Prohairetic Deontic Logic and Qualitative Decision Theory*

Leendert W.N. van der Torre
Max Planck Institut für Informatik
Im Stadtwald, Gebäude 46.1
D-66123 Saarbrücken
Germany

Yao-Hua Tan
Erasmus University Rotterdam
P.O. Box 1738
3000 DR Rotterdam
The Netherlands

Abstract

In this paper we introduce Prohairetic Deontic Logic, a preference-based dyadic deontic logic. An obligation 'α should be (done) if β is (done)' is true if (1) no $\neg\alpha \wedge \beta$ is preferred to an $\alpha \wedge \beta$ and (2) the preferred β are α. We show that this mixed representation solves several problems of deontic logic. Moreover, we discuss the relation between preference-based deontic logic and qualitative decision theory.

Introduction

Deontic logic is a modal logic developed in philosophical logic. Absolute and conditional obligations are represented by the modal formulas $O\alpha$ and $O(\alpha \mid \beta)$, where the latter is read as 'α ought to be (done) if β is (done).' The development of preference-based deontic logic is plagued by the following three (related) problems:

Strong preference problem. Preferences for α_1 and α_2 conflict for $\alpha_1 \wedge \neg\alpha_2$ and $\neg\alpha_1 \wedge \alpha_2$.

Contrary-to-duty problem. A contrary-to-duty obligation is an obligation that refers to a sub-ideal situation. Reasoning structures like 'α_1 should be (done), but if $\neg\alpha_1$ is (done) then α_2 should be (done)' must be formalized without running into the notorious contrary-to-duty paradoxes of deontic logic.

Dilemma problem. The three formulas $O\alpha \wedge O\neg\alpha$, $O(\alpha_1 \wedge \alpha_2) \wedge O\neg\alpha_1$ and $O(\alpha_1 \wedge \alpha_2 | \beta_1) \wedge O(\neg\alpha_1 | \beta_1 \wedge \beta_2)$ represent dilemmas and should therefore be inconsistent. However, $O(\alpha \mid \beta_1) \wedge O(\neg\alpha \mid \beta_2)$ does not represent a dilemma and should therefore be consistent.

In this paper we propose Prohairetic Deontic Logic. An obligation 'α should be (done) if β is (done)' is true if (1) no $\neg\alpha \wedge \beta$ is preferred to an $\alpha \wedge \beta$ and (2) the preferred β are α. The preference-based dyadic deontic logic shares the intuitive semantics of preference-based deontic logics, and solves the strong preference problem without introducing additional semantic baggage like multi-preference semantics or ceteris paribus preferences. Moreover, Prohairetic Deontic Logic shares the intuitive formalization of contrary-to-duty reasoning of dyadic deontic logics. Finally, Prohairetic Deontic Logic solves the dilemma problem by making the right set of formulas inconsistent.

This paper consists of two parts. In the first part we discuss the relation between obligations and preferences and we give an axiomatization of Prohairetic Deontic Logic in modal logic. In the second part we discuss the relation between preference-based deontic logic and the logic for qualitative decision theory. There is a structural similarity between the two logics, but they have different perspectives.

Obligations and preferences

It has been suggested (Jennings 1985; Jackson 1985; Goble 1990; Hansson 1990) that a unary operator O capable of bearing a deontic interpretation might be defined in a logic of preference by

$$O\alpha =_{def} \alpha \succ \neg\alpha$$

This is only the start of a definition of preference-based obligations, because we can accept different axioms for the preference relation $\succ$. A logic of preference can only be tested in the process of testing the more general theory in which it is embedded (Mullen 1979), in this case deontic logic. In the following discussion we assume possible worlds models with a preference relation on the worlds. Details of this semantics are given later in this paper when we discuss the axiomatization of Prohairetic Deontic Logic. It is well-known from the preference logic literature (von Wright 1963) that the preference $\alpha \succ \neg\alpha$ cannot be defined by the set of preferences of all α worlds to each $\neg\alpha$ world, because two obligations 'be polite' Op and 'be helpful' Oh would conflict when considering 'being polite and unhelpful' $p \wedge \neg h$ and 'being impolite and helpful' $\neg p \wedge h$. If a preference relation $\succ$ has left and right strengthening, then $p \succ \neg p$ and $h \succ \neg h$ derive $p \wedge \neg h \succ \neg p \wedge h$ and $\neg p \wedge h \succ p \wedge \neg h$, where $\succ$

*This work was partially supported by the ESPRIT Working Group 8319 (MODELAGE).

has the lowest binding priority. The two derived preferences seem contradictory. This is the strong preference problem.

Jackson (1985) and Goble (1990) introduce a second ordering representing degrees of 'closeness' of worlds and they define the preference $\alpha \succ \neg\alpha$ by the set of preferences of *the closest* α worlds to *the closest* $\neg\alpha$ worlds. The idea can be covered that in certain contexts the way things are in some worlds can be ignored – perhaps they are too remote from the actual world, or outside any agent's control. The obligation 'be polite' Op prefers the closest p to the closest $\neg p$. For example, the obligations Op and Oh are consistent when 'polite and unhelpful' $p \wedge \neg h$ and 'impolite and helpful' $\neg p \wedge h$ are not among the closest p, $\neg p$, h and $\neg h$ worlds. As a consequence, the preferences no longer have left and right strengthening. This solution of the strong preference problem introduces an irrelevance problem. For example, from 'be polite' $p \succ \neg p$ the preference $p \wedge h \succ \neg p \wedge h$ cannot even be derived, because $p \wedge h$ or $\neg p \wedge h$ may not be among the closest p or $\neg p$ worlds. The multi preference semantics seems a formalization of *defeasible* deontic logic (Tan & van der Torre 1995; van der Torre & Tan 1995; 1997b), sometimes called the logic of prima facie obligations. That is, 'closest' can be interpreted as 'the most normal' as used in the preferential semantics of logics of defeasible reasoning. However, it is not clear that closeness is an intuitive concept for non-defeasible obligations.

Hansson (1990) defines $\alpha \succ \neg\alpha$ by a 'ceteris paribus' preference of α to $\neg\alpha$. That is, for each pair of α and $\neg\alpha$ worlds such that both (except for α and $\neg\alpha$) describe similar circumstances, the α world is preferred to the $\neg\alpha$ world. The obligations 'be polite' Op prefers 'polite and helpful' $p \wedge h$ to 'impolite and helpful' $\neg p \wedge h$, and 'polite and unhelpful' $p \wedge \neg h$ to 'impolite and unhelpful' $\neg p \wedge \neg h$, but it does not say anything about $p \wedge h$ and $\neg p \wedge \neg h$. Conceptually, this formalization of preferences is more intuitive than the multi preference semantics of Jackson and Goble. Unfortunately, ceteris paribus preferences introduce an independence problem. At first sight, it seems that a 'ceteris paribus' preference $\alpha \succ \neg\alpha$ is a set of preferences of all $\alpha \wedge \beta$ worlds to each $\neg\alpha \wedge \beta$ world for all circumstances β such that $\alpha \wedge \beta$ and $\neg\alpha \wedge \beta$ are complete descriptions (represented by worlds). However, consider the preference $p \succ \neg p$ and circumstances $p \leftrightarrow \neg h$. The preference $p \succ \neg p$ would derive $p \wedge (p \leftrightarrow \neg h) \succ \neg p \wedge (p \leftrightarrow \neg h)$, which is logically equivalent to the problematic $p \wedge \neg h \succ \neg p \wedge h$. The exclusion of circumstances like $p \leftrightarrow \neg h$ is the independence problem. Only for independent β there is a preference of $\alpha \wedge \beta$ over $\neg\alpha \wedge \beta$ (see e.g. (Tan & Pearl 1994)).

In this paper we investigate a third approach. First, consider a preference $\alpha \succ \neg\alpha$ defined by 'every $\neg\alpha$ world is not preferred to or equivalent with any α world.' The definition is equivalent to 'all α worlds are preferred to each $\neg\alpha$ world' if the underlying preference ordering on worlds $\leq$ is strongly connected, i.e. if for each pair of worlds w_1 and w_2 we have either $w_1 \leq w_2$ or $w_2 \leq w_1$. However, the two obligations Op and Oh do not conflict when considering $p \wedge \neg h$ and $\neg p \wedge h$ *when we allow for incomparable worlds*, following (van Fraassen 1973). This solution is simpler than the first two solutions of the strong preference problem, because it does not use additional semantic baggage of the second ordering or the ceteris paribus preferences. Moreover, it does not have an irrelevance or an independence problem. In contrast to the other solutions of the strong preference problem, dilemmas like $Op \wedge O\neg p$ are consistent. In Chellas' terminology (Chellas 1974), it is a minimal deontic logic. The preference relation $\succ$ has left and right strengthening, and $p \succ \neg p$ and $h \succ \neg h$ derive $p \wedge \neg h \succ \neg p \wedge h$ and $\neg p \wedge h \succ p \wedge \neg h$. However, the latter two preferences are not logically inconsistent. The $\neg p \wedge h$ and $p \wedge \neg h$ worlds are incomparable. To formalize the no-dilemma assumption, we call Standard Deontic Logic to the rescue. We write $M \models I\alpha$ for 'the ideal worlds satisfy α.' Hence, if we ignore infinite descending chains then we have $M \models I\alpha$ if and only if *Pref* $\subseteq |\alpha|$ where *Pref* stands for the set of most preferred (ideal) worlds of M, and $|\alpha|$ stands for the set of all worlds satisfying α.

$$O\alpha =_{def} (\alpha \succ \neg\alpha) \wedge I\alpha$$

For example, the two obligations 'be polite' Op and 'be helpful' Oh are formalized by (1) p worlds are preferred to or incomparable with $\neg p$ worlds, (2) h worlds are preferred to or incomparable with $\neg h$ worlds, and (3) the ideal worlds are $p \wedge h$ worlds. In the following section we argue that this solution is not only simpler than the first two solutions of the strong preference problem, but it also gives a more intuitive solution to the contrary-to-duty problem.

Dyadic obligations and contrary-to-duty preferences

The contrary-to-duty problem is the major problem of monadic deontic logic, as shown by the notorious Good Samaritan, Chisholm and Forrester paradoxes. The formalization of these paradoxes should be consistent. For example, the formalization of the Forrester paradox (Forrester 1984) in monadic deontic logic is 'Smith should not kill Jones' $O\neg k$, 'if Smith kills Jones, then he should do it gently' $k \rightarrow Og$ and 'Smith kills Jones' k. From the three formulas $O\neg k \wedge Og$ can be derived. The derived formula should be consistent, although we have 'gentle killing implies killing' $\vdash g \rightarrow k$, see e.g. (Goble 1991). However, this formalization of the Forrester paradox does not do justice to the fact that only in *very* few cases we *seem* to have that $O\neg\alpha \wedge O(\alpha \wedge \beta)$ is not a dilemma, and should be consistent. The consistency of $O\neg k \wedge Og$ is a solution that seems like overkill.

B. Hansson (1971) and Lewis (1974) showed that the contrary-to-duty problem can be solved with dyadic obligations. A dyadic obligation $O(\alpha \mid \beta)$ is read as 'α ought to be (done) if β is (done).' The introduction of the dyadic representation was inspired by the standard way of representing conditional probability, that is, by $Pr(\alpha \mid \beta)$ which stands for 'the probability for α given β.' In a dyadic deontic logic, the Forrester paradox can be formalized by 'Smith should not kill Jones' $O(\neg k \mid \top)$, 'if Smith kills Jones, then he should do it gently' $O(g \mid k)$ and 'Smith kills Jones' k. In this formalization, $\top$ stands for any tautology like $p \vee \neg p$. The obligation $O(g \mid k)$ is a contrary-to-duty (CTD) obligation of $O(\neg k \mid \top)$, because an obligation $O(\alpha \mid \beta)$ is a CTD obligation of the primary obligation $O(\alpha_1 \mid \beta_1)$ if and only if $\alpha_1 \wedge \beta$ is inconsistent, as illustrated in Figure 1. In dyadic deontic logic, the formula $O(\neg k \mid \top) \wedge O(g \mid k)$ is consistent, whereas the formula $O(\neg k \mid \top) \wedge O(g \mid \top)$ is inconsistent given $\vdash g \rightarrow k$.

$$\begin{tikzcd} O(\alpha_1|\beta_1) \arrow[d, leftrightarrow, "\text{inconsistent}"'] \\ O(\alpha|\beta) \end{tikzcd}$$

Figure 1: $O(\alpha|\beta)$ is a CTD of $O(\alpha_1|\beta_1)$

Dyadic obligations give a solution to the contrary-to-duty problem, without making dilemmas consistent. However, the dyadic representation also introduces a new instance of the dilemma problem, that is represented by the formula $O(\alpha \mid \beta_1) \wedge O(\neg\alpha \mid \beta_1 \wedge \beta_2)$. An example is Prakken and Sergot's considerate assassin example, that consists of the two obligations 'Smith should not offer Jones a cigarette' $O(\neg c \mid \top)$ and 'Smith should offer Jones a cigarette if he kills him' $O(c \mid k)$. Figure 2 illustrates that $O(c \mid k)$ is not a CTD obligation of $O(\neg c \mid \top)$.

$$\begin{tikzcd} O(\neg c \mid \top) \arrow[d, leftrightarrow, shift right, "\text{inconsistent}"'] \\ O(c \mid k) \arrow[u, shift right, "\text{implies}"'] \end{tikzcd}$$

Figure 2: $O(c|k)$ is not a CTD of $O(\neg c|\top)$

Prakken and Sergot (1996) argue that the two sentences of the considerate assassin example represent a dilemma. Hence, $O(\neg c \mid \top) \wedge O(c \mid k)$ should be inconsistent, even when there is another premise 'Smith should not kill Jones' $O(\neg k \mid \top)$.

> 'Is this acceptable? In our opinion it is: what is crucial is that $O(c \mid k)$ is not a CTD rule of $O(\neg c \mid \top)$ but of $O(\neg k \mid \top)$, for which reason $O(c|k)$ and $O(\neg c|\top)$ are unrelated obligations. Now one may ask how this conflict should be resolved and, of course, one plausible option is to regard $O(c|k)$ as an exception to $O(\neg c|\top)$ and to formalize this with a suitable nonmonotonic defeat mechanism. However, it is important to note that this is a separate issue, which has nothing to do with the CTD aspects of the example.' (Prakken & Sergot 1996)

B.Hansson-Lewis dyadic deontic logics do not give a satisfactory solution to the dilemma problem, because the considerate assassin example is consistent. They define a dyadic obligation by $O_{HL}(\alpha \mid \beta) =_{def} I(\alpha \mid \beta)$, where we write $I(\alpha \mid \beta)$ for 'the ideal β worlds satisfy α.' Hence, if we again ignore infinite descending chains then we have $M \models I(\alpha \mid \beta)$ if and only if *Pref*$(\beta) \subseteq | \alpha |$ where *Pref*(β) stands for the preferred β worlds of M. The formula $O_{HL}(\neg c|\top) \wedge O_{HL}(c|k)$ is consistent. In Prohairetic Deontic Logic, dyadic obligations are defined in a similar spirit as the absolute obligations in the previous section.

$$O(\alpha|\beta) =_{def} (\alpha \wedge \beta \succ \neg\alpha \wedge \beta) \wedge I(\alpha|\beta)$$

The set of obligations $S = \{O(\neg c|\top), O(c|k)\}$ is inconsistent, because $(\neg c \succ c) \wedge I(c|k)$ is inconsistent, as is shown later in this paper.

Axiomatization

We axiomatize Prohairetic Deontic Logic in modal logic. Lamarre (1991) and Boutilier (1992) have shown how the expression $I(\alpha \mid \beta)$ can be formalized in modal S4 structures. As is well-known, S4 is a normal modal system **KT4** where axiom **T**: $\Box\alpha \rightarrow \alpha$ characterizes reflexivity and axiom **4**: $\Box\alpha \rightarrow \Box\Box\alpha$ characterizes transitivity. S4 is extended with the following definition **C**. Hence, the dyadic conditional is formalized in terms of the monadic operator. The conditional $I(\alpha|\beta)$ is true in a world if α is true in all the preferred β worlds (that are accessible from the actual world), because for each $\beta \wedge \neg\alpha$ world there is a $\beta \wedge \alpha$ world that is strictly preferred to it. This definition can also deal with infinite descending chains.

$$\textbf{C}: \quad I(\alpha|\beta) \quad =_{def} \quad \Box(\beta \rightarrow \Diamond(\beta \wedge \Box(\beta \rightarrow \alpha)))$$

Moreover, Boutilier shows that Humberstone's logic of inaccessible worlds can be used to axiomatize the property that every worlds sees the same preference ordering. In this paper we do not need that expressivity, because we only consider the derivation of obligations from obligations.[1] In the following definition of PDL we extend the logic with the preference relation $\succ$, see (Tan & van der Torre 1996).

[1] If we also consider facts, then we have to define the preferences by $\alpha_1 \succ \alpha_2 =_{def} \overset{\leftrightarrow}{\Box}(\alpha_1 \rightarrow \Box\neg\alpha_2)$, in which $\overset{\leftrightarrow}{\Box}$ is a second modal operator that satisfies at least $\overset{\leftrightarrow}{\Box}\alpha \rightarrow \Box\alpha$. An example of such a modal operator has been given by Boutilier.

Definition 1 (PDL) *The modal language $\mathcal{L}$ is formed from a denumerable set of propositional variables together with the connectives $\neg$, $\rightarrow$, and the normal modal connective $\Box$. The logic PDL is the smallest $S \subset \mathcal{L}$ such that S contains classical logic and the following axiom schemata, and is closed under the following rules of inference,*

K	$\Box(\alpha \rightarrow \beta) \rightarrow (\Box\alpha \rightarrow \Box\beta)$
T	$\Box\alpha \rightarrow \alpha$
4	$\Box\alpha \rightarrow \Box\Box\alpha$
Nes	*From* α *infer* $\Box\alpha$
MP	*From* $\alpha \rightarrow \beta$ *and* α *infer* β

extended with the following three definitions.

$$\begin{array}{lll} \alpha_1 \succ \alpha_2 & =_{def} & \Box(\alpha_1 \rightarrow \Box\neg\alpha_2) \\ I(\alpha|\beta) & =_{def} & \Box(\beta \rightarrow \Diamond(\beta \wedge \Box(\beta \rightarrow \alpha))) \\ O(\alpha|\beta) & =_{def} & (\alpha \wedge \beta \succ \neg\alpha \wedge \beta) \wedge I(\alpha|\beta) \end{array}$$

Definition 2 (PDL Semantics) *Possible worlds models $M = \langle W, \leq, V\rangle$ for PDL consist of W, a set of worlds, $\leq$, a binary transitive and reflexive accessibility relation, and V, a valuation of the propositional atoms in the worlds. The partial pre-ordering $\leq$ expresses preferences: $w_1 \leq w_2$ if and only if w_1 is as preferable as w_2. The truth conditions are defined as usual.*

$$M, w \models \Box\alpha \textit{ iff } \forall w' \in W \textit{ if } w' \leq w, \textit{ then } M, w' \models \alpha$$

The preference relation $\succ$ is quite weak. For example, it is not anti-symmetric ($\neg(\alpha_2 \succ \alpha_1)$ cannot be derived from $\alpha_1 \succ \alpha_2$) and it is not transitive ($\alpha_1 \succ \alpha_3$ cannot be derived from $\alpha_1 \succ \alpha_2$ and $\alpha_2 \succ \alpha_3$). The lack of these properties is the result of the fact that we do not have totally connected orderings. In the remainder of this section, we consider three properties of the dyadic obligations: lack of weakening of the consequent, lack of reasoning by cases and restricted strengthening of the antecedent. First, the logic PDL does not have closure under logical implication. This is a typical property of preference-based deontic logics, because the preference-based deontic logics discussed in (Jackson 1985; Hansson 1990; Goble 1990) do not have closure under logical implication either. Hence, the following theorem Weakening of the Consequent **WC** is *not* valid.

WC $\quad O(\alpha_1|\beta) \rightarrow O(\alpha_1 \vee \alpha_2|\beta)$

The second property we consider is the following theorem *Reasoning-By-Cases* **RBC**, sometimes called the sure-thing principle. It is *not* valid either.

RBC $\quad (O(\alpha|\beta_1) \wedge O(\alpha|\beta_2)) \rightarrow O(\alpha|\beta_1 \vee \beta_2)$

An example of **RBC** is 'if you ought to do α given β and you ought to do α given $\neg\beta$, then you ought to do α without examining β.' For example, this reading is given by Pearl (1993). The following example illustrates that the non-validity of **RBC** can be used to analyze dominance arguments. A common sense dominance argument (1) divides possible outcomes into two or more exhaustive, exclusive cases, (2) points out that in each of these alternatives it is better to perform some action than not to perform it, and (3) concludes that this action is best unconditionally. Thomason and Horty (1996) observe that, although such arguments are often used, and are convincing when they are used, they are invalid. The following example adapted from (Thomason & Horty 1996) is a classic illustration of the invalidity of the dominance argument (see also (Jeffrey 1983)).

Example 1 (Cold-war disarmament) *Either there will be a nuclear war or there will not. If there will not be a nuclear war, it is better for us to disarm because armament would be expensive and pointless. If there will be a nuclear war, we will be dead whether or not we arm, so we are better of saving money in the short term by disarming. So we would disarm. The fallacy, of course, depends on the assumption that the action of choosing whether to arm or disarm will have no effect on whether there is war or not.*

Consider the contextual obligations 'we ought to be disarmed if there will be a nuclear war' $O(d \mid w)$ and 'we ought to be disarmed if there will be no war' $O(d \mid \neg w)$. The obligation $O(d \mid w \vee \neg w)$ cannot be derived, because $d \succ \neg d$ cannot be derived from $d \wedge w \succ \neg d \wedge w$ and $d \wedge \neg w \succ \neg d \wedge \neg w$. In fact, the model in Figure 3 satisfies $\neg d \wedge \neg w \succ d \wedge w$, which represents 'we ought to be armed if we have peace if and only if we are armed' $O(\neg d|d \leftrightarrow w)$. This figure should be read as follows. A circle represents a non-empty equivalence class of worlds, that satisfy the propositions written in the circle. An arrow represents strict preferences for all worlds represented by the circle. The transitive closure is left implicit.

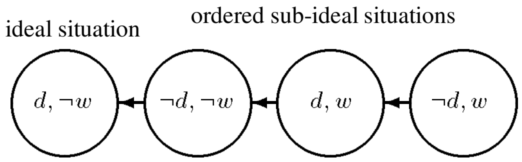

Figure 3: Model of cold-war disarmament

The third property we consider is so-called *Restricted Strengthening of the Antecedent* **RSA**, expressed by the following theorem of the logic.

RSA $\quad (O(\alpha|\beta_1) \wedge I(\alpha|\beta_1 \wedge \beta_2)) \rightarrow O(\alpha|\beta_1 \wedge \beta_2)$

We can add strengthening of the antecedent with the following notion of preferential entailment.

Definition 3 (Pref. entailment) *Let $M_1 = \langle W, \leq_1, V \rangle$ and $M_2 = \langle W, \leq_2, V \rangle$ be two PDL models. M_1 is as preferable as M_2, written as $M_1 \sqsubseteq M_2$, iff for all $w_1, w_2 \in W$ if $w_1 \leq_2 w_2$ then $w_1 \leq_1 w_2$. M_1 is preferred to M_2, written as $M_1 \sqsubset M_2$, iff $M_1 \sqsubseteq M_2$ and $M_2 \not\sqsubseteq M_1$. The formula ϕ is preferentially entailed by T, written as $T \models_{\sqsubset} \phi$, iff $M \models \phi$ for all preferred models M of T.*

The following example illustrates restricted strengthening of the antecedent.

Example 2 *Consider the set of obligations*

$$S = \{O(p|\top), O(h|\top)\}$$

The unique (given W and V) preferred model of S is represented in Figure 4. It is easily checked that we have $S \models_{\sqsubset} O(p|h)$ and $S \models_{\sqsubset} O(p|\neg h)$. Hence, the obligation $O(p|\top)$ is strengthened to $O(p|h)$ and $O(p|\neg h)$.

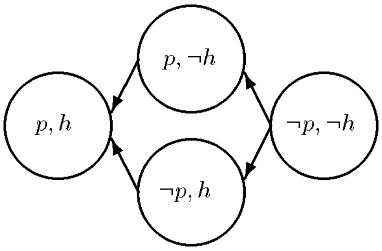

Figure 4: Preferred model of $\{O(p|\top), O(h|\top)\}$

Finally, we mention two further extensions of PDL. The first extension is to consider only full models, i.e. models that contain a world corresponding to each possible interpretation. A second possible extension is to drop the transitivity axiom **4**: $\Box\alpha \to \Box\Box\alpha$. It is easily shown that in the latter case, the maximally connected models of a set of obligations are unique (for a given W and V).

The three problems reconsidered

In the introduction of this paper we mentioned three problems: the strong preference problem, the contrary-to-duty problem and the dilemma problem. In this section we show how Prohairetic Deontic Logic solves the three problems. The strong preference problem is that preferences for α_1 and α_2 conflict for $\alpha_1 \wedge \neg\alpha_2$ and $\neg\alpha_1 \wedge \alpha_2$. The following example illustrates that the problem is solved by the 'dynamics' of preferential entailment. It also illustrates why the logic is non-monotonic.

Example 3 *Consider the sets*

$$S = \emptyset$$

$$S' = \{O(p|\top)\}$$

$$S'' = \{O(p|\top), O(h|\top)\}$$

The three unique preferred models of S, S' and S'' are represented in Figure 5.

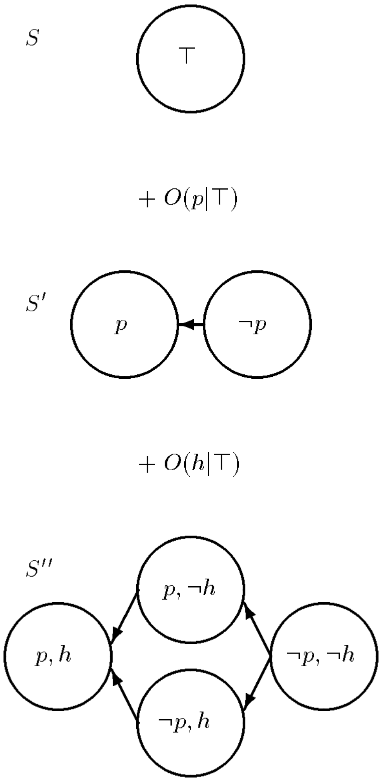

Figure 5: Dynamics

With no premises, all worlds are equally ideal. By addition of premise $O(p \mid \top)$, the p worlds are strictly preferred over $\neg p$ worlds. By addition of the second premise $O(h|\top)$, the h worlds are strictly preferred over $\neg h$ worlds, and the $p \wedge \neg h$ and $\neg p \wedge h$ worlds become incomparable. Hence, the problem is solved by representing the conflicting worlds with incomparable worlds. This solution uses techniques from non-monotonic reasoning, namely preferential entailment. We have $S' \models_{\sqsubset} O(p \mid \neg(p \wedge h))$ and $S'' \not\models_{\sqsubset} O(p|\neg(p \wedge h))$. Hence, by addition of a formula we loose conclusions.

The solution of the contrary-to-duty problem is based on the dyadic representation. The solution is illustrated by the representation of the Forrester and Chisholm paradoxes in Prohairetic Deontic Logic.

Example 4 (Forrester paradox) *Consider the set of obligations*

$$S = \{O(\neg k|\top), O(g|k)\}$$

where k can be read as 'Smith kills Jones' and g as 'Smith kills him gently,' and g logically implies k. The unique preferred model is represented in Figure 6. It is easily checked *that we have $M \models O(\neg k \vee g \mid \top)$, which expresses that $k \wedge \neg g$ is the worst state which should be evaded.*

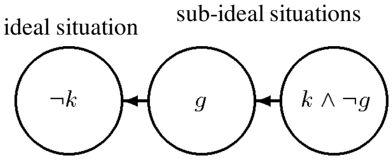

Figure 6: Preferred model of $\{O(\neg k|\top), O(g|k)\}$

Example 5 (Chisholm paradox) *Consider the set of obligations*

$$S = \{O(a|\top), O(t|a), O(\neg t|\neg a)\}$$

where a can be read as 'a certain man going to his neighbors' assistance' and t as 'telling the neighbors that he will come.' The unique preferred model of S is represented in Figure 7.

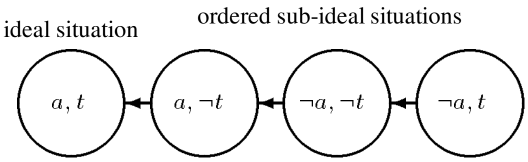

Figure 7: Preferred model of $\{O(a|\top), O(t|a), O(\neg t|\neg a)\}$

The crucial question of the Chisholm paradox is whether there is an obligation 'the man should tell the neighbors that he will come' $O(t \mid \top)$. This obligation can be derived by so-called deontic detachment (deontic transitivity), represented by the following formula **DD**.

DD $\quad (O(\alpha|\beta) \wedge O(\beta|\gamma)) \rightarrow O(\alpha|\gamma)$

DD *is not valid in Prohairetic Deontic Logic. As a consequence, the obligation 'the man should tell his neighbors that he will come' $O(t|\top)$ cannot be derived. However, the following variant of the theorem* **DD** *is valid.*

DD′ $\quad (O(\alpha|\beta) \wedge O(\beta|\gamma) \wedge I(\alpha \wedge \beta|\gamma)) \rightarrow O(\alpha \wedge \beta|\gamma)$

The obligation 'the man should go to his neighbors and tell his neighbors that he will come' $O(a \wedge t|\top)$ can be derived.

Finally, we show that Prohairetic Deontic Logic solves the dilemma problem, because it makes the considerate assassin set in Example 6 inconsistent, without making the window set in Example 7 inconsistent.

Example 6 (Considerate assassin) *Consider the set of obligations*

$$S = \{O(\neg c|\top), O(c|k)\}$$

where c can be read as 'Smith offers Jones a cigarette' and k can be read as 'Smith kills Jones.' The set S is inconsistent with $\Diamond(k \wedge c)$, as can be verified as follows. Assume there is a model of S. The obligation $O(c|k)$ implies $I(c|k)$, which means that for every world w_1 such that $M, w_1 \models \neg c \wedge k$ there is a world w_2 such that $M, w_2 \models c \wedge k$ and $w_2 < w_1$ (i.e. $w_2 \leq w_1$ and $w_1 \not\leq w_2$). However, the obligation $O(\neg c|\top)$ implies $\neg c \succ c$, which means that for all worlds w_1 such that $M, w_1 \models \neg c \wedge k$ there is not a world w_2 such that $M, w_2 \models c \wedge k$ and $w_2 \leq w_1$. These two conditions are contradictory (if there is a world w_2).

A problem of several other solutions of the dilemma problem like (von Wright 1971) is that the set of obligations $S = \{O(\alpha \mid \beta_1), O(\neg\alpha \mid \beta_2)\}$ is no longer consistent. It is argued by von Wright (1971) that S does not represent a dilemma and that it should therefore be consistent.

> 'Herewith it has been proven that, if there is a duty to see to it that α under circumstances β, then there is no duty to see to it that not-α under circumstances γ. For example: It has been proven that, if there is a duty to see to it that a certain window is closed should it start raining, then there cannot be a duty to see to it that the window is open should the sun be shining. This is manifestly absurd. Generally speaking: From a duty to see to a certain thing under certain circumstances nothing can follow logically concerning a duty or not-duty under entirely different, logically unrelated, circumstances. Least of all should one be able to prove that there is under those unrelated circumstances a duty of contradictory content.'(von Wright 1971, p.116).

The following example illustrates that the set S is consistent in Prohairetic Deontic Logic.

Example 7 (Window) *Consider the set of obligations*

$$S = \{O(c|r), O(\neg c|s)\}$$

where c can be read as 'the window is closed,' r as it starts raining' and s as 'the sun is shining.' S is consistent, and a preferred model M of S is given in Figure 8. The ideal worlds satisfy $r \rightarrow c$ and $s \rightarrow \neg c$, and the subideal worlds either $\neg c \wedge r$ or $c \wedge s$. We have $M \not\models O(c|r \wedge s)$ and thus $S \not\models_{\sqsubset} O(c|r \wedge s)$.

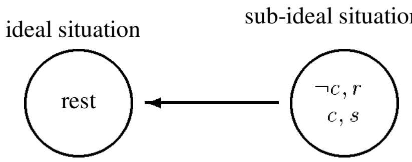

Figure 8: Preferred model of $\{O(c|r), O(\neg c|s)\}$

In this section we have shown that Prohairetic Deontic Logic solves the three problems discussed in the introduction: the strong preference problem, the contrary-to-duty problem and the dilemma problem.

Qualitative decision theory

In the remainder of this paper, we discuss the relation between preference-based deontic logic and qualitative decision theory. Deontic logic has been developed as a branch of philosophical logic, and it has recently been studied in computer science. Topics identified are legal knowledge-based systems, the specification of fault tolerant systems, the specification of security policies, the automatization of contracting and the specification of normative integrity constraints for databases (Wieringa & Meyer 1993). However, deontic logic is not sufficient for all applications that are based on normative reasoning. The problem is that deontic logic only formalizes reasoning *about* obligations, that is, which obligations follow from a set of obligations. However, there is a demand to formalize reasoning *with* obligations. For example, a legal expert system may face the diagnostic problem to determine whether a suspect has violated a legal rule, and a robot may have to solve the planning problem how to fulfill the desires of his owner. This can be done in the area of *Qualitative Decision Theory (QDT)*, a rapidly developing area within artificial intelligence to formalize, for example, robot planning. Decision theory provides for most of the basic concepts we need for rational decision making, in particular, the ability to specify arbitrary preferences over circumstances or goals (and hence appropriate behaviors) to vary with context (Boutilier 1994). Decision theory and related theories in economics concentrate on a notion of expected utility that is representable using quantitative preferences and probabilities. More recent traditions in artificial intelligence have explored qualitative decision methods including control rules, rule orderings, default preferences, and qualitative approaches to probability. The relation between deontic logic and qualitative decision theory is described by a structural similarity from different perspectives.

Structural similarity There is a structural similarity between deontic logic and logics for QDT (Boutilier 1994; Lang 1996; Thomason & Horty 1996), because both are based on preferences. The use of preferences is introduced in deontic logic to deal with contrary-to-duty type of problems. In QDT, preferences are used for context-sensitive goals (Doyle & Wellman 1991; Boutilier 1994).

Different perspectives The main purpose of a deontic logic is deriving new obligations (and permissions) from an initial specification, while QDT focuses on the search for optimal acts and decisions (Lang 1996). Deontic logic and a logic for QDT have different perspectives. Obligations are exogenous (they are imposed by a legal or moral code) while desires in logics for QDT are endogenous (coming from the agent) (Lang 1996). It is this distinction which we call *the gap between deontic logic and qualitative decision theory*.

The structural similarity suggests that deontic logic can be used in a qualitative decision theory. In particular, the three problems discussed in this paper also occur in reasoning about goals, and Prohairetic Deontic Logic can be used to reason about goals. However, as a consequence of the two different perspectives we first have to bridge the gap between deontic logic and QDT.

Structural similarities

We first consider the structural distinction between the monopolar (only bad violation pole) and bipolar (a bad violation pole and a good ideal pole) interpretation of obligations. An obligation $O\alpha$ is violated if $\neg\alpha$ is true, i.e. if $\neg\alpha \wedge O\alpha$ is true. In the *Monopolar Interpretation* (MI), an obligation describes a violation condition, and it is obligatory to minimize the set of violations. Ideally no obligation is violated, but if a violation has occurred then the remainder of the obligations should not be violated. Once you drive too fast, you should still drive on the right side of the road. In preference-based semantics, a world is preferred to another world if it represents less violations (with respect to set inclusion). MI can represent contrary-to-duty reasoning, which was the reason preferences were introduced in deontic logic. It is related to a theory of diagnosis (Reiter 1987), because the latter formalizes reasoning with violations (and additionally it also *assumes* that the set of violations is minimal). In the *Bipolar Interpretation* (BI), an obligation describes a choice between fulfilling and violating, and in case of a deontic choice, it is obligatory to fulfill obligations. The structural distinction between MI and BI can be illustrated by *conditional* obligations. MI of an obligation $O(\alpha|\beta)$ prefers the non-violation $|\beta \rightarrow \alpha|$ to violation $|\neg\alpha \wedge \beta|$ whereas BI prefers fulfillment $|\alpha \wedge \beta|$ to violation $|\neg\alpha \wedge \beta|$, and does not say anything about non-violation and non-fulfillment $| \neg\beta |$. For example, Pearl's logic of pragmatic obligation (Pearl 1993) is a decision-theoretic account of obligation statements and uses qualitative abstractions of utilities and probabilities.[2] Thus, it is based on

[2]Pearl (1993) criticizes deontic logic: 'exploratory reading of the literature reveals that philosophers hoped to develop deontic logic as a branch of conditional logic, not as a synthetic amalgam of logic of belief, action, and causation,' and that 'such an isolationistic strategy has little chance of succeeding.' This criticism is in our opinion not very convincing, because deontic logic is characterized by the distinction between actual and ideal, not by beliefs and actions. Finally Pearl concludes that 'the decision-theoretic account can be used to generate counterexamples to most of the principles suggested in the literature.' However, his logic is based on a multi preference semantics and in our opinion better understood as a defeasible deontic logic. Obviously, *any* defeasible deontic logic can be used to generate 'counterexamples' for a non-defeasible deontic logic.

utilitarian semantics (Jennings 1974). Prohairetic Deontic Logic is also based on BI. A further discussion of the distinction between MI and BI can be found in (Boutilier 1994; Tan & Pearl 1994).[3]

Different perspectives

A theory of diagnosis, like MI, only reasons about violations and it reasons about the past with incomplete knowledge (if everything is known then a diagnosis is completely known). A (qualitative) decision theory, like BI, reasons about the future. The two perspectives are represented in Figure 9 below.

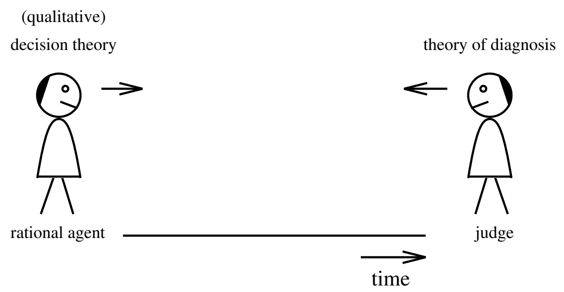

Figure 9: Reasoning with norms

The distinction between the perspective of a rational agent (qualitative decision theory) and a judge (theory of diagnosis) corresponds to Thomason's distinction between the context of deliberation and the context of justification. It is important to discriminate between the two contexts, because a sentence can sometimes be interpreted differently in each of them. For example, consider the sentence 'you should not smoke and you smoke.' In the context of justification the obligation is interpreted as the identification of the fact that you are violating a rule, whereas in the context of deliberation, it is interpreted as the deontic cue to stop smoking. The two perspectives are represented in Figure 10.

The distinction between the two interpretations of the obligation is as important as the distinction between Alchourrón-Gärdenfors-Makinson belief revision (or theory revision) and Katsuno-Mendelzon belief update in the area of logics of belief. There is a strong analogy, because belief revision is reasoning about a non-changing world and update is reasoning about a changing world. It follows directly from Figure 10 that a similar distinction is made between respectively the context of justification and the context of deliberation, because the past is fixed, whereas the future is wide open.

[3]The discussion in these papers concentrates on two schemes for preferential entailment called System Z and compactness. Both have disadvantages compared to the notion of preferential entailment introduced in this paper based on maximally connected models, because the schemes force (unique) totally connected orderings. Lack of incomparable worlds turn these schemes into violation counters, in the sense that under certain circumstances they can only discriminate between the number of violations.

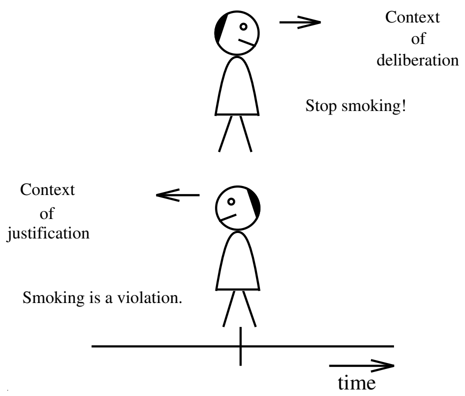

Figure 10: Contexts of deontic logic

One of the main features of deontic logic is the fact that actors do not always obey the law. Indeed, it is precisely when a forbidden act occurs, or an obligatory action does not occur, that we need deontic logic, to detect a violation and to take appropriate action. For purposes of planning, it is often useful to assume that actors *do* obey the law. McCarty (1994) calls this the *causal assumption*, since it enables us to 'predict the actions that *will* occur by reasoning about the actions that *ought* to occur.' If we adopt the causal assumption, we can use deontic logic to reason about the physical world (McCarty 1994). Moreover, preferences can be used to express the gains and losses consequent to normative decisions. Once agents know the losses and gains of their decisions, the normative decision is no longer a problem. Interesting questions are: how do agents work out norms in terms of gains and losses? What are the gains of observing norms? How do they learn the effects of norms and how do they reason about these effects? In which way does a normative decider differ from an ordinary decider, if any?

Conclusions

In this paper, we introduced Prohairetic Deontic Logic. We showed that it gives a satisfactory solution to the strong preference problem, the contrary-to-duty problem and the dilemma problem. Moreover, we discussed the relation between preference-based deontic logic and a logic for qualitative decision theory. Prohairetic Deontic Logic shows that one needs not get embroiled in questions of relevance, ceteris paribus etc. to get a good reading for preference-based obligations or goals.

We observed a dynamic interpretation of Prohairetic Deontic Logic with preferential entailment. The observations hint at a similarity with update semantics (Veltman 1996). The formalization of obligations in update semantics is subject of current research.

References

Boutilier, C. 1992. Conditional logics for default reasoning and belief revision. Technical Report 92-1, Department of Computer Science, University of British Columbia.

Boutilier, C. 1994. Toward a logic for qualitative decision theory. In *Proceedings of the KR'94*, 75–86.

Chellas, B. 1974. Conditional obligation. In *Logical Theory and Semantical Analysis*. Dordrecht, Holland: D. Reidel Publishing Company. 23–33.

Doyle, J., and Wellman, M. 1991. Preferential semantics for goals. In *Proceedings of AAAI-91*, 698–703.

Forrester, J. 1984. Gentle murder, or the adverbial Samaritan. *Journal of Philosophy* 81:193–197.

Goble, L. 1990. A logic of good, would and should, part 2. *Journal of Philosophical Logic* 19:253–276.

Goble, L. 1991. Murder most gentle: the paradox deepens. *Philosophical Studies* 64:217–227.

Hansson, B. 1971. An analysis of some deontic logics. In *Deontic Logic: Introductionary and Systematic Readings*. Dordrecht, Holland: D. Reidel Publishing Company. 121–147.

Hansson, S. 1990. Preference-based deontic logic (PDL). *Journal of Philosophical Logic* 19:75–93.

Jackson, F. 1985. On the semantics and logic of obligation. *Mind* 94:177–196.

Jeffrey, R. 1983. *The Logic of Decision*. University of Chicago Press, 2nd edition.

Jennings, R. 1974. A utilitarian semantics for deontic logic. *Journal of Philosophical Logic* 3:445–465.

Jennings, R. 1985. Can there be a natural deontic logic? *Synthese* 65:257–274.

Lamarre, P. 1991. S4 as the conditional logic of nonmonotonicity. In *Proceedings of KR'91*, 357–367.

Lang, J. 1996. Conditional desires and utilities - an alternative approach to qualitative decision theory. In *Proceedings of the ECAI'96*, 318–322.

Lewis, D. 1974. Semantic analysis for dyadic deontic logic. In *Logical Theory and Semantical Analysis*. Dordrecht, Holland: D. Reidel Publishing Company. 1–14.

McCarty, L. 1994. Modalities over actions: 1. model theory. In *Proceedings of the KR'94*, 437–448. San Francisco, CA: Morgan Kaufmann.

Mullen, J. 1979. Does the logic of preference rest on a mistake? *Metaphilosophy* 10:247–255.

Pearl, J. 1993. A calculus of pragmatic obligation. In *Proceedings of the UAI'93*, 12–20.

Prakken, H., and Sergot, M. 1996. Contrary-to-duty obligations. *Studia Logica* 57:91–115.

Reiter, R. 1987. A theory of diagnosis from first principles. *Artificial Intelligence* 32:57–95.

Tan, S.-W., and Pearl, J. 1994. Specification and evaluation of preferences under uncertainty. In *Proceedings of the KR'94*, 530–539.

Tan, Y.-H., and van der Torre, L. 1995. Why defeasible deontic logic needs a multi preference semantics. In *Proceedings of the ECSQARU'95. LNAI 946*, 412–419. Springer Verlag.

Tan, Y.-H., and van der Torre, L. 1996. How to combine ordering and minimizing in a deontic logic based on preferences. In *Deontic Logic, Agency and Normative Systems. Proceedings of the Δeon'96*, 216–232. Springer Verlag, Workshops in Computing.

Thomason, R., and Horty, R. 1996. Nondeterministic action and dominance: foundations for planning and qualitative decision. In *Proceedings of the TARK'96*, 229–250. Morgan Kaufmann.

van der Torre, L., and Tan, Y. 1995. Cancelling and overshadowing: two types of defeasibility in defeasible deontic logic. In *Proceedings of the IJCAI'95*, 1525–1532. Morgan Kaufman.

van der Torre, L., and Tan, Y. 1997a. Contextual deontic logic. In *Proceedings of the Third International conference on Modeling and Using Context (CONTEXT'97)*. Forthcoming.

van der Torre, L., and Tan, Y. 1997b. The many faces of defeasibility in defeasible deontic logic. In Nute, D., ed., *Defeasible Deontic Logic*. Kluwer. 79–121.

van Fraassen, B. 1973. Values and the heart command. *Journal of Philosophy* 70:5–19.

Veltman, F. 1996. Defaults in update semantics. *Journal of Philosophical Logic* 25:221–261.

von Wright, G. 1963. *The Logic of Preference*. Edinburgh University Press.

von Wright, G. 1971. A new system of deontic logic. In *Deontic Logic: Introductory and Systematic Readings*. Dordrecht, Holland: D. Reidel Publishing company. 105–120.

Wieringa, R., and Meyer, J.-J. 1993. Applications of deontic logic in computer science: A concise overview. In *Deontic Logic in Computer Science*. Chichester, England: John Wiley & Sons. 17–40.