An attempt to adapt a logic of conditional preferences for reasoning with Contrary-To-Duties

Laurence Cholvy

ONERA Centre de Toulouse
Toulouse, France
cholvy@cert.fr

Christophe Garion

ONERA Centre de Toulouse
Toulouse, France
garion@cert.fr

Abstract. This paper presents an attempt to adapt, in the context of deontic reasoning, a logic defined by Boutilier for reasoning with conditional preferences. The first motivation for this work is that deontic logic can be given a semantics in terms of ordered worlds like in this kind of logic: the preference relation among worlds aims at ordering worlds from the most ideal ones to the least ideal ones. The second motivation is that Boutilier introduced a model of an agent’s ability by distinguishing between controllable, influenceable and uninfluenceable propositions. And we noticed that this partition can be related to the notions introduced by Carmo and Jones for reasoning with Contrary-To-Duties. This present work shows an extension of Boutilier’s work in order to use his logic for reasoning with Contrary-To-Duties. An exhaustive study of a benchmark example of CTDs leads us to show that the results obtained with this extension coincide exactly with those obtained by Carmo and Jones.

Keywords: Conditional preferences, Deontic reasoning, Contrary-To-Duties.

1. Introduction

Deontic logic can be given a semantics in terms of ordered worlds, the preference relation ordering worlds from the most ideal ones to the least ideal ones. Hansson, in [5], was among the first to
note this. Prakken and Sergot presented in [9] an approach to adapting a semantics of ordered worlds for reasoning correctly with norms with Contrary-To-Duty (CTD). More recently, van der Torre and Tan [10] defined a preference-based deontic logic, based on Hansson’s semantics to correctly reason with Contrary-To-Duties and moral dilemmas.

The main justification is that, from a semantical point of view, one can consider that the obligations expressed by some norms induce a preference order between worlds. For instance, even in the simplest case of a deontic logic like SDL, one distinguishes in the semantics between the ideal worlds (those in which the obligatory propositions are true) and the others; the ideal worlds being more preferred than the others. Here, there are only two levels of worlds. In the more complex case of norms with CTDs, one needs to reason with primary obligations, and secondary obligations (which arise when the first ones are violated), or more in case of multi-level CTDs. In such a case, one can see this as inducing an ordering between worlds which distinguishes the ideal worlds, the sub-ideals worlds [6] and so on.

In order to illustrate this point of view, let us consider the classical dog scenario:

(a) There ought to be no dog.
(b) If there is no dog, there ought not to be a warning sign.
(c) If there is a dog, there ought to be a warning sign.

If we consider a language with two propositional variables, dog and sign, then the possible worlds are:

\[ w_1 = \{ \text{dog, sign} \}, \ w_2 = \{ \text{dog, \neg sign} \}, \ w_3 = \{ \neg \text{dog, sign} \}, \ w_4 = \{ \neg \text{dog, \neg sign} \}. \]

One meaning\(^1\), in terms of possible worlds, that we can give to the three previous sentences is the following:

Let \( \leq \) be a preference relation.

- Sentence (a) can be interpreted as follows: any world in which there is no dog is more preferred than any world in which there is a dog.

  Thus, \( w_4 \leq w_1 \) and \( w_4 \leq w_2 \) but also \( w_3 \leq w_1 \) and \( w_3 \leq w_2 \).

- Sentence (b) can be interpreted as follows: the world in which there is no dog and no sign is preferred to the world in which there is no dog but a sign, i.e., \( w_4 \leq w_3 \).

  Which leads, with the previous data, to: \( w_4 \leq w_3 \leq w_1 \) and \( w_4 \leq w_3 \leq w_2 \).

- Finally, sentence (c) can be interpreted as follows: the world in which there is a dog and a sign is preferred to the world in which there is a dog but no sign, i.e., \( w_1 \leq w_2 \).

  Which leads, with the previous data, to: \( w_4 \leq w_3 \leq w_1 \leq w_2 \).

Thus, finally, these sentences offer us the only one possibility of ordering worlds:

\[ w_4 \leq w_3 \leq w_1 \leq w_2 \]

\(^1\)In section 6, we will discuss the classical problem of natural language sentence modelisation and the difficulties raised by natural language ambiguities.
Following these ideas, we have found that the work described in [2] was quite interesting. There, Boutilier describes a logic of goals, called CO, whose semantics is defined in terms of worlds ordered by a total pre-order, ranking worlds from the most preferred ones to the least preferred ones.

And precisely, the previous model is a CO model of some sentences expressing (a) (b) and (c) in terms of preferential goals.

One can notice that Boutilier himself mentions that CO logic is based on Hansson’s deontic logic, and that a conditional goal can be given the interpretation of a conditional obligation.

Furthermore, in his work, Boutilier tries to characterize an adequate notion of ideal goals and shows the importance of taking into account a model of the agent’s ability. In doing so, Boutilier introduces the notions of controllable, influenceable and uninfluencable propositions. These notions are then used to define ideal goals. Intuitively speaking, the impact of this model on the definition of ideal goal is twofold: for characterizing ideal goals, one must reason with the propositions which are not influenceable (i.e., those whose truth value the agent cannot change) and furthermore, a goal must be a proposition which is controllable by the agent (i.e., one whose truth value the agent can change).

It seemed interesting to us to relate this model of agent ability to the one introduced by Carmo and Jones [4] who aimed to show its impact on a correct reasoning with CTDs. This relation constitutes the core of our work.

The present paper is organized as follows. In section 2, Boutilier’s logic is presented. We focus on its semantics. We also focus on the model of an agent’s ability and its impact on the definitions of ideal goals. In section 3, we recall the problem of Contrary-To-Duties (CTD) raised in deontic reasoning. We focus on the postulates introduced by Carmo and Jones who express a set of properties a deontic logic should satisfy in order to reason correctly with CTD. Carmo and Jones show the impact of a model of the agent’s ability on the definitions of two kinds of obligations in reasoning with CTD. Section 4 shows how we suggest extending Boutilier’s work in order to fulfill Carmo and Jones’ postulates about a logic for reasoning with norms with CTD. One classical example of reasoning with Contrary-to-Duties is examined in section 5. Finally, section 6 is devoted to a discussion.

2. Boutilier’s logic of goals

In this section, we present the logic CO developed by Craig Boutilier to reason with qualitative statements of preference. We only recall its semantics, the axiomatic presentation being described in [2]. Then we focus on a restriction of CO, named CO*.

2.1. Semantics of CO

Boutilier assumes a propositional bimodal language over a finite set of atomic propositional variables, with the usual connectives and two modal operators denoted □ and ◊. The semantics of CO is based on models of the form $M = \langle W, \leq, \phi \rangle$ where $W$ is a set of possible worlds, $\phi$ is
a valuation function which associates any propositional letter to a set of worlds in which it is true, and \( \leq \) is a total preorder\(^2\) on worlds. \( v \leq w \) means that \( v \) is a world at least as preferred as \( w \).

Let \( M = \langle W, \leq, \phi \rangle \) be a model. The valuation of a formula in \( M \) is given by the following definition:

**Definition 2.1.** 1. \( M \models \alpha \) iff for all \( w \) in \( W \), \( M \models_w \alpha \), for any formula \( \alpha \).

2. \( M \models_w \alpha \) iff \( w \in \phi(a) \), for any propositional letter \( a \).

3. \( M \models_w \neg \alpha \) iff \( M \not\models_w \alpha \), for any formula \( \alpha \).

4. \( M \models_w (\alpha_1 \land \alpha_2) \) iff \( M \models_w \alpha_1 \) and \( M \models_w \alpha_2 \), if \( \alpha_1 \) and \( \alpha_2 \) are formulas.

5. \( M \models_w \Box \alpha \) iff for all \( v \) such that \( v \leq w \), \( M \models_v \alpha \).

6. \( M \models_w \Diamond \alpha \) iff for all \( v \) such that \( w < v \), \( M \models_v \alpha \).

Thus, \( \Box \alpha \) is true in a world \( w \) iff \( \alpha \) is true in all the worlds at least as preferred as \( w \). And \( \Diamond \alpha \) is true in a world \( w \) iff \( \alpha \) is true in all worlds less preferred than \( w \). As usual, the dual operators \( \Diamond \) and \( \Box \) are defined by: \( \Diamond \alpha \equiv_{df} \neg \Box \neg \alpha \) and \( \Box \alpha \equiv_{df} \neg \Diamond \neg \alpha \).

Boutilier also defines \( \Diamond \alpha \equiv_{df} \Box \alpha \land \Box \alpha \) and \( \Box \alpha \equiv_{df} \alpha \lor \Box \alpha \).

**Example.** Let us consider a model \( M \) consisting in four worlds \( w_1, w_2, w_3, w_4 \) ordered as follows:\(^3\)

\[
\begin{array}{ccc}
  w_1 & \leq & w_2 \\
  a & b & a \\
  b & c & b \\
  \neg c & \neg c & \neg c \\
\end{array}
\]

\[
\begin{array}{ccc}
  w_3 & \leq & w_4 \\
  a & \neg b & a \\
  b & \neg c & b \\
  \neg c & c & \neg c \\
\end{array}
\]

Then, \( M \models \Diamond a \) since every world satisfies \( a \). Furthermore, \( M \models_w \Box b \) since any world at least as preferred as \( w_2 \) satisfies \( b \).

**Definition 2.2.** As usual, we say that \( M \) satisfies a formula \( \alpha \) iff \( M \models \alpha \).

**Definition 2.3.** Let \( E \) be a set of formulas and \( \alpha \) a formula of CO. We say that \( \alpha \) is derived (or deduced) from \( E \) iff any model \( M \) which satisfies \( E \) also satisfies \( \alpha \). We note it: \( E \models \alpha \).

---

\(^2\)This means that \( \leq \) is a reflexive, transitive and connected (i.e., total) binary relation over \( W \)

\(^3\)We adopt the usual following notation: the letter \( P \) appears as positive in the world \( w \) iff \( w \in \phi(P) \), and appears as negative if \( w \not\in \phi(P) \)
As Boutilier noted it in [1], when ranking states of affairs, we want often to take all logically possible worlds into consideration. For this reason, he considered a class of $CO$-models in which each propositional truth valuation is witnessed by some possible world.

This leads to the definition of logic $CO^*$ which is the smallest extension of $CO$ closed under all rules of $CO$ and containing the following axiom:

- $\Diamond \alpha$ for all satisfiable proposition $\alpha$.

In the rest of the paper, we will focus on $CO^*$ instead of $CO$.

### 2.3. Conditional preferences

In order to express conditional preferences, Boutilier defines a conditional connective $I(-|-)$ by the following definition:

**Definition 2.4.**

$$ I(\beta|\alpha) \equiv_{df} \neg \alpha \lor \Diamond (\alpha \land \Box (\alpha \rightarrow \beta)) $$

In a first approximation, we can interpret $I(\beta|\alpha)$ by “if $\alpha$ then the agent should ensure $\beta$”.

Let us recall some more definitions that will be useful in the following.

**Definition 2.5.** An absolute preference $\alpha$ is defined as $I(\alpha|\top)$ (or equivalently $\Diamond \Box \alpha$) and abbreviated $I(\alpha)$.

In semantical terms, this means that the most preferred worlds are $\alpha$ worlds.

**Definition 2.6.** The relative preference between two propositions is defined by:

$$ \alpha \leq_P \beta \equiv_{df} \Diamond (\beta \rightarrow \Diamond \alpha) $$

$\alpha \leq_P \beta$ means that $\alpha$ is at least as preferred as $\beta$ (i.e., the best $\alpha$-worlds are at least as good as the best $\beta$-worlds).

In order to determine his/her actual goals, an agent must have knowledge about the real world. Boutilier then considers $KB$, a consistent finite set of formulas written without any modality, in order to represent the knowledge the agent has about the real world. $KB$ is called a knowledge base. Considering $KB$ and a $CO^*$-model, the ideal situations are characterized by the most preferred worlds which satisfy all the agent’s knowledge. This is formally defined by the following definitions:
Definition 2.7. Let $L_{CP}$ be the propositional sublanguage of $CO^*$. Let $KB$ be a knowledge base. $Cl(KB)$ is defined by:

$$Cl(KB) = \{ \alpha \in L_{CP} \text{ such that } \models KB \rightarrow \alpha \text{ and } \alpha \text{ is a literal} \}$$

Definition 2.8. Let $E$ be a set of conditional preferences. Let $KB$ be a knowledge base. An ideal goal, derived from $E$, is a formula $\alpha$ of $L_{CP}$ such that: $E \models I(\alpha | Cl(KB))$

Example. Let $L$ be a language with propositional variables $r$ (the agent crosses the main road) and $b$ (the agent bikes). Consider the two conditionals $I(r)$ and $I(\neg r | b)$ which express that, generally the agent prefers to cross the main road (to go to the university for instance), but if he bikes, he prefers not to cross the main road.

The possible worlds are $w_1 = \{r, \neg b\}$, $w_2 = \{\neg r, b\}$, $w_3 = \{r, b\}$ and $w_4 = \{\neg r, \neg b\}$

Because of $I(r)$, $w_1$ and $w_3$ can be the most preferred worlds. But, due to $I(\neg r | b)$, $w_3$ cannot be the most preferred world. So $w_1$ is the most preferred world. This leads to $w_1 \leq w_2$, $w_1 \leq w_3$ and $w_1 \leq w_4$. Besides that, we cannot have $w_3 \leq w_2$ because of $I(\neg r | b)$, so we have $w_2 \leq w_3$. Thus, $I(r)$ and $I(\neg r | b)$ are only satisfied in the following $CO^*$-models:

- $M_1: w_1 \leq w_2 \leq w_3 \leq w_4$
- $M_2: w_1 \leq w_2 \leq w_4 \leq w_3$
- $M_3: w_1 \leq w_4 \leq w_2 \leq w_3$

Suppose that $KB_1 = \{\neg b\}$ (i.e., the agent does not bike). Then, $Cl(KB_1) = \{\neg b\}$. So, the ideal goals are the formulas $\alpha$ such that $I(\alpha | \neg b)$ is satisfied in all the previous models, i.e., the ideal goal is here $r$. This means that, since he/she does not bike, the agent prefers to cross the main road.

Suppose now that $KB_2 = \{b\}$ (i.e., the agent bikes). Then, we can now deduce that $\neg r$ is an ideal goal. This means that, since he/she bikes, the agent prefers not to cross the main road.

2.4. Controllable, influenceable and uninfluenceable propositions

By the previous definition of ideal goals, every formula $\alpha$ which satisfies $E \models I(\alpha | Cl(KB))$ is a goal. But, as Boutilier notes, this definition is fair only if $KB$ is fixed. If the agent can change the truth of some elements in $KB$, ideal goals as previously defined may be too restrictive. For instance, $KB_2$ expresses that the agent is biking. Assume that he may change his mind and not to use his bike (this is a quite reasonable assumption here). In this case, he may prefer not use his bike in order to fulfill the most preferred goal, i.e., crossing the main road.

So, for computing the closure of the knowledge base $Cl(KB)$, Boutilier suggests considering only the formulas whose truth cannot be changed by the agent's actions.

---

4In fact, Boutilier defines $Cl(KB)$ in a more general way: he uses a non monotonic logic to deduce the default knowledge of the agent. Here, in order to simplify the presentation, we choose to reduce his point of view to classical logic.
Furthermore, as Boutilier notes, an agent’s actions must play a role in factoring out unachievable goals. For instance, an agent might prefer that it does not rain; but this is something over which he has no control. Though it is an ideal outcome, calling this a goal is unreasonable.

To capture such distinctions, Boutilier introduces a simple model of action and ability to demonstrate its influence on conditional goals. He suggests partitioning the atomic propositions in two classes: \( P = C \cup \overline{C} \), in which \( C \) is the set of atomic propositions that the agent can control (i.e., the agent can change the truth value of those propositions) and \( \overline{C} \) is the set of atomic propositions that the agent cannot control.

For instance the atomic proposition representing the fact the agent bikes can be considered as controllable. The atomic proposition representing the fact it rains can reasonably be considered as uncontrollable.

Then Boutilier generalises this notion as follows:

**Definition 2.9.** For any set of atomic variables \( P \), let \( V(P) \) be the set of truth assignments to this set. If \( v \in V(P) \) and \( w \in V(Q) \) for disjoint sets \( P \) and \( Q \), then \( v; w \in V(P \cup Q) \) denotes the obvious extended assignment. The neutral element of \( ; \) is \( V(\emptyset) \) which is the empty truth assignment.

**Definition 2.10.** Given \( C \) and \( \overline{C} \), a proposition \( \alpha \) is controllable iff, for every \( u \in V(C) \), there is some \( v \in V(C) \) and \( w \in V(C) \) such that \( v; u \models \alpha \) and \( w; u \models \neg \alpha \).

A proposition \( \alpha \) is said uncontrollable iff it is not controllable.

A proposition \( \alpha \) is influenceable iff, for some \( u \in V(C) \), there is some \( v \in V(C) \) and \( w \in V(C) \) such that \( v; u \models \alpha \) and \( w; u \models \neg \alpha \).

A proposition \( \alpha \) is influenceable iff it is not influenceable.

For instance, if \( a \in C \) and \( b \in \overline{C} \), then \( a \wedge b \) is influenceable but not controllable. Indeed \( a \wedge b \) is not controllable because \( b \) is not controllable and if \( b \) is false the agent cannot make \( a \wedge b \) be true. But it is influenceable because if \( b \) is true then the agent can influence the truth value of \( a \wedge b \) by assigning \( a \) to true or false.

**Definition 2.11.** The uninfluenceable belief set of an agent is defined by:

\[
UI(KB) = \{ \alpha \in Cl(KB) \text{ such that } \alpha \text{ is uninfluenceable} \}
\]

In a first part of his work, Boutilier considers that \( UI(KB) \) is a complete set, i.e. the truth value of all uncontrollable (not controllable) atoms is known, which is expressed by:

\[
\forall \alpha \text{ uncontrollable, } \alpha \in UI(KB) \text{ or } \neg \alpha \in UI(KB)
\]

Under this hypothesis, Boutilier refines the notion of goals and introduces the notion of \( CK\)-goal as follows:

\footnote{In a second part of his work, Boutilier also examines the case when \( UI(KB) \) is not complete but we will not consider it in detail here.}
Definition 2.12. Let $E$ be a set of conditional preferences and $KB$ a knowledge base such that $UI(KB)$ is complete. Then a proposition $\alpha$ is a CK-goal derived from $E$ iff $\alpha$ is controllable and $E \models I(\alpha|UI(KB))$.

Example (continued). Let us again consider $E = \{I(r), I(\neg r|b)\}$. Assume here that the agent does not bike because his bike has been stolen.

First assume that crossing the road is still in the agent’s ability. Thus $b$ is uncontrollable, but $r$ is not. So, by definition 2.12, $r$ is a CK-goal derived from $E$: the agent will have to cross the main road.

Now, assume that there are works on the main road and it is impossible for anybody to cross it. This is expressed by considering that $r$ is now uncontrollable. Here, according to definition 2.12, $r$ (crossing the road) is not a CK-goal derived from $E$ due to the fact that the agent cannot cross the main road (and this is beyond his control).

3. Contrary-to-Duties

Formalisation of Contrary-to-Duties (CTD) is a central point in the study in deontic logic. In this section, we mainly sum up Carmo and Jones’s work [4].

3.1. Postulates defined by Carmo and Jones

A CTD can be modelled by a primary norm and a secondary norm which takes effect when the first norm is violated. One famous CTD example is the Chisholm’s paradox:

(a) X ought to go to help your neighbours.
(b) It ought to be that if X goes, he tells them he is coming.
(c) If X does not go, he ought not to tell them he is coming.
(d) X does not go.

If X does not go to help his neighbours, he ought not to tell them he is coming. In this case, he violated the first norm which compelled him to help his neighbours.

The first attempt to model CTD was by using SDL logic, as follows:

(a) $O(help)$
(b) $O(tell|help)$
(c) $O(\neg tell|\neg help)$
(d) $\neg help$

where the conditional $O(A|B)$ can be represented by one of the two following ways:

1. $O(B|A) \equiv_{def} A \rightarrow OB$
\( O(B|A) \equiv_{df} O(A \rightarrow B) \)

Unfortunately, the two previous ways of modelling CTD lead to inconsistencies or dependencies between the four sentences which are properties unanimously rejected.

Some attempts have been done to use a temporal deontic logic to solve the Chisholm paradox. But like Prakken and Sergot in [8], Carmo and Jones reject this formalisation, because not all examples of CTD are temporal ones.

Besides this, some tried to use nonmonotonic logics to solve CTDs problems. But this approach has been criticized by many people insisting on the fact that the rule (c) is not an exception to the rule (a) (which is a typical application of nonmonotonic logic), but is a case of violation of a prima facie obligation.

Another problem which has been discussed concerns the representation of (b) and (c). According to Prakken and Sergot [8], (b) and (c) must have different representations since (c), unlike (b), is a contrary-to-duty conditional expressing what should hold in case of violation of (a). Carmo and Jones reject this point of view because it makes the representation dependent on updates: if one wants to introduce new CTDs or remove norms, one has to modify the norms already expressed.

This leads Carmo and Jones to express a first set of postulates that must be respected by a logic for correctly reasoning with CTDs:

1. the set of formulas (a), (b), (c) and (d) must be consistent;
2. the formulas (a), (b), (c) and (d) must be logically independent, i.e. none of them is a logical consequence of some others;
3. the logic must apply to timeless and actionless CTD-examples;
4. sentences (b) and (c) must have a similar structure;

Furthermore, as Carmo and Jones point out, the CTD examples show the existence of two kinds of obligations, they call ideal and actual. For instance, in the Chisholm set, we need to be able to deduce that under ideal circumstances, the ideal obligation of the agent is to help his neighbours and to tell them he is coming, but under the certain circumstance (i.e., the violation of (a)) the actual obligation of the agent is not to tell his neighbours he is coming. So Carmo and Jones add three postulates:

1. capacity to derive actual obligations;
2. capacity to derive ideal obligations;
3. capacity to represent the fact that a violation has occurred;

Finally, Carmo and Jones address the "pragmatic oddity" problem [7], raised by some formalisations of CTD, in particular in SDL. This problem comes from the fact that some formalisations of CTD structures assume only one kind of obligation and then do not allow the derivation of different levels of obligations. For instance, in the Chisholm example, the "pragmatic oddity problem" can be illustrated by the following reasoning: the agent must help his neighbours
(this is said by the first rule). Furthermore, since he does not help his neighbours, he must not tell them he is coming. So, the agent must help his neighbours and must not tell them he is coming. Which is a bit strange. So, Carmo and Jones add a last postulate which an adequate representation of CTDs should satisfy:

(viii) capacity to avoid “pragmatic oddity”;

3.2. Carmo and Jones’s logic

In [4], Carmo and Jones define a logic to reason with CTDs and which satisfies the previous postulates. This logic is based on a dyadic conditional deontic operator $O$ which allows one to express norms and two monadic deontic operators $O_i$ and $O_a$ to express respectively ideal obligation and actual obligation.

The main question is: what are the contexts that allow one to derive the two kinds of obligation, i.e. actual and ideal obligation? To answer that question, Jones and Carmo use two kinds of necessity operators called here “agent-dependent-necessity” and “agent-independent-necessity” and respectively denoted here\(^6\) $\Box_a$ and $\Box_i$.

- The agent-dependent-necessity

  The operator of agent-dependent-necessity is denoted here $\Box_a$ and its dual connective $\Diamond_a$. Intuitively, $\Box_a \alpha$ expresses that the proposition $\alpha$ is fixed, in a certain situation, given what the agent decides to do or not to do.

  In the Chisholm scenario, the fact that the agent does not go to help his neighbours and decides not to go can be formulated by $\Box_a \neg \text{help}$. On the other hand, $\Diamond_a \text{help}$ expresses the fact that the agent did not decide not to go to help his neighbours.

  This operator is used to derive actual obligations through the following axiom:

  $$O(\beta|a) \land \Box_a \alpha \land \Diamond_a \beta \land \Diamond_a \neg \beta \rightarrow O_a \beta$$

  In the context of the Chisholm scenario, an instance of this axiom is the following: knowing that if the agent does not go to help his neighbours then he must not tell them he is coming, the agent decided not to help them, the agent did not decide to tell them he is coming nor not to tell them he is coming, then we can infer that the agent has the actual obligation not to tell his neighbours he is coming. Furthermore, this axiom does not allow one to derive the actual obligation not to tell his neighbours he is coming if the agent can change his mind and, by one of his actions, help his neighbours.

\(^6\)We change the notations because Carmo and Jones use some symbols that Boutilier already uses in his work for different purposes.
• The agent-independent necessity

The operator of agent-independent-necessity is denoted $\Box_i$ here and its dual connective $\Diamond_i$. Intuitively, $\Box_i \alpha$ expresses that the proposition $\alpha$ is not fixed by the agent’s decisions, but is fixed whatever the agent could decide.

For instance, in the Chisholm example, if the agent cannot move (he is ill for example), then it is impossible for him to help its neighbours, independently of his will.

The agent-independent-necessity is used to derive ideal obligations as shown by the following axiom:

$$O(\beta | \alpha) \wedge \Box_i \alpha \wedge \Diamond_i \beta \wedge \Diamond_i \neg \beta \rightarrow O_i \beta$$

For instance, knowing that if the agent does not help his neighbours then he must not tell them he is coming, if the agent cannot help them (because he is ill), and if he can tell them or not tell them (the telephone works) then he ideally must not tell them he is coming.

One must notice that the agent-independent-necessity and the agent-dependent-necessity are two different notions. Indeed, the fact that the agent decides not to help his neighbours ($\Box_\omega \neg \text{help}$) is different from the fact that the agent cannot help them ($\Box_i \neg \text{help}$). Notice however that $\vdash \Box_i \alpha \rightarrow \Box_\omega \alpha$, i.e. if it is necessary that $\alpha$ beyond the agent’s will, then $\alpha$ is fixed whatever the agent could decide.

Carmo and Jones define a semantics and an axiomatics of a logic for deriving actual obligations, ideal obligations (see [4] for more details). Furthermore, they define the notion of violation by:

$$\text{viol}(A) =_{df} O_i A \wedge \neg A$$

That is, $A$ is violated if $A$ is ideally obligatory but $A$ is not true.

4. Adaptation of Boutilier’s formalism to represent CTDs

Here, we show an adaptation of Boutilier’s work which fulfills the previous postulates, if, as we will see, we accept to reformulate them. This adaptation follows Carmo and Jones’s ideas.

4.1. Modelling CTDs in $CO^*$

We present the way of modelling CTDs in $CO^*$ on a particular example which is the dog scenario:

(a) There ought to be no dog.
(b) If there is no dog, there ought to be no sign.
(c) If there is a dog, there ought to be a sign.
There is a dog

In the introduction, we have described one meaning that can be given to the first three sentences. This meaning led to one possibility of ordering worlds. It happens that this order between possible worlds is the only $CO^*$-model of the four following $CO^*$ formulas:

(a) $I(\neg \text{dog})$
(b) $I(\neg \text{sign} | \neg \text{dog})$
(c) $I(\text{sign} | \text{dog})$
(e) $(\neg \text{dog} \land \text{sign}) \preceq_P (\text{dog} \land \text{sign})$

Then this is the way we suggest to model the first three sentences.
And we suggest to represent the fact that there is a dog (sentence (d)) by:

(d) $\text{dog} \in KB$

Notice that here, the first three sentences are modelled by four formulas. Thus, literally, we cannot check if this modelisation respect Carmo and Jones’s first two postulates. However, it respects them if we accept to slightly reformulate them as follows:

1. The set of formulas which model the CTD must be consistent
2. The formulas which model the CTD must be logically independent

One can check than the $CO^*$ formulas (a), (b) (c) and (e) are consistent and none of them is a logical consequence of some others. However, there is, in some sense, a dependence between these formulas. Some comments about this will be made in section 6.

Furthermore, this modelisation respects postulates (iii) and (iv).

4.2. Back to the controllability and influenceability of propositions

As already mentioned, Boutilier partitions the propositions in three classes: controllable, influenceable and uninfluenceable. Our goal in this section is to find a relation between Boutilier’s notions of controllability and influenceability, and Carmo and Jones’s notions of necessity.

**Assumption.** First of all, let us insist on the fact that comparison is done under the assumption that the knowledge on the real world is complete. That is to say, if $KB$ denotes the description of the real world, for every atomic proposition $a$, we have $\vdash KB \rightarrow a$ or $\vdash KB \rightarrow \neg a$.

Let us first notice that if a proposition $a$ is controllable (resp. influenceable, uninfluenceable), then $\neg a$ is controllable (resp. influenceable, uninfluenceable) too. The proof is immediate.

In the following, we examine the different cases that can arise for a given propositional formula $a$. 
1. \( \alpha \) is a uninfluenceable propositional formula (in the sense of Boutilier)

(a) Suppose that \( \vdash KB \to \alpha \) (to simplify, we say that \( \alpha \) is true in the real world or simply, true), then whatever the agent can do, \( \alpha \) will be true. This means that \( \alpha \) is “agent-independent-necessarily true” to use the terminology we introduce in section 3. So:

**Proposition 4.1.** \( \{ \alpha : \vdash KB \to \alpha \text{ and } \alpha \text{ uninfluenceable} \} \equiv_{df} \{ \alpha : \vdash \Box_i \alpha \} \)

(b) Suppose now that \( \not\vdash KB \to \alpha \). Then, because \( KB \) is complete, we have \( \vdash KB \to \neg \alpha \). \( \alpha \) is uninfluenceable, so \( \neg \alpha \) is uninfluenceable too. Whatever the agent will do, \( \alpha \) will be false.

**Proposition 4.2.** \( \{ \alpha : \not\vdash KB \to \alpha \text{ and } \alpha \text{ uninfluenceable} \} \equiv_{df} \{ \alpha : \vdash \Box_i \neg \alpha \} \)

2. \( \alpha \) is influenceable.

\( \alpha \) may be controllable or not.

(a) \( \alpha \) is controllable

i. Suppose that \( \vdash KB \to \alpha \).

Since \( \alpha \) is controllable, the agent can decide to maintain \( \alpha \) true or can decide to change it to false. For representing this, we introduce two new notions in addition to Boutilier’s:

A. We say that \( \alpha \), true in the actual situation, is *controllable-and-fixed* if the agent decides to keep \( \alpha \) true.

For instance, if \( \alpha \) represents the fact “the agent does not help his neighbours” and if the actual situation is such that the agent does not help his neighbours and decides not to help them, then we consider that \( \alpha \in KB \) and \( \alpha \) is controllable-and-fixed.

**Proposition 4.3.** \( \{ \alpha : \vdash KB \to \alpha \text{ and } \alpha \text{ controllable-and-fixed} \} \equiv_{df} \{ \alpha : \vdash \alpha \land \Box_i \alpha \} \)

B. We say that \( \alpha \), true in the actual situation, is *controllable-and-unfixed* if the agent can decide to change the truth of \( \alpha \).

For instance, if \( \alpha \) represents the fact “the agent does not help his neighbours” and if the actual situation is such that the agent does not help his neighbours but accepts the idea that he can help them, then we consider that \( \alpha \in KB \) and \( \alpha \) is controllable-and-unfixed.

\(^7\)Note that we could also write \((\alpha \land \Box_i \alpha)\), but this is equivalent, because the connective \( \Box_i \) is KT-type.
Proposition 4.4. \{\alpha : \vdash KB \rightarrow \alpha \text{ and } \alpha \text{ is controllable-and-unfixed}\} \equiv_{def} \{\alpha : \vdash \alpha \land \Diamond \neg \alpha\}

ii. In the case where \(\alpha\) cannot be deduced from \(KB\), we have \(\vdash KB \rightarrow \neg \alpha\) (because \(KB\) is complete).

\(\alpha\) is controllable, so is \(\neg \alpha\). We can introduce the same previous two cases (A) and (B) depending on the fact that \(\neg \alpha\) (or equivalently \(\alpha\)) is controllable-and-fixed or controllable-and-unfixed.

(b) If \(\alpha\) is not controllable (but influenceable)

i. Assume \(\vdash KB \rightarrow \alpha\).

\(\alpha\) is true in the actual situation. Since \(\alpha\) is influenceable, there are two cases: the agent can or cannot change the truth of \(\alpha\).

Let us illustrate this by the following example. Assume that \(p\) is a controllable variable and \(q\) an uncontrollable variable. Assume that \(KB = \{p, \neg q\}\). The formula \(p \land q\) is false in \(KB\) and will remain false whatever the value the agent will give to \(p\). We will say that \(p \land q\) is uninfluenceable in \(KB\). In contrast, the formula \(p \lor q\) is true in \(KB\) but the agent can change its truth value by changing \(p\) to false or can keep it to true.

The details are as follows:

A. If the actual situation is such that the agent cannot change the truth of \(\alpha\), then \(\alpha\) will be true whatever the agent can do. It is the same case as \(\alpha\) is true and uninfluenceable. In this case, we say that \(\alpha\) is uninfluenceable in \(KB\).

B. If the actual situation is such that the agent can change the truth of \(\alpha\), there are two cases: the agent can keep \(\alpha\) true (case 2.a.i.A) and \(\alpha\) is controllable-and-fixed or can change its truth (case 2.a.i.B) and \(\alpha\) is controllable-and-unfixed.

ii. Finally, assume \(\nabla KB \rightarrow \alpha\).

Then because \(KB\) is complete, we have \(\vdash KB \rightarrow \neg \alpha\), so \(\alpha\) is false in the actual situation. We can do the same study as before and deduce that \(\alpha\) can be uninfluenceable in \(KB\), controllable-and-fixed or controllable-and-unfixed.

Summary

After this comparison, we find the following relations:

- the true and uninfluenceable propositions and the true and uninfluenceable in \(KB\) propositions both correspond to the agent-independent-necessarily true propositions.
- true and controllable-and-fixed propositions correspond to the “true and agent-dependent-necessarily true” propositions.
• the notion of true and controllable-and-unfixed propositions correspond to the “true and agent-dependent-possibly false” propositions.

4.3. Definitions

Let us extend Boutilier’s definitions taking into account the new model of agent ability defined above. We define ideal obligations, violations and actual obligations as follows:

Definition 4.1. $UI_1(KB)$ is the set of propositions that are true and uninfluenceable or uninfluenceable in $KB$. $UI_2(KB)$ is the set of propositions that are true and controllable-and-fixed or true and uninfluenceable or uninfluenceable in $KB$.

Ideal obligations are defined by:

Definition 4.2. Let $E$ be a set of conditionals. The ideal obligations deduced from $E$ are defined by the following set:

$$O_i \equiv_{def} \{O_i \alpha : E \models I(\alpha | UI_1(KB)) \text{ and } \alpha \text{ is controllable} \}$$

Actual obligations are defined by:

Definition 4.3. Let $E$ be a set of conditionals. The actual obligations deduced from $E$ are defined by the set:

$$O_a \equiv_{def} \{O_a \alpha : E \models I(\alpha | UI_2(KB)) \text{ and } \alpha \text{ is controllable-and-unfixed} \}$$

Violations are defined by:

Definition 4.4.

$$\text{viol } \alpha \equiv_{def} O_i \alpha \text{ and } (\models KB \rightarrow \neg \alpha)$$

4.4. Semantical point of view

We can reformulate the previous definitions in terms of $CO^*$’s ordered models as follows:

Definition 4.5. Let $E$ be a set of preferential conditionals and $M_1...M_n$ all its possible models. Let $KB$ be the current situation.

Let $U$ denote the set of propositions which are true in $KB$ and uninfluenceable or uninfluenceable in $KB$.

Let $CF$ denote the set of propositions which are true in $KB$ and controllable-and-fixed.

For $i \in \{1...n\}$, $M^1_i$ is defined as the restriction of $M_i$ to $U$-worlds.

For $i \in \{1...n\}$, $M^2_i$ is defined as the restriction of model $M_i$ to $U \land CF$-worlds.

---

8 Recall the definition: $E \models \alpha$ iff any model $M$ which satisfies $E$ also satisfies $\alpha$

9 Remember that each model is a set of ordered worlds.

10 $M^1_i$ is obtained by considering $M_i$ and deleting any world which does not satisfy $U$

11 $M^2_i$ is obtained by considering $M_i$ and deleting any world which does not satisfy $U \land CF$
Ideal obligations are defined by:

**Definition 4.6.** \( O_i =_{df} \{ O_i \alpha : \forall i \in 1...n, M_i^1 \models I(\alpha) \text{ and } \alpha \text{ is controllable} \} \)

In other words, \( \alpha \) is ideally obligatory if \( \alpha \) is controllable (i.e. the agent can control its truth value) and \( \phi \) is satisfied in each preferred world of each \( M_i^1 \), i.e. in each preferred world of models of the norms, restricted to worlds which satisfy the propositions true in \( KB \) and uninefluenceable or uninefluenceable in \( KB \).

Actual obligations are defined by:

**Definition 4.7.** \( O_a =_{df} \{ O_a \alpha : \forall i \in 1...n, M_i^2 \models I(\alpha) \text{ and } \alpha \text{ is controllable-and-unfixed} \} \)

In other words, \( \alpha \) is actually obligatory if, \( \alpha \) is controllable-and-unfixed (i.e. the agent did not decide not to change its truth value) and \( \alpha \) is satisfied in each preferred world of each \( M_i^2 \) i.e. in each preferred world of the models of the norms, restricted to worlds which satisfy the propositions true in \( KB \), uninefluenceable or uninefluenceable in \( KB \), or controllable-and-fixed.

Violations are defined by:

**Definition 4.8.**

\[ \text{viol } \equiv_{df} O_i \alpha \text{ and } (\vdash KB \rightarrow \neg \alpha) \]

We can prove that the definitions given in section 4.3 are equivalent to the definitions given here, in terms of restricted models.

These definitions will be illustrated in the next section, in which we examine some interesting cases that can happen in the dog scenario. We will show that the last four postulates listed by Carmo and Jones are satisfied.

5. **Study of an example**

The first three sentences of the dog scenario are modelled by the four \( CO^* \)-formulas:

\[
(a) I(\neg \text{dog}) \quad (b) I(\neg \text{sign} \neg \text{dog}) \quad (c) I(\text{sign} \neg \text{dog}) \quad (\varepsilon)(\neg \text{dog} \land \text{sign}) \leq P (\text{dog} \land \text{sign})
\]

As already mentioned, the only \( CO^* \)-model which satisfies these sentences is shown in figure 1.

Let us examine some cases that can happen. In each case, we characterize what are the ideal obligations, violations and actual obligations.
5.1. First case: $KB = \{\text{dog, sign}\}$

1. The variables $\text{dog}$ and $\text{sign}$ are uncontrollable.

   In this case, the presence of the dog and the sign is imposed on the agent. So, since no variable is controllable, there is no ideal obligation, thus no violation, and no actual obligation.

2. $\text{dog}$ is uncontrollable and $\text{sign}$ is controllable-and-fixed.

   The presence of the dog is imposed. The agent put a sign and decided not to remove it.

   If we restrict the previous model to $\text{dog}$-worlds, we get only one restricted model, say $M_1^1$, in which the preferred world is $\{\text{dog, sign}\}$.

   If we restrict it to $\text{dog} \land \text{sign}$-models, we get only one model, say $M_2^1$, with only one world $\{\text{dog, sign}\}$.

   Thus, the ideal obligation is $O_i \text{sign}$.

   Since there is a sign in the current situation, there is no violation.

   Finally, since $\text{sign}$ is controllable-and-fixed, there is no actual obligation.

   Intuitively, this means that having the sign is ideally required, and, since we are in a situation where there is the sign, and the agent decided not to remove it, there is no actual obligation about that sign.

3. $\text{dog}$ is uncontrollable and $\text{sign}$ is controllable-and-unfixed.

   The presence of the dog is still imposed. The agent put a sign but here, he has not decide to keep that sign (i.e., he can still change his mind and remove it).

   If we restrict the previous model to $\text{dog}$-worlds, we get only one restricted model $M_1^1$, in which the preferred world is $\{\text{dog, sign}\}$. Furthermore, we get a restricted model $M_2^2$ which is identical to $M_1^1$.

   Thus, again, the ideal obligation is $O_i \text{sign}$.

   Since there is a sign in the current situation, there is no violation.

   But now, the actual obligation is $O_a \text{sign}$.

   This means that, in $M_2^2$, even if the world which corresponds to the current situation is the most preferred one, since the agent may decide to remove the sign, the situation can
change to a less preferred world, in which there is no sign. So, if we want the situation to be the most preferred world, we must require, actually, that the agent does not remove the sign.

4. *dog* is controllable-and-fixed and *sign* is uncontrollable

Here, the presence of the sign is imposed on the agent who has got a dog and decided not to get rid of it.

The restricted model $M^1_1$ is $\{\neg\text{dog}, \text{sign}\} \leq \{\text{dog}, \text{sign}\}$.

The most preferred world is $\{\neg\text{dog}, \text{sign}\}$, so there is an ideal obligation which is $O_1 \neg\text{dog}$.

Because $KB = \{\text{dog}, \text{sign}\}$, $O_1 \neg\text{dog}$ is violated: $\text{Viol}\neg\text{dog}$.

There is only one restricted model $M^2_1$ where the unique world is $\{\text{dog}, \text{sign}\}$. But, since *dog* is controllable-and-fixed, there is no actual obligation.

This means that the current situation does not correspond to the most preferred world, so the agent violates the ideal obligation of not having a dog. Furthermore, since the agent has decided not to get rid of the dog, there is no actual obligation.

5. *dog* is controllable-and-unfixed and *sign* is uncontrollable

Again, the presence of the sign is imposed on the agent who has got a dog, but here, he may decide to get rid of it.

Again, there is only one ideal obligation, $O_1 \neg\text{dog}$, which is violated.

The difference between this case and the former one, is that because *dog* is controllable-and-unfixed, we can derive $O_1 \neg\text{dog}$.

This means that as the agent has not decided to keep his dog, he has now the actual obligation to get rid of it.

In terms of ordered worlds, this can be reformulated as follows: given what is independent of the agent or fixed by the agent (here, there is a sign), the most ideal world is $\{\neg\text{dog}, \text{sign}\}$. So, the current situation is not the most ideal world. But here, the agent, by one of his action (getting rid of his dog), can reach this world, he has the actual obligation to do so.

6. *dog* and *sign* are controllable-and-fixed

We can deduce that $O_1 \neg\text{dog}$ and $O_1 \neg\text{sign}$.

Both are violated (since the agent has a dog and put a sign and decided not to remove them), but there is no actual obligation.

7. *dog* is controllable-and-fixed and *sign* is controllable-and-unfixed

We can deduce that $O_1 \neg\text{dog}$ and $O_1 \neg\text{sign}$.

\[\text{Notice that in Carmo and Jones logic, we can only derive the ideal obligation of } \neg\text{dog} \land \neg\text{sign}, \text{ which does not imply } O_1(\neg\text{dog}) \text{ and } O_1(\neg\text{sign}).\]
Both $O_i\neg\text{dog}$ and $O_i\neg\text{sign}$ are violated.

Furthermore, there is an actual obligation: $O_a\text{sign}$.

In other words, the agent violates the ideal obligation of not having a dog, but since he does have one and has decided not to get rid of it, he does not have the actual obligation to get rid of his dog. Besides, he violates the ideal obligation of not having a sign, but since he has not decided to keep his sign or to remove it, he has the actual obligation not to remove it.

8. $\text{dog}$ is controllable-and-unfixed and $\text{sign}$ is controllable-and-fixed

We can still deduce that $O_i\neg\text{dog}$ and $O_i\neg\text{sign}$ and that both ideal obligations are violated. Furthermore, here we can derive an actual obligation: $O_a\neg\text{dog}$. The agent has to get rid of his dog.

In terms of ordered worlds, this can be explained as it has been done in case 5.

9. $\text{dog}$ and $\text{sign}$ are controllable-and-unfixed

We still have $O_i\neg\text{dog}$ and $O_i\neg\text{sign}$.

Both ideal obligations are violated.

The actual obligations are $O_a\neg\text{dog}$ and $O_a\neg\text{sign}$.

That means that since there must be no dog and no sign, ideally, even if there were a dog and a sign, since the agent can remove both of them, he has the actual obligation to do so.

5.2. Second case: $KB = \{\text{dog, } \neg\text{sign}\}$

1. $\text{dog}$ is uncontrollable and $\text{sign}$ is controllable-and-unfixed.

The presence of a dog is imposed on the agent who did not put a sign but can change his mind and put one.

Restricting the model to $\text{dog}$-worlds leads to a model in which the world $\{\text{dog, sign}\}$ is preferred to the world $\{\text{dog, } \neg\text{sign}\}$.

So, the ideal obligation is $O_i\text{sign}$.

Here, this obligation is violated since there is no sign in the current situation. To move to the most preferred world from the actual one, the agent has to put a sign.

As the agent can decide to put a sign or not, the actual obligation is $O_a\text{sign}$. I.e, the agent is required actually to do so.

2. $\text{dog}$ is controllable-and-fixed and $\text{sign}$ is uncontrollable

The absence of a sign is imposed on the agent: for instance, there is no possibility for the agent to put such a warning sign because the service which delivers them stopped their production.
The agent has got a dog and decided not to get rid of it.

Restricting the model to \(\neg\)\textit{sign}\-worlds leads to a model in which the world \(\{\neg\text{dog}, \neg\text{sign}\}\) is preferred to the world \(\{\text{dog}, \neg\text{sign}\}\).

So, the ideal obligation is \(O_i\neg\text{dog}\). And this is violated since the current situation is not the preferred one.

There is no actual obligation.

3. \textit{dog} is controllable-and-unfixed and \textit{sign} is uncontrollable

Again, the absence of a sign is imposed on the agent. But here, the agent has got a dog but may decide to get rid of it.

The restricted model \(M_1^1\) here is the same as before. So the ideal obligation is still \(O_i\neg\text{dog}\).

This ideal obligation is violated since the current world is \(\{\text{dog}, \neg\text{sign}\}\).

The restricted model \(M_2^1\) is identical to \(M_1^1\). So the actual obligation is \(O_a\neg\text{dog}\): the agent must actually get rid of the dog.

5.3. Third case: \(KB = \{\neg\text{dog}, \text{sign}\}\)

1. \textit{dog} and \textit{sign} are uncontrollable.

As before, there is no ideal obligation, no violation and no actual obligation since everything is outside the agent’s power.

2. \textit{dog} is uncontrollable and \textit{sign} is controllable-and-fixed.

The restricted model \(M_1^1\) is such that the world \(\{\neg\text{dog}, \neg\text{sign}\}\) is preferred to the world \(\{\neg\text{dog}, \text{sign}\}\). So the ideal obligation is \(O_i\neg\text{sign}\).

This is violated since the current situation (there is no dog but a sign) is not the most preferred one.

There is no actual obligation.

3. \textit{dog} and \textit{sign} are controllable-and-unfixed.

The ideal obligations are still \(O_i\neg\text{dog}\) and \(O_i\neg\text{sign}\). Both of them are violated.

The actual obligations are \(O_a\neg\text{dog}\) and \(O_a\neg\text{sign}\) (i.e., the agent must actually not buy a dog and remove the sign).

5.4. Fourth case: \(KB = \{\neg\text{dog}, \neg\text{sign}\}\)

1. \textit{dog} is controllable-and-fixed and \textit{sign} is controllable-and-unfixed

The ideal obligations are \(O_i\neg\text{dog}\) and \(O_i\neg\text{sign}\).

None of them is violated.

The actual obligation is \(O_a\neg\text{sign}\) (i.e., the agent is actually required not to put a sign).
2. \(\text{dog}\) is controllable-and-unfixed and \(\text{sign}\) is controllable-and-fixed

The ideal obligations are \(O_i \neg \text{dog}\) and \(O_i \neg \text{sign}\).

None of them is violated.

The actual obligation is \(O_a \neg \text{dog}\) (i.e., the agent is required not to buy a dog in order to respect the primary obligation)

3. \(\text{dog}\) and \(\text{sign}\) are controllable-and-unfixed

The ideal obligations are \(O_i \neg \text{dog}\) and \(O_i \neg \text{sign}\).

None of them is violated.

The actual obligations are \(O_a \neg \text{dog}\) and \(O_a \neg \text{sign}\) (the current situation fulfills the ideal obligations, since there is no dog and no sign, and the agent is actually required not to buy a dog nor to put a sign).

6. Discussion

We have modelled the dog scenario with Carmo and Jones’s logic and we have compared all the different cases. In most cases, the two representations lead to the same results. This tends to prove that the present model is interesting.

However, some assumptions, such as the completeness of the description of the current world, should be examined.

One can notice that the modelisation we give here is the not the one presented at DEON’00 [3] but is the one which was discussed at the end of this paper.

Indeed, in the initial version of the paper, the interpretation in term of ordered models we made of the CTD was different. Consequently, the CTD was modelled only by the three formulas (a), (b) and (c). With this first modelisation, in some cases of the dog scenario, the results we get were different from the ones Carmo and Jones get. We noticed that the interpretation Carmo and Jones make was using the assumption saying that among the two \(\text{sign}\)-worlds, the one in which there is no dog is preferred to the one in which there is a dog, which is due to first sentence. Carmo and Jones convinced us that this assumption was not implicit but was due to the fact that the sentence \(\text{There ought to be no dog}\) should be interpreted by \(\text{any world in which there is no dog is preferred to any world in which there is a dog}\). This is why we changed the modelisation and add a fourth sentence, as suggested at the end of [3].

This hesitation illustrates the well-known problem of formalizing natural language sentences. Our negative results were not due to the formalism we used but due to the way we used it to model sentences. Natural language is ambiguous. And the only way to raise the ambiguities is to find the models of the sentences. Here, we claim that these models are expressed by orders on possible worlds. So, depending on the models one wants, one get one formula or another i.e., one modelisation or another. However, we insist on the fact that what is the most important for understanding a set of natural language sentences is to determine its underlying models. Here, the interpretation we give to the dog scenario leads to only one model.
What is subject to critics is the fact that, from a syntactical point of view, the three natural language sentences are modelled by four formulas and more precisely, the fact that the fourth formula depends on the others. We suspect that this fourth formula is due to the first sentence of course, but also to the other ones, and so in a sense, constraints the context of the first sentence. It is not yet entirely clear to us the kind of dependence that exists between the fourth formula and the others. The clarification of this important aspect needs further research.

However, even if far from perfect, the approach of adapting a logic of conditional preferences for reasoning with CTDs offers the advantage of reasoning correctly with norms with exceptions. This is not at all surprising since this kind of logic is designed for this purpose.

Let us consider for instance the following sentence:

*There ought to be no fence, except if the cottage is by the seaside*

Due to the ambiguity of the natural language we think that this sentence can be given two readings respectively modelled by the following set of CO* formulas:

- \( I(\neg \text{fence}), \neg I(\neg \text{fence} | \text{seaside}) \).

  Here the interpretation is: generally there ought to be no fence, but if the cottage is by the seaside, having a fence is permitted.

- \( I(\neg \text{fence}), I(\text{fence} | \text{seaside}) \).

  Here, the interpretation is: generally there ought to be no fence, but if the cottage is by the seaside then there ought to a be a fence.

Of course, these two representations lead to two different sets of CO*-models, and deductions that can be made are different from one case to the other.

As for the sentences of the kind: *There ought to be no dog unless there is a sign*, which, we think, are not CTD norms nor norms with exceptions, they also can be modelled in CO*, by the following sentence: \( I(\neg \text{dog} \lor \text{sign}) \). By this representation, we can conclude that the worlds in which there is no dog (with or without a sign) or in which there is a dog but a sign, are all equally preferred and all preferred to the world in which there is a dog but no sign.

Finally, let us notice that the same kind of approach than ours, has been recently followed by van der Torre and Tan. Indeed, in [10], they introduce a preference-based logic (PDL), based on Hansson’s semantics. Models are defined by: a set of worlds; a preference relation between worlds, not totally connected; an equivalence relation between worlds (used to characterize some possible worlds describing all the possible states of affairs); and a valuation function.

Like Boutilier, they define a dyadic operator \( I(.|.) \) and a betterness relation \( \succ \). The first operator corresponds to Boutilier’s \( I \) operator but is differently defined. The second corresponds to Boutilier’s \( \leq_P \) relation, but again is differently defined.

Finally, van der Torre and Tan define the conditional obligation \( O(\alpha|\beta) \) by:

\[
O(\alpha|\beta) \overset{\text{def}}{=} I(\alpha|\beta) \land (\alpha \land \beta) \succ (\neg \alpha \land \beta).
\]

Which means that the preferred \( \beta \) worlds are \( \alpha \) worlds and no \( \neg \alpha \land \beta \) world is as preferable as an \( \alpha \land \beta \) world.
With their definitions, the modelisation of the dog scenario is: \( O(\neg \text{dog}|T) \), \( O(\neg \text{sign}|\neg \text{dog}) \), \( O(\text{sign}|\text{dog}) \), \( \text{dog} \). This set of formulas has only one model in PDL which is the one described in figure 1. Thus, it is interesting to notice that, despite different definitions, van der Torre and Tan’s approach and the one described in this present paper characterize the same model (in terms of ordered worlds) of the dog scenario. But a formal comparison between the two approaches remains to be done. And we must, in the future, examine what are the impacts of their differences. However, the comparison will not be so easy because they have no agent’s ability model nor the distinction between ideal obligations and actual obligations.

Acknowledgments. We would like to thank Jose Carmo, Andrew Jones and Leendert van der Torre whose comments on the initial version of this paper helped us to revise it.

This work has been supported by ONERA.

References