A bimodal simulation of defeasibility in the normative domain

Tomer Libal[1], Matteo Pascucci[2], Leendert van der Torre[3], and Dov Gabbay[4]

[1] University of Luxembourg, Luxembourg
shaolintl@gmail.com
[2] Central European University, Austria
pascuccim@ceu.edu
[3] University of Luxembourg, Luxembourg
and
Zhejiang University, China
leon.vandertorre@uni.lu
[4] University of Luxembourg, Luxembourg
dov.gabbay@ext.uni.lu
and
King's College London, United Kingdom
dov.gabbay@kcl.ac.uk

Abstract. In the present work we illustrate how two sorts of defeasible reasoning that are fundamental in the normative domain, that is, reasoning about exceptions and reasoning about violations, can be simulated via monotonic propositional theories based on a bimodal language with primitive operators representing knowledge and obligation. The proposed theoretical framework paves the way to using native theorem provers for multimodal logic, such as MleanCoP, in order to automate normative reasoning.

Keywords: Bimodal Logic - Contrary-to-duty Reasoning - Deontic Logic - Exceptions - Non-monotonic Reasoning - Violations

1 Introduction

Deciding whether an obligation (permission, prohibition, etc.) applies to a given scenario ultimately depends on what one knows about that scenario. This work aims at providing a simple approach to defeasible reasoning in the normative domain based on the combination of deontic and epistemic modalities.[5] We will distinguish between two sorts of defeasibility, which are at the core of normative

[5] Several ways of exploiting epistemic concepts in the normative domain have been proposed in the literature; for instance, they are used by Aucher, Boella and van der Torre [1] to capture the dynamic character of normative systems, and by Pacuit, Parikh and Cogan [14] in the definition of deontic concepts. Our work can be located in this tradition.

reasoning. The first sort emerges in reasoning about *exceptions of norms*, namely in cases in which a norm applies to more specific circumstances than another; the second sort emerges in reasoning about *violations of norms*, namely in cases in which a norm applies when another is violated.

The two sorts of defeasibility can be illustrated with a classical example adapted from Prakken and Sergot [16] (for a more detailed presentation, see van der Torre [21] or van der Torre and Tan [22]). Consider the following set of sentences:

- you ought not to build a fence in your land;
- if there is a fence in your land, then it must be white;
- if there is a dog living in your land, then you must build a fence.

The first sentence presents a general norm; the second sentence describes what happens if the first norm is violated (a contrary-to-duty obligation becomes effective); the last sentence provides an exception to the first norm (there are special circumstances in which what is prohibited by the first norm becomes obligatory).

Many formal approaches to deal with defeasible reasoning, starting with Reiter's logic of defaults [17], rely on a set of primitive rules for non-monotonic inferences. Other approaches, such as Moore's autoepistemic logic [11] and Boutilier's conditional logics of normality [3], can be used to capture defeasible reasoning within a framework of (multi)modal logic. In the normative domain, a systematic formal treatment of defeasible reasoning is offered by dyadic deontic logic [7], whose language allows one to express norms of the form $\mathsf{O}(\phi/\psi)$ (to be read 'ϕ is obligatory under condition ψ'), where O is an operator for obligation, ψ is the antecedent of the obligation and ϕ the consequent of the obligation. In these systems the property called 'strengthening of the antecedent' (which is a form of monotonicity) typically fails: there is no guarantee that an inference from $\mathsf{O}(\phi/\psi)$ to $\mathsf{O}(\phi/\psi \wedge \chi)$ is sound; however, alternative axioms and rules for restricted monotonic inferences have been proposed (see, for instance, the survey by Goble [5]). Despite its flexibility, the framework of dyadic deontic logic does not provide a straightforward solution to the problem of detaching an actual norm from a conditional norm and the truth of its antecedent; this is extensively discussed by Straßer [20]. For a useful introduction to some additional approaches to defeasible deontic logic, we refer the reader to Nute [12].

Taking a computational stance, one notices that there are various tools implementing defeasible reasoning *in general*, such as the LogicKEY system [2]. However, LogicKEY and related approaches rely on a higher-order formalism which makes them less efficient than approaches based on a first-order or propositional formalism. Furthermore, there are tools focused on a specific type of defeasible reasoning (for the normative context, see, for instance, Governatori [6]).

Other popular implementations of defeasible reasoning are based on *logic programming*. For instance, Prolog allows for the definition of defeasible rules by means of negation-as-failure operators. The three sentences from the fence example above can be respectively encoded in Prolog as follows (where '*not*' denotes negation-as-failure, '*ob*' obligation and '*neg*' classical negation):

- $not(dog) \Rightarrow ob(neg(fence))$
- $fence \Rightarrow ob(white)$
- $dog \Rightarrow ob(fence)$

One can support this approach with Answer-Set Programming in order to deal with classical negations. Yet, Prolog does not offer a straightforward solution to characterize the difference between exceptions and violations, which is essential in the normative domain. Kowalski and Satoh [9] present a possible way of overcoming the latter problem in *abductive logic programming*. Their solution is based on the addition of *explicit sanctions* to the program. For instance,

- $neg(dog) \Rightarrow not(fence) \vee sanction_1$
- $fence \Rightarrow white$

Sanctions can thus be used to distinguish violations from exceptions of a norm.

In this paper we present a new framework for knowledge-based normative reasoning which combines the two sorts of defeasible reasoning discussed above. We describe a general method to formalize a set of norms as *bimodal conditional formulae*; furthermore, we provide criteria to *identify which norms constitute exceptions or reparations* of other norms within the formalized set. On top of this, we use an agent's information about a scenario to infer which norms can be detached among those considered. In our approach, *constants for sanctions* are used to capture the difference between (I) cases in which the violation of a norm can be *fully compensated* by complying with a reparation norm and (II) cases in which the violation of a norm leads to negative consequences regardless of its possible reparations. Thus, the underlying idea is that *reparation does not entail compensation*.

The main advantage of this framework lies in its potential computational properties. While most of the existing implementations depend either on a customized or a non-monotonic or a higher-order formalism, our approach is based on a bimodal normal propositional logic, and there are many efficient theorem provers designed specifically for normal multimodal logics, such as MleanCoP [13].

Our presentation will be arranged according to the following structure. In Section 2 we provide a detailed description of the proposed framework to combine epistemic and deontic concepts. In Section 3 we discuss how to complete the information provided by a reasoning agent on a certain scenario. In Section 4 we use possible descriptions of a scenario in terms of known and unknown facts to define the coherence of a normative theory. In Section 5 we discuss the way in which the proposed framework could be implemented. In Section 6 we illustrate how our framework deals with a very simple problem of normative reasoning. Finally, in Section 7 we summarize the essential features of our approach and specify some directions for future research.

2 Knowledge-supported normative theories

The formal framework we present here consists in a step after step construction of *deductive theories* where deontic and epistemic modalities are combined. The

sort of problem we want to address within this framework is the following: an agent α wants to understand which are the normative consequences of a scenario S that is regulated by a certain set of norms N. It might be that α has *partial information* about S. In particular, she knows that some propositions mentioned in N are true, some others are false and she is not aware of the rest. Furthermore, α might have partial information about the meaning of norms in N, in the sense that she might fail to notice conceptual dependencies among different norms. Given the epistemic boundaries mentioned, how can α determine which norms are in effect in the considered scenario and whether any sanction is applicable? In this section we lie down the theoretical grounds for a procedure of logic-assisted reasoning that in similar cases would allow an agent to derive all normative conclusions needed.

We start by introducing the formal language that will be used within the framework.

Definition 1 (Formal language) *Let PRO be a set of propositional variables and p an element in this set; furthermore, let SAN be a set of sanction constants and s an element in this set. The language $\mathcal{L}$ over PRO and SAN is specified by the grammar below:*

$$\phi ::= p \mid s \mid \neg\phi \mid \phi \Rightarrow \phi \mid \mathsf{O}\phi \mid \mathsf{K}\phi$$

Formulae of $\mathcal{L}$ not containing K or O will be said to be *purely boolean formulae*. We take $\mathsf{O}\phi$ as meaning that ϕ is obligatory and $\mathsf{K}\phi$ as meaning that ϕ is known (by the reasoning agent). Round brackets will be used as auxiliary symbols when needed. Furthermore, the additional boolean operators $\vee$ (inclusive disjunction), $\wedge$ (conjunction) and $\equiv$ (material equivalence) can be defined in the usual way. Finally, we will use the label **PC** for the Classical Propositional Calculus.

Definition 2 (Basic notation) *The following notation will be used to represent relevant subsets of the language $\mathcal{L}$:*

- $\mathcal{L}_{PRO}$ *is the set of purely boolean formulae in $\mathcal{L}$;*
- $LIT = PRO \cup \{\neg p : p \in PRO\}$ *is the set of literals and a generic element in this set will be denoted by l;*
- $\mathcal{L}^{+}_{\mathsf{K}} = \{\mathsf{K}l : l \in LIT\}$;
- $\mathcal{L}^{-}_{\mathsf{K}} = \{\neg\mathsf{K}l : l \in LIT\}$;
- $\mathcal{L}_{K} = \mathcal{L}^{+}_{\mathsf{K}} \cup \mathcal{L}^{-}_{\mathsf{K}}$;
- *for any set $X \subseteq \mathcal{L}$, $\mathsf{K}(X) = \{\mathsf{K}\phi : \phi \in X\}$.*

Definition 3 (Logical system) *The logical system that will be used as a basis for our theories is $\mathbf{KT}_{\mathsf{K}} \otimes \mathbf{KD}_{\mathsf{O}}$, that is, the result of putting together the ax-*

iomatic basis of **KT** *for the modal operator* K *and the axiomatic basis of* **KD** *for the modal operator* O.[6]

Definition 4 (K-conjunctive normal form) *A formula of* $\mathcal{L}$ *will be said to be in* K-conjunctive normal form *if and only if it has the form* $\bigwedge_{1\leq i\leq n}\psi_i$ *(for some* $n \in \mathbb{N}$*), where each* ψ_i *has the form* $\bigvee_{1\leq j\leq m}\theta_j$ *(for some* $m \in \mathbb{N}$*) and* $\theta_j \in \mathcal{L}_{\mathsf{K}}$.

We assume the following notational convention: if $n, m = 1$, then $\bigwedge_{1\leq i\leq n}\psi_i = \psi$ and $\bigvee_{1\leq j\leq m}\theta_j = \theta$.

Now we can introduce the deductive theories that constitute the skeleton of our framework for logic-assisted reasoning on normatively relevant scenarios. These will be called *knowledge-supported normative theories*, since all their components exploit epistemic concepts.

Definition 5 (Knowledge-supported normative theory) *We will say that a knowledge-supported normative theory is a deductive theory having the following five components (C.1–C.5):*

(C.1) ***Conditional norms.***
A finite set N *of formulae of the form*

$$\phi \Rightarrow (\mathsf{mod}(\chi) \wedge \xi)$$

where:

- ϕ *is a formula in* K*-conjunctive normal form;*
- mod *is an expression having one of the following forms: either* O *or* $\neg\mathsf{O}$ *or* $\mathsf{O}\neg$ *or* $\neg\mathsf{O}\neg$*, and is called a* deontic modality*;*
- $\chi \in \mathcal{L}_{PRO}$*;*
- ξ *is a conjunction of sanction symbols;*
- *one among* $\mathsf{mod}(\chi)$ *and* ξ *can be absent (so, as peculiar cases, we have norms of the form* $\phi \Rightarrow \mathsf{mod}(\chi)$ *and norms of the form* $\phi \Rightarrow \xi$*).*

Any norm in N is thus represented as a *material conditional* of a specific type: its antecedent is a conjunction of clauses, each being a *disjunctive combination of known and unknown facts*, and its consequent is a conjunction of deontic propositions and/or sanctions. To give an example: a formula in N may encode the proposition expressed by the sentence "if we know that a patient has been treated with aspirin or warfarin and has had three days of rest, then he/she can be discharged". The encoding consists in rearranging the logical structure of the proposition so as to get a conditional where the antecedent ("if we know... ...three days of rest") becomes expressible in K-conjunctive normal form, while

[6] These two components are normal modal systems: $\mathbf{KT}_{\mathsf{K}}$ is obtained by adding the axiom-schema $\mathsf{K}\phi \Rightarrow \phi$ to system $\mathbf{K}_{\mathsf{K}}$ (i.e., **K** for the operator K), $\mathbf{KD}_{\mathsf{O}}$ by adding the axiom-schema $\mathsf{O}\phi \Rightarrow \neg\mathsf{O}\neg\phi$ to $\mathbf{K}_{\mathsf{O}}$ (i.e., **K** for the operator O).

the consequent ("then... ...discharged") instantiates one of the deontic modalities specified above.

(C.2) ***Conceptual relations.***
A finite set M of formulae of the form $\mathsf{K}(\phi \Rightarrow \psi)$, where $\phi, \psi \in \mathcal{L}_{PRO}$ are neither tautologies nor contradictions of **PC**.
This can be regarded as the set of assumptions that are needed to understand the conceptual relations between the norms in N. For instance, suppose that we have a norm $n_1 \in N$ containing the proposition (p) that a patient is treated with aspirin, and a norm $n_2 \in N$ containing the proposition (q) that the patient is treated with acetylsalicylic acid: since p and q should be known to be equivalent, then the norms n_1 and n_2 should be treated as conceptually related. We will say that an agent who knows all such relations is *optimally informed* about the content of N; clearly, this is not always the case for an actual reasoning agent. Therefore, the optimally informed agent is, somehow, an idealized figure. However, since the set of norms is finite and contains a finite number of terms, their conceptual relations are finite as well; this means that the amount of knowledge that we expect an optimally informed agent to have is finite anyway. The fact that the main boolean operator in a formula associated with a conceptual relation is a material conditional indicates that we are representing a relation of *conceptual entailment* among two propositions (*conceptual equivalence* is straightforwardly obtained by putting together two material conditionals). The fact that ϕ and ψ are neither tautologies nor contradictions ensures that conceptual relations are *informative*, that is, they only concern propositions which can be true and can be false in different circumstances. We will say that $N \cup M$, for a set of norms N and a set of conceptual relations M, constitutes a *normative theory.*

(C.3) ***The reasoning agent's explicit knowledge.***
A finite set of formulae $B \subset \mathcal{L}^{+}_{\mathsf{K}}$.
The reasoning agent's description of a scenario S (e.g., the reasoning agent knows that a patient has had three days of rest), which might fail to be sufficient to properly assess which norms are in effect. Indeed, there might be propositions mentioned in N that are neither known to be true nor known to be false in S by the reasoning agent. We will say that B is an *explicit knowledge box* on S.

(C.4) ***The reasoning agent's integrated knowledge.***
A finite set of formulae B' s.t. $B \subseteq B' \subset \mathcal{L}^{+}_{\mathsf{K}}$.
This corresponds with what the reasoning agent would know about S if she were optimally informed about the meaning of norms in N, namely if she were aware of all conceptual relations between the norms. We call B' an *integrated knowledge box* on S. This set might still represent an incomplete description of S.

(C.5) ***Ignorance claims.***
A finite set U of formulae of the form $\neg\mathsf{K}l$ for some $l \in LIT$.

This constitutes the set of propositions mentioned in N on which the reasoning agent would remain ignorant even if she were optimally informed about the meaning of norms. Going back to our working example: if the reasoning agent does not know whether a patient has been treated with aspirin (p) or warfarin (w), U will contain $\neg\mathsf{K}p$ and $\neg\mathsf{K}w$.

Logic-assisted reasoning results from the interplay between the five components C.1-C.5: the conceptual relations (C.2) can be used to expand the explicit knowledge box provided by an agent (C.3) and get an integrated knowledge box (C.4); then, the addition of ignorance claims (C.5) leads to a sufficiently informative description of a scenario, and such description can be used to detach actual deontic statements and sanctions from the conditional norms (C.1). The *epistemic analysis of norms* is revealed by the fact that detachment depends only on what is known and unknown about a scenario; however, given that $\mathbf{KT}_\mathsf{K}$ is closed under the schema T, that is, $\mathsf{K}A \Rightarrow A$, then knowledge entails truth. Furthermore, an important aspect in which the present approach turns out to be simpler than other approaches used to capture non-monotonic reasoning in a modal framework (such as autoepistemic logic [11]) is that in normative contexts one does not have to take into account *introspection*: the fact that one *knows that she knows* (or that she does not know) whether something is the case is totally irrelevant with respect to norm detachment. This is the reason why we choose the weakest normal modal system suitable for knowledge ($\mathbf{KT}_\mathsf{K}$), rather than its extensions $\mathbf{S4}_\mathsf{K}$ and $\mathbf{S5}_\mathsf{K}$.

In the next section we will illustrate how to move from the explicit knowledge box provided by the reasoning agent on a scenario S to an integrated knowledge box which takes into account also the meaning of norms. Then, the addition of a set of ignorance claims to the latter box completes the picture that is needed to detach, from N, the deontic statements that are in effect and the sanctions that are applicable in S, on the basis of the available information.

3 Completing the explicit knowledge box

Our purpose is to define a procedure to complete the information explicitly provided by a reasoning agent on a scenario in such a way that it can be implemented in efficient tools for automated deduction.

The most important aspect of the knowledge-supported theories described in Section 2 is that in representing information we need to use only a fragment of the language of the bimodal system $\mathbf{KT}_\mathsf{K}\otimes\mathbf{KD}_\mathsf{O}$, namely the fragment in which K and O always have purely boolean formulae in their scope. In this part we focus on provability in $\mathbf{KT}_\mathsf{K}$. We explain how the explicit knowledge box B provided by an agent on a scenario S can be used to build a description of S that is sufficient to trigger all norms that an optimally informed subject (in the sense specified in Definition 5) would identify as applicable in S on the basis of

B; we will call the latter a *normatively exhaustive description* of S (depending on B).[7]

Definition 6 (Normatively exhaustive description) *Let M be a given set of conceptual relations, B an explicit knowledge box on a scenario S, $p \in PRO$ and $l \in LIT$. If $B^\sharp = \{\mathsf{K}l : B \cup M \vdash_{\boldsymbol{KT}_\mathsf{K}} \mathsf{K}l\}$ is the integrated knowledge box on S and $U^* = \{\neg\mathsf{K}p, \neg\mathsf{K}\neg p : \mathsf{K}p, \mathsf{K}\neg p \notin B^\sharp\}$ the set of ignorance claims, then the normatively exhaustive description of S is $B^\sharp \cup U^*$.*

Now we show that a normatively exhaustive description of a scenario is consistent whenever the union of the explicit knowledge box and the set of conceptual relations from which it was generated is consistent.

Proposition 1 (Consistency preservation) *If B is a knowledge box, M a set of conceptual relations and $B \cup M$ is a $\mathbf{KT}_\mathsf{K}$-consistent set of formulae, then $\Sigma = B^\sharp \cup U^*$ (where $B^\sharp$ and U^* are as in Definition 6) is a $\mathbf{KT}_\mathsf{K}$-consistent set of formulae as well.*

Proof. Consider the canonical model for the modal system $\mathbf{KT}_\mathsf{K}$, $\mathfrak{M} = \langle W, R, V\rangle$. If $B \cup M$ is $\mathbf{KT}_\mathsf{K}$-consistent, then there is a non-empty set $X \subseteq W$ s.t. all formulae in $B \cup M$ belong to every state (maximal $\mathbf{KT}_\mathsf{K}$-consistent set of formulae) in X. We have that $B^\sharp$ is the set of all formulae of the form $\mathsf{K}l$, for some $l \in LIT$, which belong to all states in X. Since neither $\neg\mathsf{K}p \Rightarrow \mathsf{K}\neg p$ nor $\neg\mathsf{K}\neg p \Rightarrow \mathsf{K}p$ is provable in $\mathbf{KT}_\mathsf{K}$, there will be at least one state $w \in X$ s.t., for every $p \in PRO$ having the property that $\mathsf{K}p, \mathsf{K}\neg p \notin B^\sharp$, $\neg\mathsf{K}p, \neg\mathsf{K}\neg p \in w$, whence $\Sigma \subset w$. By construction, w is a maximal $\mathbf{KT}_\mathsf{K}$-consistent set of formulae. Therefore, Σ is a $\mathbf{KT}_\mathsf{K}$-consistent set of formulae.

For the sake of brevity in the exposition, we will hereafter assume reference only to normatively exhaustive descriptions of scenarios that are $\mathbf{KT}_\mathsf{K}$-consistent.

4 Coherence of normative theories

We next illustrate how normatively exhaustive descriptions of scenarios can be used to define the coherence of a normative theory. In order to do this we first need to define some auxiliary notions. Depending on the purposes, we will alternatively make reference to provability in $\mathbf{KT}_\mathsf{K} \otimes \mathbf{KD}_\mathsf{O}$ or in some of its monomodal fragments. Furthermore, given a set of conceptual relations M, we denote by M^O the set $\{\mathsf{O}\phi : \mathsf{K}\phi \in M\}$.

[7] We observe once more that B might contain only partial information about S; hence, the notion of a normatively exhaustive description of a scenario is distinct from the notion of a *total description* of a scenario (the latter entails the former but not vice versa).

Definition 7 (Exception of a norm) *Given two norms* $\psi = (\phi \Rightarrow (\mathsf{mod}(\chi) \wedge \xi))$ *and* $\psi' = (\phi' \Rightarrow (\mathsf{mod}'(\chi') \wedge \xi'))$ *in* N *and a set of conceptual relations* M*, we say that* ψ' *is an exception of* ψ *if and only if the following holds:*

- $M^{\mathsf{O}} \vdash_{\boldsymbol{KD}_{\mathsf{o}}} \mathsf{mod}'(\chi') \rightarrow \neg\mathsf{mod}(\chi)$*;*
- *If* $X = \{\mathsf{K}l : \vdash_{\boldsymbol{KT}_{\mathsf{k}}} \phi \Rightarrow \mathsf{K}l\}$ *and* $X' = \{\mathsf{K}l : \vdash_{\boldsymbol{KT}_{\mathsf{k}}} \phi' \Rightarrow \mathsf{K}l\}$*, then* $X \subset X'$*.*

Thus, a norm ψ' is an exception of a norm ψ if and only if they allow one to detach conflicting deontic statements and the antecedent of ψ' contains more known facts than the antecedent of ψ.

Definition 8 (Norm violation) *Given a norm* $\psi \in N$*, where* ψ *has the form* $\phi \Rightarrow (\mathsf{mod}(\chi) \wedge \xi)$ *and a normatively exhaustive description of a scenario* S*, denoted by* $\Sigma = B^{\sharp} \cup U^{*}$*, we say that* ψ *is violated in* S*, in symbols* $\Sigma \vdash_{\boldsymbol{KT}_{\mathsf{k}}} \mathsf{vio}(\psi)$*, if and only if:*

- $\mathsf{mod} = \mathsf{O}$ *and* $\Sigma \vdash_{\boldsymbol{KT}_{\mathsf{k}}} \phi \wedge \mathsf{K}\neg\chi$*;*
- $\mathsf{mod} = \mathsf{O}\neg$ *and* $\Sigma \vdash_{\boldsymbol{KT}_{\mathsf{k}}} \phi \wedge \mathsf{K}\chi$*.*

Clearly, violations only concern conditional norms whose consequent includes either an obligation or a prohibition. Then, we need to associate norms to sanctions.

Definition 9 (Sanction assignment) *Given a set of norms* N*, a sanction assignment is a (possibly partial) function* $\mathfrak{f} : N \longrightarrow SAN$*. If* $\mathfrak{f}(\psi) = s$*, then we say that* s *is a sanction applicable for the violation of* ψ*.*

The function $\mathfrak{f}$ may be partial since it is not always the case that the violation of a norm ψ in N leads to a sanction. Indeed, there can be a reparation norm that fully compensates for the violation of ψ.

Definition 10 (Reparation norm) *Given two norms* $\psi = (\phi \Rightarrow (\mathsf{mod}(\chi) \wedge \xi))$ *and* $\psi' = (\phi' \Rightarrow (\mathsf{mod}'(\chi') \wedge \xi'))$ *in* N*, we say that* ψ' *is a reparation norm for the violation of* ψ *whenever the following holds for every normatively exhaustive description of a scenario* Σ *s.t.* $\neg\mathsf{K}\chi', \neg\mathsf{K}\neg\chi' \in \Sigma$*:*

- $N \cup \Sigma \vdash_{\boldsymbol{KT}_{\mathsf{k}} \otimes \boldsymbol{KD}_{\mathsf{o}}} \mathsf{vio}(\psi)$ *only if* $N \cup \Sigma \vdash_{\boldsymbol{KT}_{\mathsf{k}} \otimes \boldsymbol{KD}_{\mathsf{o}}} \mathsf{mod}'(\chi') \wedge \xi'$*.*

Thus, ψ' is a reparation norm for the violation of ψ if its consequent is detached in all circumstances in which (I) ψ is violated, and (II) we have no information about the fulfilment of ψ'. Moreover, we can say that a reparation norm ψ' *fully compensates* for the violation of ψ if and only if whenever ψ' is observed, $\mathfrak{f}(\psi)$ cannot be detached.

We finally have all ingredients needed to define the coherence of a normative theory (see Definition 5).

Definition 11 (Coherent normative theory) *Given a set of norms N and a set of conceptual relations M, a normative theory $N \cup M$ is coherent if and only if there is no normatively exhaustive description Σ of a scenario S satisfying one of the following two properties:*

- *for some norm $\psi = (\phi \Rightarrow (mod(\chi) \wedge \xi)) \in N$ we have*
  (i) $N \cup \Sigma \nvdash_{\boldsymbol{KT}_{\mathsf{k}} \otimes \boldsymbol{KD}_{\mathsf{o}}} \mathsf{vio}(\psi)$;
  (ii) $N \cup \Sigma \vdash_{\boldsymbol{KT}_{\mathsf{k}} \otimes \boldsymbol{KD}_{\mathsf{o}}} \mathsf{f}(\psi)$.
- $N \cup \Sigma \vdash_{\boldsymbol{KT}_{\mathsf{k}} \otimes \boldsymbol{KD}_{\mathsf{o}}} \bot$.

Thus, a normative theory is coherent if and only if under any normatively exhaustive description of a scenario received as an input (I) it does not allow one to detach that a sanction applies while the associated norm has not been violated, and (II) it does not produce a contradiction, which would mean an explosion of the set of deontic statements detached.

5 Remarks on implementation

Many theorem provers for deontic reasoning have been designed over the years, and some of these are also able to deal with defeasibility problems (see, e.g., Lee and Ryu [10], Governatori [6] and Steen [19]). The following part explains that we can use state-of-the-art automated reasoning tools for modal propositional logic in order to expand the explicit knowledge box provided by a reasoning agent to get a normatively exhaustive description of a scenario. First, consider the following fact.

Fact 1 (Derivability translation) *Let ϕ be a formula in $\mathcal{L}_{PRO}$ and X a set of formulae in $\mathcal{L}_{PRO}$, then $X \vdash_{\boldsymbol{PC}} \phi$ if and only if $\mathsf{K}(X) \vdash_{\boldsymbol{KT}_{\mathsf{k}}} \mathsf{K}\phi$*

One direction of the biconditional follows from the deduction theorem for normal modal logic and modal propositional reasoning. The other direction follows from the Post-completeness of the Classical Propositional Calculus and the fact that $\mathbf{KT}_{\mathsf{K}}$ is a consistent extension of it.

Fact 1 ensures that the transition from the explicit knowledge box to the integrated knowledge box can be computed within a *purely boolean level of analysis*. This level is equipped with a broader set of computational tools, due to the extensive variety of applications it has found over the years [4]. A model enumeration technique was already introduced by Smullyan [18] via the construction of a finite and closed tableaux for a finite set of formulae. However, given the NP-completeness of the Boolean satisfiability problem, we cannot expect these methods to perform efficiently in all cases.

In addition, we note that problems of normative reasoning are usually transposed in a computational setting following two phases. In the first phase, one formalizes a normative text, while in the second phase one reasons over the text in combination with a specified scenario. This suggests that the computational

issue of enumerating all relevant minimal models can be addressed at once and "offline", that is, when one formalizes the normative text. Then, given a specific scenario, one just needs to consider those minimal models which are also models of that scenario.

With reference to the process of completing a reasoning agent's knowledge, one should further note that in many formal frameworks for epistemic reasoning this process is cyclic, since one looks for a fixed point construction, taking into account also positive and negative introspection (namely, that an agent knows that she knows/ignores something). In our case, instead, we expand the explicit knowledge box of an agent B, which is a finite set, by adding only formulae of the form Kl, for some $l \in LIT$, obtained by combining B and a finite set of conceptual relations among norms M. Similarly, we then expand the integrated knowledge box with a finite amount of formulae of the form $\neg \mathsf{K} p$ and $\neg \mathsf{K} \neg p$, for some $p \in PRO$, to get the exhaustive description of a scenario needed for normative detachment. Thus, the resulting set of known and unknown facts will always be finite.

Finally, normative detachment can be automatically performed via an efficient theorem prover for propositional multimodal logic, such as MleanCoP [13].

6 A test case

In the present section we exploit the formal framework developed to represent a very simple example. Consider the following set of sentences (letters within brackets will be used below for the formalization of atomic propositions that can be obtained from parts of the sentences):

($n1$) you ought not to enclose your land **(e)**;
($n2$) if there is a fence for animals **(f)** in your land, then it must be white **(w)**;
($n3$) in case of animals **(a)** living in your land, you must build a fence;
($n4$) in case goats **(g)** are living in your land, you are obliged not to build a fence;
($n5$) in case your dogs are used for herding **(h)**, you are permitted not to build a fence;
($n6$) if you enclose your land without a justified reason for that, you have to pay a fine of €500 ($\mathbf{s}_1$);
($n7$) if you have a fence for animals and it is not white, you have to pay a fine of €200 ($\mathbf{s}_2$);
($n8$) if you only have animals that are different from goats and herding dogs, and you do not have a fence, you have to pay a fine of €600 ($\mathbf{s}_3$);
($n9$) if you have goats and you build a fence, you have to pay a fine of €100 ($\mathbf{s}_4$).

In addition, we have the following conceptual relations:

($r1$) if you build a fence for animals in your land, this entails that you enclose your land.
($r2$) dogs **(d)** are animals;
($r3$) goats are animals.

Violations are assigned via a function $\mathfrak{f}$ which is entirely described by the following list of ordered pairs: $(n1, s_1), (n2, s_2), (n3, s_3), (n4, s_4)$. A plausible formalization according to Definition 5 is illustrated below:

($n1^*$) $(\neg \mathsf{K}a \vee \mathsf{K}\neg a) \wedge \neg \mathsf{K}e \wedge \neg \mathsf{K}\neg e \Rightarrow \mathsf{O}\neg e$
($n2^*$) $\mathsf{K}f \wedge \neg \mathsf{K}w \wedge \neg \mathsf{K}\neg w \Rightarrow \mathsf{O}w$
($n3^*$) $\mathsf{K}a \wedge (\neg \mathsf{K}g \vee \mathsf{K}\neg g) \wedge (\neg \mathsf{K}h \vee \mathsf{K}\neg h) \wedge \neg \mathsf{K}f \wedge \neg \mathsf{K}\neg f \Rightarrow \mathsf{O}f$
($n4^*$) $\mathsf{K}g \wedge \neg \mathsf{K}f \wedge \neg \mathsf{K}\neg f \Rightarrow \mathsf{O}\neg f$
($n5^*$) $\mathsf{K}h \wedge \neg \mathsf{K}f \wedge \neg \mathsf{K}\neg f \Rightarrow \neg \mathsf{O}f$
($n6^*$) $(\neg \mathsf{K}a \vee \mathsf{K}\neg a) \wedge \mathsf{K}e \Rightarrow s_1$
($n7^*$) $\mathsf{K}a \wedge \mathsf{K}\neg g \wedge \mathsf{K}\neg h \wedge \mathsf{K}f \wedge \mathsf{K}\neg w \Rightarrow s_2$
($n8^*$) $\mathsf{K}a \wedge \mathsf{K}\neg g \wedge \mathsf{K}\neg h \wedge \mathsf{K}\neg f \Rightarrow s_3$
($n9^*$) $\mathsf{K}g \wedge \mathsf{K}f \Rightarrow s_4$
($r1a^*$) $\mathsf{K}(w \Rightarrow f)$
($r1b^*$) $\mathsf{K}(f \Rightarrow e)$
($r2^*$) $\mathsf{K}(d \Rightarrow a)$
($r3^*$) $\mathsf{K}(g \Rightarrow a)$

Our formalization can be generated from the set of norms by first transforming the proposition actually expressed by each norm into an equivalent proposition having the conditional structure described in Definition 5. An unconditional norm n, such as the one expressed by "you ought not to enclose your land" ($n1$), is transformed into a conditional one by taking for its antecedents claims that either (I) represent the fact that n is not already known to be fulfilled or violated (in the specific case of $n1$, these are $\neg \mathsf{K}e \wedge \neg \mathsf{K}\neg e$) or (II) are obtained from the analysis of the other norms in the set and represent conditions which complement those triggering exceptions to n (in our case $\neg \mathsf{K}a \vee \mathsf{K}\neg a$, given that the exception to $n1$ is triggered by $\mathsf{K}a$). According to Definition 7, we can recognize directly in the formalism that $n3$ is an exception to $n1$. Furthermore, we can recognize, for instance, that $n6$ is a reparation norm for the violation of $n1$, and that the proposed reparation does not compensate for the violation (indeed, s_1, which is $\mathfrak{f}(n1)$, can be detached).

Consider a scenario in which one knows that there are both goats and herding dogs in her land. From this information we can infer that she is obliged not to build a fence. In the given scenario, the reasoning agent's explicit knowledge is $\mathsf{K}g$ and $\mathsf{K}h$ and her integrated knowledge is $\mathsf{K}a$ and $\mathsf{K}d$. The ignorance closure contains $\neg \mathsf{K}f \wedge \neg \mathsf{K}\neg f$, $\neg \mathsf{K}e \wedge \neg \mathsf{K}\neg e$ and $\neg \mathsf{K}w \wedge \neg \mathsf{K}\neg w$. By putting together the exhaustive description of the scenario and the set of norms, one can derive $\mathsf{O}\neg f$ and $\neg \mathsf{O}f$. No other outcome is derivable.

Consider, instead, a scenario in which the agent's explicit knowledge contains only $\mathsf{K}g$ and $\mathsf{K}f$, i.e. she knows that there are goats and a fence. The integrated knowledge is $\mathsf{K}a$ and $\mathsf{K}e$ and the ignorance closure is $\neg \mathsf{K}d \wedge \neg \mathsf{K}\neg d$, $\neg \mathsf{K}h \wedge \neg \mathsf{K}\neg h$

and $\neg \mathsf{K} w \wedge \neg \mathsf{K} \neg w$. We can now derive only s_4, meaning that the fourth norm — saying that in case you have goats, you should not build a fence— was violated, and a sanction is applicable. Note that we are no longer able to derive that one should not build a fence.

7 Conclusion and Future Work

In this article we addressed the issue of representing normative reasoning under partial information about a scenario. We introduced a preliminary formal framework based on deductive theories of propositional bimodal logic which consist of five components: a set of norms (C.1), represented as material conditionals whose antecedent contains epistemic notions and whose consequent contains deontic notions and/or reference to sanctions; a set of entailment relations among propositions, which clarify conceptual dependencies among norms (C.2); a set of facts known by the reasoning agent, also called an explicit knowledge box (C.3); a set of facts that the reasoning agent would know if she were optimally informed about the meaning of norms, also called an integrated knowledge box (C.4); a set of facts that remain unknown (C.5).

This framework can be implemented via efficient monotonic reasoning tools, while offering the possibility of simulating non-monotonic reasoning. More precisely, the deductive theories described in this work satisfy *norm monotony* but do not satisfy *factual monotony*, two properties discussed by Parent and van der Torre [15]. Indeed, consider a set of norms N and two exhaustive descriptions of a scenario Σ and Σ' obtained by completing two explicit knowledge boxes B and B' such that $B \subset B'$; it might be that Σ and Σ' trigger different sets of norms in N. Not only this, but the framework is also able to capture the specific kind of defeasible reasoning that plays a central role in the normative domain, namely the distinction between exceptions of norms and reparation norms. Within the category of reparation norms we can further distinguish between those norms that offer a full compensation for the violation of another and those that do not. Two types of monotonic reasoning tools can be used for an implementation: a classical propositional model enumerator and a bimodal propositional theorem prover; see, for instance, Jabbour et al. [8] and Otten [13].

Finally, the envisaged future directions to extend and refine the present work include: (I) testing the computational properties of the proposed formalism in comparison with those of other approaches suitable for automated deduction that have been developed in the literature, and (II) applying our formalization procedure to rigorously represent normative theories that can be extracted from larger portions of natural language texts, such as fragments of legal codes. As far as the second direction is concerned, we plan to move from the current modal propositional setting to a modal first-order setting, which would allow for explicit reasoning on relations among different normative parties. Such an extension would be still supported by provers like MleanCoP.

References

1. G. Aucher, G. Boella, and L. van der Torre (2011). A Dynamic Logic for Privacy Compliance. *Artificial Intelligence and Law* **19**(2-3): 187–231.
2. C. Benzmüller, X. Parent and L. van der Torre (2019). Designing Normative Theories of Ethical Reasoning: Formal Framework, Methodology, and Tool Support. *Artificial Intelligence*, forthcoming.
3. C. Boutilier (1994). Toward a Logic for Qualitative Decision Theory. In J. Doyle, E. Sandewall and P. Torasso (eds.), *Proceedings of KR 1994*, pp. 75–86.
4. R.E. Bryant (1986). Graph-Based Algorithms for Boolean Function Manipulation. *IEEE Transactions on Computers*, **35**(8): 677–691.
5. L. Goble (2013). Prima Facie Norms, Normative Conflicts and Dilemmas. In D. Gabbay et al. (eds.), *Handbook of Deontic Logic and Normative Systems*, pp. 241–351.
6. G. Governatori (2018). Practical Normative Reasoning with Defeasible Deontic Logic. In C. d'Amato and M. Theobald (eds.), *Reasoning Web 2018*, pp. 1–25.
7. B. Hansson (1969). An Analysis of Some Deontic Logics. *Noûs* **3**(4): 373–398.
8. S. Jabbour, J. Lonlac, L. Sais, and Y. Salhi (2014). Extending Modern SAT Solvers for Models Enumeration. In J. Joshi, E. Bertino, B.M. Thuraisingham and L. Liu (eds.): *Proceedings of IEEE IRI 2014*, pp. 803–810.
9. R. Kowalski and K. Satoh (2018). Obligation as Optimal Goal Satisfaction. *Journal of Philosophical Logic* **47**(4): 579–609.
10. R. Lee and Y.U. Ryu (1995). DX: A Deontic Expert System. *Journal of Management Information Systems* **12**(1): 145–169.
11. R. Moore (1985). Semantical Considerations on Nonmonotonic Logic. *Artificial Intelligence* **25**(1): 75–94.
12. D. Nute (1997). *Defeasible Deontic Logic*. Dordrecht: Kluwer.
13. J. Otten (2014). MleanCoP: a Connection Prover for First-order Modal Logic. In S. Demri, D. Kapur and C. Weidenbach (eds.), *Proceedings of IJCAR 2014*, pp. 269–276.
14. E. Pacuit, R. Parikh and E. Cogan (2005). The Logic of Knowledge Based Obligation. *Synthese* **149**(2): 311–341.
15. X. Parent and L. van der Torre (2017). Detachment in Normative Systems: Examples, Inference Patterns, Properties. *IfCoLog Journals of Applied Logics* **4**(9): 2995–3038.
16. H. Prakken and M. Sergot (1996). Contrary-to-duty Obligations. in *Studia Logica*, **57**(1), 91–115.
17. R. Reiter (1980). A Logic for Default Reasoning. *Artificial Intelligence* **13**(1–2): 81–132.
18. R.M. Smullyan (1968). *First-Order Logic*. New York: Dover.
19. A. Steen (2021). Goal-Directed Decision Procedures for Input/Output Logics. In F. Liu, A. Marra, P. Portner, and F. Van De Putte (eds.), *Proceedings of DEON 2020/2021*, forthcoming.
20. C. Straßer (2011). A Deontic Logic Framework Allowing for Factual Detachment. *Journal of Applied Logic* **9**(1): 61–80.
21. L. van der Torre (1997). Reasoning About Obligations. Phd Thesis, Erasmus University Rotterdam.
22. L. van der Torre and Y.H. Tan (1997). The Many Faces of Defeasibility in Defeasible Deontic Logic. In D. Nute (ed.) *Defeasible Deontic Logic*, pp. 79–121.