$

By the completeness Theorem 2, it is enough to show that $(\alpha \wedge \mathsf{t}(x)) \rightarrow \bigvee_{\lambda \subseteq \alpha} \mathsf{PImp}(\lambda, x)$ is a theorem of the logic $\mathsf{BCL}_{\{\mathbf{Def}_Z, \mathbf{Comp}_Z, \mathbf{NTr}_Z\}}$. We use the standard symbol $\vdash$ for $\mathsf{BCL}_{\{\mathbf{Def}(Z), \mathbf{Comp}(Z), \mathbf{NTr}(Z)\}}$-theoremhood. Assume the opposite towards a contradiction, we range over $\lambda \subseteq \alpha$ by the number of its variables, with $\lambda^0 = \top$ and $\lambda^n = \alpha$:

1. $\vdash (\alpha \wedge \mathsf{t}(x)) \wedge \bigwedge_{\lambda \subseteq \alpha} \langle \emptyset \rangle (\lambda \wedge (\neg \mathsf{t}(x) \vee \bigvee_{p \in Atm(\lambda)} [Atm(\lambda) \setminus \{p\}] \mathsf{t}(x)))$ by assumption
2. When $\lambda = \top$, since $Atm(\top) =$, so $\bigvee_{p \in \emptyset} [\emptyset \setminus p] \mathsf{t}(x)$ is empty, and it has to be $\vdash \langle \emptyset \rangle \big(\top \wedge \neg \mathsf{t}(x)\big)$
3. Take any $\lambda^1 \subseteq \alpha$ s.t. $|Atm(\lambda^1)| = 1$, then $\vdash \langle \emptyset \rangle (\lambda^1 \wedge (\neg \mathsf{t}(x) \vee \bigvee_{p \in \lambda^1} [\lambda^1 \setminus \{p\}] \mathsf{t}(x)))$, notice $\bigvee_{p \in \lambda^1} [\lambda^1 \setminus \{p\}] \mathsf{t}(x)) = [\emptyset] \mathsf{t}(x)$; but if $\vdash \langle \emptyset \rangle [\emptyset] \mathsf{t}(x)$, then it contradicts 2. together with $\mathbf{B}_{[\emptyset]}, \mathbf{4}_{[\emptyset]}$, so it has to be $\vdash \langle \emptyset \rangle (\lambda^1 \wedge \neg \mathsf{t}(x))$
4. Take any $\lambda^2 \subseteq \alpha$ s.t. $|Atm(\lambda^2)| = 2$. By 3. we have $\vdash \neg \bigvee_{p \in Atm(\lambda^2)} [Atm(\lambda^2) \setminus \{p\}] \mathsf{t}(x)$, so analogously it has to be $\vdash \langle \emptyset \rangle (\lambda^2 \wedge \neg \mathsf{t}(x))$
5. Similar progress for all $\lambda^3, \lambda^4, \ldots, \lambda^{n-1}, \alpha$, hence $\vdash \langle \emptyset \rangle (\alpha \wedge \neg \mathsf{t}(x))$

6. But we have $\vdash (\alpha \wedge \mathsf{t}(x)) \rightarrow [\emptyset](\alpha \rightarrow \mathsf{t}(x))$ by $\mathbf{Def}_Z$, a contradiction.

Notice that besides the logic BCL, only $\mathbf{Def}_Z$ is used in the proof. Darwiche & Hirth's result is then a corollary of the previous validity:

$$\models_{\mathbf{NDM}_{rep(Z)}} (\alpha \wedge \mathsf{t}(x)) \rightarrow \bigvee_{\lambda \subseteq \alpha} \mathsf{AXp}(\lambda, \alpha, x).$$

5.3 Contrastive Explanation (CXp)

AXp is a minimal part of a given instance α verifying the current decision. A natural counterpart of AXp, contrastive explanation (CXp), is a minimal part of α which falsifies the current decision. In our setting it is defined as follow.

Definition 11 (Contrastive Explanation (CXp))**.** *Let $M^{nr} = (S, f) \in \mathbf{NDM}_{rep(Z)}$ and $s \in S$. We say that λ contrastively explains the decision x for an instance α at (M^{nr}, s), noted $(M^{nr}, s) \models \mathsf{CXp}(\lambda, \alpha, x)$, if $(M^{nr}, s) \models \alpha \wedge \mathsf{t}(x)$, $\exists t \in S$ such that $t \equiv_{Z \setminus Atm(\lambda)} s$ and $f(t) \neq x$, and for any other λ', if $\exists t' \in S$ such that $t' \equiv_{Z \setminus Atm(\lambda')} s$ and $f(t') \neq x$, then $\lambda' \not\subseteq \lambda$.*

Like prime implicant and abductive explanation, the language $\mathcal{L}(Atm)$ is expressive enough to capture contrastive explanation.

Proposition 7. *Let $M^{nr} = (S, f) \in \mathbf{NDM}_{rep(Z)}$ and $s \in S$. $(M^{nr}, s) \models \mathsf{CXp}(\lambda, \alpha, x)$ iff*

$$(M^{nr}, s) \models \alpha \wedge \lambda \wedge \mathsf{t}(x) \wedge \langle Atm(\alpha) \setminus Atm(\lambda) \rangle \neg \mathsf{t}(x) \wedge \bigwedge_{p \in Atm(\lambda)} [(Atm(\alpha) \setminus Atm(\lambda)) \cup \{p\}] \mathsf{t}(x).$$

We define CXp by minimizing the change from the current input, while in defining counterfactual conditional we maximize the similarity. It makes one question whether these are two paths towards the same end. Actually in XAI, many researchers consider contrastive explanations and counterfactual explanations either closely related [26], or even interchangeable [20]. Our framework agrees that CXp has a counterfactual nature in light of the next proposition.

Proposition 8. *We have the following validity:*

$$\models_{\mathbf{NDM}_{rep(Z)}} (\alpha \wedge \lambda \wedge \mathsf{t}(x)) \rightarrow \big(\mathsf{CXp}(\lambda, \alpha, x) \leftrightarrow (\bar{\lambda} \Rightarrow_Z \neg \mathsf{t}(x))\big).$$

Back to Example 2, we have $(M^{nr}, s) \models \mathsf{CXp}(\neg p_1, \alpha, y) \wedge \mathsf{CXp}(\neg q_2, \alpha, y)$, which means that both Alice's being female and not owning property contrastively explains her rejection. Moreover, since gender is hard to change, owing a property is the (relatively) *actionable* explanation for Alice,[5] if she wants to follow the rule of f. But surely Alice has another option, i.e. alleging the classifier as biased. As we will see in the next subsection, an application of CXp is to detect decision bias.

[5]For the significance of actionablility in XAI, see e.g. [20].

5.4 Decision Bias

A primary goal of XAI is to detect and avoid biases. Bias is understood as making decision with respect to some protected features, e.g. race, gender and age. For this goal we split set of atoms of Z into two parts, protected feature (PF) and non-protected feature (NF), i.e. $Z = \mathsf{PF} \cup \mathsf{NF}$.

A widely accepted notion of decision bias can be expressed in our setting as follows. Intuitively, the rejection for Alice is biased, if there is a Bob, who only differs from Alice on some protected feature, but gets accepted.

Definition 12 (Decision Bias)**.** *Let $M^{nr} = (S, f) \in \mathbf{NDM}_{rep(Z)}$ and $s \in S$. We say that decision x for α is biased at (M^{nr}, s), noted $(M^{nr}, s) \models \mathsf{Bias}(\alpha, x)$, if $(M^{nr}, s) \models \alpha \wedge \mathsf{t}(x)$, and $\exists X \subseteq \mathsf{PF}$, $\exists t \in S$ such that $s \equiv_{Z \setminus X} t$, and $(M^{nr}, t) \models \neg\mathsf{t}(x)$.*

The notion of bias is also expressible in our language.

Proposition 9. *Let $M^{nr} = (S, f) \in \mathbf{NDM}_{rep(Z)}$ and $s \in S$. Then, $(M^{nr}, s) \models \mathsf{Bias}(\alpha, x)$ iff*

$$(M^{nr}, s) \models \alpha \wedge \mathsf{t}(x) \wedge \bigvee_{X \subseteq \mathsf{PF}} \langle Atm(\alpha) \setminus X \rangle \neg\mathsf{t}(x).$$

Let us continue with the Alice example.

Example 3. *The language of the model M^{nr} divides Z into two disjoint parts: $\mathsf{PF} = \{p_1, p_2\}$ and $\mathsf{NF} = \{q_1, q_2\}$. We then have $(M^{nr}, s) \models \mathsf{Bias}(\alpha, y)$. The decision for Alice is biased. Indeed, $\{p_1\}$ is the subset of the set of protected features which is responsible for the classifying Alice as y.*

As we see from the Alice case, some CXp can detect decision bias. The following result makes the statement precise.

Proposition 10. *We have the following validity:*

$$\models_{\mathbf{NDM}_{rep(Z)}} \mathsf{Bias}(\alpha, x) \leftrightarrow \bigvee_{Atm(\lambda) \subseteq \mathsf{PF}} \mathsf{CXp}(\lambda, \alpha, x).$$

6 Extensions

In this section, we briefly discuss two interesting extensions of our logical framework and analysis of binary classifiers. Their full development is left for future work.

6.1 Dynamic Extension

The first extension we want to discuss consists in adding to the language $\mathcal{L}(Atm)$ dynamic operators of the form $[x := \varphi]$ with $x \in Dec$, where $x := \varphi$ is a kind of assignment in the sense of [21, 25] and the formula $[x := \varphi]\psi$ has to be read "ψ holds after every decision is set to x in context φ". The resulting language, noted $\mathcal{L}^{dyn}(Atm)$, is defined by the following grammar:

$$\varphi \quad ::= \quad p \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid [X]\varphi \mid [x{:=}\varphi]\psi$$

where p ranges over Atm, X ranges over $AtmSet_0$ and x ranges over Dec. The interpretation of formula $[x := \varphi]\psi$ relative to a pointed decision model (M, w) with $M = (W, (\equiv_X)_{X \in AtmSet_0}, V)$ goes as follows:

$$(M, w) \models [x{:=}\varphi]\psi \iff (M^{x:=\varphi}, w) \models \psi,$$

where $M^{x:=\varphi} = (W, (\equiv_X)_{X \in AtmSet_0}, V^{x:=\varphi})$ is the updated decision model where, for every $w \in W$:

$$V^{x:=\varphi}(w) = \begin{cases} \big(V(w) \setminus \bigcup_{y \in Val}\{\mathtt{t}(y)\}\big) \cup \{\mathtt{t}(x)\} \\ \text{if } (M, w) \models \varphi, \\ V(w) \text{ otherwise.} \end{cases}$$

Intuitively, the operation $x{:=}\varphi$ consists in globally classifying all instances satisfying φ with value x. This update operation is well-defined relative to the X-completeness property as well as to the X-definite property, under the assumption that $Atm(\varphi) \subseteq X$. Indeed, as the following proposition indicates such properties are preserved under update.

Proposition 11. *Let M be a decision model and $X \subseteq \big(Atm \setminus Dec\big)$ finite. Then, if M is X-complete then $M^{x:=\varphi}$ is also X-complete. Moreover, if $Atm(\varphi) \subseteq X$ and M is X-definite, then $M^{x:=\varphi}$ is also X-definite.*

Dynamic operators $[x := \varphi]$ are useful for modeling a classifier's revision. Specifically, new knowledge can be injected into the classifier thereby leading to a change in its classification. For example, the classifier could learn that if an object is a furniture, has one or more legs and has a flat top, then it is a table. This is captured by the following assignment:

$$\text{“}\mathtt{table}\text{”} := \mathit{objIsFurniture} \wedge \mathit{objHasLegs} \wedge \mathit{objHasFlatTop}.$$

An application of dynamic change is to model the training process of a classifier, together with counterfactual conditionals with ? in Section 4. Suppose at the beginning we have a DM M grounded on Z which is totally ignorant, i.e. $\forall w \in W, V_{Dec}(w) = \{\mathtt{t}(?)\}$. It associates with a classifier g_Z (because Section 2.3 shows any g_Z corresponds to some NDM, and any NDM has some equivalent DM). We then prepare to train the classifier. The training set consists of dynamic operators in the form $[x := \alpha]$ with $x \neq ?$ and $\alpha \in \mathit{Input}_Z$. We train the classifier by revising it with all members of the training set one by one. [6] In other words, we re-valuate decisions for some worlds. With a bit abuse of notation, let M^{train} denote the model resulting from the series of revisions. We finish training by inducing the final model $M^{\dagger}$ from M^{train}, where $V^{\dagger}_{Dec}(w) = \mathtt{t}(x)$, if $(M^{train}, w) \models \mathsf{apprDec}(x, Z)$, others being equal with M^{train}.

The logic $\mathsf{BCL}{-}\mathsf{DC}$ (BCL with Decision Change) extends the logic BCL by the dynamic operators $[x{:=}\varphi]$. It is defined as follows.

[6] The order does not matter here, for we assume for any $\alpha \in \mathit{Input}_Z$, there is at most one $x \in Dec \setminus \{?\}$, s.t. $[x := \alpha]$ is in the training set.

Definition 13 (Logic BCL−DC)**.** *We define* BCL−DC *(*BCL *with Decision Change) to be the extension of* BCL *of Definition 6 generated by the following reduction axioms for the dynamic operators* $[x{:=}\varphi]$*:*

$$\begin{aligned}
[x{:=}\varphi]\mathsf{t}(x) &\leftrightarrow \big(\varphi \vee \mathsf{t}(x)\big)\\
[x{:=}\varphi]\mathsf{t}(y) &\leftrightarrow \big(\neg\varphi \wedge \mathsf{t}(y)\big) \textit{ if } x \neq y\\
[x{:=}\varphi]p &\leftrightarrow p \textit{ if } p \notin \mathit{Dec}\\
[x{:=}\varphi]\neg\psi &\leftrightarrow \neg[x{:=}\varphi]\psi\\
[x{:=}\varphi](\psi_1 \wedge \psi_2) &\leftrightarrow \big([x{:=}\varphi]\psi_1 \wedge [x{:=}\varphi]\psi_2\big)\\
[x{:=}\varphi][X]\psi &\leftrightarrow [X][x{:=}\varphi]\psi
\end{aligned}$$

and the following rule of inference:

$$\frac{\varphi_1 \leftrightarrow \varphi_2}{\psi \leftrightarrow \psi[\varphi_1/\varphi_2]} \qquad \textbf{(RE)}$$

It is routine exercise to verify that the equivalences in Definition 13 are valid for the class $\mathbf{DM}$ and that the rule of replacement of equivalents (**RE**) preserves validity. The completeness of BCL−DC for this class of models follows from Theorem 2, in view of the fact that the reduction axioms and the rule of replacement of proved equivalents can be used to find, for any $\mathcal{L}^{dyn}$-formula, a provably equivalent $\mathcal{L}$-formula.

Theorem 4. *The logic* BCL−DC *is sound and complete relative to the class* $\mathbf{DM}$*.*

The following complexity result is a consequence of Theorem 3 and the fact that via the reduction axioms in Definition 13 we can find a polynomial reduction of satisfiability checking for formulas in $\mathcal{L}^{dyn}$ to satisfiability checking for formulas in $\mathcal{L}$.

Theorem 5. *Let $\varphi \in \mathcal{L}^{dyn}$, Atm be finite. Then, for every finite $X \subseteq \big(Atm \setminus Dec\big)$ and $P \subseteq \{def(Atm), comp(X)\}$, checking satisfiability of φ relative to $\mathbf{DM}_P$ is NP-complete.*

We recall that, by Proposition 11, properties $def(Atm)$ and $comp(X)$ are preserved under the operation $x{:=}\varphi$.

6.2 Epistemic Extension

In the second extension we consider, a classifier is conceived as an agent which has to classify what it perceives. The agent could have uncertainty about the actual instance to be classified since it cannot see all its features.

In order to represent the agent's epistemic state and uncertainty, we slightly redefine the set of atomic formulas Atm. We note Atm_0 the set of atomic propositions and assume that $Dec \subseteq Atm_0$, where Dec is the set of decision atoms defined in Section 2.2. Then, we define:

$$Atm = Atm_0 \cup \{\mathsf{s}(p) : p \in Atm_0\},$$

where $\mathsf{s}(p)$ is a 'observability' (or 'visibility') atom in the sense of [4, 23, 12, 22] which has to be read "the agent can see the truth value of p". For notational convenience, we note $ObsAtm = \{\mathsf{s}(p) : p \in Atm_0\}$.

The language for our epistemic extension is noted $\mathcal{L}^{epi}(Atm, Agt)$ and defined by the following grammar:

$$\varphi \quad ::= \quad p \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid [X]\varphi \mid \mathsf{K}\varphi,$$

where p ranges over Atm and X ranges over $AtmSet_0$.

The epistemic operator K is used to represent what agent i knows in the light of what it sees. In order to interpret this new modality, we have to enrich decision models with an epistemic component.

Definition 14 (Epistemic decision model)**.** *An epistemic decision model (EDM) is a tuple $M = (W, (\equiv_X)_{X \in AtmSet_0}, \sim, V)$ where $M = (W, (\equiv_X)_{X \in AtmSet_0}, V)$ is a decision model and $\sim$ is a binary relation on W such that, for all w, v:*

$$w \sim v \text{ if and only if } (i)\ Obs(w) = Obs(v) \text{ and}$$
$$(ii)\ V_{ObsAtm}(w) = V_{ObsAtm}(v),$$

where, for every $w \in W$, $Obs(w) = \{p \in Atm_0 : \mathsf{s}(p) \in V(w)\}$ is the set of atomic propositions that are visible to the agent at w.

The class of EDMs is noted $\mathbf{EDM}$. Moreover, for each $P \subseteq Prop$, $\mathbf{EDM}_P$ denotes the class of EDMs satisfying each property in P, as defined in Section 2.2.

According to Definition 14, the agent cannot distinguish between w and v if and only if (i) the truth values of the visible variables are the same at w and v, and (ii) what the agent can see is the same at w and v. The way the epistemic accessibility relation $\sim$ is defined guarantees that it is an equivalence relation.

Proposition 12. *Let $M = (W, (\equiv_X)_{X \in AtmSet_0}, \sim, V)$ be a EDM. Then, the relation $\sim$ is reflexive, transitive and symmetric.*

The interpretation for formulas in $\mathcal{L}^{epi}(Atm)$ extends the interpretation for formulas in $\mathcal{L}(Atm)$ given in Definition 4 by the following condition for the epistemic operator:

$$(M, w) \models \mathsf{K}\varphi \iff \forall v \in W, \text{ if } w \sim v \text{ then } (M, v) \models \varphi.$$

The following proposition offers three interesting validities of our epistemic extension.

Proposition 13. *Let $P \subseteq Prop$. Then,*

$$\models_{\mathbf{EDM}_P} \mathsf{s}(p) \to \big((p \to \mathsf{K}p) \wedge (\neg p \to \mathsf{K}\neg p)\big),$$
$$\models_{\mathbf{EDM}_P} \mathsf{s}(p) \leftrightarrow \mathsf{K}\,\mathsf{s}(p),$$
$$\models_{\mathbf{EDM}_P} \Big(\bigwedge_{p \in X} \mathsf{s}(p) \wedge \mathsf{t}(x)\Big) \to \mathsf{K}\,\mathsf{t}(x) \text{ if } def(X) \in P.$$

According to the first validity, the agent knows the truth value of each variable it can see. According to the second validity, the agent knows what it can see. Finally, according to the third validity, in a X-definite model in which the agent can see the truth value of each variable in X, then the agent has no uncertainty about the classification of the actual instance (since it can see all features that are relevant for the classification).

As the following theorem indicates, the complexity result of Section 3 generalizes to the epistemic extension. It is based on (i) the fact that for every formula in $\mathcal{L}^{epi}(Atm)$ we can find an equivalent formula in $\mathcal{L}(Atm)$ with no epistemic operators, (ii) the adaptation of the polynomial satisfiability preserving translation from $\mathcal{L}(Atm)$ to the modal logic S5 given in [10].

Theorem 6. *Let $\varphi \in \mathcal{L}^{epi}(Atm)$ and Atm be finite. Then, for every finite $X \subseteq \big(Atm \setminus Dec\big)$ and $P \subseteq \{def(X), comp(X), ntr(X)\}$, checking satisfiability of φ relative to $\mathbf{EDM}_P$ is NP-complete.*

7 Conclusion

We have introduced a modal language and a formal semantics that allow us to capture the *ceteris paribus* nature of binary classifiers. We have formalized in the language a variety of notions which are relevant for understanding a classifier's behavior including counterfactual conditional, abductive and contrastive explanation, bias. We have provided two extensions that support reasoning about classifier change and a classifier's uncertainty about the actual instance to be classified. We have also offered axiomatization and complexity results for our logical setting.

We believe that the complexity results given in Theorems 3 and 5 are exploitable in practice. We have shown that, under the assumption that the set of atomic propositions Atm is finite, satisfiability checking in the basic *ceteris paribus* setting and in its epistemic extension are NP-complete. Indeed, both problems are polynomially reducible to satisfiability checking in the modal logic S5 for which a polynomial satisfiability preserving translation into propositional logic exists [3]. This opens up the possibility of using SAT solvers for automated verification and generation of explanation and bias in binary classifiers. We plan to focus on this topic in future research.

Another direction of future research is the generalization of the epistemic extension given in Section 6.2 to the multi-agent case. The idea is to conceive classifiers as agents and to be able to represent both the agents' uncertainty about the instance to be classified and their knowledge and uncertainty about other agents' knowledge and uncertainty (i.e., higher-order knowledge and uncertainty).

Finally, we plan to investigate more in depth classifier dynamics we briefly discussed in Section 6.1. The idea is to see them as learning dynamics. Based on this idea, we plan to study the problem of finding a sequence of update operations guaranteeing that the classifier will be able to make approximate decisions for a given set of instances.

References

[1] Or Biran and Courtenay Cotton. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, volume 8(1), pages 8–13, 2017.

[2] Alexander Borgida. Language features for flexible handling of exceptions in information systems. *ACM Transactions on Database Systems (TODS)*, 10(4):565–603, 1985.

[3] Thomas Caridroit, Jean-Marie Lagniez, Daniel Le Berre, Tiago de Lima, and Valentin Montmirail. A sat-based approach for solving the modal logic s5-satisfiability problem. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, pages 3864–3870. AAAI Press, 2017.

[4] Tristan Charrier, Andreas Herzig, Emiliano Lorini, Faustine Maffre, and François Schwarzentruber. Building epistemic logic from observations and public announcements. In *Proceedings of the Fifteenth International Conference on Principles of Knowledge Representation and Reasoning (KR 2016)*, pages 268–277. AAAI Press, 2016.

[5] Mukesh Dalal. Investigations into a theory of knowledge base revision: preliminary report. In *Proceedings of the Seventh National Conference on Artificial Intelligence*, volume 2, pages 475–479. Citeseer, 1988.

[6] Adnan Darwiche and Auguste Hirth. On the reasons behind decisions. In *ECAI 2020 - 24th European Conference on Artificial Intelligence*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 712–720. IOS Press, 2020.

[7] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in neural information processing systems*, pages 592–603, 2018.

[8] Ronald Fagin, Yoram Moses, Joseph Y Halpern, and Moshe Y Vardi. *Reasoning about Knowledge*. MIT Press, 1995.

[9] Patrick Girard and Marcus Anthony Triplett. Ceteris paribus logic in counterfactual reasoning. In *TARK 2015*, pages 176–193, 2016.

[10] Davide Grossi, Emiliano Lorini, and François Schwarzentruber. The ceteris paribus structure of logics of game forms. *Journal of Artificial Intelligence Research*, 53:91–126, 2015.

[11] Carl G Hempel and Paul Oppenheim. Studies in the logic of explanation. *Philosophy of science*, 15(2):135–175, 1948.

[12] Andreas Herzig, Emiliano Lorini, and Faustine Maffre. A poor man's epistemic logic based on propositional assignment and higher-order observation. In *Proceedings of the 5th International Workshop on Logic, Rationality, and Interaction*,

volume 9394 of *Lecture Notes in Computer Science*, pages 156–168. Springer, 2015.

[13] Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. Abduction-based explanations for machine learning models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1511–1519, 2019.

[14] Boris Kment. Counterfactuals and explanation. *Mind*, 115(458):261–310, 2006.

[15] David. Lewis. *Counterfactuals*. Harvard University Press, 1973.

[16] David Martens and Foster Provost. Explaining data-driven document classifications. *Mis Quarterly*, 38(1):73–100, 2014.

[17] Brent Mittelstadt, Chris Russell, and Sandra Wachter. Explaining explanations in ai. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 279–288, 2019.

[18] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617, 2020.

[19] Weijia Shi, Andy Shih, Adnan Darwiche, and Arthur Choi. On tractable representations of binary neural networks. *arXiv preprint arXiv:2004.02082*, 2020.

[20] Kacper Sokol and Peter A Flach. Counterfactual explanations of machine learning predictions: opportunities and challenges for ai safety. In *SafeAI@ AAAI*, 2019.

[21] Johan Van Benthem, Jan Van Eijck, and Barteld Kooi. Logics of communication and change. *Information and Computation*, 204(11):1620–1662, 2006.

[22] Wiebe van der Hoek, Petar Iliev, and Michael J Wooldridge. A logic of revelation and concealment. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems, (AAMAS 2012)*, pages 1115–1122. IFAAMAS, 2012.

[23] Wiebe Van Der Hoek, Nicolas Troquard, and Michael J Wooldridge. Knowledge and control. In *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021)*, pages 719–726. IFAAMAS, 2011.

[24] Hans Van Ditmarsch, Wiebe van Der Hoek, and Barteld Kooi. *Dynamic Epistemic Logic*, volume 337 of *Synthese Library*. Springer, 2007.

[25] Hans P van Ditmarsch, Wiebe van der Hoek, and Barteld P Kooi. Dynamic epistemic logic with assignment. In *Proceedings of the 4th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2005)*, pages 141–148. ACM, 2005.

[26] Sahil Verma, John Dickerson, and Keegan Hines. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596*, 2020.

A Tecnical annex

This technical annex contains a selection of proofs of the results given in the paper.

A.1 Proof of Proposition 1

Proof. The first two are obvious and we just sketch them.

For 1., notice that since $X \subseteq X'$, for any $w, v \in W, V_{X'}(w) = V_{X'}(v)$ implies $V_X(w) = V_X(v)$. Then apply X-definiteness of M we obtain the result.

For 2., by X-completeness and transitivity of $\subseteq$, we have $\forall X'' \subseteq X', X'' \subseteq X$. Then apply X-completeness we obtain the result.

For 3., suppose $M = (W, (\equiv_Y)_{Y \in AtmSet_0}, V)$ which is X-definite. We build an $M' = (W', (\equiv_Y)_{Y \in AtmSet_0}, V')$ s.t. $W \subseteq W', V \subseteq V'$, and $\forall X' \subseteq X, \exists w' \in W', V_X(w') = X'$. Additionally, for any $w' \in W' \setminus W$, if $\exists w \in W, V_X(w') = V_X(w)$ then $V_{Dec}(w') = V_{Dec}(w)$, otherwise simply let $V_{Dec}(w') = \mathtt{t}(x)$ for some $\mathtt{t}(x)$.

For 4., given a decision model $M = (W, (\equiv_Y)_{Y \in AtmSet_0}, V)$, $M \in \mathbf{DM}_{\{def(X), ntr(X), ntr(X'), comp(X \cup X')\}}$ and $X \cap X' = \emptyset$, we shall prove $M \notin \mathbf{DM}_{\{def(X')\}}$. Suppose $M \in \mathbf{DM}_{\{def(X')\}}$ towards a contradiction. Since $M \in \mathbf{DM}_{\{ntr(X)\}}$, we shall have $\exists w, v \in M$ in M, s.t. $V_X(w) \neq V_X(v), V_{Dec}(w) \neq V_{Dec}(v)$. We claim that it cannot be $V_{X'}(w) = V_{X'}(v)$, otherwise $M \notin \mathbf{DM}_{\{def(X')\}}$ and we have the contradiction . However, it can also not be $V_{X'}(w) \neq V_{X'}(v)$. Since $M \in \mathbf{DM}_{\{comp(X \cup X')\}}$ and $X \cap X' = \emptyset$, we can find a $w' \in W$, s.t. $V_X(w') = V_X(w), V_{X'}(w') = V_{X'}(v)$. But then by X- and X'-definiteness we have $V_{Dec}(w) = V_{Dec}(w') = V_{Dec}(v)$, contradicts the X-non triviality. □

A.2 Proof of Proposition 2

Proof. For any $X \in AtmSet_0$, we build a decision model $M = (W, (\equiv_Y)_{Y \in AtmSet_0}, V)$ as follows:

- W is a set of worlds
- every $\equiv_Y$ is a binary relation on W that satisfies **C1 - C3**
- $V : W \to 2^{Atm}$ is a valuation function,
- particularly, there exists $x, y \in Dec, x \neq y$
  - $\forall w \in W, V_X(w) = X$ implies $V_{Dec}(w) = \mathtt{t}(x)$
  - $\forall w \in W, V_X(w) \neq X$ implies $V_{Dec}(w) = \mathtt{t}(y)$
  - $\forall X' \subseteq X, \exists w \in W$ s.t. $V_X(w) = X'$

It is easy to see that M is indeed a decision model, and X-complete. To see X-definiteness, since $\forall w, v \in W, V_X(w) = V_X(v)$, then either $V_{Dec}(w) = V_{Dec}(v) = \mathtt{t}(x)$, if $V_X(w) = V_X(v) = X$; or $V_{Dec}(w) = V_{Dec}(v) = \mathtt{t}(y)$, if $V_X(w) = V_X(v) \neq X$. X-completeness guarantees that there always are two w, v s.t. $V_X(w) = X$ and $V_X(v) \neq X$, therefore X-non triviality is satisfied. Hence we proved that $\mathbf{DM}_P$ is non-empty.

Since there are infinite many Xs, s.t. $X \in AtmSet_0$ when Atm is infinite, $\mathbf{DM}_P$ is also infinite. □

A.3 Proof of Theorem 1

Proof. We prove the X-representative case, namely $P = \{def(X), comp(X), ntr(X)\}$ for a given X.

$\Rightarrow$ Suppose a decision model $M = (W, (\equiv_Y)_{Y \in AtmSet_0}, V) \in \mathbf{DM}_P$ and $(M, w) \models \varphi$, we construct a non-redundant decision model $M^{nr} = (S, f)$ as follow

- $S = \{t : t = V_{Atm \setminus Dec}(v) \text{ for some } v \in W\}$
- $f : S \to Val$ is a decision function s.t. $f(t) = y$ if $t = V(v)$ and $\mathtt{t}(y) \in V(v)$ for some $v \in W$

It is obviously a non-redundant decision model. To show $def(X)$, $\forall t, t' \in S$, notice if $t \cap X = t' \cap X$, we have $V(v) \cap X = V(v') \cap X$ for some $V(v) = t, V(v') = t'$. Thus by **C4**, $V_{Dec}(v) = f(t) = f(t') = V_{Dec}(v')$. To show $comp(X)$, notice by **C5** we have $\forall X' \subseteq X, \exists v \in W, V_X(v) = X'$ iff $\exists t \in S, V(v) = t, v \cap X = X'$. For $ntr(X)$, by **C6** $\exists v, v' V_X(v) \neq V_X(v')$ and $V_{Dec}(v) \neq V_{Dec}(v')$, then let $V(v) = t, V(t') = t'$ we shall have $t \cap X \neq t' \cap X$ and $f(t) \neq f(t')$.

Then let a state $s \in S$ s.t. $s = V_{Atm \setminus Dec}(w)$. We prove by induction that $(M, w) \models \varphi$ then $(M^{nr}, s) \models \varphi$. For inductive bases, if φ is some $p \in Atm \setminus Dec$, then $(M, w) \models p \iff p \in V(w) \iff p \in s \iff (M^{nr}, s) \models p$. If φ is $\mathtt{t}(x) \in Dec$, then $(M, w) \models \mathtt{t}(x) \iff \mathtt{t}(x) \in V(w) \iff f(s) = x \iff (M^{nr}, s) \models \mathtt{t}(x)$. Negation and conjunction cases are trivial. When φ takes the form $[X]\psi$, notice that $w \equiv_X v \in W \iff V(w) \cap X = V(v) \cap X \iff s \cap X = t \cap X$, where $V_{Atm \setminus Dec}(v) = t$ for some $v \in W, t \in S$. Hence that $\forall v \in W, w \equiv_X v$ implies $v \models \psi$, if and only if $\forall t \in S, s \cap X = t \cap X$ implies $t \models \psi$.

$\Leftarrow$ Given $M^{nr} = (S, f) \in \mathbf{NDM}_P$ and $(M^{nr}, s) \models \varphi$, we construct a decision model $M^\flat = (W^\flat, (\equiv_Y)_{Y \in AtmSet_0}, V)$ as follow

- $W^\flat = \{s : s \in S\}$
- $s \equiv_Y t$ if $s \cap Y = t \cap Y$
- $V : W^\flat \to 2^{Atm}$ s.t. $V(s) = s \cup \{\mathtt{t}(x) : f(s) = x\}$

We need to check $M^\flat$ is indeed a decision model, namely it satisfies **C1-C6**. Obviously **C1** holds. **C2-C3** hold by virtue of the functionality of f. For **C4**, we show for any $s, s' \in W$, $V_X(s) = V_X(s')$, we have $f(s) = f(s')$, therefore $V_{Dec}(s) = V_{Dec}(s')$. Similar for **C5-C6**. Finally we prove by induction that $(M^{nr}, s) \models \varphi$ then $(M^\flat, s) \models \varphi$. The proof is trivial and omitted.

As for non-representative cases things are trickier. As Proposition 1 shows some certain classes $\mathbf{DM}_P$ and $\mathbf{NDM}_P$ are even empty. However, the ways of constructing an NDM for DM, and a DM for NDM remain the same. After all, only $def(X) \in P$ for some $X \in AtmSet_0$ is crucial to the corresponding model construction. □

A.4 Proof of Theorem 2

Proof. We prove the basic case, namely BCL. Proofs for the other cases of $\mathsf{BCL}_{\{cp(pr):pr\in P\}}$ with respect to corresponding model $\mathbf{DM}_P$ are analogous and omitted, since the one-one corresponding of semantic constraints $def(X), comp(X)$ and $ntr(X)$ to axioms $\mathbf{Def}_X, \mathbf{Comp}_X, \mathbf{NTr}_X$ is obvious. The proof is conducted by constructing the canonical model.

Definition 15. *The canonical decision model* $\mathfrak{M} = (W^c, \{\equiv_X^c\}_{X\in AtmSet_0}, V^c)$ *is defined as follows*

- $W^c = \{\Gamma : \Gamma$ *is a maximal consistent* BCL *theory.*$\}$
- $\Gamma \equiv_X^c \Delta \iff \{[X]\varphi : [X]\varphi \in \Gamma\} = \{[X]\varphi : [X]\varphi \in \Delta\}$
- $V^c(\Gamma) = \{p : p \in \Gamma\} \cup \{\neg p : p \notin \Gamma\}$

We omit the superscript ^{c} whenever there is no misunderstanding.

Lemma 1. *For any set of non-decision value atoms* X, $\varphi \in \mathcal{L}(Atm)$ *and* $\Gamma \in \mathfrak{M}$, $\Gamma \vdash_{\mathsf{BCL}} [X]\varphi \to \varphi$.

Proof. According to $\mathbf{Red}_{[\emptyset]}$ we want to prove $\Gamma \vdash_{\mathsf{BCL}} \bigwedge_{Y\subseteq X}(\mathsf{con}_{Y,X} \to [\emptyset](\mathsf{con}_{Y,X} \to \varphi))$. By deduction theorem it is enough to show $\Gamma \cup \bigwedge_{Y\subseteq X}(\mathsf{con}_{Y,X} \to [\emptyset](\mathsf{con}_{Y,X} \to \varphi)) \vdash_{\mathsf{BCL}} \varphi$. There is exactly one $Z \subseteq X$ s.t. $Z \subseteq \Gamma$. Hence from $\Gamma \vdash_{\mathsf{BCL}} \mathsf{con}_{Z,X}$ and $\Gamma \vdash_{\mathsf{BCL}} \mathsf{con}_{Z,X} \to [\emptyset](\mathsf{con}_{Z,X} \to \varphi)$ by $\mathbf{K}_{[\emptyset]}$ and $\mathbf{MP}$ we derive $\Gamma \vdash_{\mathsf{BCL}} \varphi$. Then by logical monotonicity we obtain $\Gamma \vdash_{\mathsf{BCL}} \bigwedge_{Y\subseteq X}(\mathsf{con}_{Y,X} \to [\emptyset](\mathsf{con}_{Y,X} \to \varphi))$. □

Lemma 2. $\mathfrak{M}$ *is indeed a decision model.*

Proof. Check the conditions one by one. For **C1**, we need show $\Gamma \equiv_X^c \Delta$, if $\forall p, p \in V(\Gamma) \cap X$ implies $p \in V(\Delta)$. Suppose not, then w.l.o.g. we have some $q \in V(\Gamma) \cap X, q \notin V(\Delta)$, by maximality of Δ namely $\neg q \in \Delta$. However, we have $[q]q \in \Gamma$, for $\vdash_{\mathsf{BCL}} q \to [q]q$, and by definition of $\equiv_X^c, [q]q \in \Delta$, hence $q \in \Delta$, since $\Delta \vdash_{\mathsf{BCL}} [q]q \to q$. But now we have a contradiction. **C2-3** hold obviously due to axioms $\mathbf{AtLeast}_{\mathsf{t}(x)}, \mathbf{Atmost}_{\mathsf{t}(x)}$ respectively. □

Lemma 3. $\mathfrak{M}, \Gamma \models \varphi \iff \varphi \in \Gamma$.

Proof. By induction on φ. We only show the interesting case when φ takes the form $[X]\psi$.

For $\Longleftarrow$ direction, if $[X]\psi \in \Gamma$, since for any $\Delta \equiv_X \Gamma$, $[X]\psi \in \Delta$, then thanks to $\Gamma \vdash_{\mathsf{BCL}} [X]\psi \to \psi$ we have $\psi \in \Delta$. By induction hypothesis this means $\Delta \models \psi$, therefore $\Gamma \models [X]\psi$.

For $\Longrightarrow$ direction, suppose not, namely $[X]\psi \notin \Gamma$. Then consider a theory $\Gamma' = \{\neg\psi\} \cup \{[X]\chi : [X]\chi \in \Gamma\}$. It is consistent since $\psi \notin \Gamma$. Then take any $\Delta \in W$ s.t. $\Gamma' \subseteq \Delta$. We have $\Delta \equiv_X \Gamma$, but $\Delta \not\models \psi$ by induction hypothesis. However, this contradicts $\Gamma \models [X]\psi$. □

Now the completeness of $\mathbf{DM}$ with respect to BCL is a corollary of Lemma 2 and 3.

□

A.5 Proof of Theorem 3

SKETCH OF PROOF. The satisfiability problem of formulas in $\mathcal{L}(Atm)$ relative to the class $\mathbf{DM}_P$ is clearly NP-hard since there exists a polynomial-time reduction of SAT to it.

As for membership, for every formula $\psi \in \mathcal{L}(Atm)$, we have $\models_{\mathbf{DM}_P} \psi$ iff $\models_{\mathbf{M}} [\emptyset]\big(\varphi_1 \wedge \varphi_2 \wedge \bigwedge_{prop \in P} \gamma(prop)\big) \rightarrow \psi$, where $\mathbf{M}$ is the class of ceteris paribus models of Definition 1,

$$\varphi_1 =_{def} \bigvee_{x \in Val} \mathsf{t}(x),$$
$$\varphi_2 =_{def} \bigwedge_{x,y \in Val: x \neq y} \big(\mathsf{t}(x) \rightarrow \neg \mathsf{t}(y)\big),$$

γ is the bijection

$$\gamma : Prop \longrightarrow \bigcup_{X \in AtmSet_0} \{\varphi_{def(X)}, \varphi_{comp(X)}, \varphi_{ntr(X)}\}$$

such that, for every $X \in AtmSet_0$:

$$\gamma\big(def(X)\big) = \varphi_{def(X)},$$
$$\gamma\big(comp(X)\big) = \varphi_{comp(X)},$$
$$\gamma\big(ntr(X)\big) = \varphi_{ntr(X)},$$

with

$$\varphi_{def(X)} =_{def} \bigwedge_{x \in Val, Y \subseteq X} \Big(\big(\mathsf{con}_{Y,X} \wedge \mathsf{t}(x)\big) \rightarrow [\emptyset]\big(\mathsf{con}_{Y,X} \rightarrow \mathsf{t}(x)\big)\Big),$$
$$\varphi_{comp(X)} =_{def} \bigwedge_{Y \subseteq X} \langle\emptyset\rangle \mathsf{con}_{Y,X},$$
$$\varphi_{ntr(X)} =_{def} \bigvee_{\substack{Y,Y' \subseteq X, x,y \in Val: \\ Y \neq Y' \text{ and } x \neq y}} \Big(\langle\emptyset\rangle\big(\mathsf{con}_{Y,X} \wedge \mathsf{t}(x)\big) \wedge \langle\emptyset\rangle\big(\mathsf{con}_{Y',X} \wedge \mathsf{t}(y)\big)\Big).$$

This means that satisfiability checking relative to the class $\mathbf{DM}_P$ can be reduced to satisfiability checking relative to the class $\mathbf{M}$. In [10] it is shown that the latter problem is in NP. Thus, since the size of $\varphi_1 \wedge \varphi_2 \wedge \bigwedge_{prop \in P} \gamma(prop)$ is constant and does not depend on the size of ψ, we can conclude that satisfiability checking relative to the class $\mathbf{DM}_P$ is also in NP.

■

A.6 Proof of Proposition 3

Proof. Consider any decision model $M = (W, (\equiv_X)_{X \in AtmSet_0}, V), w \in W$ and $X \subseteq (Atm \backslash Dec)$ finite. Assume $(M, w) \not\models \varphi \Rightarrow_X \psi$, i.e. $\exists v \in closest_M(w,\varphi,X), v \models \varphi \wedge \neg\psi$. We show thus $(M, w) \not\models \bigwedge_{0\leq k\leq |X|}(\mathsf{maxSim}(\varphi, X, k) \rightarrow \bigwedge_{Y\subseteq X:|Y|=k}[Y](\varphi \rightarrow \psi))$, namely $(M, w) \models \bigvee_{0\leq k\leq |X|}(\mathsf{maxSim}(\varphi, X, k) \wedge \bigvee_{Y\subseteq X:|Y|=k}\langle Y\rangle(\varphi \wedge \neg\psi))$. Let $Z = \{p : p \in X \ \& \ w \equiv_{\{p\}} v\}$, and $k = |Z|$. It is easy to see $(M, w) \models \langle Z\rangle(\varphi \wedge \neg\psi)$, for $(M, v) \models \varphi \wedge \neg\psi$. To prove $(M, w) \models \mathsf{maxSim}(\varphi, X, k)$ it is enough to show $w \models \langle Z\rangle\varphi \wedge \bigwedge_{Y\subseteq X:k<|Y|}[Y]\neg\varphi$. The first conjunct is obvious. For the second, suppose the opposite, then $\exists Z' \subseteq X, |Z'| > k$, $\exists u \in W, w \equiv_{Z'} u$ and $u \models \varphi$. However, that means $v \notin closest_M(w,\varphi,X)$, since $|Z| < |Z'|$, a contradiction.

For the other direction, similarly if $(M, w) \not\models \bigwedge_{0\leq k\leq |X|}(\mathsf{maxSim}(\varphi, X, k) \rightarrow \bigwedge_{Y\subseteq X:|Y|=k}[Y](\varphi \rightarrow \psi))$, then there is a $Z \subseteq X$, s.t. $\exists v \in W, w \equiv_Z v$ and $v \models \varphi \wedge \neg\psi$. We show that $v \in closest_M(w,\varphi,X)$. Suppose not, then $\exists u \in closest_M(w,\varphi,X)$ s.t. $|\{p : p \in X \ \& \ w \equiv_p u\}| > Z = k$ and $u \models \varphi$. But this contradicts $(M, w) \models \bigwedge_{Y\subseteq X:k<|Y|}[Y]\neg\varphi$. □

A.7 Proof of Proposition 10

Proof. We show that for any $s \in S$, $(M^{nr}, s) \models \mathsf{Bias}(\alpha, \mathsf{t}(x))$ iff $(M^{nr}, s) \models \mathsf{CXp}(\lambda, \alpha, x)$ for some $Atm(\lambda) \subseteq \mathsf{PF}$. For the left direction, by definition the antecedent gives $(M^{nr}, s) \models \alpha \wedge \mathsf{t}(x) \wedge \langle Atm(\alpha) \setminus Atm(\lambda)\rangle\neg\mathsf{t}(x)$, which means the decision $\mathsf{t}(x)$ for α is biased regarding $Atm(\lambda)$. For the right direction, by antecedent we know some $X \subseteq Z \cap \mathsf{PF}$, s.t. $(M^{nr}, s) \models \alpha \wedge \mathsf{t}(x) \wedge \langle Atm(\alpha) \setminus X\rangle\neg\mathsf{t}(x)$. We find a minimal X, and let $\lambda = \{p : p \in X \cap s\} \cup \{\neg p : p \in X \ \& \ p \notin s\}$. Then we shall have $(M^{nr}, s) \models \mathsf{CXp}(\lambda, \alpha, x)$. □

A.8 Proof of Theorem 6

SKETCH OF PROOF. Under the assumption that Atm is finite, the following equivalence is valid for every class $\mathbf{EDM}_P$:

$$\mathsf{K}\varphi \leftrightarrow \bigwedge_{Y\subseteq Obs} \big(\mathsf{con}_{Y,Obs} \rightarrow [Y \cup \{p : \mathsf{s}(p) \in Y\}]\varphi\big).$$

Thanks to the previous equivalence we can adapt the polynomial satisfiability-preserving translation from $\mathcal{L}(Atm)$ to the language of the modal logic S5 given in [10] to obtain a polynomial satisfiability-preserving translation from $\mathcal{L}^{epi}(Atm)$ to the S5 language. ■