Contrary-to-duty reasoning with preference-based dyadic obligations

Leendert van der Torre[a] and Yao-Hua Tan[b]

[a] *Department of Artificial Intelligence, Faculty of Sciences, Vrije Universiteit Amsterdam,*
De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands
E-mail: torre@cs.vu.nl

[b] EURIDIS, *Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands*
E-mail: ytan@fac.fbk.eur.nl

In this paper we introduce Prohairetic Deontic Logic (PDL), a preference-based dyadic deontic logic. In our preference-based interpretation of obligations "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 this representation solves different problems of deontic logic. The first part of the definition is used to formalize contrary-to-duty reasoning, which, for example, occurs in Chisholm's and Forrester's notorious deontic paradoxes. The second part is used to make deontic dilemmas inconsistent.

1. Introduction

Deontic logic is a modal logic in which absolute and conditional obligations are represented by the modal formulas $O\alpha$ and $O(\alpha \mid \beta)$, where the latter is either read as "α ought to be (done) if β is (done)" or as "α ought to be (done) in the context where 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 [87]). 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, observed when deontic logic was still a purely philosophical enterprise, become more pressing. Of particular practical interest are so-called contrary-to-duty and dilemma reasoning discussed below. For example, Jones and Sergot argue in [32,33] that contrary-to-duty reasoning is necessary to represent certain aspects of the legal code in legal expert systems. Unfortunately, this contrary-to-duty reasoning leads to notorious paradoxes of deontic logic.

The most notorious paradoxes of deontic logic are caused by the *Contrary-to-duty Problem* (CP). 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 [9,13]. 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, then 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.

Moreover, notorious paradoxes of deontic logic are caused by what we call the *Dilemma Problem* (DP). A dilemma is a deontic conflict, as for example experienced by Sartre’s famous soldier that has the obligation to kill and the (moral) obligation not to kill. In this paper we adopt the perspective that deontic logic formalizes the reasoning of an authority issuing norms, and we assume that such an authority does not intentionally create dilemmas. Consequently, dilemmas only occur in incoherent systems and they should thus be inconsistent. However, some deontic logicians, most notably in computer science, try to model obligations in practical reasoning, and in daily life dilemmas exist (see, e.g., [6,82]). Consequently, according to this alternative perspective dilemmas should be consistent. We call deontic logics with this property conflict tolerant deontic logics. Besides, other deontic logicians try to formalize prima facie obligations [53] that can be overridden by other prima facie obligations, for example based on the specificity principle or on a prioritisation of the norms. In this paper we follow the mainstream perspective and we therefore do not discuss conflict-tolerant or prima facie obligations. According to the deontic benchmark examples the two formulas $O(\alpha_1 \wedge \alpha_2) \wedge O\neg\alpha_1$ and $O(\alpha_1 \wedge \alpha_2 \mid \beta_1 \wedge \beta_2) \wedge O(\neg\alpha_1 \mid \beta_1)$ represent dilemmas (if $\beta_1 \wedge \beta_2$ does not imply α_1 [63]), but the formula $O(\alpha \mid \beta_1) \wedge O(\neg\alpha \mid \beta_2)$ does not represent a dilemma if neither β_1 implies β_2 nor vice versa. Consequently the first two should be inconsistent and the last one should be consistent. The conceptual problem of these paradoxes is to determine the coherence conditions of a normative system, because a coherent system does not contain dilemmas. In several applications it is relevant to know whether a normative system is coherent. For example, when drafting regulations that should be coherent a consistency check on the formalization of the regulations in deontic logic indicates whether they have this desired property, or whether they should be further modified.

Several solutions for either CP and DP have been proposed, but a solution that satisfactorily solves both problems at the same time has proven to be very hard. This was shown in particular by Prakken and Sergot, who introduced several examples of the combination of the CP and DP – which we call CDP – in [49,50]. For example, consider the following three sentences, in which $\top$ stands for any tautology like $p \vee \neg p$

$$\begin{tikzcd} O(\neg d \mid \top) \arrow[d, leftrightarrow, "\text{inconsistent}"'] \\ O(p \wedge d \mid d) \end{tikzcd}$$

a. CP

$$\begin{tikzcd} O(\neg(p \wedge d) \mid \top) \arrow[d, leftrightarrow, shift right=2, "\text{inconsistent}"'] \\ O(p \wedge d \mid d) \arrow[u, shift right=2, "\text{more specific}"'] \end{tikzcd}$$

b. DP

Figure 1. Combined CDP3: the poodle.

and for representational convenience the conjunction $p \wedge d$ stand for a poodle:

(1) $O(\neg d \mid \top)$: There must be no dog.
(2) $O(p \wedge d \mid d)$: There must be a poodle, if there is a dog.
(3) $O(\neg(p \wedge d) \mid \top)$: There must be no poodle.

Prakken and Sergot argue extensively that the first two sentences do not represent a dilemma and should be consistent, whereas the latter two represent a dilemma and therefore should be inconsistent (in sections 2.3 and 4.4 we further discuss the latter assumption). The distinction between the CP interpretation and the DP interpretation of the poodle example is visualized in figure 1. This figure should be read as follows. Each arrow is a condition: a two-headed arrow is an inconsistency, and a single-headed arrow is a logical implication. In this particular example, the dilemma is that an obligation $O(\alpha \mid \beta)$ conflicts with $O(\alpha' \mid \beta')$ because the conclusions are contradictory (the double-headed arrow), and in addition the condition of one of the obligations is more specific (β' logically implies β). Case (a) represents the CP interpretation of the CDP example, and case (b) represents its DP interpretation. Consequently, case (a) should be consistent in any deontic logic, and case (b) should be inconsistent. A test of this benchmark example later in this paper shows that in most deontic logics either the first two sentences are inconsistent (e.g., in variants of so-called Standard Deontic Logic [83]; these logics are too strong) or the three sentences together are consistent (e.g., in variants of the Hansson's dyadic deontic logic DSDL3 [19,38]; these logics are too weak).

In this paper we introduce *Prohairetic* (*i.e., Preference-based*) *Deontic Logic* (PDL).[1] 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 formalization of contrary-to-duty reasoning of dyadic deontic logic [19,38]. It shares the intuitive semantics of preference-based deontic logics, without introducing additional semantic machinery like bi-ordering semantics [15,28] or ceteris paribus preferences [22]. Moreover, PDL solves the dilemma problem by making the right set of formulas inconsistent.

[1] Not to be confused with the preference-based deontic logic PDL proposed in [22].

This paper is organized as follows. In section 2 we first give an informal motivation for PDL by discussing a set of deontic benchmark examples. We then give an axiomatization of PDL in a modal preference logic (section 3) and we reconsider the problems in PDL (section 4). Finally, we discuss related research (section 5) and how the logic can be extended to incorporate other desirable properties such as, for example, reasoning by cases (section 6).

2. Prohairetic Deontic Logic – informal motivation

In this paper we present a scenic tour through the land of deontic logic. During our tour we discuss the most important benchmark examples of deontic logic. We focus on so-called *preference-based* deontic logic, which can be interpreted in either one of the following two ways.

- The semantics of the deontic logics contains a binary relation which is transitive, as first introduced by Hansson [19] and studied by Lewis [38] in the framework of dyadic deontic logic. This property of transitivity makes it possible to define minimal or ideal elements of this transitive relation, which have a special status. In preferential semantics this transitive relation is called a preference relation, and the minimal elements are called the preferred elements (see, e.g., [34,55]). However, there is no direct link between this first type of preference-based deontic logic and either the logic of preference [84] or the logic of decision [29], although utilitarian semantics have been proposed too [5,30,48].
- In so-called prohairetic deontic logic the obligations of the deontic logic are defined in terms of an underlying logic of preference by $O\alpha =_{\text{def}} \alpha \succ \neg\alpha$ and $O(\alpha \mid \beta) =_{\text{def}} (\alpha \wedge \beta) \succ (\neg\alpha \wedge \beta)$, in which $\succ$ stands for an operator from the logic of preference [6,15,22,25,28,31]. For Hansson's logic later underlying logics of preference have been defined too [5,86], because "the preferred β are α" is equivalent to "the preferred $\alpha \wedge \beta$ are preferred to the preferred $\neg\alpha \wedge \beta$". However, it has been questioned whether there exists a general logic of preference which can be used in different domains – including normative reasoning [46].

We only consider preference-based obligations, because thus far they have proven to be the most appropriate to solve CP. However, preference-based obligations have the disadvantage that they lead to the so-called Strong preference Problem (SP) discussed in section 2.2 below. The most important deontic benchmark examples related to CP, DP and SP have been summarized in table 1, in which $\perp$ stands for a contradiction like $p \wedge \neg p$.

2.1. Contrary-to-duty problem (CP)

CP is the major problem of monadic deontic logic, as shown by the notorious Good Samaritan [2], Chisholm [9] and Forrester [13] paradoxes. The formalization of these paradoxes should be consistent. For example, the formalization of Forrester's

Table 1
Deontic benchmark examples.

		Premises	Desired	Undesired
[13]	CP1	$O(\neg k \mid \top)$, $O(g \mid k)$, k, $\vdash g \rightarrow k$		$\bot$
[49]	CP1a	$O(\neg f \mid \top)$, $O(w \wedge f \mid f)$, f		$\bot$
[9]	CP2	$O(a \mid \top)$, $O(t \mid a)$, $O(\neg t \mid \neg a)$, $\neg a$		$\bot$
[49]	CP3	$O(p \mid \top)$, $O(a' \mid \neg p)$, $\neg p$		$O(p \wedge a')$
[84]	SP	$O(p' \mid \top)$	$O(p' \mid \neg(p' \wedge h))$	
		$O(p' \mid \top)$, $O(h \mid \top)$		$O(p' \wedge \neg h \mid \neg(p' \wedge h))$
[83]	DP1	$O(p \mid \top)$, $O(\neg p \mid \top)$	$\bot^*$	
	DP2	$O(p \mid \top)$, $O(\neg p \wedge q \mid \top)$	$\bot^*$	
[85]	DP3	$O(c \mid r)$	$O(c \mid r \wedge s)$	
		$O(c \mid r)$, $O(\neg c \mid s)$		$O(c \mid r \wedge s)$, $\bot$
[49]	DP4	$O(\neg c' \mid \top)$, $O(c' \mid k)$	$\bot^{**}$	
[24]	DP4a	$O(\neg f' \mid \top), O(f' \mid a'')$	$\bot^{**}$	
[49]	CDP1	$O(\neg f \mid \top)$, $O(w \wedge f \mid f)$, $O(f \mid d)$	$\bot^{**}$	
[50]	CDP2	$O(\neg d \mid \top)$, $O(s' \mid d)$, $O(\neg s' \mid \top)$	$\bot^{**}$	
[49]	CDP2a	$O(\neg k \mid \top)$, $O(c' \mid k)$, $O(\neg c' \mid \top)$	$\bot^{**}$	
[50]	CDP3	$O(\neg d \mid \top)$, $O(p'' \wedge d \mid d)$, $O(\neg(p'' \wedge d) \mid \top)$	$\bot^{**}$	
[50]	CDP3a	$O(r' \mid \top)$, $O(u \mid \neg r')$, $O(\neg u \mid \top)$, $\vdash u \rightarrow \neg r'$	$\bot^{**}$	

The first column contains references to the papers in which the example was first described, and the second column contains the names used in this paper. The third row contains a formalization of the example in dyadic deontic logic, although many of them were first introduced in monadic deontic logic. The last two columns contain desired and undesired consequences.

k	kill	a	assistance	p'	polite	s	sun	d	dog	u	unpainted
g	gently	t	tell	h	honest	c'	cigarettes	s'	sign		
f	fence	p	promise	c	close	f'	fingers	p''	poodle		
w	white	a'	apologize	r	rain	a''	asparagus	r'	red		

$\bot^*$: in a conflict tolerant deontic logic this derivation is undesired.
$\bot^{**}$: in a logic of prima facie obligations and a conflict tolerant logic it is undesired.

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., [16]. 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 DP2 in table 1 is not a dilemma, and should be consistent. The consistency of $O\neg k \wedge Og$ is a solution that seems like overkill. For example, in Standard Deontic Logic (SDL)[2] $O\neg k \wedge Og$ is inconsistent. Deontic logicians therefore tried to formalize contrary-to-duty reasoning by introducing, for example, temporal and preferential notions [73].

[2] SDL is a modal system of type KD according to the Chellas classification [8]. It is the smallest set that contains the propositional theorems and the axioms K: $O(\beta \rightarrow \alpha) \rightarrow (O\beta \rightarrow O\alpha)$ and D: $\neg(O\alpha \wedge O\neg\alpha)$, and that is closed under the inference rules modus ponens and necessitation.

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

Figure 2. $O(\alpha \mid \beta)$ is a contrary-to-duty obligation of $O(\alpha_1 \mid \beta_1)$.

Hansson [19] and Lewis [38] showed how dyadic obligations can solve CP without making DP2 consistent. In the following discussion we assume possible worlds models with a deontic preference ordering on the worlds, i.e., Kripke models $\langle W, \leqslant, V\rangle$ which consist of a set of worlds W, a binary reflexive, transitive and connected accessibility relation $\leqslant$ and a valuation function V. They define a dyadic obligation by $O_{HL}(\alpha \mid \beta) =_{\text{def}} I(\alpha \mid \beta)$, where we write $I(\alpha \mid \beta)$ for "the preferred (or ideal) β worlds satisfy α". Hence, if we ignore infinite descending chains,[3] then we can define $M \models I(\alpha \mid \beta)$ 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 of representing conditional probability, that is, by $Pr(\alpha \mid \beta)$ which stands for "the probability that α is the case given β". Forrester's paradox can be formalized by CP1 below.

CP1-1	$O(\neg k \mid \top)$:	Smith should not kill Jones.
CP1-2	$O(g \mid k)$:	Smith should kill Jones gently, if he kills him.
CP1-3	k:	Smith kills Jones.

The obligation $O(g \mid k)$ is a contrary-to-duty (CTD or secondary) obligation of $O(\neg k \mid \top)$, because $O(\alpha \mid \beta)$ is a CTD obligation of the primary obligation $O(\alpha_1 \mid \beta_1)$ if and only if $\alpha_1 \wedge \beta$ is inconsistent, see figure 2. 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 when we have $\vdash g \rightarrow k$. The consistency follows from the fact that the preferred $\top$ worlds may be different from the preferred k worlds, and the first set of worlds thus can satisfy $\neg k$ when the second set satisfies g and, therefore, also k. The inconsistency of the latter follows directly from the fact that both obligations have the same antecedent. Hence, Hansson's logic gives the desired representation of the paradox.

Hansson's logic has been criticized (see, e.g., [39]), because it does not derive absolute obligations from dyadic ones, as represented by the factual detachment formula FD: $O(\alpha \mid \beta) \wedge \beta \rightarrow O\alpha$.[4] However, there are good reasons not to accept FD.

[3] 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 [4,37] for a discussion.

[4] It is beyond the scope of this paper to discuss the problematic relation between conditional and absolute obligations, see, e.g., [1]. However, it is important to note that an absolute obligation should *not* be defined as an conditional obligation with a tautological antecedent, i.e., by $O\alpha =_{\text{def}} O(\alpha \mid \top)$. The two sentences formalize two completely different things: the first represents a detached obligation, the second represents a universally applicable norm.

If the logic would have FD, then it would reinstate Forrester's paradox, because again $O\neg k \wedge Og$ would be derivable from $O(\neg k \mid \top) \wedge O(g \mid k) \wedge k$. To explicate the difference with dyadic obligations which do have FD and therefore cannot represent Forrester's paradox, we prefer to call Hansson's obligations contextual obligations [76]. Instead of FD we may have $O(\alpha \mid \beta \wedge O(\alpha \mid \beta))$ as a theorem, see [86].

2.2. Strong preference problem (SP)

Hansson's obligations do not have strengthening of the antecedent, and the underlying logic of preference does not have left or right strengthening. On the one hand this property seems only adequate for prima facie obligations, see the discussion in [1]. On the other hand, it is well known from the preference logic literature [84] that the preference $\alpha \succ \neg\alpha$ cannot be defined by the set of preferences of all α worlds to each $\neg\alpha$ world. For example, consider paradox SP of table 1.

SP-1	$O(p \mid \top) =_{\text{def}} p \succ \neg p$:	You must be polite.
SP-2	$O(h \mid \top) =_{\text{def}} h \succ \neg h$:	You must be honest.

The two unrelated obligations 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 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, because DP1 represents a dilemma and should be contradictory.

DP1-1	$O(p \mid \top) =_{\text{def}} p \succ \neg p$:	You must be polite.
DP1-2	$O(\neg p \mid \top) =_{\text{def}} \neg p \succ p$:	You must be impolite.

The technical problem of the logic of preference is how to lift preferences between worlds to preferences between sets of worlds or propositions, written as $\alpha_1 \succ \alpha_2$. This problem is analogous to the lifting problem of an utility function on worlds to an utility function on sets of worlds in decision theory, without using a probability distribution [29,80]. SP is only a problem of preference-based logics, because from $O(\alpha \mid \beta) =_{\text{def}} (\alpha \wedge \beta) \succ (\neg\alpha \wedge \beta)$ follows directly the theorem $O(\alpha \mid \beta) \leftrightarrow O(\alpha \wedge \beta \mid \beta)$. If we have strengthening of the antecedent, then we can derive $O(p \mid \neg(p \wedge h))$ from the obligation $O(p \mid \top)$, and therefore we can derive $O(p \wedge \neg h \mid \neg(p \wedge h))$. The latter obligation conflicts with a second premise $O(h \mid \top)$, because from the latter we can similarly derive $O(\neg p \wedge h \mid \neg(p \wedge h))$.

To solve SP, we accept the consistency of the preferences $(p \wedge \neg h) \succ (\neg p \wedge h)$ and $(\neg p \wedge h) \succ (p \wedge \neg h)$. Moreover, to make DP1 inconsistent, obligations are defined as a combination of a strong preference and an ideal preference. Thus, we do not use $I(\alpha \mid \beta)$ for CTD and dilemma reasoning, as Hansson does, but we only use it for dilemma reasoning.

$$O(\alpha \mid \beta) =_{\text{def}} \big((\alpha \wedge \beta) \succ (\neg\alpha \wedge \beta)\big) \wedge I(\alpha \mid \beta).$$

The formula $O(p \mid \top) \wedge O(\neg p \mid \top)$ is inconsistent, because $I(p \mid \top) \wedge I(\neg p \mid \top)$ is inconsistent. The two obligations $O(p \mid \top)$ and $O(h \mid \top)$ are formalized by (1) p worlds

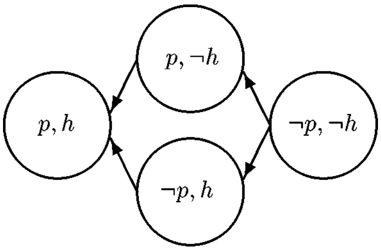

Figure 3. Model of $\{O(p \mid \top), O(h \mid \top)\}$.

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. A typical model is represented in figure 3. The transitive closure of the preference relation is left implicit.

In the following subsection we argue that this solution – in contrast to Hansson's logic – solves the dilemma problem.

2.3. Dilemma problem

Hansson's logic solves CP, without making dilemmas like DP1 and DP2 consistent. However, the dyadic representation also introduces a new instance of the dilemma problem, represented by $O(\alpha \mid \beta_1) \wedge O(\neg\alpha \mid \beta_1 \wedge \beta_2)$. An example is Prakken and Sergot's considerate assassin example DP4 [49].

DP4-1	$O(\neg c \mid \top)$:	A person should not offer someone else a cigarette.
DP4-2	$O(c \mid k)$:	An assassin should offer a victim a cigarette, if he kills him.

Prakken and Sergot argue that DP4 represents a dilemma, because the obligation $O(c \mid k)$ is not a CTD obligation of $O(\neg c \mid \top)$.

> "Is this acceptable? In our opinion it is: what is crucial is that $O(c \mid k)$ is not a contrary-to-duty rule of $O(\neg c \mid \top)$ but of $O(\neg k \mid \top)$, for which reason $O(c \mid k)$ and $O(\neg c \mid \top)$ are unrelated obligations. Now one may ask how this conflict should be resolved and, of course, one plausible option is to regard $O(c \mid k)$ as an exception to $O(\neg c \mid \top)$ and to formalize this with a suitable non-monotonic defeat mechanism. However, it is important to note that this is a separate issue, which has nothing to do with the contrary-to-duty aspects of the example". [49]

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" $O(\neg k \mid \top)$ (combined CDP2a). Hansson's logic does not give a satisfactory solution for DP4, because $O_{HL}(\neg c \mid \top) \wedge O_{HL}(c \mid k)$ is consistent. In PDL the set of obligations $S = \{O(\neg c \mid \top), O(c \mid k)\}$ is inconsistent, because the formula $(\neg c \succ c) \wedge I(c \mid k)$ is inconsistent, as is shown in section 4.

2.4. Combined problems

Prakken and Sergot introduced several combined CDP, as represented in table 1. The new feature of the poodle example discussed in the introduction is that (3) is implied by (1) if the logic has consequential closure. The example suggests that we have to abandon consequential closure, because this is the only way to make (1)–(3) logically stronger than (1)–(2), such that we can add constraints that make only the latter set consistent. For this reason PDL does not have consequential closure. However, Prakken and Sergot do not accept this conclusion. They also argue that if a normgiver forbids having dogs he also implicitly forbids having any particular kind of dog. Therefore they look for other methods to deal with the example. In this paper we accept that consequential closure cannot hold *for dyadic deontic logic*, and we discuss alternatives in section 6. Moreover, in section 4.1 we present more problems with consequential closure.

3. Axiomatization

Prohairetic Deontic Logic (PDL) is a logic of dyadic obligations (we do not discuss absolute obligations) defined in a modal preference logic. The standard Kripke models $M = \langle W, R, \leqslant, V\rangle$ of PDL contain a binary accessibility relation $\leqslant$, that is interpreted as a (reflexive and transitive) *deontic preference ordering*. The advantages of our formalization in a modal framework are twofold. First, if 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.

It is important here to distinguish clearly between a modal *preference* logic and the classic modal approach. Both use modal logic, but they use it in a completely different way. This becomes clear when different axiomatizations of Hansson's DSDL3 [19] are compared. Hansson himself characterized his system only in purely semantical terms, without making an attempt to characterize it in axiomatic or proof-theoretic terms. Axiomatizations were given later by, for example, Spohn [57] and, more recently, Åqvist [3]. The latter has to introduce a lot of additional machinery, such as infinitary hierarchies of systems and systematic frame constants. The underlying idea is that in the semantics there is an explicit representation of each equivalence class of the preference ordering. Boutilier's modal preference logic of DSDL3 [5], like the logic developed here, does not have to explicitly represent each equivalence class, because it is represented implicitly by the preference relation $\leqslant$.

However, there is also a (minor) drawback to modal preference logic. To define the preferred worlds of a model, one has to be able to refer to all possible worlds of this model. Unfortunately, a bimodal logic with one universal relation (as in [72]) is not as straightforward as it seems at first sight, see, e.g., [18,62]. In previous work we implicitly defined a universal relation via a logic of inaccessible worlds [5, 27,64]. However, this leads to a less transparent axiomatization and to additional expressive power which is not exploited. We, therefore, use in this paper a more standard approach, based on two accessibility relations. The models contain, besides the preferences relation $\leqslant$, also a binary equivalence relation R, that is used to interpret a necessity–possibility operator. This is analogous to, for example, the use of the $\Box$-operator in Prakken and Sergot's systems [49,50].

In this section we axiomatize PDL in the following three steps in terms of a monadic modal logic and a deontic betterness relation. See [15,21] for an analogous stepwise construction of "good" in terms of "better" and [5,86] 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 always worse", and $\overleftrightarrow{\Box}\alpha$ can be read as "α is true in all possible worlds".

$$
\begin{array}{lll}
M, w \models \Box\alpha & \text{iff} & \forall w' \in W \text{ if } w' \leqslant w, \text{ then } M, w' \models \alpha \\
M, w \models \overleftrightarrow{\Box}\alpha & \text{iff} & \forall w' \in W \text{ if } w' \in R(w), \text{ then } M, w' \models \alpha
\end{array}
$$

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 [8,26]. 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 deontically preferred to (better than) α_2", is defined in terms of the monadic operators. The following betterness relation obeys von Wright's expansion principle [84], because a preference of α_1 over α_2 only compares the two formulas $\alpha_1 \wedge \neg\alpha_2$ and $\neg\alpha_1 \wedge \alpha_2$:

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

We have $M, w \models \alpha_1 \succ \alpha_2$ if we have $w_2 \not\leqslant w_1$ for all worlds $w_1, w_2 \in R(w)$ such that $M, w_1 \models \alpha_1 \wedge \neg\alpha_2$ and $M, w_2 \models \alpha_2 \wedge \neg\alpha_1$. 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, where we write as usual $\Diamond\alpha =_{\text{def}} \neg\Box\neg\alpha$:

$$\begin{aligned} I(\alpha \mid \beta) &=_{\text{def}} \overset{\leftrightarrow}{\Box}\big(\beta \rightarrow \Diamond\big(\beta \wedge \Box(\beta \rightarrow \alpha)\big)\big), \\ O(\alpha \mid \beta) &=_{\text{def}} \big((\alpha \wedge \beta) \succ (\neg\alpha \wedge \beta)\big) \wedge I(\alpha \mid \beta). \end{aligned}$$

We have $M, w \models I(\alpha \mid \beta)$ if the preferred β worlds of $R(w)$ are α worlds, and α eventually becomes true in all infinite descending chains of β worlds [4,35]. Finally, we have $M, w \models O(\alpha \mid \beta)$ if we have $w_2 \nleqslant w_1$ for all $w_1, w_2 \in R(w)$ such that $M, w_1 \models \alpha \wedge \beta$ and $M, w_2 \models \neg\alpha \wedge \beta$, and $M, w \models I(\alpha \mid \beta)$.

The logic PDL is the definition of 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 $\overset{\leftrightarrow}{\Box}$. Dual "possibility" connectives $\Diamond$ and $\overset{\leftrightarrow}{\Diamond}$ are defined as usual by $\Diamond\alpha =_{\text{def}} \neg\Box\neg\alpha$ and $\overset{\leftrightarrow}{\Diamond}\alpha =_{\text{def}} \neg \overset{\leftrightarrow}{\Box} \neg\alpha$.

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,

$$\begin{array}{llll} \text{K} & \Box(\alpha \rightarrow \beta) \rightarrow (\Box\alpha \rightarrow \Box\beta), & \text{K}' & \overset{\leftrightarrow}{\Box}(\alpha \rightarrow \beta) \rightarrow (\overset{\leftrightarrow}{\Box}\alpha \rightarrow \overset{\leftrightarrow}{\Box}\beta), \\ \text{T} & \Box\alpha \rightarrow \alpha, & \text{T}' & \overset{\leftrightarrow}{\Box}\alpha \rightarrow \alpha, \\ 4 & \Box\alpha \rightarrow \Box\Box\alpha, & 4' & \overset{\leftrightarrow}{\Box}\alpha \rightarrow \overset{\leftrightarrow}{\Box}\overset{\leftrightarrow}{\Box}\alpha, \\ \text{R} & \overset{\leftrightarrow}{\Box}\alpha \rightarrow \Box\alpha, & 5' & \neg \overset{\leftrightarrow}{\Box} \alpha \rightarrow \overset{\leftrightarrow}{\Box}\neg \overset{\leftrightarrow}{\Box} \alpha, \\ \text{Nec} & \text{From } \alpha \text{ infer } \overset{\leftrightarrow}{\Box}\alpha, & & \\ \text{MP} & \text{From } \alpha \rightarrow \beta \text{ and } \alpha \text{ infer } \beta, & & \end{array}$$

extended with the following three definitions:

$$\begin{aligned} \alpha_1 \succ \alpha_2 &=_{\text{def}} \overset{\leftrightarrow}{\Box}\big((\alpha_1 \wedge \neg\alpha_2) \rightarrow \Box\neg(\alpha_2 \wedge \neg\alpha_1)\big), \\ I(\alpha \mid \beta) &=_{\text{def}} \overset{\leftrightarrow}{\Box}\big(\beta \rightarrow \Diamond\big(\beta \wedge \Box(\beta \rightarrow \alpha)\big)\big), \\ O(\alpha \mid \beta) &=_{\text{def}} \big((\alpha \wedge \beta) \succ (\neg\alpha \wedge \beta)\big) \wedge I(\alpha \mid \beta). \end{aligned}$$

Definition 2 (PDL semantics). Kripke models $M = \langle W, R, \leqslant, V \rangle$ for PDL consist of W, a set of worlds, R, a binary equivalence relation, $\leqslant$, a binary transitive and reflexive accessibility relation, and V, a valuation of the propositional atoms in the worlds. The partial pre-ordering $\leqslant$ expresses preferences: $w_1 \leqslant w_2$ if and only if w_1 is at least as preferable as w_2. The set of better worlds is a subset of the R-accessible worlds: if $w_1 \leqslant w_2$ then $w_1 \in R(w_2)$. The modal connective $\Box$ refers to better worlds and the modal connective $\overset{\leftrightarrow}{\Box}$ to R-accessible worlds.

$$\begin{array}{lll} M, w \models \Box\alpha & \text{iff} & \forall w' \in W \text{ if } w' \leqslant w, \text{ then } M, w' \models \alpha, \\ M, w \models \overset{\leftrightarrow}{\Box}\alpha & \text{iff} & \forall w' \in W \text{ if } w' \in R(w), \text{ then } M, w' \models \alpha. \end{array}$$

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 3 (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 $\vdash_{\text{PDL}} \alpha$ if and only if $\models_{\text{PDL}} \alpha$.

Proof. Follows directly from standard modal soundness and completeness proofs, see, e.g., [8,26,51]. The axiom **R** corresponds to the restriction that the better worlds are a subset of the R-accessible worlds. □

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 $O(\alpha \mid \beta) =_{\text{def}} (\alpha \wedge \beta) \succ (\neg\alpha \wedge \beta)$ (for any betterness relation $\succ$), for example, by $O_{HL}(\alpha \mid \beta) =_{\text{def}} I(\alpha \mid \beta)$.

$$\text{Id} \quad O(\alpha \mid \beta_1 \wedge \beta_2) \leftrightarrow O(\alpha \wedge \beta_1 \mid \beta_1 \wedge \beta_2).$$

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 [6,15,22,25,28] do not have closure under logical implication either. The following theorem *Weakening of the Consequent* (WC) is *not* valid in PDL. This theorem is further discussed in section 6.

$$\text{WC} \quad O(\alpha_1 \mid \beta) \rightarrow O(\alpha_1 \vee \alpha_2 \mid \beta).$$

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, not even when $\beta_2 = \neg\beta_1$. This theorem is also further discussed in section 6.

$$\text{OR} \quad \big(O(\alpha \mid \beta_1) \wedge O(\alpha \mid \beta_2)\big) \rightarrow O(\alpha \mid \beta_1 \vee \beta_2).$$

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

$$\text{RSA} \quad \big(O(\alpha \mid \beta_1) \wedge I(\alpha \mid \beta_1 \wedge \beta_2)\big) \rightarrow O(\alpha \mid \beta_1 \wedge \beta_2).$$

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 $\leqslant$ contains more elements. In our terminology $\{(w_1, w_2), (w_2, w_1)\}$ is therefore more connected than $\{(w_1, w_2)\}$.

Definition 4 (Preferential entailment). Let the two possible worlds models $M_1 = \langle W, R, \leqslant_1, V\rangle$ and $M_2 = \langle W, R, \leqslant_2, V\rangle$ be two PDL models with the same set of worlds W, accessibility relation R and valuation function V. M_1 is at least as connected as M_2 in $w \in W$, written as $M_1 \sqsubseteq_w M_2$, if and only if for all $w_1, w_2 \in R(w)$ if $w_1 \leqslant_2 w_2$, then $w_1 \leqslant_1 w_2$. M_1 is more connected than M_2 in w, written as $M_1 \sqsubset_w M_2$, if and only if $M_1 \sqsubseteq_w M_2$ and $M_2 \not\sqsubseteq_w M_1$. The formula ϕ is preferentially entailed by T, written as $T \models_{\sqsubset} \phi$, if and only if $M, w \models \phi$ for all w and maximally connected models M in w of T.

The fact that the logic of obligations that cannot be overridden (i.e., which is *not* a logic of prima facie obligations) is non-monotonic may seem strange at first sight. However, it follows directly from the desired consequences of the benchmark examples in table 1. In particular, the desired consequences of DP3 show that the logic has to be non-monotonic. Moreover, for preference-based deontic logics (or other logics satisfying theorem Id) it also follows from SP. These examples are discussed in detail in the following section.

We end with two technical observations. First, the maximally connected models of a consistent set of obligations are unique (for a given W, R and V) if the transitivity axiom 4: $\Box\alpha \to \Box\Box\alpha$ is omitted from the axiomatization. The unique maximally connected model of the set of obligations $S = \{O(\alpha_i \mid \beta_i) \mid 1 \leqslant i \leqslant n\}$ has the accessibility relation

$$\bigl\{w_1 \leqslant w_2 \mid \text{no } O(\alpha \mid \beta) \in S \text{ such that } M, w_1 \models \neg\alpha \wedge \beta \text{ and } M, w_2 \models \alpha \wedge \beta\bigr\}.$$

However, if axiom 4 is omitted, then ‘preferred’ in $I(\alpha \mid \beta)$ no longer has a natural meaning. Second, if R is universal and 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 $\emptyset \models_{\sqsubset} \neg O(p \mid \top)$ for non-tautological p, because there cannot be a model with only p worlds. In the following section preferential entailment is illustrated by several examples.

4. A guided tour through deontic logic

In this section we present a tour through the landscape of deontic logic and its benchmark examples, with PDL as our guide. We show how PDL solves the problems CP, SP and DP discussed in this paper.

4.1. Contrary-to-duty problem (CP): Forrester’s paradox

CTD obligations are obligations which are only in force if another obligation has been violated. Thus they refer to sub-ideal circumstances. We discuss CP1 in SDL and dyadic deontic logic, and we discuss CP2 only in dyadic deontic logic. In SDL, a conditional obligation $\beta \rightarrow O\alpha$ is a CTD (or secondary) obligation of the (primary) obligation $O\alpha_1$ if and only if $\beta \wedge \alpha_1$ is inconsistent. The following example is the original version of the notorious gentle murderer paradox [13], a strengthened version of the Good Samaritan paradox [2].

Example 5 (Forrester's paradox in SDL). Consider the following sentences of an SDL theory T: "Smith should not kill Jones" $O\neg k$, "if Smith kills Jones, then he should do it gently" $k \rightarrow Og$, "Smith kills Jones" k, and "killing someone gently logically implies killing him" $\vdash g \rightarrow k$. The second obligation is a CTD obligation of the first obligation, because $\neg k$ and k are contradictory. SDL allows so-called factual detachment, i.e.,

$$\models_{\mathrm{SDL}} \big(\beta \wedge (\beta \rightarrow O\alpha)\big) \rightarrow O\alpha$$

and, therefore, we have $T \models_{\mathrm{SDL}} Og$ from the second and third sentence of T. From the CTD obligation Og the obligation Ok can be derived with the K axiom of SDL (weakening). Hence, we have $T \models_{\mathrm{SDL}} O\neg k$ and $T \models_{\mathrm{SDL}} Ok$. The main problem of this paradox is that $O\neg k$ and Ok are inconsistent in SDL, although the set of premises is intuitively consistent.

Forrester's paradox raised an extensive discussion in the deontic logic literature. We first mention several consistent formalizations that have been proposed.

4.1.1. Scope

Scope distinctions, which have been proposed (by, e.g., Castañeda [7]) to solve the Good Samaritan paradox, seem to be absent from Forrester's paradox. However, Sinnot-Armstrong [56] argues that also Forrester's paradox rests on scope confusions. According to Davidson's account of the logical form of action statements [11] adverbial modifiers like gently in the consequent of $k \rightarrow Og$ are represented as predicates of action-events. Hence, the obligation is translated to "there is an event e, which is a murdering event, and it, e, is gentle" – $\exists e(Me \wedge Ge)$. Because of the conjunction, we can distinguish between wide scope $O\exists e(Me \wedge Ge)$ and narrow scope $\exists e(Me \wedge OGe)$. The narrow scope representation consistently formalizes the paradox, because we cannot derive "Smith ought to kill Jones" from "the event e ought to be gentle".

A drawback of this solution [16,17,40] is that not every adverb of action is amenable to treatment as a predicate. For example, Goble [16] gives the example "Jones ought not to wear red to school" and "if Jones wears red to school, then Jones ought to wear scarlet to school". Goble observes that the relation between scarlet and red is not such that we can say scarlet is "red *and* ...", which might allow us to pull the term red away from the deontic operator in the manner of Sinnot-Armstrong and Castañeda, leaving the operator to apply only to whatever fills the blank. Scarlet is just a determinate shade of red; that is all we can say. Prakken and Sergot [49] mention "fences should be white" and "if they are not white, then they should be black". Another argument against the scope distinction is that it does not block the so-called pragmatic oddity [49].

CP3-1 Op: You should keep your promise.
CP3-2 $\neg p \rightarrow Oa$: If you do not keep you promise, then you should apologize.
CP3-3 $\neg p$: You do not keep your promise.

From the three sentences we can derive the counterintuitive "you should keep your promise and apologize for not keeping it" $O(p \wedge a)$.

4.1.2. Weakening

Goble [16] argues that Forrester's paradox is caused by weakening, following a suggestion of Forrester [13, p. 196]. His consistent formalization is based on rejection of the property weakening. The most convincing argument that weakening is invalid is the paradox of the knower [2], represented by the sentence "if you ought to know p, then p ought to be (done)" $OKp \rightarrow Op$. In his monadic logic, $O\neg k \wedge Ok$ is inconsistent whereas $O\neg k \wedge Og$ is consistent.

A drawback of this solution is that only in a *few* cases it *seems* that the formula $O\alpha \wedge O(\neg\alpha \wedge \beta)$ is not a dilemma and should therefore be consistent, see example DP2. This solution seems like overkill. Moreover, it is unclear how to deal with the pragmatic oddity CP3.

4.1.3. Defeasibility

Non-monotonic techniques can be used to consistently formalize the paradox [44, 47,54]. The problem of the paradox is that it is inconsistent, whereas intuitively it is consistent. Hence, a pragmatic formalization of the paradox can make use of "restoring consistency" techniques in case of a paradox.

However, this solution is ad hoc. Restoring consistency is like treating symptoms without treating the disease. The term hack comes to mind! Moreover, restoring consistency techniques cannot deal with the pragmatic oddity CP3, see section 4.1.1, because in these paradoxes there is no inconsistency.

Non-monotonic techniques were already used by Loewer and Belzer [39,40], who solve the Forrester paradox in their temporal deontic logic "Dyadic Deontic Detachment" (3D). In 3D a dyadic obligation $O(\alpha \mid \beta)$ is read as "if it is settled that β will be (done), then α ought to be (done)". Moreover, there is an operator $S\alpha$ in 3D that represents that a proposition α is settled. A fact can be settled to become true, without factually being true. Loewer and Belzer [40] also discuss the relation between their solution and Castañeda's approach to the CTD paradoxes [7]. Moreover, it is observed in [76] that approaches based on contextual reasoning use non-monotonic techniques, when "α ought to be (done) in context γ" is defined by "α ought to be (done), unless $\neg\gamma$".

4.1.4. Dyadic operators

Dyadic deontic logics were developed to formalize CTD reasoning [19,38] and to analyze the Good Samaritan paradox, and they can also be used for the formalization of

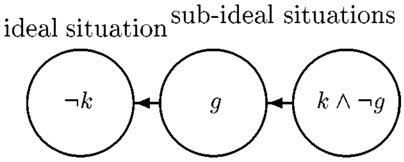

Figure 4. Unique maximally connected model of $\{O(\neg k \mid \top), O(g \mid k), k\}$.

Forrester's paradox. The two obligations are simply represented by the dyadic obligations $O(\neg k \mid \top)$ and $O(g \mid k)$. The problem of dyadic deontic logic is which inference rules can be accepted. For example, we already observed that they cannot accept the factual detachment rule FD: $O(\alpha \mid \beta) \wedge \beta \rightarrow O\alpha$, or the paradox is immediately reinstated. Moreover, the logic cannot have strengthening of the antecedent and weakening of the consequent, as shown by the counterintuitive derivation below of "I must get coffee or no tea if there is no coffee" $O(c \vee \neg t \mid \neg c)$ from "I must get coffee" $O(c \mid \top)$, even in the context of "I must get tea if there is no coffee" $O(t \mid \neg c)$ [65,66]. This is in particular problematic in the context of the rule Id, because the derived obligation is then equivalent to $O(\neg t \wedge \neg c \mid \neg c)$, which conflicts with $O(t \mid \neg c)$.

$$\cfrac{\cfrac{O(c \mid \top)}{O(c \vee \neg t \mid \top)}\ \text{WC}}{O(c \vee \neg t \mid \neg c)}\ \text{SA}$$

4.1.5. Prohairetic Deontic Logic

The solution of CP in PDL is based on the dyadic representation.

Example 6 (CP1: Forrester's paradox). Consider the set of PDL sentences $S = \{O(\neg k \mid \top), O(g \mid k), k\}$, where k can be read as "Smith kills Jones" and g as "Smith kills him gently", and g logically implies k.[5] The unique maximally connected model of S is represented in figure 4. In this figure, as well as in the following ones, we only show the relevant worlds R-accessible from the actual world. The actual world is any of the k worlds. The formalization of S is unproblematic and the semantics reflects the three states that seem to be implied by the paradox.

4.2. The contrary-to-duty problem (CP): Chisholm's paradox and according-to-duty reasoning

The second CTD paradox we consider is Chisholm's paradox [9]. It consists of the three obligations of a certain man "to go to his neighbors' assistance", "to tell them that he comes if he goes", and "not to tell them that he comes if he does not go", together with the fact "he does not go". In particular, Chisholm shows that in

[5] In PDL, the relation between g and k can also be formalized by $\overleftrightarrow{\Box}(g \rightarrow k)$. We do not discuss or use the distinction between logical implication and necessary implication.

SDL the sentences are either inconsistent or logically dependent. There is no example in deontic logic literature that provoked so much discussion as Chisholm's paradox. The formalizations we mentioned already at the discussion of Forrester's paradox were also proposed for Chisholm's paradox. Moreover, monadic modal logic was extended with additional semantic features, such as time and actions.

4.2.1. Time

Variants of Chisholm's paradox have been formalized in temporal deontic logic [39,81], which usually assume a temporal lag between antecedent and consequent. However, additional machinery has to be introduced to represent the paradox itself [73].

The drawback of the temporal solution is that the expressive power of the temporal solution is limited. For example, temporal deontic logics that make a distinction between antecedent and consequent cannot represent the set of premises of Forrester's paradox in example 6, see also the discussion in [49,88].

4.2.2. Action

A related formalization distinguishes two (propositional) base languages, one for the antecedent and one for the consequent [1,45], following Castañeda's distinction between assertions and actions [7].

Hence, they do not allow that a proposition occurs in one formula in the antecedent and in another formula in the consequent, and they thus cannot formalize the Forrester or Chisholm set without introducing additional machinery.

4.2.3. Dyadic operators

Chisholm's paradox is more complicated than Forrester's paradox, because it also contains an *According-To-Duty* (ATD) obligation. Figure 5 illustrates that a conditional obligation $O(\alpha \mid \beta)$ is an ATD obligation of $O(\alpha_1 \mid \beta_1)$ if and only if β logically implies α_1. The condition of an ATD obligation is satisfied only if the primary obligation is fulfilled. The definition of ATD is analogous to the definition of CTD (see figure 2) in the sense that an ATD obligation is an obligation conditional to a fulfillment of an obligation and a CTD obligation is an obligation conditional to a violation.

See, e.g., [61] for a discussion of the paradox in several dyadic deontic logics.

4.2.4. Prohairetic Deontic Logic

The consistent representation of the paradox in PDL is again based on the dyadic operators.

$$\begin{tikzcd} O(\alpha_1|\beta_1) \\ O(\alpha|\beta) \arrow[u, "\text{implies}"] \end{tikzcd}$$

Figure 5. $O(\alpha \mid \beta)$ is an according-to-duty obligation of $O(\alpha_1 \mid \beta_1)$.

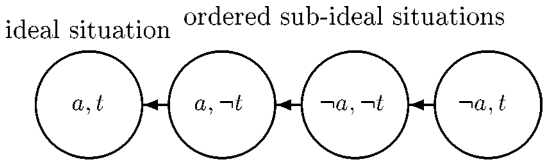

Figure 6. Unique maximally connected model of $\{O(a \mid \top), O(t \mid a), O(\neg t \mid \neg a), \neg a\}$.

Example 7 (CP2: Chisholm's paradox). Consider the set of PDL sentences $S = \{O(a \mid \top), O(t \mid a), O(\neg t \mid \neg a), \neg a\}$, 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 6. The crucial question of Chisholm's 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. It is well known (see, e.g., [49,73]) that the main problem of Chisholm's paradox is caused by so-called deontic detachment (also called deontic transitivity). This obligation is derived from S by any deontic logic that has deontic detachment, represented by the following formula DD^0:

$$\mathrm{DD}^0 \quad \big(O(\alpha \mid \beta) \wedge O(\beta \mid \gamma)\big) \rightarrow O(\alpha \mid \gamma).$$

However, 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 DD^0 is valid in PDL:

$$\mathrm{RDD} \quad \big(O(\alpha \mid \beta \wedge \gamma) \wedge O(\beta \mid \gamma) \wedge I(\alpha \wedge \beta \mid \gamma)\big) \rightarrow O(\alpha \wedge \beta \mid \gamma).$$

The obligation that "the man should go to his neighbors and tell his neighbors that he will come" $O(a \wedge t \mid \top)$ can (preferentially) be derived from S.

The derivation of $O(a \wedge t \mid \top)$ captures the spirit of deontic detachment, without deriving the counterintuitive consequence $O(t \mid \top)$. This example is another reason that indicates that consequential closure should not hold, because with this proof rule we would derive the counterintuitive latter from the intuitive former.

4.3. Strong preference problem

The strong preference problem is that preferences for α_1 and α_2 conflict for $\alpha_1 \wedge \neg\alpha_2$ and $\neg\alpha_1 \wedge \alpha_2$.

SP-1	$O(p \mid \top)$:	You must be polite.
SP-2	$O(h \mid \top)$:	You must be honest.

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 logic of prima facie obligations, in which obligations can be overridden by other obligations. The reason is that $(p \wedge \neg h) \succ (\neg p \wedge h)$ is overridden by $(\neg p \wedge h) \succ (p \wedge \neg h)$, and $O(p \mid \top)$ 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-overridable obligations. For such logics, the following three solutions have been considered.

4.3.1. Bi-ordering

Jackson [28] and Goble [15] introduce a second ordering representing degrees of "closeness" of worlds to solve SP. 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 $O(p \mid \top)$ and $O(h \mid \top)$ 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.

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. The "multi preference" semantics is a formalization of *defeasible* deontic logic [70,71]. However, it is not clear that closeness is an intuitive concept for non-defeasible obligations.

4.3.2. Ceteris paribus

Sven Ove Hansson [20–22] defines $\alpha \succ \neg\alpha$ by a "ceteris paribus" preference of α to $\neg\alpha$, see [6] for a discussion. Informally, 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" $O(p \mid \top)$ 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" $O(h \mid \top)$ 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 $O(p \mid \top) \wedge O(h \mid \top)$ that prefers $p \wedge h$ worlds to all other worlds, and that prefers all worlds to $\neg p \wedge \neg h$ worlds. Moreover, Hansson defines obligations by the property of negativity. According to this principle, what is worse than something wrong is itself wrong. See [6,22] for a discussion on this assumption.

In a later paper [23] 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 the preference $(p \wedge (p \leftrightarrow \neg h)) \succ (\neg p \wedge (p \leftrightarrow \neg h))$, 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., [58] for an ad hoc solution of the problem).

4.3.3. 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 $\leqslant$ is strongly connected, i.e., if for each pair of worlds w_1 and w_2 in a model M we have either $w_1 \leqslant w_2$ or $w_2 \leqslant w_1$. However, the two obligations $O(p \mid \top)$ and $O(h \mid \top)$ of SP do not conflict when considering $p \wedge \neg h$ and $\neg p \wedge h$ *when we allow for incomparable worlds*, following [82]. In contrast to the other solutions of the strong preference problem, DP1 is consistent. 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.

It was already argued by von Wright [84] 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 neither has an irrelevance nor an independence problem.

4.3.4. Prohairetic Deontic Logic

PDL proposed in this paper is an extension of the third approach. The following example illustrates that SP is solved by the dynamics of preferential entailment. It also illustrates one of the reasons why the logic is non-monotonic (the other reason is DP3 discussed in example 10).

Example 8 (SP: Polite and helpful). Consider the three sets of obligations $S = \emptyset$, $S' = \{O(p \mid \top)\}$ and $S'' = \{O(p \mid \top), O(h \mid \top)\}$. The three unique maximally connected models of S, S' and S'' are represented in figure 7. With no premises, all worlds are equally ideal. By addition of the premise $O(p \mid \top)$, the p worlds are strictly preferred over $\neg p$ worlds. Moreover, by addition of the second premise $O(h \mid \top)$, 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 $S' \models_{\sqsubset} O(p \mid \neg(p \wedge h))$ and $S'' \not\models_{\sqsubset} O(p \mid \neg(p \wedge h))$. By addition of a formula we loose conclusions.

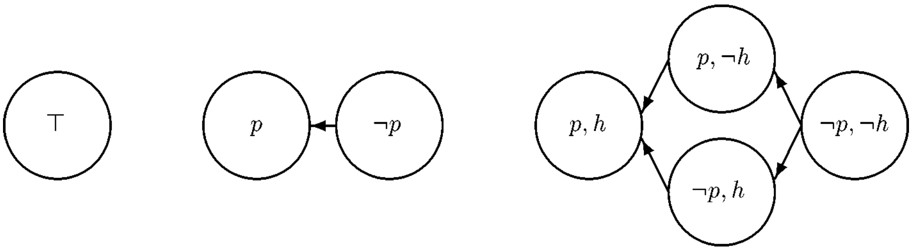

Figure 7. Unique maximally connected models of $\emptyset$, $\{O(p \mid \top)\}$ and $\{O(p \mid \top), O(h \mid \top)\}$.

4.4. Dilemma problem

There are many logics in which DP4= $\{O(\neg c \mid \top), O(c \mid k)\}$ of section 2.3 is inconsistent, as desired, or in which the set becomes inconsistent once the fact k is added to it. However, a problem of most solutions of the dilemma problem is that the set of obligations $S = \{O(\alpha \mid \beta_1), O(\neg\alpha \mid \beta_2)\}$ is no longer consistent (e.g., [85]), or that can be derived that $\beta_1 \wedge \beta_2$ is impossible (e.g., [1]). It is illustrated with DP3 by von Wright [85] 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". [85, p. 116]

4.4.1. Deontic independence

In [69] it is argued that DP4 is only inconsistent if the two obligations are, in a suitable sense, independent. That is, the formalization in dyadic deontic logic does not contain enough information to decide whether the set is inconsistent or not. It is also shown how this solves the mixed problem CDP3 discussed in the introduction. If the obligations are independent then the set is inconsistent, but if they are dependent (for example because the third obligation has been derived from the first one) then they are consistent. A drawback of this solution is that we have to explicitly present independence information, which is often difficult to state.

4.4.2. Prohairetic Deontic Logic

Finally, we show that PDL solves DP, because it makes DP4 in example 9 inconsistent, without making DP3 in example 10 inconsistent.

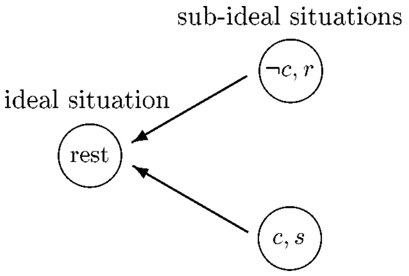

Figure 8. Preferred model of $\{O(c \mid r), O(\neg c \mid s)\}$.

Example 9 (DP4: Considerate assassin). Consider $S = \{O(\neg c \mid \top), O(c \mid k)\}$. 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 $w_2 \in R(w_1)$ such that $M, w_2 \models c \wedge k$ and $w_2 < w_1$ (i.e., $w_2 \leqslant w_1$ and $w_1 \not\leqslant w_2$). However, the obligation $O(\neg c \mid \top)$ 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 $w_2 \in R(w_1)$ such that $M, w_2 \models c \wedge k$ and $w_2 \leqslant w_1$. These two conditions are contradictory (if there is such a world w_1).

The following example illustrates that DP3 is consistent in PDL.

Example 10 (DP3: 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”. In PDL the set S is consistent, and a maximally connected model M of S is given in figure 8. The ideal worlds satisfy $r \rightarrow c$ and $s \rightarrow \neg c$, and the sub-ideal worlds either $\neg c \wedge r$ or $c \wedge s$. The $r \wedge s \wedge c$ and $r \wedge s \wedge \neg c$ worlds are incomparable, for similar reasons as those discussed in example 8. We have $M \not\models O(c \mid r \wedge s)$ and thus $S \not\models_{\sqsubset} O(c \mid r \wedge s)$, as desired. Moreover, since we obviously also have $\{O(c \mid r)\} \models_{\sqsubset} O(c \mid r \wedge s)$, this example clearly illustrates why PDL is non-monotonic.

Note that there are many maximally connected models.

4.5. Combined problems

From the previous results it follows that the combined problems are represented as desired. For example, consider the combined CDP3. The first two sentences are consistent, as shown by the consistency of CP1 in example 6. The three sentences together are inconsistent, which follows from the inconsistency of DP4 in example 9, because the last two sentences of CDP3 are structurally equivalent to DP4 in the context of the theorem Id. By analogy the other combined problems are formalized according to the desiderata too. For example, the first two sentences of CDP1 are again analogous to CP1, and the first and third sentence are analogous to DP4.

In this paper we follow the mainstream perspective and we therefore do not discuss conflict-tolerant or prima facie obligations. It seems that in these other types

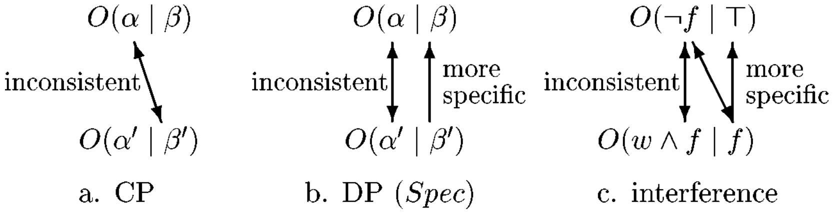

Figure 9. The interference problem of the logic of prima facie obligations (from [71]).

of deontic logics the combined problems can easily be solved by developing a logic which is very weak, such as for example Hansson's logic. However, it is important to note that here the combined problems are also problematic, because they give rise to the interference problem observed and discussed in [71]. For example, consider CDP1.

CDP1-1	$O(\neg f \mid \top)$:	There must be no fence.
CDP1-2	$O(w \wedge f \mid f)$:	There must be a white fence, if there is a fence.
CDP1-3	$O(f \mid d)$:	There must be a fence, if there is a dog.

In CDP1 the dilemma interpretation coincides with a so-called overriding interpretation, because according to the specificity principle more specific obligations override conflicting more general ones. What is most striking about the fence example of CDP1 is the observation that when the premise $O(\neg f \mid \top)$ is violated by f, then the obligation for $\neg f$ should be derivable, but not when $O(\neg f \mid \top)$ is overridden by the exception $f \wedge d$. In general, the sentence "α is not done but prima facie α should be done" can be diagnosed (for example by a judge) as either a violation or an exception. The CTD and overriding interpretations of $O(\neg f \mid \top)$ are quite different in the sense that they lead to different conclusions. This interference problem is visualized in figure 9, where the DP interpretation is based on the specificity principle. Case (c) shows that the pair of obligations $O(\neg f \mid \top)$ and $O(w \wedge f \mid f)$ is ambiguous, because it can be given a CP as well as a DP interpretation.

5. Related research

Besides the deontic logics already discussed in this paper, there is a preference-based deontic logic proposed by Brown, Mantha and Wakayama [6]. We write the obligations in their logic as O_{BMW}. At first sight it seems that the logic is closely related to Prohairetic Deontic Logic, because $O_{\text{BMW}}\alpha$ 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

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

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, $O_{\text{BMW}}\alpha$ 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" [6, p. 200]. The first distinction is that in the logic, dilemmas are consistent (which they consider an advantage, following [82]). Secondly, the motivation for the mixed representation is different. Whereas we introduced the mixed representation to solve both CP (for which we use $\alpha \succ \neg\alpha$) and DP (for which we use $I\alpha$), they use the mixed representation to block the derivation of $O_{\mathrm{BMW}}(\alpha_1 \vee \alpha_2) \rightarrow O_{\mathrm{BMW}}\alpha_1$, which they consider as "somewhat unreasonable, since the premise is weaker than the conclusion". However, it is easily checked that the logic validates the theorem $O_{\mathrm{BMW}}(\alpha_1 \vee \alpha_2) \wedge A_m\alpha_1 \rightarrow O_{\mathrm{BMW}}\alpha_1$. Hence, under certain circumstances stronger obligations can be derived. This counterintuitive formula is not a theorem of PDL.

6. Further developments

The logic PDL is the basis of the logics developed in [64]. It is the only deontic logic we know of – besides the extensions discussed below – that gives the desired properties for the benchmark examples in table 1. However, other benchmark examples have been proposed more recently, for which PDL does not give the desired conclusions. For example, in [42] it is argued that it is desired to derive $O(a \mid \top)$ from $O(a \wedge b \mid c)$ and $O(a \wedge \neg b \mid \neg c)$. This example illustrates two drawbacks of PDL: it does not have weakening of the consequent (or consequential closure), and it does not have the disjunction rule for the antecedent. These two properties are a consequence of the fact that we *only* use dyadic deontic logic. We already saw several examples where the addition of consequential closure to PDL would have catastrophic consequences: in the combined CDP3 and CDP3a, in Forrester's CP1 if combined with strengthening of the antecedent, in Chisholm's CP2 if combined with our new type of deontic detachment, and in Åqvist's paradox of the knower. Moreover, counterintuitive consequence can be derived in conflict tolerant logics that contain the restricted conjunction rule RAND (called consistent aggregation in [82]), see [65,66]. With this proof rule $O(\alpha_1 \wedge \alpha_2 \mid \top)$ can only be derived from $O(\alpha_1 \mid \top)$ and $O(\alpha_2 \mid \top)$ if $\alpha_1 \wedge \alpha_2$ is consistent. Consequently, the counterintuitive $O(p \wedge \neg p \mid \top)$ cannot be derived from $O(p \mid \top)$ and $O(\neg p \mid \top)$, as desired in conflict tolerant logics, but as shown below the counterintuitive $O(q \mid \top)$ (for any q!) can be derived.

$$\dfrac{\dfrac{\dfrac{O(p \mid \top)}{O(p \vee q \mid \top)}\,\mathrm{WC} \qquad O(\neg p \mid \top)}{O(q \wedge \neg p \mid \top)}\,\mathrm{RAND}}{O(q \mid \top)}\,\mathrm{WC}$$

Moreover, in natural language consequential closure is often counterintuitive, as shown by Ross: "you should mail the letter implies that you should mail or burn it" [52]. The following example is an example which may be problematic for the disjunction rule OR. It illustrates that the non-validity of OR in PDL can be used to analyze dominance

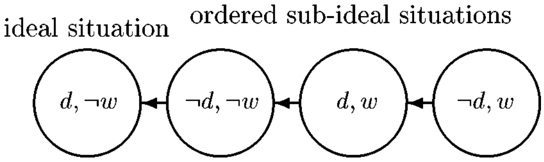

Figure 10. Model of cold-war disarmament.

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 [60] observe that, although such arguments are often used, and are convincing when they are used, they are invalid. The following example adapted from [60] is a classic illustration of the invalidity of the dominance argument (see also [29]).

Example 11 (Cold-war disarmament). Either there will be a nuclear war or there will not. If there will be no 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 conditional obligations "we ought to disarm, if there will be a nuclear war" $O(d \mid w)$ and "we ought to disarm, 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 the preferences $(d \wedge w) \succ (\neg d \wedge w)$ and $(d \wedge \neg w) \succ (\neg d \wedge \neg w)$. In fact, the model in figure 10 satisfies $(\neg d \wedge \neg w) \succ (d \wedge w)$, which represents "we ought to arm if we have peace if and only if we are armed" $O(\neg d \mid d \leftrightarrow w)$.

Nevertheless, PDL has been extended such that WC and OR can be incorporated, as discussed below. For these extensions we had to reach beyond the framework of dyadic deontic logic. Moreover, two other drawbacks of PDL are that it is based on truth values and it has the identity Id as a theorem. The latter two properties are a consequence of the fact that we use preference-based logic. New results have been obtained here too.

CDL Contextual deontic logic [69,76] contains ternary operators $O_\gamma(\alpha \mid \beta)$, to be read as "$\alpha$ ought to be (done) if β is (done) in the context where γ is (done)". In [76] contextual obligations have been defined as a kind of defeasible obligations $O_\gamma(\alpha \mid \beta) =_{\text{def}} O(\alpha \mid \beta \setminus \neg\gamma)$, to be read as "$\alpha$ ought to be (done) if β is (done) unless $\neg\gamma$ is (done)". The logic has a kind of weakening of the consequent as a theorem $O_\gamma(\alpha_1 \wedge \alpha_2 \mid \beta) \rightarrow O_{\gamma \wedge \alpha_2}(\alpha_1 \mid \beta)$.

2DL Two-phase deontic logic [59] has two distinct dyadic operators. The premises are phase-1 obligations O_1 (e.g., PDL obligations), the conclusions are phase-2 obligations O_2 (e.g., Hansson's obligations [19]), and the two are related by the axiom $O_1(\alpha \mid \beta) \rightarrow O_2(\alpha \mid \beta)$. The proof rules WC and OR are valid, but they can only be used after PDL proof rules have been applied. In 2DL we can derive $O_2(a \mid \top)$ from $O_1(a \wedge c \mid b)$ and $O_1(a \wedge \neg c \mid \neg b)$.

DUS The dynamic update semantics for obligations [74,75,78,79] is related to the dynamic interpretation of PDL with preferential entailment observed in example 8. In DUS the definition of logical validity of obligations is not based on truth values but on action dynamics. You know the meaning of a normative sentence if you know the change it brings about in the ideality relation of anyone the news conveyed by the norm applies to. A drawback of this logic, as well as of LDL below, is that it is only defined for a restricted fragment of the PDL language: disjunctions and negations of obligations are not allowed. In [74] an implementation is envisaged for this fragment of PDL.

LDL Labeled deontic logic [70,71] analyses inference patterns of PDL and 2DL – SA, DD, WC and OR – in Gabbay's labelled deductive systems. It is not preference-based, and it does not have the identity. In PLLG [66] it is shown how the two phases of 2DL naturally arise in LDL with the additional check that "it must always be possible to fulfill the derived obligations together with the premises it is derived from". LDL has been generalized to the logic of reusable propositional output (joint work with David Makinson) [43,68].

The further developments have been summarized in the following table.

		WC	OR	noTV	noId	Related
	PDL	–	–	–	–	
[76]	CDL	+	–	–	–	
[59]	2DL	+	+	–	–	[66]
[74]	DUS	–	–	+	–	[75,78,79]
[70]	LDL	+	+	+	+	[42,43,65–68]

For applications, the logic PDL clearly needs additional expressive power. Makinson observed in 1993 [41] that "at the present state of play, it would not seem advisable to try to cover all complicating factors [of deontic logic] at once, but rather to get an initial appreciation of them few at a time, only subsequently putting them together and investigating their interactions". In this paper, like in, e.g., [42], obligations have been studied in a simple propositional setting. The language of PDL now can be extended with permissions (following the suggestions in [78]) and prohibitions, a first-order base language, nested conditionals (following [78,86]), background knowledge [36,63], authorities, agents [78], actions, time [73] and exceptions [75,79].

7. Conclusions

The contrary-to-duty and dilemma 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 introduced Prohairetic

Deontic Logic. We showed that it gives a satisfactory solution to the contrary-to-duty problem, the strong preference problem and the dilemma problem. We now study the use of PDL for legal expert systems and to specify intelligent agents for the Internet [10,12] like the drafting, negotiation and processing of trade contracts in electronic commerce, the relevance for logics of desires and goals as these are developed in qualitative decision theory [5,36,48,77,80], and its use for multi-agent systems [78].

Acknowledgements

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

References

[1] C.E. Alchourrón, Philosophical foundations of deontic logic and the logic of defeasible conditionals, in: *Deontic Logic in Computer Science: Normative System Specification*, eds. J.-J. Meyer and R. Wieringa (Wiley, Chichester, UK, 1993) pp. 43–84.

[2] L. Åqvist, Good Samaritans, contrary-to-duty imperatives, and epistemic obligations, Noûs 1 (1967) 361–379.

[3] L. Åqvist, Systematic frame constraints in defeasible deontic logic, in: *Defeasible Deontic Logic*, ed. D. Nute, Synthese Library, Vol. 263 (Kluwer, Dordrecht, 1997) pp. 59–77.

[4] C. Boutilier, Conditional logics of normality: a modal approach, Artificial Intelligence 68 (1994) 87–154.

[5] C. Boutilier, Toward a logic for qualitative decision theory, in: *Proceedings of the Fourth International Conference on Principles of Knowledge Representation and Reasoning (KR'94)* (Morgan Kaufmann, San Francisco, CA, 1994) pp. 75–86.

[6] 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 (1993) 167–203.

[7] H. Castañeda, The paradoxes of deontic logic: the simplest solution to all of them in one fell swoop, in: *New Studies in Deontic Logic: Norms, Actions and the Foundations of Ethics*, ed. R. Hilpinen (D. Reidel, 1981) pp. 37–85.

[8] B.F. Chellas, *Modal Logic: An Introduction* (Cambridge University Press, 1980).

[9] R.M. Chisholm, Contrary-to-duty imperatives and deontic logic, Analysis 24 (1963) 33–36.

[10] R. Conte and R. Falcone, ICMAS'96: Norms, obligations, and conventions, AI Magazine 18(4) (1997) 145–147.

[11] D. Davidson, The logical form of action sentences, in: *The Logic of Decision and Action*, ed. N. Rescher (University of Pittsburg Press, 1967).

[12] B.S. Firozabadhi and L. van der Torre, Towards a formal analysis of control systems, in: *Proceedings of the Thirteenth European Conference on Artificial Intelligence (ECAI'98)*, ed. H. Prade (1998) pp. 317–318.

[13] J.W. Forrester, Gentle murder, or the adverbial Samaritan, Journal of Philosophy 81 (1984) 193–197.

[14] L. Goble, A logic of good, would and should, part 1, Journal of Philosophical Logic 19 (1990) 169–199.

[15] L. Goble, A logic of good, would and should, part 2, Journal of Philosophical Logic 19 (1990) 253–276.

[16] L. Goble, Murder most gentle: the paradox deepens, Philosophical Studies 64 (1991) 217–227.

[17] L. Goble, "Ought" and extensionality, Noûs 30 (1996) 330–355.

[18] V. Goranko and S. Passy, Using the universal modality: gains and questions, Journal of Logic and Computation 2 (1992) 5–30.

[19] B. Hansson, An analysis of some deontic logics, in: *Deontic Logic: Introductionary and Systematic Readings*, ed. R. Hilpinen (D. Reidel, Dordrecht, Holland, 1971) pp. 121–147. Reprint from Noûs 3 (1969) 373–398.

[20] S.O. Hansson, A new semantical approach to the logic of preference, *Erkenntnis* 31 (1989) 1–42.

[21] S.O. Hansson, Defining "good" and "bad" in terms of "better", Notre Dame Journal of Formal Logic 31 (1990) 136–149.

[22] S.O. Hansson, Preference-based deontic logic (PDL), Journal of Philosophical Logic 19 (1990) 75–93.

[23] S.O. Hansson, Situationist deontic logic, Journal of Philosophical Logic 26 (1997) 423–448.

[24] J.F. Horty, Deontic logic as founded in nonmonotonic logic, Annals of Mathematics and Artificial Intelligence 9 (1993) 69–91.

[25] 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 (1997) 149.

[26] H.G. Hughes and M.J. Creswell, *A Companion to Modal Logic* (Methuen, London, 1984).

[27] I.L. Humberstone, Inaccessible worlds, Notre Dame Journal of Formal Logic 24 (1983) 346–352.

[28] F. Jackson, On the semantics and logic of obligation, Mind 94 (1985) 177–196.

[29] R. Jeffrey, *The Logic of Decision* (University of Chicago Press, 2nd edition, 1983).

[30] R.E. Jennings, A utilitarian semantics for deontic logic, Journal of Philosophical Logic 3 (1974) 445–465.

[31] R.E. Jennings, Can there be a natural deontic logic? Synthese 65 (1985) 257–274.

[32] A.J.I. Jones and M. Sergot, Deontic logic in the representation of law: Towards a methodology, Artificial Intelligence and Law 1 (1992) 45–64.

[33] A.J.I. Jones and M. Sergot, On the characterisation of law and computer systems: The normative systems perspective, in: *Deontic Logic in Computer Science. Normative System Specification*, eds. J. Meyer and R. Wieringa (Wiley, Chichester, UK, 1993) pp. 275–307.

[34] S. Kraus, D. Lehmann and M. Magidor, Nonmonotonic reasoning, preferential models and cumulative logics, Artificial Intelligence 44 (1990) 167–207.

[35] P. Lamarre, S4 as the conditional logic of nonmonotonicity, in: *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning (KR'91)* (Morgan Kaufmann, 1991) pp. 357–367.

[36] J. Lang, Conditional desires and utilities - an alternative approach to qualitative decision theory, in: *Proceedings of the 12th European Conference on Artificial Intelligence (ECAI'96)* (1996) pp. 318–322.

[37] D. Lewis, *Counterfactuals* (Blackwell, Oxford, 1973).

[38] D. Lewis, Semantic analysis for dyadic deontic logic, in: *Logical Theory and Semantical Analysis*, ed. S. Stunland (D. Reidel, Dordrecht, Holland, 1974) pp. 1–14.

[39] B. Loewer and M. Belzer, Dyadic deontic detachment, Synthese 54 (1983) 295–318.

[40] B. Loewer and M. Belzer, Help for the good Samaritan paradox, Philosophical Studies 50 (1986) 117–127.

[41] D. Makinson, Five faces of minimality, Studia Logica 52 (1993) 339–379.

[42] D. Makinson, On a fundamental problem of deontic logic, in: *Norms, Logics and Information Systems. New Studies on Deontic Logic and Computer Science*, eds. P. McNamara and H. Prakken (IOS Press, 1999) pp. 29–54.

[43] D. Makinson and L. van der Torre, Input-output logics, in: *Proceedings of the* Δ*EON-2000* (2000, to appear).

[44] L.T. McCarty, Defeasible deontic reasoning, Fundamenta Informaticae 21 (1994) 125–148.

[45] J.-J.Ch. Meyer, A different approach to deontic logic: deontic logic viewed as a variant of dynamic logic, Notre Dame Journal of Formal Logic 29 (1985) 109–136.

[46] J.D. Mullen, Does the logic of preference rest on a mistake? Metaphilosophy 10 (1979) 247–255.

[47] D. Nute and X. Yu, Introduction, in: *Defeasible Deontic Logic*, ed. D. Nute, Synthese Library, Vol. 263 (Kluwer, Dordrecht, 1997) pp. 1–16.

[48] J. Pearl, From conditional oughts to qualitative decision theory, in: *Proceedings of the 9th Conference on Uncertainty in Artificial Intelligence (UAI'93)* (Morgan Kaufmann, 1993) pp. 12–20.

[49] H. Prakken and M.J. Sergot, Contrary-to-duty obligations, Studia Logica 57 (1996) 91–115.

[50] H. Prakken and M.J. Sergot, Dyadic deontic logic and contrary-to-duty obligations, in: *Defeasible Deontic Logic*, ed. D. Nute, Synthese Library, Vol. 263 (Kluwer, 1997) pp. 223–262.

[51] Y. Moses R. Fagin, J.Y. Halpern and M.Y. Vardi, *Reasoning About Knowledge* (MIT Press, Cambridge, MA, 1995).

[52] A. Ross, Imperatives and logic, Theoria 7 (1941) 53–71.

[53] D. Ross, *The Right and the Good* (Oxford University Press, 1930).

[54] Y. Ryu and R. Lee, Defeasible deontic reasonig: A logic programming model, in: *Deontic Logic in Computer Science: Normative System Specification* (Wiley, Chichester, UK, 1993) pp. 225–241.

[55] Y. Shoham, *Reasoning About Change* (MIT Press, Cambridge, MA, 1988).

[56] W. Sinnot-Armstrong, A solution to Forrester's paradox of gentle murder, Journal of Philosophy 82 (1985) 162–168.

[57] W. Spohn, An analysis of Hansson's dyadic deontic logic, Journal of Philosophical Logic 4 (1975) 237–252.

[58] S. Tan and J. Pearl, Specification and evaluation of preferences under uncertainty, in: *Proceedings of the Fourth International Conference on Principles of Knowledge Representation and Reasoning (KR'94)* (Morgan Kaufmann, San Francisco, CA, 1994) pp. 530–539.

[59] Y. Tan and L. van der Torre, 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*, Workshops in Computing (Springer, 1996) pp. 216–232.

[60] R. Thomason and J.F. Horty, Nondeterministic action and dominance: foundations for planning and qualitative decision, in: *Proceedings of the Conference on Theoretical Aspects of Reasoning about Agents (TARK'96)* (Morgan Kaufmann, 1996) pp. 229–250.

[61] J.E. Tomberlin, Contrary-to-duty imperatives and conditional obligation, Noûs 16 (1981) 357–375.

[62] D. Vakarelov, Modal characterization of the classes of finite and infinite quasi-ordered sets, in: *Mathematical Logic*, ed. P. Petkov (Plenum Press, New-York, 1990) pp. 373–387.

[63] L. van der Torre, Violated obligations in a defeasible deontic logic, in: *Proceedings of the Eleventh European Conference on Artificial Intelligence (ECAI'94)* (Wiley, 1994) pp. 371–375.

[64] L. van der Torre, Reasoning about obligations: defeasibility in preference-based deontic logics, Ph.D. thesis, Erasmus University Rotterdam (1997).

[65] L. van der Torre, Labeled logics of conditional goals, in: *Proceedings of the Thirteenth European Conference on Artificial Intelligence (ECAI'98)*, ed. H. Prade (1998) pp. 368–369.

[66] L. van der Torre, Phased labeled logics of conditional goals, in: *Logics in Artificial Intelligence*, Lecture Notes in Artificial Intelligence, Vol. 1489 (Springer, 1998) pp. 92–106.

[67] L. van der Torre, Defeasible goals, in: *Symbolic and Quantitative Approaches to Reasoning and Uncertainty. Proceedings of the ECSQARU'99*, Lecture Notes in Artificial Intelligence, Vol. 1638 (Springer, 1999) pp. 374–385.

[68] L. van der Torre, The logic of reusable propositional output with the fulfilment constraint, in: *Labelled Deduction*, eds. D. Basin et al., Applied Logic Series (Kluwer, 1999) pp. 247–268.

[69] L. van der Torre, Violation contexts and deontic independence, in: *Modeling and Using Context*, Lecture Notes in Artificial Intelligence, Vol. 1688 (Springer, 1999) pp. 361–374.

[70] L. van der Torre and Y. Tan, Cancelling and overshadowing: two types of defeasibility in defeasible deontic logic, in: *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI'95)* (Morgan Kaufman, 1995) pp. 1525–1532.

[71] L. van der Torre and Y. Tan, The many faces of defeasibility in defeasible deontic logic, in: *Defeasible Deontic Logic*, ed. D. Nute, Synthese Library, Vol. 263 (Kluwer, 1997) pp. 79–121.

[72] L. van der Torre and Y. Tan, Prohairetic Deontic Logic (PDL), in: *Logics in Artificial Intelligence*, Lecture Notes in Artificial Intelligence, Vol. 1489 (Springer, 1998) pp. 77–91.

[73] L. van der Torre and Y. Tan, The temporal analysis of Chisholm's paradox, in: *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI'98)* (1998) pp. 650–655.

[74] L. van der Torre and Y. Tan, An update semantics for deontic reasoning, in: *Norms, Logics and Information Systems. New Studies on Deontic Logic and Computer Science*, eds. P. McNamara and H. Prakken (1998) pp. 73–90.

[75] L. van der Torre and Y. Tan, An update semantics for prima facie obligations, in: *Proceedings of the Thirteenth European Conference on Artificial Intelligence (ECAI'98)* ed. H. Prade (1998) pp. 38–42.

[76] L. van der Torre and Y. Tan, Contextual deontic logic: Violation contexts and factual defeasibility, in: *Formal Aspects in Context*, ed. M. Cavalcanti, Applied Logic Series (Kluwer, 1999).

[77] L. van der Torre and Y. Tan, Diagnosis and decision making in normative reasoning, Artificial Intelligence and Law 7 (1999) 51–67.

[78] L. van der Torre and Y. Tan, Rights, duties and commitments between agents, in: *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI'99)* (Morgan Kaufmann, 1999) pp. 1239–1244.

[79] L. van der Torre and Y. Tan, An update semantics for defeasible obligations, in: *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI'99)* (Morgan Kaufmann, 1999) pp. 631–638.

[80] L. van der Torre and E. Weydert, Parameters for utilitarian desires in a qualitative decision theory, Applied Intelligence (2000, to appear).

[81] J. van Eck, A system of temporally relative modal and deontic predicate logic and its philosophical application, Logique et Analyse 100 (1982) 249–381.

[82] B.C. van Fraassen, Values and the heart command, Journal of Philosophy 70 (1973) 5–19.

[83] G.H. von Wright, Deontic logic, Mind 60 (1951) 1–15.

[84] G.H. von Wright, *The Logic of Preference* (Edinburgh University Press, 1963).

[85] G.H. von Wright, A new system of deontic logic, in: *Deontic Logic: Introductory and Systematic Readings*, ed. R. Hilpinen (D. Reidel, Dordrecht, Holland, 1971) pp. 105–120.

[86] E. Weydert, Hyperrational conditionals. Monotonic reasoning about nested default conditionals, in: *Foundations of Knowledge Representation and Reasoning*, Lecture Notes in Artificial Intelligence, Vol. 810 (Springer, 1994) pp. 310–332.

[87] R.J. Wieringa and J.-J.Ch. Meyer, Applications of deontic logic in computer science: A concise overview, in: *Deontic Logic in Computer Science: Normative System Specification*, eds. J.-J. Meyer and R. Wieringa (Wiley, Chichester, UK, 1993) pp. 17–40.

[88] X. Yu, Deontic logic with defeasible detachment, Ph.D. thesis, University of Georgia (1995).