How to Combine Ordering and Minimizing in a Deontic Logic based on Preferences*†

Yao-Hua Tan‡ Leendert W.N. van der Torre‡

Abstract

In this paper we propose a semantics for dyadic deontic logic with an explicit preference ordering between worlds, representing different degrees of ideality. We argue that this ideality ordering can be used in two ways to evaluate formulas, which we call *ordering* and *minimizing*. Ordering uses all preference relations between relevant worlds, whereas minimizing uses the most preferred worlds only. We show that ordering corresponds to strengthening of the antecedent, and minimizing to weakening of the consequent. Moreover, we show that in some cases ordering and minimizing have to be combined to obtain certain desirable conclusions, and that this can only be done in a so-called two-phase deontic logic. In the first phase, the preference ordering is constructed, and in the second phase the ordering is used for minimization. If these two phases are not distinguished, then counterintuitive conclusions follow.

1 Introduction

Preference-based deontic logics are deontic logics of which the semantics contains a preference ordering (usually on worlds of a Kripke style possible world model). This preference ordering reflects different degrees of 'ideality': a world is preferred over another world if it is, in some sense, better than the other world. For example, in some logics (e.g. [18, 28]) a value is associated with each world; in such cases, the ordering is connected (for all w_1 and w_2 we have $w_1 \leq w_2$ or $w_2 \leq w_1$). However, in general the preference ordering can be any partial pre-ordering. Hence, only reflexivity and transitivity are assumed. In such preference orderings there can be incomparable worlds. Incomparable worlds can be used to formalize moral dilemmas like $\mathcal{O}(p) \wedge \mathcal{O}(\neg p)$ in a consistent way, see [29].

In this paper we argue that a preference ordering can be used in two different ways to evaluate formulas. One way, which we call *Minimizing*, is to use the ordering to select the minimal elements that satisfy a formula. This minimizing approach

*This research was partially supported by the ESPRIT III Basic Research Project No.6156 DRUMS II and the ESPRIT III Basic Research Working Group No.8319 MODELAGE.

†To appear in: *Third International Workshop on Deontic Logic in Computer Science (∆eon'96)*, Springer Verlag, 1996.

‡Erasmus University Research Institute for Decision and Information Systems (EURIDIS). Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands. E-Mail: {ytan,ltorre}@euridis.fbk.eur.nl. Tel: (+31)10-4082601. Fax: (+31)10-4526134. Http://www.euridis.fbk.eur.nl/Euridis/welcome.html.

is commonly taken in preferential semantics for non-monotonic logics, see for example [21, 11, 14, 3]. The other way, which we call *Ordering*, is to use the whole ordering to evaluate a formula. One could explain the intuition behind the distinction between ordering and minimizing with the following metaphor. Ordering is what a person does when he envisions the message of the law, issued by the legislator, by determining the preference relations between the possible deontic states. In this envisionment process bad states are as important as good states. Minimizing is what the person does when he also tries to realize the best states. These two things are completely separated. A person might very well know how he should act, without acting accordingly.

The first deontic logic based on a preference ordering was introduced by Bengt Hansson [8]. It is a dyadic logic and it belongs to the first category. A dyadic obligation $\mathcal{O}(\alpha|\beta)$ can be read as 'if β (the antecedent) is the case, then α (the consequent) ought to be the case'. Dyadic logics were developed to represent *Contrary-To-Duty* (CTD) obligations and to solve notorious CTD paradoxes like the Chisholm and Forrester paradox. For example, the Forrester paradox [7] 'you should not kill ($\neg k$), but if you kill you should kill gently ($k \wedge g$)', can be represented in a dyadic deontic logic by the two dyadic obligations $\mathcal{O}(\neg k|\top)$ and $\mathcal{O}(k \wedge g|k)$. The second obligation is a CTD obligation of the first one and, as B. Hansson put it, it instructs you 'how to make the best of your sad circumstances'. In B. Hansson's logic, a dyadic obligation $\mathcal{O}(\alpha|\beta)$ is true iff α is true in the *minimal* (the preferred) β world. Hence, the first dyadic obligation says that k worlds are sub-ideal, and the CTD obligation says that optimal sub-ideal worlds are $k \wedge g$ worlds. B. Hansson's logic has been criticized because it lacks strengthening of the antecedent. For example, Alchourrón argues in [1] that lack of strengthening of the antecedent is acceptable for logics of defeasible reasoning (though it gives rise to the so-called irrelevance problem) or logics of defeasible obligations (traditionally called prima facie obligations, see [20]), but not for undefeasible obligations.

Recently, Sven Ove Hansson [9] introduced a preference ordering in a (monadic) deontic logic for a completely different reason. His preference-based deontic logic (PDL) avoids the Ross paradox by making $\mathcal{O}(p) \rightarrow \mathcal{O}(p \vee q)$ invalid. Hence, PDL is weaker than so-called 'standard' deontic logic (SDL), a normal monadic modal system of type KD according to the Chellas classification [6]. S.O. Hansson's PDL belongs to the second category of preference-based deontic logics, because the truth of $\mathcal{O}(\alpha)$ depends on the whole ordering. In this paper we generalize the ordering approach to the dyadic case and we show that ordering validates strengthening of the antecedent, whereas minimizing validates weakening of the consequent. The following example shows that combining strengthening of the antecedent and weakening of the consequent is problematic for *any* deontic logic that can represent a-temporal Contrary-To-Duty obligations. It is not a problem for deontic logics that make a temporal distinction between antecedent and consequent (see e.g. [14, 1]). However, such logics cannot represent several paradoxes, for example the Forrester paradox [7], see [16, 27].

Example 1 (SA+WC problem) *Consider the obligations that (1) you should either serve in the army or do alternative service, and (2) you should not serve in the army. It is counterintuitive to derive that you should do alternative service, given* *that you already serve in the army. Still this counterintuitive conclusion follows in several dyadic deontic logics. Assume a dyadic deontic logic that validates at least substitution of logical equivalents and the following inference patterns* Strengthening of the Antecedent *(*SA*),* Weakening of the Consequent *(*WC*) and* Conjunction *(*AND*).*

$$\mathrm{SA}: \frac{\mathcal{O}(\alpha|\beta_1)}{\mathcal{O}(\alpha|\beta_1 \wedge \beta_2)}$$

$$\mathrm{WC}: \frac{\mathcal{O}(\alpha_1|\beta)}{\mathcal{O}(\alpha_1 \vee \alpha_2|\beta)}$$

$$\mathrm{AND}: \frac{\mathcal{O}(\alpha_1|\beta), \mathcal{O}(\alpha_2|\beta)}{\mathcal{O}(\alpha_1 \wedge \alpha_2|\beta)}$$

Furthermore, assume as premises the obligations $\mathcal{O}(a \vee s|\top)$ and $\mathcal{O}(\neg a|\top)$, where a can be read as 'serving in the army' and s as 'doing alternative service' [10], and $\top$ stands for any tautology. The intuitive obligation $\mathcal{O}(\neg a \wedge s \mid \top)$ can be derived by AND*. From this obligation, the obligation $\mathcal{O}(\neg a \wedge s \mid a)$ can be derived by* SA*. Unfortunately, from this obligation, the counterintuitive obligation $\mathcal{O}(s \mid a)$ can be derived by* WC*. This obligation is considered to be counterintuitive, because it is not grounded in the premises. If a is true, then the first premise is fulfilled and the second one is violated. This inference can be blocked by replacing unrestricted strengthening of the antecedent by the following version of restricted strengthening of the antecedent, in which $\Diamond$ is a modal operator and $\Diamond\phi$ is true for all consistent propositional formulas ϕ.*

$$\mathrm{RSA}: \frac{\mathcal{O}(\alpha|\beta_1), \Diamond(\alpha \wedge \beta_1 \wedge \beta_2)}{\mathcal{O}(\alpha|\beta_1 \wedge \beta_2)}$$

The obligation $\mathcal{O}(\neg a \wedge s \mid a)$ cannot be derived from the obligation $\mathcal{O}(\neg a \wedge s \mid \top)$ by RSA*. Unfortunately, the counterintuitive $\mathcal{O}(s|a)$ can still be derived in another way. From the intuitive obligation $\mathcal{O}(\neg a \wedge s|\top)$ the intuitive $\mathcal{O}(s|\top)$ can be derived by* WC*. From this latter obligation, the counterintuitive $\mathcal{O}(s \mid a)$ can be derived by* RSA*.*

This problem can be solved by a technique, which might look odd at first sight, but which turns out to work well, namely to forbid application of RSA after WC has been applied. This means that in derivations first RSA has to be applied, and only afterwards WC may be applied. We call this the *two-phase approach* in deontic logic. Such a sequencing in derivations is rather unnatural and cumbersome from a proof-theoretic point of view. Surprisingly, the two-phase approach can be obtained very intuitively from a semantic point of view. In this paper we show that the two-phase approach can be obtained by combining the two usages of the preference ordering in a preference-based semantics of a deontic logic. The first phase corresponds to ordering, and the second phase corresponds to minimizing. In semantic terms the two-phase approach simply means that first a preference ordering has to be constructed by ordering worlds, and subsequently the constructed ordering can be used for minimization.

This paper is organized as follows. In Section 2 we give the preference logic in which we formalize the ordering (Section 2) as well as the minimizing (Section 3) aspects of preference orderings. It is based on a standard modal system, in which the accessibility relation of the Kripke models is interpreted as a preference ordering on the worlds, like in [5, 4]. In Section 4 we show how ordering and minimization can be combined in a two-phase deontic logic and how this solves the problems of Example 1.

2 Ordering

The logic $\mathcal{L}$ discussed in this paper is a modal logic with a preference-based semantics. In the logic, we abstract from actions, time and individuals.

Definition 1 (Syntax of $\mathcal{L}$) *The logic $\mathcal{L}$ is a bimodal system with the two normal modal operators $\Box$ and $\mathcal{I}$. It satisfies the inference rules modus ponens $\frac{\alpha, \alpha \rightarrow \beta}{\beta}$ and necessitation for both modal operators $\frac{\vdash \alpha}{\vdash \Box \alpha}$, $\frac{\vdash \alpha}{\vdash \mathcal{I} \alpha}$, and contains all propositional tautologies and the following axioms.*

$$\mathrm{K} : \Box(\alpha \rightarrow \beta) \rightarrow (\Box\alpha \rightarrow \Box\beta) \tag{1}$$

$$\mathrm{T} : \Box\alpha \rightarrow \alpha \tag{2}$$

$$4 : \Box\alpha \rightarrow \Box\Box\alpha \tag{3}$$

$$5 : \neg\Box\alpha \rightarrow \Box\neg\Box\alpha \tag{4}$$

$$\mathrm{K} : \mathcal{I}(\alpha \rightarrow \beta) \rightarrow (\mathcal{I}\alpha \rightarrow \mathcal{I}\beta) \tag{5}$$

$$\mathrm{T} : \mathcal{I}\alpha \rightarrow \alpha \tag{6}$$

$$4 : \mathcal{I}\alpha \rightarrow \mathcal{I}\mathcal{I}\alpha \tag{7}$$

$$\Box\alpha \rightarrow \mathcal{I}\alpha \tag{8}$$

The modal operator $\Box$ expresses necessity and satisfies the S5 axioms. $\mathcal{I}\alpha$ expresses that α is true in all states that are 'at least as ideal as the actual state' and $\mathcal{I}$ satisfies the S4 axioms. Furthermore, the axiom $\Box\alpha \rightarrow \mathcal{I}\alpha$ gives the relation between the two modal operators.

(Semantics of $\mathcal{L}$) *Kripke models $M = \langle W, R, \leq, V \rangle$ for $\mathcal{L}$ consist of W, a set of worlds, R and $\leq$, binary accessibility relations, and V, a valuation of the propositions in the worlds.*

$M, w \models \Box\alpha$ iff $\forall w' \in W$ with $R(w, w')$, it is true that $M, w' \models \alpha$, and
$M, w \models \mathcal{I}\alpha$ iff $\forall w' \in W$ with $w \leq w'$, it is true that $M, w' \models \alpha$.

$\Box$ defines an equivalence relation R on the worlds and $\mathcal{I}$ defines a reflexive and transitive ordering $\leq$ on the worlds of each equivalence class of $\Box$. The partial pre-ordering $\leq$ expresses preferences: $w_1 \leq w_2$ iff w_2 is at least as ideal as w_1.

In this section, we only consider the ordering approach to deontic logic. In evaluating formulas, the whole ordering is taken into account. This idea was first introduced in a monadic logic by S.O. Hansson [9] and formalized in a modal preference structure by Brown and Mantha [5]. In the following definition, we generalize this idea to dyadic deontic logics.

Definition 2 *Dyadic obligation and permission are defined as follows.*

$$\mathcal{O}(\alpha|\beta) =_{def} \Box((\alpha \wedge \beta) \rightarrow \mathcal{I}(\beta \rightarrow \alpha)) \tag{9}$$

$$\mathcal{P}(\alpha|\beta) =_{def} \Box((\alpha \wedge \beta) \rightarrow \mathcal{I}((\beta \rightarrow \alpha) \vee \neg\mathcal{I}\neg(\alpha \wedge \beta))) \tag{10}$$

Intuitively, an obligation $\mathcal{O}(p\,|\,q)$ expresses a strict preference of all $p \wedge q$ over $\neg p \wedge q$. This preference is represented by a negative condition: no $\neg p \wedge q$ is preferred over some $p \wedge q$. Formally, $M, w \models \mathcal{O}(p\,|\,q)$ iff $\forall w_1, w_2 \in W$ such that $R(w, w_1)$, $R(w, w_2)$, $M, w_1 \models p \wedge q$ and $M, w_2 \models \neg p \wedge q$, it is true that $w_1 \not\leq w_2$.[1] Notice that the ideality ordering is global (in the sense that the ideality ordering is *not* relative to a world) and nested operators therefore do *not* have an intuitive reading, although they could have a formal meaning in $\mathcal{L}$. The following example illustrates the definition of obligation.

Example 2 *Let $|\alpha|$ denote a world that satisfies α. Assume models that consist of four worlds $|p \wedge q|$, $|\neg p \wedge q|$, $|p \wedge \neg q|$ and $|\neg p \wedge \neg q|$, which are in the same equivalence class of R. A model satisfies the obligation $\mathcal{O}(p \mid \top)$ when neither $|\neg p \wedge q|$ nor $|\neg p \wedge \neg q|$ is preferred over $|p \wedge q|$ or $|p \wedge \neg q|$. For example, the Kripke model $|\neg p \wedge q| < |\neg p \wedge \neg q| < |p \wedge \neg q| < |p \wedge q|$ is a model for $\mathcal{O}(p|\top)$ but not for $\mathcal{O}(q|\top)$. Note that $\mathcal{O}(q|\top)$ is not true, because $|\neg p \wedge q| < |\neg p \wedge \neg q|$ and $|\neg p \wedge q| < |p \wedge \neg q|$. This shows how in the ordering approach the whole ordering is taken into account in the evaluation of a formula, and not just the most preferred $|p \wedge q|$ worlds.*

The following proposition gives several properties of the dyadic obligations.

Proposition 1 *The logic $\mathcal{L}$ validates the following theorems of Strengthening of the Antecedent (SA), Conjunction (AND) and Disjunction (OR), a version of Deontic Detachment (DD′) and the standard relation between obligations and permissions.*

$$\mathrm{SA}: \mathcal{O}(\alpha|\beta_1) \rightarrow \mathcal{O}(\alpha|\beta_1 \wedge \beta_2) \tag{11}$$

$$\mathrm{AND}: \mathcal{O}(\alpha_1|\beta) \wedge \mathcal{O}(\alpha_2|\beta) \rightarrow \mathcal{O}(\alpha_1 \wedge \alpha_2|\beta) \tag{12}$$

$$\mathrm{OR}: \mathcal{O}(\alpha_1|\beta) \wedge \mathcal{O}(\alpha_2|\beta) \rightarrow \mathcal{O}(\alpha_1 \vee \alpha_2|\beta) \tag{13}$$

$$\mathrm{DD}': \mathcal{O}(\alpha|\beta) \wedge \mathcal{O}(\beta|\gamma) \rightarrow \mathcal{O}(\alpha \wedge \beta|\gamma) \tag{14}$$

[1] In this article, we do not discuss permissions and their relations with obligations due to space limitations. We would have liked to define permission such that $M, w \models \mathcal{P}(p|q)$ iff $\forall w_1, w_2 \in W$ such that $R(w, w_1)$, $R(w, w_2)$, $M, w_1 \models p \wedge q$ and $M, w_2 \models \neg p \wedge q$, it is true that $w_1 \not< w_2$. This, however, is not possible in this simple setup. The definition of permission in Definition 2 is weaker.

$$\mathcal{O}(\alpha|\beta) \rightarrow \mathcal{P}(\alpha|\beta) \tag{15}$$

The logic $\mathcal{L}$ does not *validate the following theorems of Weakening of the Consequent (*WC *and* WC$'$*), Deontic Detachment (or transitivity) (*DD*) and the Deontic axiom (*D*).*

$$\text{WC} : \mathcal{O}(\alpha_1|\beta) \rightarrow \mathcal{O}(\alpha_1 \vee \alpha_2|\beta) \tag{16}$$

$$\text{WC}' : \mathcal{O}(\alpha_1 \wedge \alpha_2|\beta) \rightarrow \mathcal{O}(\alpha_1|\beta) \wedge \mathcal{O}(\alpha_2|\beta) \tag{17}$$

$$\text{DD} : \mathcal{O}(\alpha|\beta) \wedge \mathcal{O}(\beta|\gamma) \rightarrow \mathcal{O}(\alpha|\gamma) \tag{18}$$

$$\text{D} : \neg(\mathcal{O}(\alpha|\beta) \wedge \mathcal{O}(\neg\alpha|\beta)) \tag{19}$$

Proof *Follows directly from the definitions.*

The validity of strengthening of the antecedent follows from the fact that a preference of p over $\neg p$ implies a preference of $p \wedge q$ over $\neg p \wedge q$. The most remarkable property of the logic is the invalidity of weakening of the consequent. Intuitively, the lack of weakening of $\mathcal{O}(p|\top)$ to $\mathcal{O}(p \vee q|\top)$ is the consequence of the fact that $\mathcal{O}(p|\top)$ expresses a preference of *all* p over $\neg p$, because from such a preference does not follow that $p \vee q$ is *always* preferred over $\neg p \wedge \neg q$. This is illustrated by the following example.

Example 3 *Reconsider the Kripke model $|\neg p \wedge q| < |\neg p \wedge \neg q| < |p \wedge \neg q| < |p \wedge q|$ of the previous example. The model satisfies $\mathcal{O}(p|\top)$ but not $\mathcal{O}(p \vee q|\top)$.*

The lack of weakening of the consequence is considered to be an advantage by [9, 5]. They quote Ross, who gave the following counterintuitive example of weakening of the consequent, called the Ross paradox: 'If you ought to mail the letter, then you ought to mail or burn the letter' [19]. The following example illustrates the intuitive theorem DD$'$.

Example 4 *Let S be the set of dyadic obligations $\{\mathcal{O}(a|\top), \mathcal{O}(t|a)\}$, 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'. Hence, the two obligations can be read as 'a certain man should go to the assistance of his neighbors', and 'if he goes, then he should tell them he is coming'. The obligation $\mathcal{O}(a \wedge t|\top)$ can be derived from S with* DD$'$*, which expresses that ideally, the man goes to the assistance of his neighbors* and *he tells them he is coming.*

Unfortunately, the logic $\mathcal{L}$ validates the counterintuitive theorems $\mathcal{O}(\perp|\alpha)$ and $\mathcal{O}(\alpha|\alpha)$. For this reason, we define various subtle other definitions of obligations in the preference logic. In the following definition, $\mathcal{O}^c(\alpha|\beta)$ has an additional condition which works like a 'consistency check' to test whether $\alpha \wedge \beta$ is possible. $\mathcal{O}^{cc}(\alpha|\beta)$ also has another additional condition which tests whether $\neg\alpha \wedge \beta$ is possible.

Definition 3 *Two alternative definitions of dyadic obligations and permissions are defined as follows. The modal operator* $\Diamond$ *is defined as usually as the dual of the* $\Box$ *operator:* $\Diamond\phi =_{def} \neg\Box\neg\phi$.

$$\mathcal{O}^c(\alpha|\beta) =_{def} \mathcal{O}(\alpha|\beta) \wedge \Diamond(\alpha \wedge \beta) \tag{20}$$

$$\mathcal{P}^c(\alpha|\beta) =_{def} \mathcal{P}(\alpha|\beta) \wedge \Diamond(\alpha \wedge \beta) \tag{21}$$

$$\mathcal{O}^{cc}(\alpha|\beta) =_{def} \mathcal{O}(\alpha|\beta) \wedge \Diamond(\alpha \wedge \beta) \wedge \Diamond(\neg\alpha \wedge \beta) \tag{22}$$

$$\mathcal{P}^{cc}(\alpha|\beta) =_{def} \mathcal{P}(\alpha|\beta) \wedge \Diamond(\alpha \wedge \beta) \wedge \Diamond(\neg\alpha \wedge \beta) \tag{23}$$

Proposition 2 *The logic* $\mathcal{L}$ *validates the following theorems.*

$$\mathcal{O}(\bot|\alpha), \neg\mathcal{O}^c(\bot|\alpha), \neg\mathcal{O}^{cc}(\bot|\alpha) \tag{24}$$

$$\mathcal{O}(\alpha|\alpha), \Diamond\alpha \rightarrow \mathcal{O}^c(\alpha|\alpha), \neg\mathcal{O}^{cc}(\alpha|\alpha) \tag{25}$$

Proof *Follows directly from the definitions.*

The obligations $\mathcal{O}^c$ and $\mathcal{O}^{cc}$ validate weaker versions of the theorems of Proposition 1, like for example the following Restricted Strengthening of the Antecedent (RSA) and Restricted Conjunction (RAND):

$$\text{RSA} : \mathcal{O}^c(\alpha|\beta) \wedge \Diamond(\alpha \wedge \beta \wedge \gamma) \rightarrow \mathcal{O}^c(\alpha|\beta \wedge \gamma) \tag{26}$$

$$\text{RAND} : \mathcal{O}^c(\alpha_1|\beta) \wedge \mathcal{O}^c(\alpha_2|\beta) \wedge \Diamond(\alpha_1 \wedge \alpha_2 \wedge \beta) \rightarrow \mathcal{O}^c(\alpha_1 \wedge \alpha_2|\beta) \tag{27}$$

We already saw restricted strengthening of the antecedent in Example 1 in the introduction. In [27] this lack of unrestricted strengthening of the antecedent is dubbed 'factual defeasibility' and it is argued that this factual defeasibility is necessary to represent a-temporal CTD obligations. We elaborate on Example 1 in Section 4.

The logic can, like [5] but in contrast to SDL, represent moral dilemmas like $\mathcal{O}(p|\top) \wedge \mathcal{O}(\neg p|\top)$ in a consistent way. This follows from the lack of the theorem $\neg(\mathcal{O}(\alpha \mid \beta) \wedge \mathcal{O}(\neg\alpha \mid \beta))$. In the semantics, a moral dilemma $\mathcal{O}(p \mid \top) \wedge \mathcal{O}(\neg p \mid \top)$ corresponds to incomparable $|p|$ and $|\neg p|$ worlds. Brown and Mantha [5] consider this property a major advantage of their logic. However, in some applications of deontic logic, this is not a desirable property. In such cases, we can add an axiom to the logic which makes moral dilemmas inconsistent. For example, we can accept the D axiom of SDL, $\neg(\mathcal{O}^c(\alpha \mid \top) \wedge \mathcal{O}^c(\neg\alpha \mid \top))$, in our logic. However, this would still leave $\mathcal{O}^c(\alpha \mid \top) \wedge \mathcal{O}^c(\neg\alpha \wedge \beta \mid \top)$ consistent. We therefore define the following version of the D axiom.

$$\text{D}' : \Box\neg(\alpha_1 \wedge \alpha_2) \rightarrow \neg(\mathcal{O}^c(\alpha_1|\top) \wedge \mathcal{O}^c(\alpha_2|\top)) \tag{28}$$

In some deontic logics, incomparable worlds are only used to model moral dilemmas, see e.g. [29]. However, in our logic this is not the only use of incomparable worlds. This follows from the fact that even when the D$'$ axiom is accepted, there can be incomparable worlds. This is illustrated by the following example.

Example 5 *Let M be a model of L that satisfies the two obligations $\mathcal{O}^c(p_1|\top)$ and $\mathcal{O}^c(p_2|\top)$. The $|p_1 \wedge \neg p_2|$ and $|\neg p_1 \wedge p_2|$ worlds of M are incomparable, because the first is better than the latter with respect to obligation $\mathcal{O}^c(p_1|\top)$, but worse with respect to obligation $\mathcal{O}^c(p_2|\top)$.*

Unfortunately, the D$'$ axiom cannot be generalized for arbitrary antecedents. The following example illustrates that the combination of RSA and a generalized D$'$ axiom is problematic.

Example 6 (SA+D problem) *Consider the following axiom:*

$$\mathrm{D}'' : \Box\neg(\alpha_1 \wedge \alpha_2 \wedge \beta) \rightarrow \neg(\mathcal{O}^c(\alpha_1|\beta) \wedge \mathcal{O}^c(\alpha_2|\beta))$$

Let S be the set of obligations $\{\mathcal{O}^c(p_1|\top), \mathcal{O}^c(p_2|\top)\}$. With RSA *(and $\Diamond(p_1 \wedge \neg p_2)$ and $\Diamond(\neg p_1 \wedge p_2)$), we can derive the obligations $\mathcal{O}^c(p_1 \mid \neg(p_1 \wedge p_2))$ and $\mathcal{O}^c(p_2 \mid \neg(p_1 \wedge p_2))$ from S. The two derived obligations are inconsistent with the* D$''$ *axiom when β is equivalent to $\neg(p_1 \wedge p_2)$.*[2]

The inconsistency is derived by only using RSA *and* D$''$. *Hence, there are two solutions for this problem: weakening* D$''$ *or weakening* RSA. *An attempt of the first solution is to weaken the* D$''$ *axiom to the following axiom.*

$$\mathrm{D}''' : \Box\neg(\alpha_1 \wedge \alpha_2) \rightarrow \neg(\mathcal{O}^c(\alpha_1|\beta) \wedge \mathcal{O}^c(\alpha_2|\beta))$$

Unfortunately, S is also inconsistent with the D$'''$ *axiom. From S we can derive $\mathcal{O}^c(p_1 \wedge \neg(p_1 \wedge p_2)|\neg(p_1 \wedge p_2))$ and $\mathcal{O}^c(p_2 \wedge \neg(p_1 \wedge p_2)|\neg(p_1 \wedge p_2))$ with $\Diamond\alpha \rightarrow \mathcal{O}^c(\alpha|\alpha)$ and the* RAND *theorem. They are logically equivalent to $\mathcal{O}^c(p_1 \wedge \neg p_2|\neg(p_1 \wedge p_2))$ and $\mathcal{O}^c(p_2 \wedge \neg p_1|\neg(p_1 \wedge p_2))$, which are inconsistent with* D$'''$ *when β is equivalent to $\neg(p_1 \wedge p_2)$. Obviously, if $\Diamond\alpha \rightarrow \mathcal{O}^c(\alpha \mid \alpha)$ would not be a theorem of the logic, then the* D$'''$ *axiom would be a solution for the* SA+D *problem (see for example the conditional logic discussed in [1]). However, in the logic $\mathcal{L}$ the formula $\Diamond\alpha \rightarrow \mathcal{O}^c(\alpha| \alpha)$ is a theorem, so this solution does not work for $\mathcal{L}$.*

The second solution of the SA+D problem is to weaken RSA. *This solution implies that the obligation $\mathcal{O}^c(p_1|\neg(p_1 \wedge p_2))$ should not be derivable from the obligation $\mathcal{O}^c(p_1|\top)$ in our example. In the next section, we show that the minimizing approach does not have any strengthening of the antecedent, thus it gives the second solution to the* SA+D *problem. However, in Example 9 we will see that this lack of strengthening of the antecedent makes another type of moral dilemma consistent. A satisfactory solution for the SA+D problem is given in the two-phase logic in Section 4.*

3 Minimizing

In the dyadic logic of Bengt Hansson, an obligation $\mathcal{O}(\alpha|\beta)$ is true iff α is true in all minimal β worlds [8]. We therefore say that his logic (and related deontic logics,

[2] Semantically, the D$'$ axiom states that the most preferred worlds are comparable, and the axiom D$''$ states that all worlds are comparable. Example 5 illustrates that this condition is too strong.

see [13] for an overview) is based on minimizing. In the following definition, we give several definitions of obligations which are closely related to the reconstruction of Boutilier [4] of B. Hansson's logic in a modal preference structure. In this logic, the maximal worlds are the preferred worlds instead of the minimal worlds. Maximizing in our logic is equivalent to minimizing in B. Hansson's logic. We give several definitions.[3]

Definition 4 *Dyadic obligation and permission are defined as follows.*

$$\mathcal{O}_\exists(\alpha|\beta) =_{def} \Diamond(\beta \wedge \mathcal{I}(\beta \to \alpha)) \tag{29}$$

$$\mathcal{O}_\forall(\alpha|\beta) =_{def} \Box(\beta \to \neg\mathcal{I}\neg(\beta \wedge \mathcal{I}(\beta \to \alpha))) \tag{30}$$

$$\mathcal{P}_\exists(\alpha|\beta) =_{def} \neg\mathcal{O}_\forall(\neg\alpha|\beta) = \Diamond(\beta \wedge \mathcal{I}(\beta \to \neg\mathcal{I}\neg(\beta \wedge \alpha))) \tag{31}$$

$$\mathcal{P}_\forall(\alpha|\beta) =_{def} \neg\mathcal{O}_\exists(\neg\alpha|\beta) = \Box(\beta \to \neg\mathcal{I}\neg(\beta \wedge \alpha)) \tag{32}$$

$$\mathcal{O}^c_\forall(\alpha|\beta) =_{def} \mathcal{O}_\forall(\alpha|\beta) \wedge \Diamond\beta \tag{33}$$

$$\mathcal{P}^c_\forall(\alpha|\beta) =_{def} \mathcal{P}_\forall(\alpha|\beta) \wedge \Diamond\beta \tag{34}$$

The obligation $\mathcal{O}_\exists(p|q)$ is true in a model if p is true in an equivalence class of most preferred $|q|$ worlds of the model and $\mathcal{O}_\forall(p|q)$ is true if p is true in all most preferred $|q|$ worlds.[4] Hence, both obligations refer to the optimal worlds where q is true, and $\mathcal{O}_\exists(p|\top)$ and $\mathcal{O}_\forall(p|\top)$ refer to the ideal worlds.[5]

Example 7 *Reconsider the Kripke model $|\neg p \wedge q| < |\neg p \wedge \neg q| < |p \wedge \neg q| < |p \wedge q|$ of Example 2. The model satisfies $\mathcal{O}_\forall(p|\top)$, $\mathcal{O}_\forall(q|\top)$, $\mathcal{O}_\exists(p|\top)$ and $\mathcal{O}_\exists(q|\top)$, whereas $\mathcal{O}(q|\top)$ is not true in this model. Since $\mathcal{O}_\exists(q|\top)$ is equivalent with $\Diamond\mathcal{I}q$ it is clear that q has to be true in some most preferred $|\top|$ worlds, and also that less preferred $|\top|$ worlds do not effect the truth of $\Diamond\mathcal{I}q$. Hence, in the evaluation of $\mathcal{O}_\exists(q|\top)$ only preferred elements are taken into account and not the whole ordering.*

The main properties of the logic are given by the following proposition.

Proposition 3 *The logic $\mathcal{L}$ validates the following theorems.*

$$\mathrm{WC}_\exists : \mathcal{O}_\exists(\alpha_1|\beta) \to \mathcal{O}_\exists(\alpha_1 \vee \alpha_2|\beta) \tag{35}$$

$$\mathrm{WC}_\forall : \mathcal{O}_\forall(\alpha_1|\beta) \to \mathcal{O}_\forall(\alpha_1 \vee \alpha_2|\beta) \tag{36}$$

[3] Obviously, we can define $\mathcal{O}^{cc}_\exists$ etc by adding conditions to the definitions.

[4] Moreover, the definition can deal with infinite chains, see [3].

[5] Again, we will not discuss permissions. Just notice that the definition of permission is closely related to the definition of permission in SDL.

$$\mathrm{AND}_\forall : \mathcal{O}_\forall(\alpha_1|\beta) \wedge \mathcal{O}_\forall(\alpha_2|\beta) \rightarrow \mathcal{O}_\forall(\alpha_1 \wedge \alpha_2|\beta) \tag{37}$$

$$\mathrm{D}_\forall : \neg(\mathcal{O}^c_\forall(\alpha|\beta) \wedge \mathcal{O}^c_\forall(\neg\alpha|\beta)) \tag{38}$$

$$\mathcal{O}_\exists(\alpha|\beta) \rightarrow \Diamond(\alpha \wedge \beta) \tag{39}$$

$$\mathcal{P}_\exists(\alpha|\beta) \rightarrow \Diamond(\alpha \wedge \beta) \tag{40}$$

$$\mathcal{O}_\exists(\alpha|\beta) \rightarrow \mathcal{P}_\exists(\alpha|\beta) \tag{41}$$

$$\mathcal{O}_\forall(\alpha|\beta) \rightarrow \mathcal{P}_\forall(\alpha|\beta) \tag{42}$$

$$\neg\mathcal{O}_\exists(\perp|\alpha) \tag{43}$$

$$\neg\mathcal{O}^c_\forall(\perp|\alpha) \tag{44}$$

$$\Diamond\alpha \rightarrow \mathcal{O}_\exists(\alpha|\alpha) \tag{45}$$

$$\mathcal{O}_\forall(\alpha|\alpha) \tag{46}$$

$$\mathcal{O}^c_\forall(\alpha|\beta) \rightarrow \mathcal{O}_\exists(\alpha|\beta) \tag{47}$$

$$\mathcal{P}^c_\forall(\alpha|\beta) \rightarrow \mathcal{P}_\exists(\alpha|\beta) \tag{48}$$

The logic $\mathcal{L}$ does not validate the following theorems.

$$\mathrm{SA}_\exists : \mathcal{O}_\exists(\alpha|\beta_1) \rightarrow \mathcal{O}_\exists(\alpha|\beta_1 \wedge \beta_2) \tag{49}$$

$$\mathrm{SA}_\forall : \mathcal{O}_\forall(\alpha|\beta_1) \rightarrow \mathcal{O}_\forall(\alpha|\beta_1 \wedge \beta_2) \tag{50}$$

$$\mathrm{AND}_\exists : \mathcal{O}_\exists(\alpha_1|\beta) \wedge \mathcal{O}_\exists(\alpha_2|\beta) \rightarrow \mathcal{O}_\exists(\alpha_1 \wedge \alpha_2|\beta) \tag{51}$$

$$\mathrm{DD}_\exists : \mathcal{O}_\exists(\alpha|\beta) \wedge \mathcal{O}_\exists(\beta|\gamma) \rightarrow \mathcal{O}_\exists(\alpha|\gamma) \tag{52}$$

$$\mathrm{DD}_\forall : \mathcal{O}_\forall(\alpha|\beta) \wedge \mathcal{O}_\forall(\beta|\gamma) \rightarrow \mathcal{O}_\forall(\alpha|\gamma) \tag{53}$$

$$\mathrm{D}_\exists : \neg(\mathcal{O}_\exists(\alpha|\beta) \wedge \mathcal{O}_\exists(\neg\alpha|\beta)) \tag{54}$$

Proof *Follows directly from the definitions.*

The logic validates weakening of the consequent, because the most preferred world that satisfies p_1 also satisfies $p_1 \vee p_2$. However, the logic does not have strengthening of the antecedent of $\mathcal{O}_\forall(p \mid \top)$ to $\mathcal{O}_\forall(p \mid q)$, because the preferred $|\top|$ worlds may be different from the preferred $|q|$ worlds.

Example 8 *Reconsider the Kripke model* $|\neg p \wedge q| < |\neg p \wedge \neg q| < |p \wedge \neg q| < |p \wedge q|$ *of the previous example. The model satisfies* $\mathcal{O}_\exists(q|\top)$ *and* $\mathcal{O}_\forall(q|\top)$*, but not* $\mathcal{O}_\exists(q|\neg p)$ *or* $\mathcal{O}_\forall(q|\neg p)$*. For example,* $\mathcal{O}_\exists(q|\neg p)$ *is false because the preferred* $|\neg p|$ *worlds are the* $|\neg p \wedge \neg q|$ *worlds. Hence,* $\mathcal{O}_\exists$ *does not have strengthening of the antecedent.*

Obviously, $\mathcal{O}_\exists$ can represent moral dilemmas like $\mathcal{O}_\exists(p \mid \top) \wedge \mathcal{O}_\exists(\neg p \mid \top)$ in a consistent way, and $\mathcal{O}^c_\forall$ cannot. However, the following example shows that $\mathcal{O}^c_\forall$ does not give a satisfactory solution for the SA+D problem discussed in Example 6. $\mathcal{O}^c_\forall$ still represents some moral dilemmas in a consistent way, as a result of the total lack of strengthening of the antecedent.

Example 9 (SA+D Problem, continued) *The following theorem of the logic* $\mathcal{L}$ *is the axiom* D$''$ *for* $\mathcal{O}^c_\forall$*. The validity of this theorem follows directly from the theorems* $\mathrm{D}_\forall$ *and* $\mathrm{WC}_\forall$*.*

$$\mathrm{D}''_\forall : \Box\neg(\alpha_1 \wedge \alpha_2 \wedge \beta) \rightarrow \neg(\mathcal{O}^c_\forall(\alpha_1|\beta) \wedge \mathcal{O}^c_\forall(\alpha_2|\beta))$$

The obligations $\mathcal{O}^c_\forall(p_1 \mid \top)$ *and* $\mathcal{O}^c_\forall(p_2 \mid \top)$ *are consistent. Hence,* $\mathcal{O}^c_\forall$ *solves the problem that was discussed in Example 6. However,* $\mathcal{O}^c_\forall$ *does not solve all of the* SA+D *problems. Consider the following example, adapted from [16]. Let* S *be the set of obligations* $S = \{\mathcal{O}^c_\forall(\neg c|\top), \mathcal{O}^c_\forall(c|k)\}$*, where* k *can be read as 'killing someone (the witness)' and* c *can be read as 'offering someone a cigarette'.* S *is consistent and from* S *we can derive the obligation* $\mathcal{O}^c_\forall(\neg k|\top)$*. Prakken and Sergot argue in [16] that* S *should be inconsistent, even when* S *is extended with the obligation* $\mathcal{O}^c_\forall(\neg k|\top)$*. In Section 4 we show how this inconsistency can be obtained in a two-phase logic.*

We can impose the D axiom for $\mathcal{O}_\exists$ by demanding that the partial ordering $\leq$ is connected: for all w_1, w_2, we have $w_1 \leq w_2$ or $w_2 \leq w_1$. If the ordering is connected, then the modal operator $\mathcal{I}$ satisfies the S4.3 axioms instead of the S4 axioms (like, for example, Boutilier's logic CO [4]) and $\mathcal{O}_\exists(\alpha|\beta)$ becomes equivalent to $\mathcal{O}^c_\forall(\alpha|\beta)$. The last property follows from the fact that truth in an equivalence class of preferred worlds and truth in all preferred worlds has become equivalent.

A relation between ordering and minimizing is given by the following proposition.

Proposition 4 *The logic* $\mathcal{L}$ *validates the following theorem.*

$$\mathcal{O}^c(\alpha|\beta) \rightarrow \mathcal{O}_\exists(\alpha|\beta) \tag{55}$$

Proof *The formula is equivalent with the following theorem of* $\mathcal{L}$*:*

$$(\Box(\beta \wedge \alpha \rightarrow \mathcal{I}(\beta \rightarrow \alpha)) \wedge \Diamond(\alpha \wedge \beta)) \rightarrow \Diamond(\beta \wedge \mathcal{I}(\beta \rightarrow \alpha))$$

4 Combining Ordering and Minimizing

4.1 Combining

The idea of combining ordering and minimizing is to combine formulas with $\mathcal{O}$, $\mathcal{O}_\exists$ and $\mathcal{O}_\forall$ operators. However, this is not possible in the logic $\mathcal{L}$. Before we can make these combinations there are some technical problems which have to be solved first. The following example shows the two problems.

Example 10 *Consider the set $S = \{\mathcal{O}(p \mid \top)\}$. Hence, all we know is that p is preferred over $\neg p$. Given this knowledge, we would expect to derive $\mathcal{O}_\exists(p \mid \top)$ and $\mathcal{O}_\forall(p \mid \top)$. This, however, is not the case. There are two problems, the first one blocks both derivations, and the second one only the derivation of $\mathcal{O}_\forall(p|\top)$.*

1. *Worlds do not have to exist in a model.*

   Let M be a model with only $|\neg p|$ worlds. M satisfies $\mathcal{O}(p|\top)$, but it does not satisfy $\mathcal{O}_\forall(p|\top)$ or $\mathcal{O}_\exists(p|\top)$. Hence, neither of these two formulas are entailed by $\mathcal{O}(p|\top)$.

2. *The ordering of worlds can be too weak.*

   Again, consider the obligation $\mathcal{O}(p \mid \top)$. All models that satisfy $|p| \not\leq |\neg p|$ are models of $\mathcal{O}(p|\top)$. Hence, $|p|$ worlds and $|\neg p|$ worlds are either incomparable, or $|p|$ worlds are strictly preferred over $|\neg p|$ worlds. Let M be a model in which all $|p|$ and $|\neg p|$ worlds are incomparable. M satisfies $\mathcal{O}(p \mid \top)$, but it does not satisfy $\mathcal{O}_\forall(p \mid \top)$. Hence, $\mathcal{O}_\forall(p \mid \top)$ is not entailed by $\mathcal{O}(p \mid \top)$. For minimization, we only want the models in which $|p|$ worlds are strictly preferred over $|\neg p|$.

The solution of the first problem is to consider only models in which all propositionally satisfiable formulas ϕ are true in *some* world. This can be 'axiomatized' with Boutilier's axiom scheme LP, see [12, 3] for a discussion. The axiom scheme LP states that every formula ϕ without any occurrences of modal operators, which is propositionally satisfiable, is true in some world.

Definition 5 *The logic $\mathcal{L}^*$ is $\mathcal{L}$ extended with the following axiom scheme:*

$$\text{LP} : \Diamond\phi \text{ for all satisfiable propositional } \phi \tag{56}$$

We write $\models$ for logical entailment in $\mathcal{L}^*$.

A solution of the second problem is to define a preference ordering on models, which prefers models which are maximally connected with respect to the partial pre-ordering $\leq$, i.e. with the most binary relations of $\leq$. The preferred models of this ordering are the only models which are used for minimization.

Definition 6 *Let $M_1 = \langle W_1, R_1, \leq_1, V_1\rangle$ and $M_2 = \langle W_2, R_2, \leq_2, V_2\rangle$ be two models of $\mathcal{L}^*$. M_1 is preferred over M_2 for mapping τ, written as $M_1 \sqsubseteq_\tau M_2$, iff:*

1. *τ is a one-to-one mapping of the worlds of W_2 to the worlds of W_1 such that the worlds satisfy the same propositions, and*

2. If $w_1 \leq_2 w_2$ for $w_1, w_2 \in W_2$ then $\tau(w_1) \leq_1 \tau(w_2)$.

We write $M_1 \sqsubset_\tau M_2$ iff $M_1 \sqsubseteq_\tau M_2$ and $M_2 \not\sqsubseteq_{\tau^{-1}} M_1$.

The ordering on models ($\sqsubseteq$) should not be confused with the ordering on worlds ($\leq$). The ordering on models is a technical trick to ensure that the worlds within a model are maximally connected, whereas the ordering on worlds expresses the ideality ordering. Given the preference ordering on models, we can define a notion of preferential entailment, see [21, 11].

Definition 7 *Let $M = \langle W, R, \leq, V\rangle$ be a model and S be a set of sentences. A world $w \in W$ of M preferentially satisfies S, written as $M, w \models_\sqsubset S$ iff $M, w \models S$ and there is not a model M' and a mapping τ such that $M', \tau(w) \models S$ and $M' \sqsubset_\tau M$ (M is a preferred model of S). S preferentially entails ϕ, written as $S \models_\sqsubset \phi$, iff for all M and w, if $M, w \models_\sqsubset S$ then $M, w \models \phi$.*

We will illustrate the notion of preferential entailment in $\mathcal{L}^*$ by the Reykjavic paradox [2]. See [26] for several interpretations of this paradox.

Example 11 *Let S be the set of obligations $\{\mathcal{O}(\neg r \wedge \neg g|\top), \mathcal{O}(r|g), \mathcal{O}(g|r)\}$, where r can be read as telling a secret to Reagan and g as telling the secret to Gorbatsjov. The preferred model is given in Figure 1. We have $S \not\models_\sqsubset \mathcal{O}(\neg r|\top)$, $S \not\models_\sqsubset \mathcal{O}(\neg g|\top)$, $S \models_\sqsubset \mathcal{O}_\forall(\neg r|\top)$ and $S \models_\sqsubset \mathcal{O}_\forall(\neg g|\top)$.*

Without the LP *axiom scheme, the $|\neg r \wedge \neg g|$ worlds do not have to exist in a model, and such a model M would not satisfy $M \models \mathcal{O}_\forall(\neg r|\top)$ and $M \models \mathcal{O}_\exists(\neg r|\top)$. Without the preference ordering on models, the $|\neg r \wedge \neg g|$ and $|r \wedge g|$ worlds could be incomparable. Such a model M would still satisfy $M \models \mathcal{O}_\exists(\neg r|\top)$ but it would not satisfy $M \models \mathcal{O}_\forall(\neg r|\top)$.*

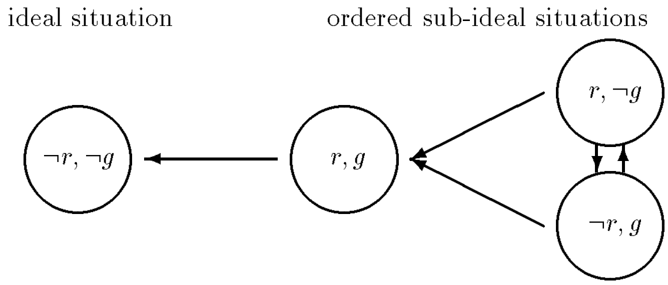

Figure 1: Preference relation of the Reykjavic paradox

Preferential entailment is a typical mechanism from non-monotonic reasoning. The combination of ordering and minimizing is non-monotonic, as the following example illustrates.

Example 12 *Let $S = \{\mathcal{O}(p|\top)\}$ and $S' = \{\mathcal{O}(p|\top), \mathcal{O}(\neg p|\top)\}$. We have $S \models_\sqsubset \mathcal{O}_\forall(p|\top)$ and $S' \not\models_\sqsubset \mathcal{O}_\forall(p|\top)$. Hence, by addition of a formula we loose conclusions.*

4.2 The two phases in a deontic logic

The two phases in a deontic logic correspond to the two different kinds of obligations $\mathcal{O}^c$ and $\mathcal{O}_\exists$. Semantically, the first phase corresponds to ordering ($\mathcal{O}^c$) and the second phase to minimizing ($\mathcal{O}_\exists$). From a proof theoretic point of view, the first phase corresponds to applying valid inferences of $\mathcal{O}^c$ like RSA, RAND etc, and the second phase corresponds to applying valid inferences of $\mathcal{O}_\exists$ like WC. The basic technique of deontic logic as a two-phase logic is that a conclusion of the form $\mathcal{O}_\exists(\alpha|\beta)$ can be derived either with or without $\mathcal{O}^c(\alpha|\beta)$. In the first case $\mathcal{O}_\exists(\alpha|\beta)$ can be derived via $\mathcal{O}^c(\alpha|\beta)$ with Proposition 4, which says that the latter formula implies the first one. If so, we say that $\mathcal{O}_\exists(\alpha\mid\beta)$ is derived in the first phase. In the second case we say that $\mathcal{O}_\exists(\alpha|\beta)$ is second phase derived. The important difference is that in the first phase we can apply RSA to $\mathcal{O}_\exists(\alpha|\beta)$, because of the simultaneous occurrence of $\mathcal{O}^c(\alpha\mid\beta)$. We apply RSA to $\mathcal{O}^c(\alpha\mid\beta)$ to obtain, for example, $\mathcal{O}^c(\alpha|\beta\wedge\gamma)$, and then due to Proposition 4 we also obtain $\mathcal{O}_\exists(\alpha|\beta\wedge\gamma)$. If $\mathcal{O}^c(\alpha|\beta)$ does not occur simultaneously with $\mathcal{O}_\exists(\alpha|\beta)$, then there is no way we can apply RSA to this formula. Being a minimizing formula it lacks RSA. Hence, once $\mathcal{O}_\exists(\alpha|\beta)$ has been derived in the second phase, we loose RSA permanently for subsequent derivations of this formula. Analogously, we can say that $\mathcal{O}_\exists(\alpha\mid\beta)$ is first phase or second phase entailed by a set of premises, depending on whether S does or does not entail $\mathcal{O}^c(\alpha|\beta)$.

The following example shows that the two-phase approach solves the SA+WC problem of Example 1 in the introduction. In this example, we only consider $\mathcal{O}_\exists$ and not $\mathcal{O}_\forall$. We can therefore use standard logical entailment $\models$ instead of the more complicated preferential entailment $\models_\sqsubset$.

Example 13 (SA+WC problem, continued) *Let $S=\{\mathcal{O}^c(a\vee s|\top),\mathcal{O}^c(\neg a|\top)\}$, where $\neg a$ does not entail the negation of s. We have $S\models\Diamond(\neg a\wedge s)$, $S\models\mathcal{O}^c(\neg a\wedge s|\top)$ and $S\models\mathcal{O}_\exists(\neg a\wedge s|\top)$, $S\not\models\mathcal{O}^c(s|\top)$ and $S\models\mathcal{O}_\exists(s|\top)$. The crucial observation is that $\mathcal{O}_\exists(s|a)$ is not entailed by S. First of all, $\mathcal{O}_\exists(s|a)$ is not first phase entailed by S via $\mathcal{O}_\exists(s\mid\top)$, because $\mathcal{O}_\exists(s\mid\top)$ is not first phase entailed by S. Secondly, $\mathcal{O}_\exists(s\mid a)$ is not second phase entailed by S via $\mathcal{O}_\exists(s\mid\top)$ either, because in second phase entailment $\mathcal{O}_\exists$ does not have strengthening of the antecedent at all. Thirdly, it is not second phase entailed by S via a first phase derivation of $\mathcal{O}_\exists(\neg a\wedge s\mid a)$, because $\mathcal{O}^c(\neg a\wedge s|a)$ is not entailed by $\mathcal{O}^c(\neg a\wedge s|\top)$ due to the restriction in* RSA.

The derived obligations $O_\exists$ can represent moral dilemmas in a consistent way. If we want to make moral dilemmas inconsistent, we can consider the two-phase approach with $\mathcal{O}^c_\forall$ instead of $\mathcal{O}_\exists$. For the first phase, we define a new kind of obligation.

$$\mathcal{O}_D(\alpha|\beta) =_{def} \mathcal{O}^c(\alpha|\beta)\wedge\mathcal{O}^c_\forall(\alpha|\beta)$$

The two-phase approach with $\mathcal{O}_D(\alpha|\beta)$ and $\mathcal{O}^c_\forall(\alpha|\beta)$ works similar to the two-phase approach with $\mathcal{O}^c(\alpha|\beta)$ and $\mathcal{O}_\exists(\alpha|\beta)$, only we have to use preferential entailment $\models_\sqsubset$ instead of $\models$. Notice that $\mathcal{O}_D(\alpha\mid\beta)\rightarrow\mathcal{O}^c_\forall(\alpha\mid\beta)$ is a (trivial) theorem of $\mathcal{L}$. This theorem is the counterpart of Proposition 4 for this second example of the two-phase approach. In this case, we say that a derived obligation $\mathcal{O}^c_\forall(\alpha|\beta)$ is first phase derived if $\mathcal{O}_D(\alpha\mid\beta)$ is derivable, and second phase derived otherwise. The following example shows that this two-phase approach solves the SA+D problem of Example 6 and 9 by weakening RSA.

Example 14 (SA+D problem, continued) *First of all, the following theorem of the logic $\mathcal{L}$ is the counterpart of axiom D'' for $\mathcal{O}_D$. This theorem follows from the fact that $\mathcal{O}_D(\alpha|\beta)$ entails $\mathcal{O}^c_\forall(\alpha|\beta)$.*

$$\mathrm{D}''_D : \Box\neg(\alpha_1 \wedge \alpha_2 \wedge \beta) \rightarrow \neg(\mathcal{O}_D(\alpha_1|\beta) \wedge \mathcal{O}_D(\alpha_2|\beta))$$

Let $S = \{\mathcal{O}_D(p_1 \mid \top), \mathcal{O}_D(p_2 \mid \top)\}$, hence $S = \{\mathcal{O}^c(p_1 \mid \top) \wedge \mathcal{O}^c_\forall(p_1 \mid \top), \mathcal{O}^c(p_2 \mid \top) \wedge \mathcal{O}^c_\forall(p_2|\top)\}$. We have $S \models_\sqsubset \mathcal{O}^c_\forall(p_1 \vee p_2|\neg(p_1 \wedge p_2))$ and $S \not\models_\sqsubset \mathcal{O}^c_\forall(p_1|\neg(p_1 \wedge p_2))$. Hence, the obligation $\mathcal{O}^c_\forall(p_1|\top)$ is not strengthened to $\mathcal{O}^c_\forall(p_1|\neg(p_1 \wedge p_2))$, and it was observed in Example 6 that this is a solution for the SA+D *problem. Furthermore, we have $S \models_\sqsubset \mathcal{O}_\exists(p_2|\neg(p_1 \wedge p_2))$ and $S \models_\sqsubset \mathcal{O}_\exists(\neg p_2|\neg(p_1 \wedge p_2))$, which explains the absence of* RSA*: there is not a unique most preferred obligation for the antecedent $\neg(p_1 \wedge p_2)$.*

Finally, let S be the set of obligations $\{\mathcal{O}_D(\neg c \mid \top), \mathcal{O}_D(c \,|\, k)\}$ where $\neg c$ does not entail the negation of k. S is inconsistent, and we can derive an inconsistency as follows. The premise $\mathcal{O}_D(\neg c|\top)$ entails the obligation $\mathcal{O}^c(\neg c|\top)$, which entails $\mathcal{O}^c(\neg c\,|\,k)$. From Proposition 4 follows that the latter obligation entails $\mathcal{O}_\exists(\neg c\,|\,k)$. $\mathcal{O}_\exists(\neg c|k)$ entails $\mathcal{P}_\exists(\neg c|k)$, which is logically equivalent to $\neg\mathcal{O}_\forall(c|k)$. The premise $\mathcal{O}_D(c|k)$ entails $\mathcal{O}^c_\forall(c|k)$, which is inconsistent with $\neg\mathcal{O}_\forall(c|k)$.

4.3 Related approaches

Thus far, we have shown that the two-phase approach to deontic logic can solve the SA+WC problem of Example 1 and the SA+D problem of Example 6 and 9. However, we have said little about the intuition behind the distinction between the two phases. In this section, we discuss the diagnostic and decision theoretic perspectives on deontic logic that illustrate aspects of these two phases.

The aim of *diagnostic reasoning* (see e.g. [17]) is to derive a conjecture about faulty components, given a system (a set of components and a description of the behavior of these components) and facts (a description of the world). In diagnostic reasoning, two phases can be distinguished. The first phase is reasoning about *all* possible conjectures, i.e. reasoning about the possible sets of broken components. The second phase is reasoning about *minimal* sets of broken components, the so-called diagnoses. This second phase reasoning is based on the so-called principle of parsimony, which states that it is expected that the set of broken components is minimal. The *diagnostic perspective* on deontic logic is based on the diagnostic problem of determining the set of violated obligations, given a normative system and facts. Hence, the analogy between deontic and diagnostic reasoning is based on the fact that violated obligations are treated in a similar way as broken components. The two phases are clearly distinguished in the DIagnostic framework for DEontic reasoning DIODE [24, 22, 25]. In DIODE there is a distinction between (phase one) *constructing* and *zooming-in on* a preference ordering of deontic states and (phase two) *finding the optimal sub-ideal state*. By zooming-in we mean that the substructure of the ordering is considered that only contains states that satisfy the actual facts. The first phase of zooming-in ignores more ideal states, since they are not reachable anymore because of the violating facts. Moreover, the second phase of finding the optimal sub-ideal state ignores the less ideal states. In DIODE, the derived obligations are called *contextual obligations* to express that they are only valid in the context of the antecedent, because they do not refer to either more ideal or less ideal states.

The aim of a *(qualitative) decision theory* is to find the optimal goal, and how to reach this goal, given (qualitative) preferences and knowledge about the world (usually containing uncertain beliefs). In decision theoretic reasoning, also two phases can be distinguished. The first phase is reasoning about the preferences and *all* (reachable) goals. The second phase is optimal goal planning. The second phase is based on the additional assumption that the agent optimizes his behavior, and hence wants to realize only maximally preferred goals. The *decision theoretic perspective* on deontic logic considers ideal states as goals. For example, Pearl gives a qualitative, decision-theoretic account of the *pragmatic* obligation $\mathcal{O}(A \mid B)$: 'If you observe, believe, or know B, then the expected utility resulting from doing A is much higher than that of resulting from not doing A' [15]. It is one thing to determine the maximum utiliy, but quite another to act accordingly.

5 Further research

In this paper, we have investigated dyadic obligations in a modal preference logic. As further research, we will investigate the formalization of *defeasible* dyadic obligations, obligations which are subject to exceptions, in the modal framework introduced in this paper. Another well-known problem of dyadic deontic logic is the relation between facts and dyadic obligations. In [27] a notion of factual detachment is proposed based on the so-called retraction test. In further research we will investigate the formalization of this test in $\mathcal{L}^*$.

Acknowledgement

Thanks to Patrick van der Laag for several discussions on the issues raised in 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*, pages 43–84. John Wiley & Sons, 1994.

[2] M. Belzer. A logic of deliberation. In *Proceedings of the Fifth National Conference on Artificial Intelligence*, pages 38–43, 1986.

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

[4] 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)*, pages 75–86, 1994.

[5] A.L. Brown and S. Mantha. Preferences as normative knowledge: Towards declarative obligations. In *Proceedings of the First Workshop on Deontic Logic in Computer Science (Deon'91)*, pages 142–163, Amsterdam, 1991.

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

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

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

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

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

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

[12] H. Levesque. All I know: a study in autoepistemic logic. *Artificial Intelligence*, 42:263–309, 1990.

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

[14] D. Makinson. Five faces of minimality. *Studia Logica*, 52:339–379, 1993.

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

[16] H. Prakken and M.J. Sergot. Contrary-to-duty imperatives, defeasibility and violability. In *Proceedings of the Second Workshop on Deontic Logic in Computer Science (Deon'94)*, Oslo, 1994. To appear in: *Studia Logica.*

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

[18] N. Rescher. The logic of preference. In *Topics in Philosophical Logic.* D. Reidel Publishing Company, Dordrecht, Holland, 1967.

[19] A. Ross. Imperatives and logic. *Theoria*, 7:53–71, 1941.

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

[21] Y. Shoham. *Reasoning About Change*. MIT Press, 1988.

[22] Y.-H. Tan and L.W.N. van der Torre. DIODE: Deontic logic based on diagnosis from first principles. In *Proceedings of the Workshop 'Artificial normative reasoning' of the Eleventh European Conference on Artificial Intelligence (ECAI'94)*, Amsterdam, 1994.

[23] Y.-H. Tan and L.W.N. van der Torre. Multi preference semantics for a defeasible deontic logic. In *Proceedings of the JURIX'94*, Amsterdam, 1994.

[24] Y.-H. Tan and L.W.N. van der Torre. Representing deontic reasoning in a diagnostic framework. In *Proceedings of the Workshop on Legal Applications of Logic Programming of the Eleventh International Conference on Logic Programming (ICLP'94)*, Genoa, Italy, 1994.

[25] Y.-H. Tan and L.W.N. van der Torre. Why defeasible deontic logic needs a multi preference semantics. In *Proceedings of the ECSQARU'95. Lecture Notes in Artificial Intelligence 946*. Springer Verlag, 1995.

[26] L.W.N. van der Torre. Violated obligations in a defeasible deontic logic. In *Proceedings of the Eleventh European Conference on Artificial Intelligence (ECAI'94)*, pages 371–375. John Wiley & Sons, 1994.

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

[28] B.C. van Fraassen. The logic of conditional obligation. *Journal of Philosophical Logic*, 1:417–438, 1972.

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