Prohairetic Deontic Logic (PDL)

Leendert W.N. van der Torre[1] and Yao-Hua Tan[2]

[1] IRIT, Paul Sabatier University
118 Route de Narbonne, Toulouse, France
http://abduction.euridis.fbk.eur.nl/~torre/
torre@irit.fr
[2] EURIDIS, Erasmus University Rotterdam
PO Box 1738, 3000 DR Rotterdam
ytan@euridis.fbk.eur.nl

Abstract. In this paper we introduce Prohairetic Deontic Logic (PDL), a preference-based dyadic deontic logic. An obligation 'α should be (done) if β is (done)' is true if (1) no $\neg\alpha \wedge \beta$ state is as preferable as an $\alpha \wedge \beta$ state and (2) the preferred β states are α states. We show that the different elements of this mixed representation solve different problems of deontic logic. The first part of the definition is used to formalize contrary-to-duty reasoning, that for example occurs in Chisholm's and Forrester's notorious paradoxes. The second part is used to make dilemmas inconsistent. PDL shares the intuitive semantics of preference-based deontic logics without introducing additional semantic machinery such as bi-ordering semantics or ceteris paribus preferences.

1 Introduction

Deontic logic is a modal logic, in which absolute and conditional obligations are represented by the modal formulas $O\alpha$ and $O(\alpha|\beta)$, where the latter is read as 'α ought to be (done) if β is (done).' It can be used for the formal specification and validation of a wide variety of topics in computer science (for an overview and further references see [44]). For example, deontic logic can be used to formally specify soft constraints in planning and scheduling problems as norms. The advantage is that norms can be violated without creating an inconsistency in the formal specification, in contrast to violations of hard constraints. With the increasing popularity and sophistication of applications of deontic logic the fundamental problems of deontic logic become more pressing.

From the early days, when deontic logic was still a purely philosophical enterprise, it is known that it suffers from certain paradoxes. The majority of the paradoxes evaporates under close scrutiny yet some of them – most notably the contrary-to-duty paradoxes – persist. The conceptual issue of these paradoxes is how to proceed once a norm has been violated. Clearly, this issue is of great practical relevance, because in most applications norms are violated frequently. Usually it is stipulated in the fine print of a contract what has to be done if a term in the contract is violated. If the violation is not too serious, or was not intended by the violating party, the contracting parties usually do not want to consider this as a breach of contracts, but simply as a disruption in the execution of the contract that has to be repaired. Hence, the contrary-to-duty paradoxes are important benchmark examples of deontic logic, and deontic logics incapable of dealing with them are considered insufficient tools to analyze deontic reasoning. In this paper we restrict ourselves to the area of *preference-based* deontic logic [12, 23, 19, 10, 15, 17, 4], i.e. deontic logics based on a preference relation in the semantics, because they have proven to be the most appropriate to solve the paradoxes. Still, preference-based deontic logic has the following three 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 is only in force in a sub-ideal situation. For example, the obligation to apologize for a broken promise is only in force in the sub-ideal situation where the obligation to keep promises is violated. 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 like Chisholm's and Forrester's paradoxes [6, 9].

Dilemma problem. Most deontic logicians take the perspective that deontic logic formalizes the reasoning of an authority issuing norms, and such an authority does not intentionally create dilemmas. Consequently, dilemmas should be inconsistent. However, some deontic logicians, most notably in computer science, try to model obligations in practical reasoning, and in real life dilemmas exist (see e.g. [40]). Consequently, according to this alternative perspective on deontic logic dilemmas should be consistent. In this paper we follow the traditional and mainstream perspective. 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 they should therefore be inconsistent. However, the formula $O(\alpha|\beta_1) \wedge O(\neg\alpha|\beta_2)$ does not represent a dilemma and should be consistent.

In this paper we propose Prohairetic (i.e. Preference-based) Deontic Logic (PDL), a logic of non-defeasible obligations in which dilemmas are inconsistent. The basic idea of this logic is that an obligation 'α should be (done) if β is (done)' is true if (1) no $\neg\alpha \wedge \beta$ state is as preferable as an $\alpha \wedge \beta$ state and (2) the preferred β states are α states. The first part of the definition is used to formalize contrary-to-duty reasoning, and the second part is used to make dilemmas inconsistent. PDL shares the intuitive semantics of preference-based deontic logics, and solves the strong preference problem without introducing additional semantic machinery like bi-ordering semantics or ceteris paribus preferences. Moreover, PDL shares the intuitive formalization of contrary-to-duty reasoning of dyadic deontic logic [12, 23]. Finally, PDL solves the dilemma problem by making the right set of formulas inconsistent.

This paper is organized as follows. We first discuss the strong preference problem (Section 2) and the contrary-to-duty and dilemma problems (Section 3). We then give an axiomatization of PDL in a modal preference logic (Section 4), and reconsider the three problems in PDL (Section 5).

2 Obligations and preferences

It has been suggested [20, 19, 10, 15, 17, 4] that a unary deontic operator O might be defined in a logic of preference by $O\alpha =_{def} \alpha \succ \neg\alpha$. In the following discussion we assume possible worlds models with a deontic preference relation on the worlds, i.e. Kripke models $\langle W, \leq, V\rangle$ which consist of a set of worlds W, a binary reflexive and transitive accessibility relation $\leq$ and a valuation function V. The problem is how to lift the preferences between worlds to preferences between sets of worlds or propositions, written as $\alpha_1 \succ \alpha_2$. It is well-known from the preference logic literature [41] 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 unrelated obligations like 'be polite' Op and 'be helpful' Oh would conflict when considering 'being polite and unhelpful' $p \wedge \neg h$ and 'being impolite and helpful' $\neg p \wedge h$. Proof-theoretically, if a preference relation $\succ$ has left and right strengthening, then the two preferences $p \succ \neg p$ and $h \succ \neg h$ derive $(p \wedge \neg h) \succ (\neg p \wedge h)$ and $(\neg p \wedge h) \succ (p \wedge \neg h)$. The two derived preferences seem contradictory.

The conflict can be resolved with additional information. For example, politeness may be less important than helpfulness, such that $(p \wedge \neg h) \succ (\neg p \wedge h)$ is less important than $(\neg p \wedge h) \succ (p \wedge \neg h)$. This relative importance of obligations can only be formalized in a defeasible deontic logic, in which obligations can be overridden by other obligations (for an analysis of the conceptual distinctions, see [33, 35]), because $(p \wedge \neg h) \succ (\neg p \wedge h)$ is overridden by $(\neg p \wedge h) \succ (p \wedge \neg h)$, and Op is not in force when only $(p \wedge \neg h) \vee (\neg p \wedge h)$ worlds are considered. However, in this paper we only consider non-defeasible or non-overridable obligations. For such logics, the following three solutions have been considered.

Bi-ordering. Jackson [19] and Goble [10] introduce a second ordering representing degrees of 'closeness' of worlds to solve the strong preference problem. 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 underlying idea is 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 an agent's control. 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.[1]

[1] This solution of the strong preference problem introduces an *irrelevance* problem, because the preferences no longer have left and right strengthening. For example, the preference $(p \wedge h) \succ (\neg p \wedge h)$ cannot even be derived from 'be polite' $p \succ \neg p$, because $p \wedge h$ or $\neg p \wedge h$ may not be among the closest p or $\neg p$ worlds. There is another interpretation of the closeness ordering. 'The closest' could also be interpreted as 'the most normal' as used in the preferential semantics of logics of defeasible reasoning [30]. The 'multi preference' semantics is a formalization of *defeasible* deontic logic [33, 35], sometimes called the logic of prima facie obligations. However, it is not clear that closeness is an intuitive concept for non-defeasible obligations.

Ceteris paribus. Hansson [15, 14, 13] defines $\alpha \succ \neg\alpha$ by a 'ceteris paribus' preference of α to $\neg\alpha$, see [4] for a discussion.[2] That is, for each pair of α and $\neg\alpha$ worlds that are identical except for the evaluation of α, the α world is preferred to the $\neg\alpha$ world. The obligation '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 $\neg p \wedge \neg h$, and neither about $p \wedge \neg h$ and $\neg p \wedge h$. Likewise, the obligation 'be helpful' Oh prefers 'polite and helpful' $p \wedge h$ to 'polite and unhelpful' $p \wedge \neg h$, and 'impolite and helpful' $\neg p \wedge h$ to 'impolite and unhelpful' $\neg p \wedge \neg h$. These preferences can be combined in a single preference ordering for $Op \wedge Oh$ that prefers $p \wedge h$ worlds to all other worlds, and that prefers all worlds to $\neg p \wedge \neg h$ worlds.[3]

Consistent dilemmas. Finally, a preference $\alpha \succ \neg\alpha$ can be defined by 'every $\neg\alpha$ world is not as preferable as any α world' (or, maybe more intuitively, $\alpha \not\succ \neg\alpha$ is defined as 'there is an $\neg\alpha$ world w_1 which is at least as preferable as an α world w_2'). The definition is equivalent to the problematic '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 in a model M 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 [40]. In contrast to the other solutions of the strong preference problem, dilemmas like $Op \wedge O\neg p$ are consistent. We say that the logic does not have the no-dilemma assumption. The preference relation $\succ$ has left and right strengthening, and $p \succ \neg p$ and $h \succ \neg h$ imply $(p \wedge \neg h) \succ (\neg p \wedge h)$ 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.[4]

---

[2] Moreover, Hansson defines obligations by the property of negativity. According to this principle, what is worse than something wrong is itself wrong. See [15, 4] for a discussion on this assumption.

[3] In [16] Hansson rejects the use of ceteris paribus preferences for obligations (in contrast to, for example, desires). Moreover, 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. [29] for an ad hoc solution of the problem).

[4] It was already argued by von Wright [41] that this latter property is highly implausible for preferences. On the other hand, this solution is simpler than the first two solutions of the strong preference problem, because it does not use additional semantic machinery such as the second ordering or the ceteris paribus preferences. Moreover, it does not have an irrelevance or an independence problem.

Prohairetic Deontic Logic (PDL) proposed in this paper is an extension of the third approach. To formalize the no-dilemma assumption, we write $M \models I\alpha$ for 'the ideal worlds in M satisfy α.' Hence, if we ignore infinite descending chains[5] then we can define that $M \models I\alpha$ if and only if $Pref \subseteq \mid \alpha \mid$ where $Pref$ stands for the set of most preferred (ideal) worlds of M, and $\mid \alpha \mid$ stands for the set of all worlds satisfying α. Obligations are defined as a combination of a strong preference and an ideal preference.

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

The formula $Op \wedge O\neg p$ is inconsistent, because the formula $Ip \wedge I\neg p$ is inconsistent. 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.

3 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 [2], Chisholm [6] and Forrester [9] paradoxes. The formalization of these paradoxes should be consistent. For example, the formalization of the Forrester paradox 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, even if we have 'gentle killing implies killing,' i.e. $\vdash g \rightarrow k$, see e.g. [11]. 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. Deontic logicians therefore tried to formalize contrary-to-duty reasoning by introducing temporal and preferential notions [37].

B. Hansson [12] and Lewis [23] argued that the contrary-to-duty problem can be solved by introducing dyadic obligations. A dyadic obligation $O(\alpha|\beta)$ is read as 'α ought to be (done) if β is (done).' They define a dyadic obligation by $O_{HL}(\alpha|\beta) =_{def} I(\alpha|\beta)$, where we write $I(\alpha|\beta)$ for 'the ideal β worlds satisfy α.' Hence, if we again ignore infinite descending chains, then we define $M \models I(\alpha|\beta)$ if and only if $Pref(\beta) \subseteq |\alpha|$, where $Pref(\beta)$ stands for the preferred β worlds of M. The introduction of the dyadic representation was inspired by the standard way

[5] The problems caused by infinite descending chains are illustrated by the following example. Assume a model that consists of one infinite descending chain of $\neg\alpha$ worlds. It seems obvious that the model should not satisfy $I\alpha$. However, the most preferred worlds (which do not exist!) satisfy α. See [22, 3] for a discussion.

of representing conditional probability, that is, by $Pr(\alpha|\beta)$ which stands for 'the probability that α is the case given β.' In a dyadic deontic logic, the Forrester paradox can be formalized by 'Smith should not kill Jones' $O(\neg k|\top)$, 'if Smith kills Jones, then he should do it gently' $O(g|k)$ and 'Smith kills Jones' k. In this formalization, $\top$ stands for any tautology, e.g. $p \vee \neg p$. The obligation $O(g|k)$ is a contrary-to-duty (CTD) obligation of $O(\neg k|\top)$, because an obligation $O(\alpha|\beta)$ is a CTD obligation of the primary obligation $O(\alpha_1|\beta_1)$ if and only if $\alpha_1 \wedge \beta$ is inconsistent. In dyadic deontic logic, the formula $O(\neg k|\top) \wedge O(g|k)$ is consistent, whereas the formula $O(\neg k|\top) \wedge O(g|\top)$ is inconsistent when we have $\vdash g \rightarrow k$.

The Hansson-Lewis dyadic deontic logics have been criticized (see e.g. [24]), because they do not have factual detachment, represented by the formula **FD** $O(\alpha\,|\,\beta) \wedge \beta \rightarrow O\alpha$, i.e. the derivation of absolute obligations from dyadic ones. However, there are good reasons not to accept **FD**. If the logic would have **FD**, then they would reinstate the Forrester paradox, because we would derive $O\neg k \wedge Og$ from $O(\neg k\,|\top) \wedge O(g\,|\,k) \wedge k$. To explicate the difference with dyadic obligations which do have factual detachment and therefore cannot represent the Forrester paradox, we prefer to call the Hansson-Lewis obligations contextual obligations, see also [34]. Instead of **FD** we may have $O(\alpha\,|\,\beta \wedge O(\alpha\,|\,\beta))$ as a theorem, see [43].

Dyadic obligations formalize contrary-to-duty reasoning, without making dilemmas like $O\alpha_1 \wedge O(\neg\alpha_1 \wedge \alpha_2)$ consistent. However, the dyadic representation also introduces a new instance of the dilemma problem, represented by the formula $O(\alpha\,|\,\beta_1) \wedge O(\neg\alpha\,|\,\beta_1 \wedge \beta_2)$. An example is Prakken and Sergot's considerate assassin example, that consists of the two obligations 'Smith should not offer Jones a cigarette' $O(\neg c|\top)$ and 'Smith should offer Jones a cigarette, if he kills him' $O(c\,|\,k)$. Prakken and Sergot [26] argue that the two sentences of the considerate assassin example represent a dilemma, because the obligation $O(c|k)$ is not a CTD obligation of $O(\neg c|\top)$. Hence, $O(\neg c|\top) \wedge O(c|k)$ should be inconsistent, even when there is another premise 'Smith should not kill Jones' $O(\neg k|\top)$.

B.Hansson-Lewis dyadic deontic logics do not give a satisfactory solution for the dilemma problem, because $O_{HL}(\neg c|\top) \wedge O_{HL}(c|k)$ 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 $S = \{O(\neg c \mid \top), O(c \mid k)\}$ is inconsistent, because the formula $(\neg c \succ c) \wedge I(c|k)$ is inconsistent, as is shown in Section 5.

4 Axiomatization

Prohairetic Deontic Logic (PDL) is defined in a modal preference logic. The standard Kripke models $M = \langle W, \leq, V\rangle$ of PDL contain a binary accessibility relation $\leq$, that is interpreted as a reflexive and transitive *deontic preference relation.* The advantages of our formalization in a modal framework are twofold.

First, when a dyadic operator is given by a definition in an underlying logic, then we get an axiomatization for free! We do not have to look for a sound and complete set of inference rules and axiom schemata, because we simply take the axiomatization of the underlying logic together with the new definition. In other words, the problem of finding a sound and complete axiomatization is replaced by the problem of finding a definition of a dyadic obligation in terms of a monadic modal preference logic. The second advantage of a modal framework in which all operators are defined, is that $I(\alpha \mid \beta)$ and $\alpha_1 \succ \alpha_2$ can be defined separately. In this section we axiomatize PDL in the following three steps in terms of a monadic modal logic and a deontic betterness relation. See [10, 14] for an analogous stepwise construction of 'good' in terms of 'better' and [43] for a stepwise construction of minimizing conditionals analogous to $O_{HL}(\alpha \mid \beta)$ in terms of a 'betterness' relation.

Ideality (deontic preference) ordering. We start with two monadic modal operators $\Box$ and $\overleftrightarrow{\Box}$. The formula $\Box\alpha$ can be read as 'α is true in all worlds at least as good (as the actual world)' or '$\neg\alpha$ is necessarily worse,' and $\overleftrightarrow{\Box}\alpha$ can be read as 'α is true in all worlds.'

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

The $\Box$ operator will be treated as an S4 modality and the $\overleftrightarrow{\Box}$ operator as an S5 modality. As is well-known, the standard system S4 is characterized by a partial pre-ordering: the axiom **T**: $\Box\alpha \rightarrow \alpha$ characterizes reflexivity and the axiom **4**: $\Box\alpha \rightarrow \Box\Box\alpha$ characterizes transitivity [18, 5]. Moreover, the standard system S5 is characterized by S4 plus the axiom **5**:$\neg\overleftrightarrow{\Box}\alpha \rightarrow \overleftrightarrow{\Box}\neg\overleftrightarrow{\Box}\alpha$. The relation between the modal operators is given by $\overleftrightarrow{\Box}\alpha \rightarrow \Box\alpha$. This is analogous to the well known relation between the modal operators for knowledge K and belief B given by $Kp \rightarrow Bp$.

Deontic betterness relation. A binary betterness relation $\alpha_1 \succ \alpha_2$, to be read as 'α_1 is preferred to α_2,' is defined in terms of the monadic operators. The following betterness relation obeys von Wright's expansion principle [41], because a preference of α_1 over α_2 only compares the two formulas $\alpha_1 \wedge \neg\alpha_2$ and $\neg\alpha_1 \wedge \alpha_2$.

$$\alpha_1 \succ \alpha_2 =_{def} \overleftrightarrow{\Box}(\alpha_1 \wedge \neg\alpha_2 \rightarrow \Box\neg(\alpha_2 \wedge \neg\alpha_1))$$

We have $M, w \models \alpha_1 \succ \alpha_2$ if we have $w_2 \not\leq w_1$ for all worlds $w_1, w_2 \in W$ such that $M, w_1 \models \alpha_1$ and $M, w_2 \models \alpha_2$, where we write as usual $\Diamond\alpha =_{def} \neg\Box\neg\alpha$. The betterness relation $\succ$ is quite weak. For example, it is not anti-symmetric (i.e. $\neg(\alpha_2 \succ \alpha_1)$ cannot be derived from $\alpha_1 \succ \alpha_2$) and it is not transitive (i.e. $\alpha_1 \succ \alpha_3$ cannot be derived from $\alpha_1 \succ \alpha_2$ and $\alpha_2 \succ \alpha_3$). It is easily checked that the lack of these properties is the result of the fact that we do not have totally connected orderings.

Obligatory. What is obligatory is defined in terms of deontic betterness.

$$I(\alpha|\beta) \;=_{def} \overleftrightarrow{\Box}(\beta \rightarrow \Diamond(\beta \wedge \Box(\beta \rightarrow \alpha)))$$
$$O(\alpha|\beta) =_{def} (\alpha \wedge \beta \succ \neg\alpha \wedge \beta) \wedge I(\alpha|\beta)$$

We have $M, w \models I(\alpha \mid \beta)$ if the preferred β worlds are α worlds, and α eventually becomes true in all infinite descending chains of β worlds [21, 3]. Finally, we have $M, w \models O(\alpha|\beta)$ if we have $w_2 \not\leq w_1$ for all $w_1, w_2 \in W$ such that $M, w_1 \models \alpha \wedge \beta$ and $M, w_2 \models \neg\alpha \wedge \beta$, and $M, w \models I(\alpha|\beta)$.

The logic PDL is defined by defining these three layers in a modal preference logic.

Definition 1 (PDL). *The bimodal language $\mathcal{L}$ is formed from a denumerable set of propositional variables together with the connectives $\neg$, $\rightarrow$, and the two normal modal connectives $\Box$ and $\overleftrightarrow{\Box}$. Dual 'possibility' connectives $\Diamond$ and $\overleftrightarrow{\Diamond}$ are defined as usual by $\Diamond\alpha =_{def} \neg\Box\neg\alpha$ and $\overleftrightarrow{\Diamond}\alpha =_{def} \neg \overleftrightarrow{\Box} \neg\alpha$.*

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	$\Box(\alpha \rightarrow \beta) \rightarrow (\Box\alpha \rightarrow \Box\beta)$	**K'**	$\overleftrightarrow{\Box}(\alpha \rightarrow \beta) \rightarrow (\overleftrightarrow{\Box}\alpha \rightarrow \overleftrightarrow{\Box}\beta)$
T	$\Box\alpha \rightarrow \alpha$	**T'**	$\overleftrightarrow{\Box}\alpha \rightarrow \alpha$
4	$\Box\alpha \rightarrow \Box\Box\alpha$	**4'**	$\overleftrightarrow{\Box}\alpha \rightarrow \overleftrightarrow{\Box}\overleftrightarrow{\Box}\alpha$
R	$\overleftrightarrow{\Box}\alpha \rightarrow \Box\alpha$	**5'**	$\neg\overleftrightarrow{\Box}\alpha \rightarrow \overleftrightarrow{\Box}\neg\overleftrightarrow{\Box}\alpha$

Nes *From α infer $\overleftrightarrow{\Box}\alpha$*
MP *From $\alpha \rightarrow \beta$ and α infer β*

extended with the following three definitions.

$$O(\alpha|\beta) \;=_{def} (\alpha \wedge \beta \succ \neg\alpha \wedge \beta) \wedge I(\alpha|\beta)$$
$$\alpha_1 \succ \alpha_2 =_{def} \overleftrightarrow{\Box}(\alpha_1 \wedge \neg\alpha_2 \rightarrow \Box\neg(\alpha_2 \wedge \neg\alpha_1))$$
$$I(\alpha|\beta) \;=_{def} \overleftrightarrow{\Box}(\beta \rightarrow \Diamond(\beta \wedge \Box(\beta \rightarrow \alpha)))$$

Definition 2 (PDL Semantics). *Kripke 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: $w_1 \leq w_2$ iff w_1 is at least as preferable as w_2. The modal connective $\Box$ refers to accessible worlds and the modal connective $\overleftrightarrow{\Box}$ to all worlds.*

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

The following proposition shows that, as a consequence of the definition in a standard bimodal logic, the soundness and completeness of PDL are trivial.

Proposition 1 (Soundness and completeness of PDL). *Let $\vdash_{\text{PDL}}$ and $\models_{\text{PDL}}$ stand for derivability and logical entailment in the logic PDL. We have $\vdash_{PDL} \alpha$ if and only if $\models_{PDL} \alpha$.*

Proof. *Follows directly from standard modal soundness and completeness proofs, see e.g. [18, 5, 28].*

We now consider several properties of the dyadic obligations. First, the logic PDL has the following theorem, which is valid for any preference-based deontic logic defined by $O(\alpha|\beta) =_{def} \alpha \wedge \beta \succ \neg\alpha \wedge \beta$ (for any betterness relation $\succ$) or by $O(\alpha|\beta) =_{def} I(\alpha|\beta)$.

R $O(\alpha|\beta_1 \wedge \beta_2) \leftrightarrow O(\alpha \wedge \beta_1|\beta_1 \wedge \beta_2)$,

Second, the logic PDL does not have closure under logical implication. This is a typical property of preference-based deontic logics. For example, the preference-based deontic logics discussed in [19, 15, 10, 4, 17] do not have closure under logical implication either. The following theorem *Weakening of the Consequent* **WC** is *not* valid in PDL.

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

The third property we consider is the following *disjunction rule* **OR**, related to *Reasoning-By-Cases* and Savage's sure-thing principle. It is *not* valid either.

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

The fourth property we consider is so-called *Restricted Strengthening of the Antecedent* **RSA**, expressed by the following theorem of the logic. In can easily be shown that $O(\alpha|\beta_1 \wedge \beta_2)$ can only be derived in PDL from $O(\alpha|\beta_1)$ when we have $I(\alpha|\beta_1 \wedge \beta_2)$ as well.

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

We can add strengthening of the antecedent with the following notion of preferential entailment, that prefers maximally connected models. We say that a model is more connected if its binary relation contains more elements (in our terminology $\{(w_1, w_2), (w_2, w_1)\}$ is therefore more connected than $\{(w_1, w_2)\}$).

Definition 3 (Preferential entailment). *Let the two possible worlds models $M_1 = \langle W, \leq_1, V\rangle$ and $M_2 = \langle W, \leq_2, V\rangle$ be two PDL models. M_1 is at least as connected 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 more connected than 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 maximally connected models M of T.*

The maximally connected models of a set of obligations are unique (for a given W and V) if the transitivity axiom **4**: $\Box\alpha \rightarrow \Box\Box\alpha$ is omitted from the axiomatization. The unique maximally connected model of the set of obligations $S = \{O(\alpha_i \,|\, \beta_i) \mid 1 \leq i \leq n\}$ has the accessibility relation $\{w_1 \leq w_2 \mid$ there is no $O(\alpha|\beta) \in S$ such that $M, w_1 \models \neg\alpha \wedge \beta$ and $M, w_2 \models \alpha \wedge \beta\}$. However, if axiom **4** is omitted, then 'preferred' in $I(\alpha \,|\, \beta)$ no longer has a natural meaning. Finally, if only full models are considered in the semantics, i.e. models that contain a world for each possible interpretation, then we can derive for example $\emptyset \models_{\sqsubset} \neg O(p|\top)$, because there cannot be a model with only p worlds. In the following section preferential entailment is illustrated by several examples.

5 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 1 (Polite and helpful, continued). Consider the three sets of obligations $S = \emptyset$, $S' = \{O(p|\top)\}$ and $S'' = \{O(p|\top), O(h|\top)\}$. The three unique maximally connected models of S, S' and S'' are represented in Figure 1. With no premises,

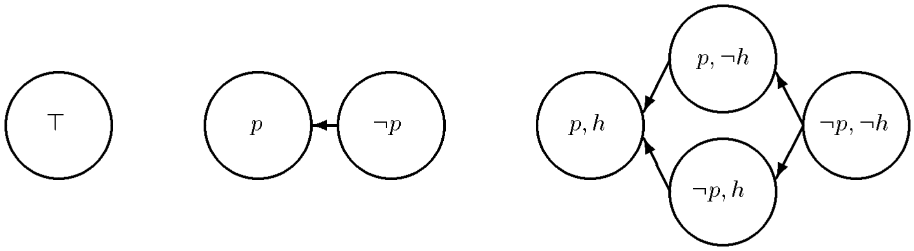

Fig. 1. Unique maximally connected models of $\emptyset$, $\{O(p|\top)\}$ and $\{O(p|\top), O(h|\top)\}$.

all worlds are equally ideal. By addition of the premise $O(p|\top)$, the p worlds are strictly preferred over $\neg p$ worlds. Moreover, 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 strong preference problem is solved by representing conflicts with incomparable worlds. This solution uses preferential entailment, a technique from non-monotonic reasoning, because for the preferred models we have that all incomparable worlds refer to some conflict. We have $S' \models_{\sqsubset} O(p|\neg(p \wedge h))$ and $S'' \not\models_{\sqsubset} O(p|\neg(p \wedge h))$. 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 Forrester's [9] and Chisholm's [6] paradoxes in Prohairetic Deontic Logic. The paradoxes were originally formulated in monadic deontic logic.

Example 2 (Forrester's paradox). Consider $S = \{O(\neg k \mid \top), O(g \mid k), k\}$, where k can be read as 'Smith kills Jones' and g as 'Smith kills him gently,' and g logically implies k. The unique maximally connected model of S is represented in Figure 2. The actual world is any of the k worlds. The formalization of S is unproblematic and the semantics reflect the three states that seem to be implied by the paradox. We have $S \models O(\neg k \vee g \mid \top)$ as a consequence of the theorem $(O(\alpha \mid \neg\beta) \wedge O(\beta \mid \gamma)) \rightarrow O(\alpha \vee \beta \mid \gamma)$, which expresses that $k \wedge \neg g$ is the worst state that should be avoided.

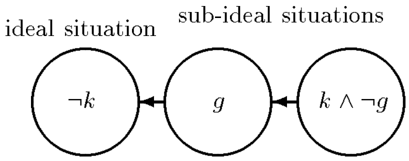

Fig. 2. Unique maximally connected model of $\{O(\neg k|\top), O(g|k), k\}$

Example 3 (Chisholm's paradox). Consider $S = \{O(a|\top), O(t|a), O(\neg t|\neg a), \neg a\}$, where a can be read as 'a certain man going to the assistance of his neighbors' and t as 'telling the neighbors that he will come.' The unique maximally connected model of S is represented in Figure 3. The crucial question of Chisholm's

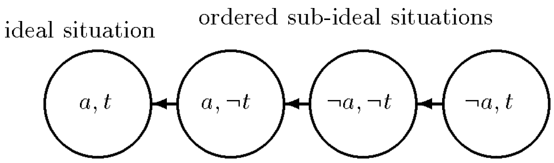

Fig. 3. Unique maximally connected model of $\{O(a|\top), O(t|a), O(\neg t|\neg a), \neg a\}$

paradox is whether there is an obligation that 'the man should tell the neighbors that he will come' $O(t \mid \top)$. This obligation is counterintuitive, given that a is false. This obligation is derived from S by any deontic logic that has so-called deontic detachment (also called deontic transitivity), represented by the following formula $\mathbf{DD}^0$.

$\mathbf{DD}^0$ $(O(\alpha|\beta) \wedge O(\beta|\gamma)) \rightarrow O(\alpha|\gamma)$

However, $\mathbf{DD}^0$ is not valid in Prohairetic Deontic Logic. The obligation 'the man should tell his neighbors that he will come' $O(t \mid \top)$ cannot be derived in PDL from S. The following weaker version of $\mathbf{DD}^0$ is valid in PDL.

$$\mathbf{DD}\ (O(\alpha|\beta \wedge \gamma) \wedge O(\beta|\gamma) \wedge I(\alpha \wedge \beta|\gamma)) \rightarrow O(\alpha \wedge \beta|\gamma)$$

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

Finally, we show that Prohairetic Deontic Logic solves the dilemma problem, because it makes the considerate assassin set in Example 4 inconsistent, without making the window set in Example 5 inconsistent.

Example 4 (Considerate assassin). Consider $S = \{O(\neg c \mid \top), O(c \mid k)\}$, 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 $\overset{\leftrightarrow}{\Diamond}(k \wedge \neg c)$, as can be verified as follows. Assume there is a model of S. The obligation $O(c \mid k)$ implies $I(c \mid k)$, 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 such a world w_1).

Moreover, consider $S' = \{O(\neg d \mid \top), O(d \wedge p \mid d), O(\neg p \mid \top)\}$, where d can be read as 'there is a dog' and p as 'there is a poodle.' Prakken and Sergot argue [27] that S' should be inconsistent based on its analogy with S. For similar reasons as S above, the set S' is inconsistent in PDL.

Example 5 (Window). Consider $S = \{O(c \mid r), O(\neg c \mid s)\}$, where c can be read as 'the window is closed,' r as 'it starts raining' and s as 'the sun is shining.' It is argued by von Wright [42] that S does not represent a dilemma and that it should therefore be consistent, see also [1]. In PDL the set S is consistent, and a maximally connected model M of S is given in Figure 4. The ideal worlds satisfy $r \rightarrow c$ and $s \rightarrow \neg c$, and the sub-ideal worlds either $\neg c \wedge r$ or $c \wedge s$. We have $M \not\models O(c|r \wedge s)$ and thus $S \not\models_{\sqsubset} O(c|r \wedge s)$.

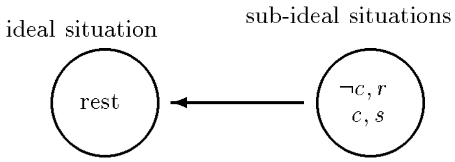

Fig. 4. Preferred model of $\{O(c|r), O(\neg c|s)\}$

Note that there are many maximally connected models. For example, the $c \wedge r$ worlds can be the only preferred worlds, when the $\neg c \wedge s$ worlds are equivalent with the $\neg c \wedge r$ worlds. Alternatively, the $\neg c \wedge s$ worlds can be the only most preferred worlds.

6 Related research

Besides the deontic logics already discussed in this paper, there is a preference-based deontic logic proposed by Brown, Mantha and Wakayama [4]. We write the obligations in their logic as O_{BMW}. At first sight it seems that the logic is closely related to our Prohairetic Deontic Logic, because $O_{BMW}\alpha$ also has a mixed representation. However, a further inspection of the definitions reveals that their proposal is quite different from ours. Obligations are defined by

$$O_{BMW}\alpha =_{def} P_f\alpha \wedge A_m\alpha = \overleftarrow{\Box}\neg\alpha \wedge \Diamond\alpha$$

where $P_f\alpha$ is read as 'α is preferred,' $A_m\alpha$ is read as 'α is admissible,' and $\overleftarrow{\Box}\alpha$ is read as 'α is true in all inaccessible worlds.' Hence, $O_{BMW}\alpha$ means 'the truth of α takes us to a world at least as good as the current world and there exists a world at least as good as the current world where α is true' [4, p.200]. The first distinction is that in the logic, dilemmas are consistent (which they consider an advantage, following [40]). Secondly, the motivation for the mixed representation is different. Whereas we introduced the mixed representation to solve both the contrary-to-duty problem (for which we use $\alpha \succ \neg\alpha$) and the dilemma problem (for which we use $I\alpha$), they use the mixed representation to block the derivation of the theorem $O_{BMW}(\alpha_1 \vee \alpha_2) \rightarrow O_{BMW}\alpha_1$, which they consider as 'somewhat unreasonable, since the premise is weaker than the conclusion.' However, it is easily checked that the logic validates $O_{BMW}(\alpha_1 \vee \alpha_2) \wedge A_m\alpha_1 \rightarrow O_{BMW}\alpha_1$. Hence, under certain circumstances stronger obligations can be derived. This counterintuitive formula is not a theorem of PDL.[6]

7 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. We now study the use of PDL for legal expert systems and to specify intelligent agents for the Internet [7, 8] like the drafting, negotiation and processing of trade contracts in electronic commerce, and the relevance for logics of desires and goals as these are developed in qualitative decision theory [25, 36].

Acknowledgements

Thanks to David Makinson for comments on an earlier version of this paper.

[6] The logic PDL is the basis of the logics developed in [32]. A previous version appeared as a part of the two-phase system [30]. Moreover, the logic PDL has been presented at the AAAI 1997 spring symposium on Qualitative Approaches to Deliberation and Reasoning. Finally, the dynamic interpretation of PDL with preferential entailment is related to more recent work on dynamic update semantics for obligations [38, 39] though the formal construction is radically different and different issues have been discussed (in particular the different notions of defeasibility, following [33, 35]).

References

1. C.E. Alchourrón. Philosophical foundations of deontic logic and the logic of defeasible conditionals. In J.-J. Meyer and R. Wieringa, editors, *Deontic Logic in Computer Science: Normative System Specification*, pages 43–84. John Wiley & Sons, 1993.
2. L. Åqvist. Good Samaritans, contrary-to-duty imperatives, and epistemic obligations. *Noûs*, 1:361–379, 1967.
3. C. Boutilier. Conditional logics of normality: a modal approach. *Artificial Intelligence*, 68:87–154, 1994.
4. A.L. Brown, S. Mantha, and T. Wakayama. Exploiting the normative aspect of preference: a deontic logic without actions. *Annals of Mathematics and Artificial Intelligence*, 9:167–203, 1993.
5. B.F. Chellas. *Modal Logic: An Introduction.* Cambridge University Press, 1980.
6. R.M. Chisholm. Contrary-to-duty imperatives and deontic logic. *Analysis*, 24:33–36, 1963.
7. R. Conte and R. Falcone. ICMAS'96: Norms, obligations, and conventions. *AI Magazine*, 18,4:145–147, 1997.
8. B.S. Firozabadhi and L.W.N. van der Torre. Towards an analysis of control systems. In H. Prade, editor, *Proceedings of the ECAI'98*, pages 317–318, 1998.
9. J.W. Forrester. Gentle murder, or the adverbial Samaritan. *Journal of Philosophy*, 81:193–197, 1984.
10. L. Goble. A logic of good, would and should, part 2. *Journal of Philosophical Logic*, 19:253–276, 1990.
11. L. Goble. Murder most gentle: the paradox deepens. *Philosophical Studies*, 64:217–227, 1991.
12. B. Hansson. An analysis of some deontic logics. In R. Hilpinen, editor, *Deontic Logic: Introductionary and Systematic Readings*, pages 121–147. D. Reidel Publishing Company, Dordrecht, Holland, 1971. Reprint from *Noûs*, 1969.
13. S.O. Hansson. A new semantical approach to the logic of preference. *Erkenntnis*, 31:1–42, 1989.
14. S.O. Hansson. Defining "good" and "bad" in terms of "better". *Notre Dame Journal of Formal Logic*, 31:136–149, 1990.
15. S.O. Hansson. Preference-based deontic logic (PDL). *Journal of Philosophical Logic*, 19:75–93, 1990.
16. S.O. Hansson. Situationist deontic logic. *Journal of Philosophical Logic*, 26:423–448, 1997.
17. Z. Huang and M. Masuch. The logic of permission and obligation in the framework of ALX3: how to avoid the paradoxes of deontic logic. *Logique et Analyse*, 149, 1997.
18. H.G. Hughes and M.J. Creswell. *A Companion to Modal Logic.* Methuen, London, 1984.
19. F. Jackson. On the semantics and logic of obligation. *Mind*, 94:177–196, 1985.
20. R.E. Jennings. Can there be a natural deontic logic? *Synthese*, 65:257–274, 1985.
21. P. Lamarre. S4 as the conditional logic of nonmonotonicity. In *Proceedings of the KR'91*, pages 357–367, 1991.
22. D. Lewis. *Counterfactuals.* Blackwell, Oxford, 1973.
23. D. Lewis. Semantic analysis for dyadic deontic logic. In S. Stunland, editor, *Logical Theory and Semantical Analysis*, pages 1–14. D. Reidel Publishing Company, Dordrecht, Holland, 1974.

24. B. Loewer and M. Belzer. Dyadic deontic detachment. *Synthese*, 54:295–318, 1983.
25. J. Pearl. From conditional oughts to qualitative decision theory. In *Proceedings of the 9th Conference on Uncertainty in Artificial Intelligence (UAI'93)*, pages 12–20, 1993.
26. H. Prakken and M.J. Sergot. Contrary-to-duty obligations. *Studia Logica*, 57:91–115, 1996.
27. H. Prakken and M.J. Sergot. Dyadic deontic logic and contrary-to-duty obligations. In *Defeasible deontic logic*. Kluwer, 1996.
28. Y. Moses R. Fagin, J.Y. Halpern and M.Y. Vardi. *Reasoning About Knowledge*. MIT press, 1995.
29. S.-W. Tan and J. Pearl. Specification and evaluation of preferences under uncertainty. In *Proceedings of the KR'94*, pages 530–539, 1994.
30. Y.-H. Tan and L.W.N. van der Torre. Why defeasible deontic logic needs a multi preference semantics. In *Symbolic and Quantitative Approaches to Reasoning and Uncertainty (ECSQARU'95)*, LNAI 946, pages 412–419. Springer, 1995.
31. L.W.N. van der Torre. Violated obligations in a defeasible deontic logic. In *Proceedings of the ECAI'94*, pages 371–375, 1994.
32. L.W.N. van der Torre. *Reasoning About Obligations: Defeasibility in Preference-based Deontic Logics*. PhD thesis, Erasmus University Rotterdam, 1997.
33. L.W.N. van der Torre and Y.-H. Tan. Cancelling and overshadowing: two types of defeasibility in defeasible deontic logic. In *Proceedings of the IJCAI'95*, pages 1525–1532, 1995.
34. L.W.N. van der Torre and Y.-H. Tan. Contextual deontic logic. In *Proceedings of the First International Conference on Modeling and Using Context (CONTEXT'97)*, pages 1–12, Rio de Janeiro, Brazil, 1997.
35. L.W.N. van der Torre and Y.-H. Tan. The many faces of defeasibility in defeasible deontic logic. In D. Nute, editor, *Defeasible Deontic Logic*, pages 79–121. Kluwer, 1997.
36. L.W.N. van der Torre and Y.-H. Tan. Diagnosis and decision making in normative reasoning. *Artificial Intelligence and Law*, 1998.
37. L.W.N. van der Torre and Y.-H. Tan. The temporal analysis of Chisholm's paradox. In *Proceedings of the AAAI'98*, 1998.
38. L.W.N. van der Torre and Y.-H. Tan. An update semantics for deontic reasoning. In P. McNamara and H. Prakken, editors, *Norms, Logics and Information Systems. New Studies on Deontic Logic and Computer Science*. IOS Press, 1998.
39. L.W.N. van der Torre and Y.-H. Tan. An update semantics for prima facie obligations. In H. Prade, editor, *Proceedings of the ECAI'98*, pages 38–42, 1998.
40. B.C. van Fraassen. Values and the heart command. *Journal of Philosophy*, 70:5–19, 1973.
41. G.H. von Wright. *The Logic of Preference*. Edinburgh University Press, 1963.
42. G.H. von Wright. A new system of deontic logic. In R. Hilpinen, editor, *Deontic Logic: Introductory and Systematic Readings*, pages 105–120. D. Reidel Publishing company, Dordrecht, Holland, 1971.
43. E. Weydert. Hyperrational conditionals, reasoning about nested default conditionals. In *Proceedings of the European Workshop on Theoretical Foundations of Knowledge Representation and Reasoning of the ECAI'92*. Springer, 1994.
44. R.J. Wieringa and J.-J.Ch. Meyer. Applications of deontic logic in computer science: A concise overview. In *Deontic Logic in Computer Science*, pages 17–40. John Wiley & Sons, Chichester, England, 1993.