Two-Phase Deontic Logic

Leendert van der Torre
Department of Artificial Intelligence
Vrije Universiteit Amsterdam
De Boelelaan 1081a
1081 HV Amsterdam
The Netherlands
torre@cs.vu.nl

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

(Received ..... ; Accepted in final form .....)

Abstract. We show that for the adequate representation of some examples of normative reasoning a combination of different operators is needed, where each operator validates different inference rules. The combination of different modal operators imposes the restriction on the proof theory of the logic that a proof rule can be blocked in a derivation due to the fact that another proof rule has been used earlier in the derivation. In this paper we only use two operators and therefore we call the restriction the two-phase approach in the proof theory, which we formalize in two-phase labeled deontic logic (2LDL) and in two-phase dyadic deontic logic (2DL). The preference-based semantics of 2DL is based on an explicit deontic preference ordering between worlds, representing different degrees of ideality. The two different modal operators represent two different usages of the preference ordering, called minimizing and ordering.

1. Why deontic logic derivations must consist of two phases

1.1. Van Fraassen's paradox

Van Fraassen (1973) presents a logical analysis of dilemmas. In a logic that formalizes reasoning about dilemmas we cannot accept the conjunction rule, because it derives $\bigcirc(p \wedge \neg p)$ from the dilemma $\bigcirc p \wedge \bigcirc \neg p$, whereas 'ought implies can' $\neg \bigcirc (p \wedge \neg p)$. On the other hand we do not want to reject the conjunction rule in all cases. For example, we want to derive $\bigcirc(p \wedge q)$ from $\bigcirc p \wedge \bigcirc q$ when p and q are distinct propositional atoms. That is, we have to add a restriction on the conjunction rule such that we only derive $\bigcirc(\alpha_1 \wedge \alpha_2)$ from $\bigcirc\alpha_1$ and $\bigcirc\alpha_2$ if $\alpha_1 \wedge \alpha_2$ is consistent. Van Fraassen calls the latter inference pattern *Consistent Aggregation*, which we write as the restricted conjunction rule (RAND). He encounters a problem in the formalization of obligations, and wonders if he needs a language in which he can talk directly about the imperatives as well. A variant of this problem is illustrated in the following example.

Example 1 (Van Fraassen's paradox). Assume a monadic deontic logic without nested modal operators[1] in which dilemmas like $\bigcirc p \wedge \bigcirc \neg p$ are consistent, but which validates $\neg \bigcirc \bot$, where $\bot$ stands for any contradiction like $p \wedge \neg p$. Moreover, assume that it satisfies replacement of logical equivalents and at least the following two inference patterns *Restricted Conjunction* rule (RAND), also called consistent aggregation, and *Weakening* (W), where $\overset{\leftrightarrow}{\Diamond}\phi$ can loosely be read as ϕ is possible (or propositionally consistent).

$$\text{RAND:}\frac{\bigcirc\alpha_1, \bigcirc\alpha_2, \overset{\leftrightarrow}{\Diamond}(\alpha_1 \wedge \alpha_2)}{\bigcirc(\alpha_1 \wedge \alpha_2)} \qquad \text{W:}\frac{\bigcirc\alpha_1}{\bigcirc(\alpha_1 \vee \alpha_2)}$$

Moreover, assume the two premises 'Honor thy father or thy mother!' $\bigcirc(f \vee m)$ and 'Honor not thy mother!' $\bigcirc\neg m$. The derivation of Figure 1 illustrates how the desired conclusion 'thou shalt honor thy father'

$$\cfrac{\cfrac{\bigcirc(f \vee m) \quad \bigcirc\neg m}{\bigcirc(f \wedge \neg m)}\,\text{RAND}}{\bigcirc f}\,\text{W}$$

Figure 1. Van Fraassen's paradox (1)

$\bigcirc f$ can be derived from the premises. Unfortunately, the derivation of Figure 2 illustrates that we cannot accept restricted conjunction and

$$\cfrac{\cfrac{\cfrac{\bigcirc p}{\bigcirc(f \vee p)}\,\text{W} \quad \bigcirc\neg p}{\bigcirc(f \wedge \neg p)}\,\text{RAND}}{\bigcirc f}\,\text{W}$$

Figure 2. Van Fraassen's paradox (2)

weakening in a monadic deontic logic, because we can derive the counterintuitive obligation $\bigcirc f$ from a deontic dilemma $\bigcirc p \wedge \bigcirc\neg p$. The point of this paradox is that *every* $\bigcirc(\beta)$, of which $\bigcirc(f)$ is a special case, would be derivable.

Van Fraassen asks himself whether restricted conjunction can be formalized, and he observes interesting technical questions. In this paper we pursue some of these technical questions.

> 'But can this happy circumstance be reflected in the logic of the ought-statements alone? Or can it be expressed only in a language in which we can talk directly about the imperatives as well? This is an important question, because it is the question whether the inferential structure of the 'ought' language game can be stated in so simple a manner that it can be grasped in and by itself. Intuitively, we want to say: there are simple cases, and in the simple cases the axiologist's logic is substantially correct even if it is not in general – but can we state precisely when we find ourselves in such a simple case? These are essentially technical questions for deontic logic, and I shall not pursue them here.' (van Fraassen, 1973)

As far as we know, there is no discussion on Van Fraassen's paradox in the deontic logic literature.[2] We analyze Van Fraassen's paradox in Example 1 by forbidding application of RAND after W has been applied. This blocks the counterintuitive derivation in Figure 2 and it does not block the intuitive derivation in Figure 1, as we show below. Our formalization of two-phase reasoning works as follows. In the logic, the two phases are represented by two different types of obligations, written as phase-1 obligations ① and phase-2 obligations ②. The premises are phase-1 obligations, the conclusions are phase-2 obligations and the two phases are linked to each other with the following inference pattern REL.

$$\text{REL} : \frac{①(\alpha)}{②(\alpha)}$$

The two-phase approach blocks the derivation of the obligation $\bigcirc f$ in Figure 1 by introducing sequencing of the derivations RAND and W, such that the former is only valid in phase-1 (i.e. for ①) and the latter only in phase-2 (for ②). First of all, Figure 3 illustrates that $②f$ is entailed by $①(f \vee m)$ and $①\neg m$.

$$\cfrac{\cfrac{\cfrac{①(f \vee m) \quad ①\neg m}{①(f \wedge \neg m)}\,\text{RAND}_1}{②(f \wedge \neg m)}\,\text{REL}}{②f}\,\text{W}_2$$

Figure 3. Analysis of Van Fraassen's paradox (1)

Second, the counterintuitive obligation $②f$ is not entailed from a dilemma $①p \wedge ①\neg p$. The blocked derivations are represented in Figure 4, where blocked derivation steps are represented by dashed lines. The counterintuitive obligation $②f$ is not entailed via the obligation $①(f \vee p)$, because in the first phase there is no weakening. Moreover, the obligation $\textcircled{2}f$ is not entailed via $\textcircled{2}(f \vee p)$ either, because in second-phase entailment $\textcircled{2}$ does not have restricted conjunction.

$$\dfrac{\dfrac{\dfrac{\dfrac{\textcircled{1}p}{\textcircled{1}(f \vee p)}\,(\mathrm{W}_2) \qquad \textcircled{1}\neg p}{\textcircled{1}(f \wedge \neg p)}\,\mathrm{RAND}_1}{\textcircled{2}(f \wedge \neg p)}\,\mathrm{REL}}{\textcircled{2}f}\,\mathrm{W}_2 \qquad\qquad \dfrac{\dfrac{\dfrac{\dfrac{\textcircled{1}p}{\textcircled{2}p}\,\mathrm{REL}}{\textcircled{2}(f \vee p)}\,\mathrm{W}_2 \quad \dfrac{\textcircled{1}\neg p}{\textcircled{2}\neg p}\,\mathrm{REL}}{\textcircled{2}(f \wedge \neg p)}\,(\mathrm{RAND}_1)}{\textcircled{2}f}\,\mathrm{W}_2$$

Figure 4. Analysis of Van Fraassen's paradox (2)

1.2. Contrary-to-duty paradoxes

The distinction between two phases can also be used to analyze the notorious contrary-to-duty (CTD) paradoxes in dyadic deontic logic. These paradoxes contain so-called contrary-to-duty obligations, which are obligations which are only in force if another obligation has been violated. Contrary-to-duty obligations refer to sub-ideal circumstances. We discuss one paradox in both Standard Deontic Logic (SDL)[3] and dyadic deontic logic, and we discuss another one only in dyadic deontic logic. In SDL, a conditional obligation $\beta \to \bigcirc\alpha$ is a contrary-to-duty (or secondary) obligation of the (primary) obligation $\bigcirc\alpha_1$ if and only if $\beta \wedge \alpha_1$ is inconsistent. The following example is the notorious gentle murderer paradox (Forrester, 1984), a strengthened version of the Good Samaritan paradox (Åqvist, 1967).

Example 2 (Forrester's paradox). Consider the following sentences of an SDL theory T: 'Smith should not kill Jones' $\bigcirc\neg k$, 'if Smith kills Jones, then he should do it gently' $k \to \bigcirc g$, 'Smith kills Jones' k, and 'killing someone gently logically implies killing him' $\vdash g \to k$. 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.

$$\models_{\text{SDL}} (\beta \wedge (\beta \to \bigcirc\alpha)) \to \bigcirc\alpha$$

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

In contrast to Van Fraassen's paradox in the previous section, Forrester's paradox raised an extensive discussion in the deontic logic literature. We first mention several consistent formalizations that have been proposed.

Scope. Scope distinctions, which have been proposed (see e.g. (Castañeda, 1981)) to solve the Good Samaritan paradox, seem to be absent from Forrester's paradox. However, Sinnot-Armstrong (1985) argues that also Forrester's paradox rests on scope confusions. He invokes Davidson's account of the logical form of action statements (Davidson, 1967), according to which adverbial modifiers like gently in the consequent of $k \rightarrow \bigcirc g$ 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 $\bigcirc\exists e(Me \wedge Ge)$ and narrow scope $\exists e(Me \wedge \bigcirc Ge)$. 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.'[4]

Weakening. Goble (1991) argues that Forrester's paradox is caused by weakening, following a suggestion of Forrester (1984, p.196). His consistent formalization is based on rejection of the property weakening.[5] In his logic, $\bigcirc\neg k \wedge \bigcirc k$ is inconsistent whereas $\bigcirc\neg k \wedge \bigcirc g$ is consistent.[6]

Defeasibility. Non-monotonic techniques can be used to consistently formalize the paradox (Ryu & Lee, 1993; McCarty, 1994; Nute & Yu, 1997). 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.[7] Non-monotonic techniques were already used by Loewer and Belzer (1983; 1986), who solve the Forrester paradox in their temporal deontic logic 'Dyadic Deontic Detachment' (3D).[8] Moreover, it is observed in (van der Torre & Tan, 2000) that approaches based on contextual reasoning (e.g. (Prakken & Sergot, 1996)) use non-monotonic techniques, when 'α ought to be (done) in context γ' is defined by 'α ought to be (done), unless $\neg\gamma$.'

Dyadic operators. Dyadic deontic logics were developed to formalize contrary-to-duty reasoning (Hansson, 1971; Lewis, 1974) and to analyze the Good Samaritan paradox, and they can also be used for the formalization of the Forrester paradox (van der Torre & Tan, 1999a; Prakken & Sergot, 1997). The two obligations are simply represented by the dyadic obligations $\bigcirc(\neg k|\top)$ and $\bigcirc(g|k)$. The second obligation is a CTD obligation of the first one, because in dyadic deontic logic an obligation $\bigcirc(\alpha|\beta)$ is a CTD obligation of the primary obligation $\bigcirc(\alpha_1|\beta_1)$ if and only if $\alpha_1 \wedge \beta$ is inconsistent, see Figure 5. However, the problem of this representation is what

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

Figure 5. $\bigcirc(\alpha|\beta)$ is a contrary-to-duty obligation of $\bigcirc(\alpha_1|\beta_1)$

properties the dyadic obligations have. For example, we obviously cannot accept the dyadic variant of factual detachment rule FD: $\bigcirc(\alpha \mid \beta) \wedge \beta \rightarrow \bigcirc\alpha$, or the paradox is immediately reinstated. From the two premises $\bigcirc(\neg k\,|\,\top)$ and $\bigcirc(g\,|\,k)$ and the fact k we can derive the contradictory $\bigcirc\neg k$ and $\bigcirc g$. Moreover, the logic cannot have strengthening of the antecedent and weakening of the consequent, as shown below.

The following formalization of Forrester's paradox in dyadic deontic logic illustrates how two-phase reasoning can be used to analyze it. The example illustrates that combining strengthening of the antecedent and weakening of the consequent is problematic. However, both properties are desirable for a dyadic deontic logic. For example, strengthening of the antecedent is used to derive 'Smith should not kill Jones in the morning' $\bigcirc(\neg k|m)$ from the obligation 'Smith should not kill Jones' $\bigcirc(\neg k\,|\,\top)$ and weakening of the consequent is used to derive 'Smith should not kill Jones' $\bigcirc(\neg k|\top)$ from the obligation 'Smith should drive on the right side of the street and not kill Jones' $\bigcirc(r \wedge \neg k|\top)$.

Example 3 (Forrester paradox, continued). Assume a dyadic deontic logic without nested modal operators that has at least substitution of logical equivalents and the following inference patterns *Strengthening of the Antecedent* (SA), the *Conjunction rule for the Consequent* (ANDC) and *Weakening of the Consequent* (WC) .

$$\text{SA}: \frac{\bigcirc(\alpha|\beta_1)}{\bigcirc(\alpha|\beta_1 \wedge \beta_2)} \qquad \text{ANDC}: \frac{\bigcirc(\alpha_1|\beta), \bigcirc(\alpha_2|\beta)}{\bigcirc(\alpha_1 \wedge \alpha_2|\beta)} \qquad \text{WC}: \frac{\bigcirc(\alpha_1|\beta)}{\bigcirc(\alpha_1 \vee \alpha_2|\beta)}$$

Furthermore, assume the following premise set with background knowledge $\vdash g \rightarrow k$.

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

The set S represents the Forrester paradox when k is read as 'Smith kills Jones' and g as 'Smith kills Jones gently.' Figure 6 below illustrates how the counterintuitive obligation $\bigcirc(\neg k \wedge g|k)$, i.e. $\bigcirc(\bot|k)$, can be derived from S by SA and ANDC. The derivation is blocked when SA is

$$\dfrac{\dfrac{\bigcirc(\neg k|\top)}{\bigcirc(\neg k|k)}\,\text{SA} \qquad \bigcirc(g|k)}{\bigcirc(\neg k \wedge g|k)}\,\text{ANDC}$$

Figure 6. Forrester's paradox (1)

replaced by the following inference pattern *Restricted Strengthening of the Antecedent* (RSA).

$$\text{RSA}: \dfrac{\bigcirc(\alpha|\beta_1), \overset{\leftrightarrow}{\Diamond}(\alpha \wedge \beta_1 \wedge \beta_2)}{\bigcirc(\alpha|\beta_1 \wedge \beta_2)}$$

Unfortunately, the counterintuitive obligation $\bigcirc(\bot \mid k)$ can still be derived from S by WC, RSA and ANDC. This paradoxical derivation from the set of obligations is represented in Figure 7. Moreover, in many

$$\dfrac{\dfrac{\dfrac{\bigcirc(\neg k|\top)}{\bigcirc(\neg g|\top)}\,\text{WC}}{\bigcirc(\neg g|k)}\,\text{RSA} \qquad \bigcirc(g|k)}{\bigcirc(\neg g \wedge g|k)}\,\text{ANDC}$$

Figure 7. Forrester's paradox (2)

dyadic deontic logics the obligation $\bigcirc(\bot \mid k)$ is inconsistent because 'ought implies can' $\neg\bigcirc(\bot|\alpha)$, whereas the premise set S is intuitively consistent.

Forrester's paradox in Example 3 shows that combining strengthening of the antecedent and weakening of the consequent is problematic for *any* deontic logic. The underlying problem of the counterintuitive derivation in Figure 7 is the derivation of $\bigcirc(\neg g \mid k)$ from the first premise $\bigcirc(\neg k \mid \top)$ by WC and RSA, because it derives a contrary-to-duty obligation from its own primary obligation. Note that the fulfillments of the two obligations are respectively $\neg k$ and $\neg g \wedge k$. Hence, the derived obligation cannot be fulfilled together with the premise it is derived from, which is counterintuitive. The two-phase approach blocks the derivation of the obligation $\bigcirc(\bot \mid k)$ in Figure 6 and 7 by introducing sequencing of the derivations RSA and WC, such that the former is only valid in phase-1 and the latter is only valid in phase-2. In dyadic deontic logic the two phases are linked to each other with the following inference pattern REL.

$$\text{REL} : \frac{\textcircled{1}(\alpha|\beta)}{\textcircled{2}(\alpha|\beta)}$$

The blocked derivations are represented in Figure 8. First of all, the obligation $\textcircled{1}(\neg k|k)$ is not entailed by $\textcircled{1}(\neg k|\top)$ due to the restriction in RSA. Secondly, $\textcircled{2}(\neg g|k)$ is not entailed via the obligation $\textcircled{1}(\neg g|\top)$, because in the first phase there is no weakening of the consequent. Finally, the obligation $\textcircled{2}(\neg g\,|\,k)$ is not entailed via $\textcircled{2}(\neg g\,|\,\top)$ either, because in second-phase entailment $\textcircled{2}$ does not have strengthening of the antecedent.

$$\begin{array}{c} \textcircled{1}(\neg k|\top) \\ \hdashline \textcircled{1}(\neg k|k) \end{array}\ (\text{RSA}_1) \qquad \begin{array}{cl} \textcircled{1}(\neg k|\top) & \\ \hdashline \textcircled{1}(\neg g|\top) & (\text{WC}_2) \\ \hline \textcircled{1}(\neg g|k) & \text{RSA}_1 \\ \hline \textcircled{2}(\neg g|k) & \text{REL} \end{array} \qquad \begin{array}{cl} \textcircled{1}(\neg k|\top) & \\ \hline \textcircled{2}\neg k|\top) & \text{REL} \\ \hline \textcircled{2}(\neg g|\top) & \text{WC}_2 \\ \hdashline \textcircled{2}(\neg g|k) & (\text{RSA}_1) \end{array}$$

Figure 8. Analysis of Forrester's paradox

The second CTD paradox we consider is Chisholm's paradox (Chisholm, 1963). 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 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. Monadic modal logic was extended with additional semantic features, such as time and actions.

Time. *Variants* of Chisholm's paradox have been formalized in temporal deontic logic (van Eck, 1982; Loewer & Belzer, 1983), which usually assume a temporal lag between antecedent and consequent. However, additional machinery has to be introduced to represent the paradox itself (van der Torre & Tan, 1998).[9]

Action. A related formalization distinguishes two (propositional) base languages, one for the antecedent and one for the consequent (Meyer, 1988; Alchourrón, 1993), following Castañeda's distinction between assertions and actions (Castañeda, 1981).[10]

Moreover, the formalizations we mentioned already at the discussion of Forrester's paradox were also proposed for Chisholm's paradox. In this paper we only consider the paradox in dyadic deontic logic, see e.g. (Tomberlin, 1981) for a discussion. Example 4 illustrates that the unrestricted combination of strengthening of the antecedent and weakening of the consequent again causes problems. Chisholm's paradox is more complicated than Forrester's paradox, because it also contains an *According-To-Duty* (ATD) obligation. Figure 9 illustrates that a conditional obligation $\bigcirc(\alpha|\beta)$ is an ATD obligation of $\bigcirc(\alpha_1|\beta_1)$ if and only if β logically implies α_1. The condition of an ATD obligation is satis-

$$\begin{array}{c} \bigcirc(\alpha_1|\beta_1) \\ \text{implies} \uparrow \\ \bigcirc(\alpha|\beta) \end{array}$$

Figure 9. $\bigcirc(\alpha|\beta)$ is an according-to-duty obligation of $\bigcirc(\alpha_1|\beta_1)$

fied only if the primary obligation is fulfilled. The definition of ATD is analogous to the definition of CTD 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.

It is well known (see e.g. (Prakken & Sergot, 1996; van der Torre & Tan, 1998)) that the main problem of Chisholm's paradox is caused by deontic detachment, or deontic transitivity, formalized by the following inference pattern DD^0.

$$\mathrm{DD}^0 : \frac{\bigcirc(\alpha|\beta), \bigcirc(\beta|\gamma)}{\bigcirc(\alpha|\gamma)}$$

We split the inference pattern in the three derivation steps SA, DD and WC.

$$\mathrm{SA} : \frac{\bigcirc(\alpha|\beta)}{\bigcirc(\alpha|\beta \land \gamma)} \qquad \mathrm{DD} : \frac{\bigcirc(\alpha|\beta \land \gamma), \bigcirc(\beta|\gamma)}{\bigcirc(\alpha \land \beta|\gamma)} \qquad \mathrm{WC} : \frac{\bigcirc(\alpha \land \beta|\gamma)}{\bigcirc(\alpha|\gamma)}$$

Notice that ANDC can be derived from SA and DD as follows. RANDC can be derived analogously from RSA and DD.

$$\dfrac{\dfrac{\bigcirc(\alpha_1|\beta)}{\bigcirc(\alpha_1|\beta \land \alpha_2)}\,\mathrm{SA} \qquad \bigcirc(\alpha_2|\beta)}{\bigcirc(\alpha_1 \land \alpha_2|\beta)}\,\mathrm{DD}$$

Now we have split deontic detachment in three inference steps, we can use our two-phase technique to block the counterintuitive derivation of Chisholm's paradox in the following example.

Example 4 (Chisholm's Paradox). Assume a dyadic deontic logic that validates at least substitution of logical equivalents and the (intuitively valid) inference patterns RSA, ANDC, WC and DD. Furthermore, consider the following premise set S.

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

The set S formalizes Chisholm's paradox (Chisholm, 1963) when a is read as 'a certain man goes to the assistance of his neighbors' and t as 'the man tells his neighbors that he will come.' The second obligation is an ATD obligation and the third obligation is a CTD obligation of the first obligation, see Figure 10. Figure 11 illustrates how the counterintuitive

$$\begin{tikzcd} \bigcirc(a|\top) \\ \bigcirc(t|a) \arrow[u, "\text{implies}"] \end{tikzcd} \qquad \begin{tikzcd} \bigcirc(a|\top) \arrow[d, leftrightarrow, "\text{inconsistent}"'] \\ \bigcirc(\neg t|\neg a) \end{tikzcd}$$

Figure 10. $\bigcirc(t|a)$ is an ATD of $\bigcirc(a|\top)$ and $\bigcirc(\neg t|\neg a)$ is a CTD of $\bigcirc(a|\top)$

$\bigcirc(\bot|\neg a)$ can be derived from S.

$$\dfrac{\dfrac{\dfrac{\dfrac{\bigcirc(t|a) \quad \bigcirc(a|\top)}{\bigcirc(a\wedge t|\top)}\,\text{DD}}{\bigcirc(t|\top)}\,\text{WC}}{\bigcirc(t|\neg a)}\,\text{RSA} \qquad \bigcirc(\neg t|\neg a)}{\bigcirc(t\wedge\neg t|\neg a)}\,\text{AND}$$

Figure 11. Chisholm's paradox

The blocked derivation in Figure 12 shows how the two-phase approach analyzes Chisholm's paradox. The problematic proof rule DD^0 is split into SA, DD and WC, where SA and DD are phase-1 rules and WC is a phase-2 rule. Hence, we can first apply DD and then WC, but not vice versa. Moreover, if we have applied DD and WC, then we can no longer use other phase-1 rules like SA. This blocks the counterintuitive derivation of $\textcircled{2}(t|\neg a)$ from $\textcircled{2}(t|\top)$.

When we compare the two derivations of the contrary-to-duty paradoxes in dyadic deontic logic, we find the following similarity. The underlying problem of the counterintuitive derivations is the derivation of the obligation $\bigcirc(\alpha_1|\neg\alpha_2)$ from $\bigcirc(\alpha_1\wedge\alpha_2|\top)$ by WC and RSA. It is respectively the derivation of $\bigcirc(\neg g\mid k)$ from $\bigcirc(\neg k\mid\top)$ in Figure 7 and $\bigcirc(t\mid\neg a)$ from $\bigcirc(a\wedge t\mid\top)$ in Figure 11. Moreover, similar

$$\dfrac{\dfrac{\dfrac{\dfrac{①(t|a) \quad ①(a|\top)}{①(a \wedge t|\top)}\text{DD}_1}{②(a \wedge t|\top)}\text{REL}}{②(t|\top)}\text{WC}_2}{②(t|\neg a)}\ (\text{RSA}_1)$$

Figure 12. Analysis of Chisholm's paradox

derivations of $\bigcirc(\neg(r \wedge g) \,|\, r)$ from $\bigcirc(\neg r \wedge \neg g \,|\, \top)$ and $\bigcirc(p \,|\, a)$ from $\bigcirc(\neg a \wedge p|\top)$ can be made from the following two sets of premises.

$$S = \{\bigcirc(\neg r \wedge \neg g|\top), \bigcirc(r \wedge g|r), \bigcirc(r \wedge g|g)\}$$

$$S' = \{\bigcirc(\neg a|\top), \bigcirc(a \vee p|\top), \bigcirc(\neg p|a)\}$$

The set S formalizes a variant of the Reykjavik Scenario (Belzer, 1986), when r is read as 'telling the secret to Reagan' and g as 'telling the secret to Gorbatchov,' see e.g. (van der Torre, 1994). S' formalizes an extension of the apples-and-pears example introduced in (Tan & van der Torre, 1996), when a is read as 'buying apples' and p as 'buying pears.'

The underlying problem of the contrary-to-duty paradoxes is that a contrary-to-duty obligation can be derived from its primary obligation. It is no surprise that this derivation causes paradoxes. The derivation of a secondary obligation from a primary obligation clearly confuses the different contexts found in contrary-to-duty reasoning.[11] The context of primary obligation is the ideal state, whereas the context of a contrary-to-duty obligation is a violation state. Preference-based deontic logics were developed to semantically distinguish the different violation contexts in a preference ordering. In this paper we show that in the preference-based two-phase framework 2DL it is not possible to derive secondary obligations from primary obligations. However, first we discuss a third phenomena – besides dilemmas and contrary-to-duty reasoning – which can be analyzed with a two-phase approach.

1.3. Reasoning by cases

Reasoning by cases is a desirable property of reasoning with conditionals. In this reasoning scheme, a certain fact is proven by proving it for a set of mutually exclusive and exhaustive circumstances. For example, assume that you want to know whether you want to go to the beach. If you desire to go to the beach when it rains, and you desire to go to the beach when it does not rain, then you may conclude by this scheme 'reasoning by cases' that you desire to go to the beach under all circumstances. The two cases considered here are rain and no rain. This kind of reasoning schemes can be formalized by the following derivation: *If* 'α if β' *and* 'α if not β,' *then* 'α regardless of β.' Formally, if we write the conditional 'α if β' by $\beta > \alpha$, then it is represented by the following disjunction rule for the antecedent.

$$\text{ORA:}\frac{\beta > \alpha, \neg\beta > \alpha}{\top > \alpha}$$

The following example illustrates that the disjunction rule for the antecedent combined with strengthening of the antecedent derives counterintuitive consequences in dyadic deontic logic.[12]

Example 5 (Disarmament paradox). Assume a dyadic deontic logic that validates at least substitution of logical equivalents and the two inference patterns RSA and the *Disjunction rule for the Antecedent* (ORA),

$$\text{ORA} : \frac{\bigcirc(\alpha|\beta_1), \bigcirc(\alpha|\beta_2)}{\bigcirc(\alpha|\beta_1 \vee \beta_2)}$$

and assume as premises the obligations 'we ought to be disarmed if there will be a nuclear war' $\bigcirc(d|w)$, 'we ought to be disarmed if there will be no war' $\bigcirc(d|\neg w)$, and 'we ought to be armed if we have peace if and only if we are armed' $\bigcirc(\neg d|d \leftrightarrow w)$. The derivation in Figure 13 shows how we can derive the counterintuitive $\bigcirc(d \wedge \neg d|d \leftrightarrow w)$. The derived

$$\dfrac{\dfrac{\dfrac{\bigcirc(d|w) \quad \bigcirc(d|\neg w)}{\bigcirc(d|\top)}\,\text{ORA}}{\bigcirc(d|d \leftrightarrow w)}\,\text{RSA} \qquad \bigcirc(\neg d|d \leftrightarrow w)}{\bigcirc(d \wedge \neg d|d \leftrightarrow w)}\,\text{AND}$$

Figure 13. The disarmament paradox

obligation is inconsistent in most deontic logics, whereas intuitively the set of premises is consistent. The derivation of $\bigcirc(d\,|\,d \leftrightarrow w)$ is counterintuitive, because it is not possible to fulfill this obligation together with the obligation $\bigcirc(d \mid \neg w)$ it is derived from. The contradictory fulfillments are respectively $d \wedge w$ and $d \wedge \neg w$.[13]

Reasoning by cases has not been discussed in deontic logic, but it has received some attention in conditional logic and default logic. However, as far as we know the problem above has not been discussed before.

The blocked derivations in Figure 14 show how the two-phase approach analyzes the paradox. We can first apply RSA and then ORA, but not vice versa.

$$\dfrac{\dfrac{①(d|w)}{②(d|w)}\text{REL} \quad \dfrac{①(d|\neg w)}{②(d|\neg w)}\text{REL}}{\dfrac{②(d|\top)}{②(d|d \leftrightarrow w)}\,(\text{RSA}_1)}\text{ORA}_2$$

Figure 14. Analysis of the disarmament paradox

Table I below summarizes the distinctions we made in the four examples above. If strengthening of the antecedent and the conjunction rule for the consequent are formalized as properties of phase-1 obligations, and weakening of the consequent and the disjunction rule for the antecedent are formalized as properties of phase-2 obligations, then all counterintuitive derivations discussed so far are blocked.

Table I. Inference patterns

Phase-1	Phase-2
Strengthening of the antecedent SA: $\dfrac{\bigcirc(\alpha\|\beta_1)}{\bigcirc(\alpha\|\beta_1 \wedge \beta_2)}$	Weakening of the consequent WC: $\dfrac{\bigcirc(\alpha_1\|\beta)}{\bigcirc(\alpha_1 \vee \alpha_2\|\beta)}$
Conjunction rule for the consequent ANDC: $\dfrac{\bigcirc(\alpha_1\|\beta), \bigcirc(\alpha_2\|\beta)}{\bigcirc(\alpha_1 \wedge \alpha_2\|\beta)}$	Disjunction rule for the antecedent ORA: $\dfrac{\bigcirc(\alpha\|\beta_1), \bigcirc(\alpha\|\beta_2)}{(\bigcirc(\alpha\|\beta_1 \vee \beta_2)}$

SDL has all the properties shown in the table, which suggests that all inference patterns are intuitive and in principle an ideal deontic logic has to validate all of them. However, it is well-known that SDL has only one phase and that it has many paradoxes. The standard approach to formalize the paradoxes is to argue that SDL is too strong. Hence, in dyadic deontic logic not all of SDL's inference patterns like SA and WC, ORA and ANDC are accepted, but some of them are rejected. However, we think that SDL has so many paradoxes *because* it only has one phase. In this paper we show that the inference patterns can all be accepted if the distinction between two phases is introduced.

In this paper we consider two two-phase deontic logics. First we discuss phased labeled deontic logic (PLDL). This logic illustrates that two phases are necessary to ensure that it is always possible to fulfill an obligation together with the obligations it is derived from. Moreover, the preference-based two-phase deontic logic (2DL) shows that the two phases correspond to two different uses of a deontic preference ordering. The logic 2DL has a possible worlds semantics that, in contrast to PLDL, can also represent disjunctions and negations of obligations, as well as facts and therefore violations.

2. Phased labeled deontic logic

In this section we introduce phased labeled deontic logic (PLDL). We only use logics in which dilemmas like $\bigcirc p \wedge \bigcirc \neg p$ are consistent, because Van Fraassen's paradox can only be analyzed in such logics. However, the PLDL-analyses of the contrary-to-duty paradoxes and the disarmament paradox is completely analogous to the analyses in a two-phase logic in which dilemmas are inconsistent.[14] Drawbacks of PLDL are that it does not have a (possible worlds) semantics, which has been very useful in the development of deontic logic, and that its language does not contain facts (and it therefore cannot represent violations) and negations and disjunctions of obligations.

Phased labeled deontic logics are versions of a labeled deductive system as it was introduced by Gabbay (1996) and extensions of the labeled deontic logics introduced in (van der Torre & Tan, 1995; van der Torre & Tan, 1997; Makinson, 1999). Labels are used to impose restrictions on the proof theory of the logic. In PLDL, a proof rule can be blocked in a derivation due to the fact that another proof rule has been used earlier in the derivation. We call a set of proof rules that may be used simultaneously a phase in the proof theory. Roughly speaking, the label L of an obligation $\bigcirc(\alpha|\beta)_L$ consists of a record of the fulfillments (F) of the premises that are used in the derivation of $\bigcirc(\alpha|\beta)$, and the phase (p) in which it is derived. Where there is no application of reasoning by cases, F can be taken to be a set of boolean formulas, that grows by joining sets as premises are combined. But in general, to cover the parallel tracks created through reasoning by cases, we need to consider sets of sets of boolean formulas (Makinson, 1999).

Definition 1 (Language). Let $\mathcal{L}$ be a propositional base logic. The language of PLDL consists of the labeled dyadic obligations $\bigcirc(\alpha|\beta)_L$, with α and β sentences of $\mathcal{L}$, and L a pair (F, p) that consists of a set of sets of sentences of $\mathcal{L}$ (fulfillments) and an integer (the phase). We write $\models$ for entailment in $\mathcal{L}$.

Each formula occurring as a premise has a label that consists of its own fulfillment and phase 0.

Definition 2 (Premise). A formula $\bigcirc(\alpha \mid \beta)_{(\{\{\alpha\wedge\beta\}\},0)}$ is called a premise of PLDL when $\alpha \wedge \beta$ is consistent in $\mathcal{L}$.

The phase of an obligation is determined by the proof rule used to derive the obligation, and the set of fulfillments is the union (ORA) or the product (SA, DD) of the labels of the premises used in this inference rule, where the product is defined by $\{S_1, \ldots, S_n\} \times \{T_1, \ldots, T_m\} = \{S_1 \cup T_1, \ldots, S_1 \cup T_m, \ldots, S_n \cup T_m\}$. The labels are used to check that fulfillments are consistent and that the phase of reasoning is non-decreasing. The consistency check realizes a variant of the *Kantian principle* that '*ought implies can.*'

*Definition 3 (*PLDL*).* Let ρ be a phasing function that associates with each proof rule below an integer called its phase. The phased labeled deontic logic PLDL for ρ consists of the inference rules below, extended with the following two conditions $R = R_F + R_p$.

R_F: $\bigcirc(\alpha \mid \beta)_{(F,p)}$ may only be derived if each $F_i \in F$ is consistent: it must always be possible to fulfill a derived obligation and each of the obligations it is derived from, though not necessarily all of them at the same time.

R_p: $\bigcirc(\alpha \mid \beta)_{(F,p)}$ may only be derived if $p \geq p_i$ for all obligations $\bigcirc(\alpha_i|\beta_i)_{(F_i,p_i)}$ it is derived from.

The inference rules of PLDL are replacements by logical equivalents and the following four rules.

$$\mathrm{RSA}_L : \frac{\bigcirc(\alpha \mid \beta_1)_{(F,p)}, R}{\bigcirc(\alpha \mid \beta_1 \wedge \beta_2)_{(F\times\{\beta_2\},\rho(\mathrm{SA}))}}$$

$$\mathrm{RDD}_L : \frac{\bigcirc(\alpha|\beta \wedge \gamma)_{(F_1,p_1)}, \bigcirc(\beta|\gamma)_{(F_2,p_2)}, R}{\bigcirc(\alpha \wedge \beta \mid \gamma)_{(F_1\times F_2,\rho(\mathrm{TRANS}))}}$$

$$\mathrm{WC}_L : \frac{\bigcirc(\alpha_1 \mid \beta)_{(F,p)}, R}{\bigcirc(\alpha_1 \vee \alpha_2 \mid \beta)_{(F,\rho(\mathrm{WC}))}}$$

$$\mathrm{ORA}_L : \frac{\bigcirc(\alpha \mid \beta_1)_{(F_1,p_1)}, \bigcirc(\alpha \mid \beta_2)_{(F_2,p_2)}, R}{\bigcirc(\alpha \mid \beta_1 \vee \beta_2)_{(F_1\cup F_2,\rho(\mathrm{ORA}))}}$$

We say $\{\bigcirc(\alpha_i \mid \beta_i) \mid 1 \leq i \leq n\} \vdash_{\text{PLDL}} \bigcirc(\alpha \mid \beta)$ if there is a labeled obligation $\bigcirc(\alpha \mid \beta)_L$ that can be derived from the set of obligations $\{\bigcirc(\alpha_i|\beta_i)_{(\{\{\alpha_i\wedge\beta_i\}\},0)} \mid 1 \leq i \leq n\}$.

In this paper, we are interested in the following two phased labeled deontic logics.

*Definition 4 (*LDL*, 2*LDL*).* Two labeled deontic logics LDL and 2LDL (with $\vdash_{\text{LDL}}$ and $\vdash_{\text{2LDL}}$) are defined as follows.

- The logic LDL is the PLDL with the phasing function ρ defined by $\rho(\text{RSA}) = 1$, $\rho(\text{RDD}) = 1$, $\rho(\text{WC}) = 1$, $\rho(\text{ORA}) = 1$.
- The logic 2LDL is the PLDL with the phasing function ρ defined by $\rho(\text{RSA}) = 1$, $\rho(\text{RDD}) = 1$, $\rho(\text{WC}) = 2$, $\rho(\text{ORA}) = 2$.

Notice that the following phase-1 conjunction rule for the consequent

$$\text{RANDC}_L : \frac{\bigcirc(\alpha_1 \mid \beta)_{(F_1,p_1)}, \bigcirc(\alpha_2 \mid \beta)_{(F_2,p_2)}, R}{\bigcirc(\alpha_1 \wedge \alpha_2 \mid \beta)_{(F_1 \times \{\alpha_2\} \times F_2,1)}}$$

is implied by LDL and 2LDL, because we can first strengthen $\bigcirc(\alpha_1|\beta)$ to $\bigcirc(\alpha_1|\beta \wedge \alpha_2)$, and then apply RDD to derive $\bigcirc(\alpha_1 \wedge \alpha_2|\beta)$.

$$\dfrac{\dfrac{\bigcirc(\alpha_1|\beta)_{(F_1,p_1)}}{\bigcirc(\alpha_1|\beta \wedge \alpha_2)_{(F_1 \times \{\alpha_2\},1)}}\text{RSA}_L \quad \bigcirc(\alpha_2|\beta)_{(F_2,p_2)}}{\bigcirc(\alpha_1 \wedge \alpha_2|\beta)_{(F_1 \times \{\alpha_2\} \times F_2,1)}}\text{RDD}_L$$

The following example illustrates that the two-phase technique can be seen as a device to effect a complex inductive definition with several arguments. It discusses the PLDL-variant of the notorious SDL theorem $\bigcirc p \wedge \bigcirc q \leftrightarrow \bigcirc(p \wedge q)$, see e.g. (von Wright, 1971).

Example 6. Let $S_1 = \{\bigcirc(p|\top), \bigcirc(q|\top)\}$ and $S_2 = \{\bigcirc(p \wedge q|\top)\}$. We have $S_1 \vdash_{\text{2LDL}} \bigcirc(p \wedge q|\top)$, $S_2 \vdash_{\text{2LDL}} \bigcirc(p|\top)$ and $S_2 \vdash_{\text{2LDL}} \bigcirc(q|\top)$. Hence, S_1 derives the obligations from S_2 and vice versa, but S_1 and S_2 are not equivalent. The first set derives $\bigcirc(p|\neg q)$, whereas the latter does not.[15]

The following example illustrates the PLDL analysis of the disarmament paradox. For further examples see (van der Torre & Tan, 1995; van der Torre & Tan, 1997; Makinson, 1999; van der Torre, 1998a; van der Torre, 1998b).

Example 7 (Disarmament, continued). The derivation in Figure 15 illustrates why we have $\bigcirc(d|w), \bigcirc(d|\neg w) \not\vdash_{\text{2LDL}} \bigcirc(d|d \leftrightarrow w)$. It is not possible to fulfill $\bigcirc(d|\neg w)$ and $\bigcirc(d|d \leftrightarrow w)$ at the same time, and the latter can therefore not be derived from the former.

$$\dfrac{\bigcirc(d|w)_{(\{\{d\wedge w\}\},0)} \quad \bigcirc(d|\neg w)_{(\{\{d\wedge \neg w\}\},0)}}{\bigcirc(d|\top)_{(\{\{d\wedge w\},\{d\wedge\neg w\}\},2)}}\ \mathrm{ORA}_L$$
$$\text{- - - - - - - - - - - - - - - - - - -}\ (\mathrm{RSA}_L)$$
$$\bigcirc(d|d\leftrightarrow w)_{(\{\{d\wedge w\},\{d\wedge\neg w\}\},1)}$$

Figure 15. Analysis of the disarmament paradox

In (van der Torre, 1998b) the following Theorem 1 and 2 are proven. The first theorem shows that for each LDL derivation there is an equivalent 2LDL derivation. In other words, the two phases are already implicit in LDL due to the condition R_F and the construction of new sets of fulfillments F by the proof rules. Consequently, deontic logic derivations must consist of two phases to ensure that it is always possible to fulfill an obligation together with the obligations it is derived from. The derivation in Van Fraassen's paradox is blocked, because it is not possible to fulfill the two obligations of a dilemma. Moreover, in the contrary-to-duty paradoxes it is not possible to fulfill a primary obligation together with one of its secondary obligations. Finally, in the disarmament paradox it is not possible to fulfill the derived obligation together with one of the cases it is derived from. We start with a lemma.

Lemma 1. For each obligation $\bigcirc(\alpha|\beta)_{(F,p)}$ derived in PLDL we have for each $F_i \in F$ that $F_i \models \alpha \wedge \beta$.

Proof By induction on the structure of the proof tree. The property trivially holds for the premises, and it is easily seen that the proof rules retain the property.

Theorem 1 (Equivalence LDL *and* 2LDL*).* Let S be a set of conditional obligations. We have $S \vdash_{\mathrm{LDL}} \bigcirc(\alpha|\beta)$ if and only if $S \vdash_{\mathrm{2LDL}} \bigcirc(\alpha|\beta)$.

Proof (outline) It is shown in (van der Torre, 1998b) that we can take any LDL derivation and construct an equivalent 2LDL derivation, by iteratively replacing two subsequent steps in the wrong order by several steps in the right order. The six relevant replacements are given in Figure 16. From Lemma 1 follows that the replacements do not violate the consistency check R_F.[16]

The second theorem shows that in 2LDL we can replace the consistency check on the fulfillments F by a consistency check on antecedent and consequent. Consequently, the set F is superfluous in the label of

$$\dfrac{\dfrac{\bigcirc(\alpha_1|\beta_1)}{\bigcirc(\alpha_1\vee\alpha_2|\beta_1)}\,\mathrm{WC}}{\bigcirc(\alpha_1\vee\alpha_2|\beta_1\wedge\beta_2)}\,\mathrm{SA} \qquad \dfrac{\dfrac{\bigcirc(\alpha_1|\beta_1)}{\bigcirc(\alpha_1|\beta_1\wedge\beta_2)}\,\mathrm{SA}}{\bigcirc(\alpha_1\vee\alpha_2|\beta_1\wedge\beta_2)}\,\mathrm{WC}$$

$$\dfrac{\dfrac{\bigcirc(\alpha_1|\beta\wedge\gamma)}{\bigcirc(\alpha_1\vee\alpha_2|\beta\wedge\gamma)}\,\mathrm{WC} \quad \bigcirc(\beta|\gamma)}{\bigcirc((\alpha_1\vee\alpha_2)\wedge\beta|\gamma)}\,\mathrm{DD} \qquad \dfrac{\dfrac{\bigcirc(\alpha_1|\beta\wedge\gamma)\quad\bigcirc(\beta|\gamma)}{\bigcirc(\alpha_1\wedge\beta|\gamma)}\,\mathrm{DD}}{\bigcirc((\alpha_1\vee\alpha_2)\wedge\beta|\gamma)}\,\mathrm{WC}$$

$$\dfrac{\bigcirc(\alpha|(\beta_1\vee\beta_2)\wedge\gamma)\quad \dfrac{\bigcirc(\beta_1|\gamma)}{\bigcirc(\beta_1\vee\beta_2|\gamma)}\,\mathrm{WC}}{\bigcirc((\alpha\wedge(\beta_1\vee\beta_2)|\gamma)}\,\mathrm{DD} \qquad \dfrac{\dfrac{\dfrac{\bigcirc(\alpha|(\beta_1\vee\beta_2)\wedge\gamma)}{\bigcirc(\alpha|\beta_1\wedge\gamma)}\,\mathrm{SA} \quad \bigcirc(\beta_1|\gamma)}{\bigcirc(\alpha\wedge\beta_1|\gamma)}\,\mathrm{DD}}{\bigcirc((\alpha\wedge(\beta_1\vee\beta_2)|\gamma)}\,\mathrm{WC}$$

$$\dfrac{\dfrac{\bigcirc(\alpha|\beta_1)\quad\bigcirc(\alpha|\beta_2)}{\bigcirc(\alpha|\beta_1\vee\beta_2)}\,\mathrm{OR}}{\bigcirc(\alpha|(\beta_1\vee\beta_2)\wedge\beta_3)}\,\mathrm{SA} \qquad \dfrac{\dfrac{\bigcirc(\alpha|\beta_1)}{\bigcirc(\alpha|\beta_1\wedge\beta_3)}\,\mathrm{SA}\quad\dfrac{\bigcirc(\alpha|\beta_2)}{\bigcirc(\alpha|\beta_2\wedge\beta_3)}\,\mathrm{SA}}{\bigcirc(\alpha|(\beta_1\wedge\beta_3)\vee(\beta_2\wedge\beta_3))}\,\mathrm{OR}$$

$$\dfrac{\dfrac{\bigcirc(\alpha|\beta_1\wedge\gamma)\quad\bigcirc(\alpha|\beta_2\wedge\gamma)}{\bigcirc(\alpha|(\beta_1\vee\beta_2)\wedge\gamma)}\,\mathrm{OR}\quad\bigcirc(\beta_1\vee\beta_2|\gamma)}{\bigcirc(\alpha\wedge(\beta_1\vee\beta_2)|\gamma)}\,\mathrm{DD}$$

$$\dfrac{\dfrac{\bigcirc(\alpha|\beta_1\wedge\gamma)\quad\dfrac{\bigcirc(\beta_1\vee\beta_2|\gamma)}{\bigcirc(\beta_1\vee\beta_2|\gamma\wedge(\beta_1\vee\neg\beta_2))}\,\mathrm{SA}}{\bigcirc(\alpha\wedge(\beta_1\vee\beta_2)|\gamma\wedge(\beta_1\vee\neg\beta_2))}\,\mathrm{DD} \quad \dfrac{\bigcirc(\alpha|\beta_2\wedge\gamma)\quad\dfrac{\bigcirc(\beta_1\vee\beta_2|\gamma)}{\bigcirc(\beta_1\vee\beta_2|\gamma\wedge(\beta_2\vee\neg\beta_1))}\,\mathrm{SA}}{\bigcirc(\alpha\wedge(\beta_1\vee\beta_2)|\gamma\wedge(\beta_2\vee\neg\beta_1))}\,\mathrm{DD}}{\bigcirc(\alpha\wedge(\beta_1\vee\beta_2)|\gamma)}\,\mathrm{OR}$$

$$\dfrac{\bigcirc(\alpha|\beta\wedge(\gamma_1\vee\gamma_2))\quad\dfrac{\bigcirc(\beta|\gamma_1)\quad\bigcirc(\beta|\gamma_2)}{G(\beta|\gamma_1\vee\gamma_2)}\,\mathrm{OR}}{\bigcirc(\alpha\wedge\beta|\gamma_1\vee\gamma_2)}\,\mathrm{DD}$$

$$\dfrac{\dfrac{\dfrac{\bigcirc(\alpha|\beta\wedge(\gamma_1\vee\gamma_2))}{\bigcirc(\alpha|\beta\wedge\gamma_1)}\,\mathrm{SA}\quad\bigcirc(\beta|\gamma_1)}{\bigcirc(\alpha\wedge\beta|\gamma_1)}\,\mathrm{DD}\quad \dfrac{\dfrac{\bigcirc(\alpha|\beta\wedge(\gamma_1\vee\gamma_2))}{\bigcirc(\alpha|\beta\wedge\gamma_2)}\,\mathrm{SA}\quad\bigcirc(\beta|\gamma_2)}{\bigcirc(\alpha\wedge\beta|\gamma_2)}\,\mathrm{DD}}{\bigcirc(\alpha\wedge\beta|\gamma_1\vee\gamma_2)}\,\mathrm{OR}$$

Figure 16. Reversing the order

2LDL obligations. The theorem also explains why we restrict ourselves to a consistency check on the conjunction of the antecedent and consequent in the preference-based deontic logic 2DL developed later in this paper.

Theorem 2. Consider any potential derivation of 2LDL, satisfying the condition R_p but not necessarily R_F. Then the following four conditions are equivalent:

1. The derivation satisfies condition R_F throughout phase 1,

2. The derivation satisfies R_F everywhere,

3. Each consequent is consistent with its antecedent throughout phase 1,

4. Each consequent is consistent with its antecedent everywhere.

Proof (outline) Clearly (2) $\Rightarrow$ (1) and (4) $\Rightarrow$ (3). Through phase 1, for each formula the conjunction of consequent and antecedent is equivalent to the unique element of its label. Hence (1) $\Leftrightarrow$ (3). In phase 2 the rules preserve the consistency of consequent and antecedent, and that they also preserve the property that each element of the label is consistent. From this we have (3) $\Rightarrow$ (4) and (1) $\Rightarrow$ (2). Putting this together gives us (1) $\Leftrightarrow$ (2) $\Leftrightarrow$ (3) $\Leftrightarrow$ (4) and we are done.

Phasing has been studied in a more general setting in input/output logics (Makinson & van der Torre, 2000; Makinson & van der Torre, 2001). One of the drawbacks of labelled deontic logic is the lack of a semantics. In the following section we consider phasing in preference-based semantics.

3. Preference-based two-phase deontic logic 2DL

In this section we give a preference-based semantics for the two-phase deontic logic 2DL. A preference 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 the inference pattern strengthening of the antecedent and the conjunction rule for the consequent, and minimizing to the inference pattern weakening of the consequent and the disjunction rule for the antecedent. In the first phase the preference ordering is constructed, and in the second phase the ordering is used for minimization.

In preference-based deontic logics dyadic obligations are defined by $\bigcirc(\alpha|\beta) =_{\text{def}} \alpha \wedge \beta \succ \neg\alpha \wedge \beta$. An example of a preference-based deontic logic is the well-known Hansson-Lewis dyadic deontic logic (Hansson, 1971; Lewis, 1974), in which an obligation $\bigcirc(\alpha|\beta)$ is defined by (a variant of) 'the (deontically) preferred β worlds are α worlds' in a suitably defined preference-based logic, as discussed below. This is equivalent to 'the preferred $\beta \wedge \alpha$ worlds are preferred to the preferred $\beta \wedge \neg\alpha$ worlds.' It is easily checked that the definition of preference-based obligations implies the theorem $\bigcirc(\alpha|\beta \wedge \gamma) \leftrightarrow \bigcirc(\alpha \wedge \beta|\beta \wedge \gamma)$, regardless of the interpretation and properties of the preference operator $\succ$. The theorem is counterintuitive on first reading, and it has also been discussed in the deontic logic literature following the Hansson-Lewis logics. For example, Hansson (1971) argues that the theorem represents that circumstances are fixed. As a consequence of this theorem, the logic 2DL developed in this section is stronger than the logic 2LDL developed in the previous section.

3.1. Preference-based semantics

We first discuss how obligations can be defined in terms of deontic preferences. Preference-based deontic logics are deontic logics of which the semantics contains a deontic preference ordering (usually on worlds of a Kripke style possible worlds model). This preference ordering reflects different degrees of 'ideality': a world is deontically preferred to another world if it is, in some sense, more ideal than the other world. We can distinguish three different levels of preference-based deontic logics.

Ideality (deontic preference) ordering on worlds.
The semantics of a preference-based logic contains a preference ordering, representing different degrees of ideality. For example, consider possible worlds models $M = \langle W, \leq, V \rangle$ that consist of a set of worlds W, a binary accessibility relation $\leq$ on the worlds of W and a valuation function V for the atomic propositions relative to the worlds. The expression $w_1 \leq w_2$ expresses that world $w_1 \in W$ is at least as ideal as world $w_2 \in W$, or that world w_2 is not more ideal than world w_1. The two most discussed properties of preference relations are transitivity and connectedness.

- *Transitivity.* Probably the most popular relation is a partial pre-ordering or quasi-ordering, which has a reflexive (for all worlds w we have $w \leq w$) and transitive (for all worlds $w_1, w_2, w_3 \in W$ with $w_1 \leq w_2$ and $w_2 \leq w_3$ we have $w_1 \leq w_3$) accessibility relation. Transitivity is a necessary property to define 'most preferred' worlds.[17]
- *Connectedness.* A partial pre-ordering is totally connected if for all worlds w_1 and w_2 we have $w_1 \leq w_2$ or $w_2 \leq w_1$. With transitive totally connected orderings, we have that there is no world preferred to a certain world w if and only if world w is at least as preferred as all other worlds. It represents that there are no dilemmas (van Fraassen, 1973), because the agent can always choose between any two alternatives. For example, the dilemma $\bigcirc p \wedge \bigcirc \neg p$ is inconsistent.

An example of a preference-based semantics is an utilitarian semantics (Jennings, 1974; Pearl, 1993), in which a real number (its utility) is associated with each world. Connectedness is a property of the preference ordering associated with probability theory and utility theory (von Neumann & Morgenstern, 1944; Keeney & Raiffa, 1976). If forced, the rational agent can choose (based on probability and utility) between each two possibilities.

Deontic betterness relation on propositions. Different kinds of deontic betterness relations between propositions – sets of worlds – can be derived from the deontic preferences between worlds. We write $\alpha_1 \succ \alpha_2$ for 'α_1 is better than α_2.' There are many different ways to lift preferences between worlds to preferences between sets of worlds. We say that a betterness relation $\succ$ formalizes strong preferences when a preference $\alpha_1 \succ \alpha_2$ logically implies $\alpha'_1 \succ \alpha'_2$ when α'_1 logically implies α_1 and α'_2 logically implies α_2. We call them weak preferences otherwise.[18] In this paper, ordering obligations are defined by strong preferences, and minimizing obligations are defined by weak preferences. It is generally accepted that the deontic betterness relation, in contrast to the ideality relation, is *not* transitive, see e.g. (Goble, 1989; Goble, 1993).

Obligations defined in terms of the betterness relation. The deontic betterness relations between propositions are used to formalize different kinds of obligations: $\bigcirc\alpha$ is some kind of deontic preference of α over $\neg\alpha$, and $\bigcirc(\alpha\,|\,\beta)$ is some kind of deontic preference of $\alpha \wedge \beta$ over $\neg\alpha \wedge \beta$.

$$\bigcirc(\alpha|\beta) =_{\text{def}} \alpha \wedge \beta \succ \neg\alpha \wedge \beta$$

A crucial idea in the formalization of the two-phase deontic logic is that the two modal operators ① and ② represent two different ways in which the deontic betterness relation is defined in the underlying deontic preference relation on worlds. One way, which we call *ordering*, is to use the whole ordering to evaluate a formula. The other way, which we call *minimizing*, is to use the ordering to select the minimal elements that satisfy a formula. The crucial distinction between the two logics is that ordering obligations have different properties than minimizing obligations. Ordering obligations have the inference pattern strengthening of the antecedent and the conjunction rule for the consequent,[19] and minimizing have the inference pattern weakening of the consequent and the disjunction rule for the antecedent.[20] We loosely say that phase-1 obligations construct a deontic preference ordering and phase-2 obligations minimize in the constructed preference ordering.

In this paper we show that typical 2DL-models of the examples are given in Figure 17. This figure represents preference models $\langle W, \leq, V\rangle$ that should be read as follows. A circle represents a nonempty set of worlds, that satisfy the propositions written within them. An arrow from a circle to another one represent that the worlds represented by the first circle are strictly preferred to the worlds represented by the second circle. The transitive closure is left implicit. The first phase of 2DL constructs the orderings and the second phase of 2DL uses the ordering for minimization.

Van Fraassen's paradox. The set $S = \{①p, ①\neg p\}$ represents a dilemma, because the consequents of the obligations are contradictory. The model in Figure 17.a illustrates that the dilemma S corresponds semantically to incomparable p and $\neg p$ worlds. The ordering consists of two states. The optimal states consist of respectively p or $\neg p$ worlds, both associated with a violation of a premise.

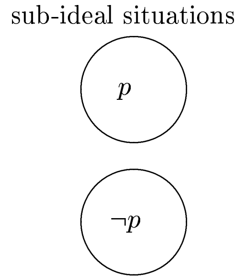
Forrester's paradox. The ordering of the model of the set of 2DL-formulas $S = \{①(\neg k \mid \top), ①(g \mid k), k, \vdash g \rightarrow k\}$ in Figure 17.b consists of three states. The ideal state consists of $\neg k$ worlds and represents that ideally Smith does not kill Jones. The sub-ideal states are ordered such that gentle killing is preferred to non-gentle killing.

Chisholm's paradox. The ordering of the model of the set of 2DL-formulas $S = \{①(a \mid \top), ①(t \mid a), ①(\neg t \mid \neg a), \neg a\}$ in Figure 17.c consists of four states. The ideal state consists of $a \wedge t$ worlds and represents that ideally the man goes to the assistance of his neighbors and he tells them that he will come. The sub-ideal states are ordered such that not going and not telling $\neg a \wedge \neg t$ is preferred to not going and telling $\neg a \wedge t$.

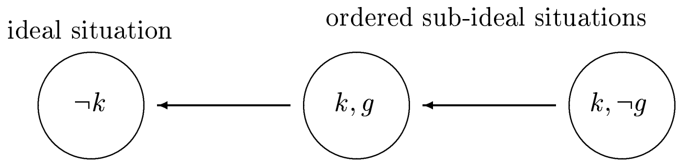
Disarmament paradox. The ordering of the model of the set of 2DL-formulas $S = \{①(d|w), ①(d|\neg w), ①(\neg d|d \leftrightarrow w)\}$ in Figure 17.d consists of four states. The ideal state consists of $d \wedge \neg w$ worlds and represents that ideally you are disarmed and there is no war. (Unfortunately, this is an unlikely state!) The three sub-ideal states are ordered such that not disarmed and no war is preferred to disarmed and war.

3.2. The two-phase deontic logic 2DL

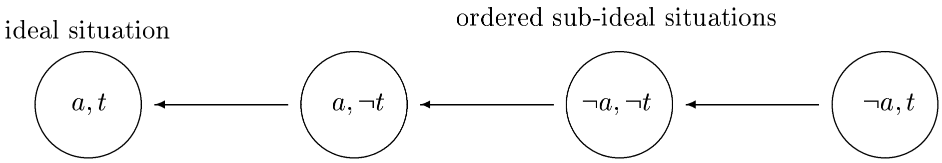
In this section we introduce the modal preference-based deontic logic 2DL. The binary accessibility relation of the Kripke models of modal



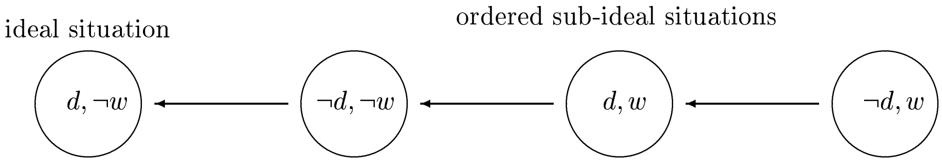
a. Van Fraassen's paradox



b. Forrester's paradox



c. Chisholm's paradox



d. the disarmament paradox

Figure 17. Preference-based models

logic is interpreted as a deontic preference relation. In (van der Torre & Tan, 1999a) we showed in Prohairetic Deontic Logic how ordering conditionals can be formalized in modal logic and in (Lamarre, 1991; Boutilier, 1994b) it has been shown how minimizing conditionals can be formalized in modal logic.[21]

3.2.1. *Modal preference logic*

Dyadic obligations are formalized in a bimodal logic, that contains an S5 operator $\overset{\leftrightarrow}{\Box}$, and an S4 operator $\Box$, where the relation between the two operators is axiomatized by $\overset{\leftrightarrow}{\Box}\,\alpha \to \Box\alpha$. 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 (Hughes & Creswell, 1984; Chellas, 1980). Moreover, S5 is characterized by an equivalence relation and additionally contains the axiom $\neg \overleftrightarrow{\Box}\alpha \rightarrow \overleftrightarrow{\Box}\neg\overleftrightarrow{\Box}\alpha$. The Kripke models $M = \langle W, \leq, V\rangle$ contain a binary accessibility relation $\leq$, that is interpreted as a reflexive and transitive *preference relation*.

*Definition 5 (*2DL*).* The bimodal language $\mathcal{L}$ is formed from a denumerable set of propositional variables together with the connectives $\neg$, $\rightarrow$, and the two normal modal connectives $\Box$ and $\overleftrightarrow{\Box}$. Dual 'possibility' connectives $\Diamond$ and $\overleftrightarrow{\Diamond}$ are defined as usual by $\Diamond\alpha =_{\text{def}} \neg\Box\neg\alpha$ and $\overleftrightarrow{\Diamond}\alpha =_{\text{def}} \neg\overleftrightarrow{\Box}\neg\alpha$.

The logic 2DL is the smallest $S \subset \mathcal{L}$ such that S contains classical logic and the following axiom schemata, and is closed under the following rules of inference.

K	$\Box(\alpha \rightarrow \beta) \rightarrow (\Box\alpha \rightarrow \Box\beta)$	**K′**	$\overleftrightarrow{\Box}(\alpha \rightarrow \beta) \rightarrow (\overleftrightarrow{\Box}\alpha \rightarrow \overleftrightarrow{\Box}\beta)$
T	$\Box\alpha \rightarrow \alpha$	**T′**	$\overleftrightarrow{\Box}\alpha \rightarrow \alpha$
4	$\Box\alpha \rightarrow \Box\Box\alpha$	**4′**	$\overleftrightarrow{\Box}\alpha \rightarrow \overleftrightarrow{\Box}\overleftrightarrow{\Box}\alpha$
R	$\overleftrightarrow{\Box}\alpha \rightarrow \Box\alpha$	**5′**	$\neg\overleftrightarrow{\Box}\alpha \rightarrow \overleftrightarrow{\Box}\neg\overleftrightarrow{\Box}\alpha$
Nes	From α infer $\overleftrightarrow{\Box}\alpha$		
MP	From $\alpha \rightarrow \beta$ and α infer β		

*Definition 6 (*2DL *Semantics).* Kripke models $M = \langle W, \leq, V\rangle$ for 2DL consist of W, a set of worlds, $\leq$, a binary transitive and reflexive accessibility relation, and V, a valuation of the propositional atoms in the worlds. The partial pre-ordering $\leq$ expresses preferences: $w_1 \leq w_2$ if and only if w_1 is at least as preferable as w_2. The modal connective $\Box$ refers to accessible worlds and the modal connective $\overleftrightarrow{\Box}$ to all worlds.

$$M, w \models \Box\alpha \text{ iff } \forall w' \in W \text{ if } w' \leq w, \text{ then } M, w' \models \alpha$$
$$M, w \models \overleftrightarrow{\Box}\alpha \text{ iff } \forall w' \in W \text{ we have } M, w' \models \alpha \qquad \text{(i.e. iff } M \models \alpha)$$

As a consequence of the definition in a standard bimodal logic, the soundness and completeness of 2DL are trivial.

Proposition 1 (Soundness and completeness of 2DL*).* Let $\vdash_{\text{2DL}}$ and $\models_{\text{2DL}}$ stand for derivability and logical entailment in the logic 2DL. We have $\vdash_{\text{2DL}} \alpha$ if and only if $\models_{\text{2DL}} \alpha$.

Proof. Follows directly from standard modal soundness and completeness proofs (Hughes & Creswell, 1984; Chellas, 1980; Fagin *et al.*, 1995).

The advantages of our formalization in a modal framework are twofold. First, when a dyadic operator is given by a definition in an underlying logic, then we get an axiomatization for free! We do not have to look for a sound and complete set of inference rules and axiom schemata, because we simply take the axiomatization of the underlying logic together with the new definition. In other words, the problem of finding a sound and complete axiomatization is translated into 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 relations between operators are expressed as theorems of the logic, which can thus be analyzed in the preference-based semantics. Finally, it facilitates the combining of the operators in the two-phase approach. We formalize ordering and minimizing in a two-phase deontic logic that consists of a weakened version of Prohairetic Deontic Logic (van der Torre & Tan, 1999a) and a weakened version of Hansson-Lewis logic (Hansson, 1971; Lewis, 1974). We call the modal preference logic with the definitions of the different types of conditionals the two-phase deontic logic 2DL. [22]

3.2.2. *Ordering: a weak variant of Prohairetic Deontic Logic*
In this subsection we only consider the ordering approach to deontic logic. In evaluating formulas, the whole ordering is taken into account. The ordering obligations are defined in two steps. First, we define a preference ordering on propositions. We write $\alpha_1 \succ \alpha_2$ for 'α_1 is deontically better than α_2,' and according to von Wright's expansion principle, a preference $\alpha_1 \succ_1 \alpha_2$ is only defined for $\alpha_1 \wedge \neg\alpha_2$ and $\neg\alpha_1 \wedge \alpha_2$. We have $\alpha_1 \succ_1 \alpha_2$ if and only if for each pair of $\alpha_1 \wedge \neg\alpha_2$ and $\alpha_2 \wedge \neg\alpha_1$ worlds we have either that the $\alpha_1 \wedge \neg\alpha_2$ world is preferred to the $\alpha_2 \wedge \neg\alpha_1$ world, or that the two are incomparable. In other words, for every $\alpha_1 \wedge \neg\alpha_2$ world there is not an $\alpha_2 \wedge \neg\alpha_1$ world that is as preferable.

We first show how a certain, predictable, definition for conditional obligation will not work. Assume for this purpose the following provitional notation $①^-(\alpha \mid \beta) =_{\text{def}} (\alpha \wedge \beta) \succ_1 (\neg\alpha \wedge \beta)$. The logic 2DL has the two counterintuitive theorems $①^-(\perp \mid \alpha)$ and $①^-(\alpha \mid \alpha)$. For this reason, we define two other types of obligations in the preference logic. In the following definition, $①(\alpha \mid \beta)$ has an additional condition that tests whether the obligation can be fulfilled, i.e. whether $\alpha \wedge \beta$ is logically possible ('ought implies can'). The obligation $①^c(\alpha \mid \beta)$ also has the additional condition that tests whether the obligation can be violated, i.e. whether $\neg\alpha \wedge \beta$ is logically possible.[23] The two conditions formalize von Wright's contingency principle (von Wright, 1981), and we therefore write 'c' in the obligation $\textcircled{1}^c$.

Definition 7 (Dyadic ordering obligation).
The dyadic ordering obligations 'α should be (done) if β is (done)', written as $\textcircled{1}(\alpha \mid \beta)$ and $\textcircled{1}^c(\alpha \mid \beta)$, are defined as strong preferences of $\alpha \wedge \beta$ over $\neg\alpha \wedge \beta$. A strong preference of α_1 over α_2, written as $\alpha_1 \succ_1 \alpha_2$, is defined as follows.

$$\begin{array}{lll}
\alpha_1 \succ_1 \alpha_2 & =_{\mathrm{def}} & \overleftrightarrow{\Box}(\alpha_1 \wedge \neg\alpha_2 \rightarrow \Box\neg(\alpha_2 \wedge \neg\alpha_1)) \\
\textcircled{1}(\alpha|\beta) & =_{\mathrm{def}} & (\alpha \wedge \beta) \succ_1 (\neg\alpha \wedge \beta) \wedge \overleftrightarrow{\Diamond}(\alpha \wedge \beta) \\
 & = & \overleftrightarrow{\Box}((\alpha \wedge \beta) \rightarrow \Box\neg(\neg\alpha \wedge \beta)) \wedge \overleftrightarrow{\Diamond}(\alpha \wedge \beta) \\
 & \leftrightarrow & \overleftrightarrow{\Box}((\alpha \wedge \beta) \rightarrow \Box(\beta \rightarrow \alpha)) \wedge \overleftrightarrow{\Diamond}(\alpha \wedge \beta) \\
\textcircled{1}^c(\alpha|\beta) & =_{\mathrm{def}} & (\alpha \wedge \beta) \succ_1 (\neg\alpha \wedge \beta) \wedge \overleftrightarrow{\Diamond}(\alpha \wedge \beta) \wedge \overleftrightarrow{\Diamond}(\neg\alpha \wedge \beta)
\end{array}$$

We conjecture that the logic can be axiomatized in terms of dyadic systems as follows.

*Definition 8 (*PDL$'$*).* The logic PDL$'$ is the smallest set closed under modus ponens and necessitation that satisfies the propositional theorems and the following axiom schemata specialize, generalize 1 and 2, and introducing exceptions:

$$\begin{array}{ll}
 & \alpha \succ_1 \bot \\
 & \bot \succ_1 \alpha \\
spec & \alpha_1 \succ_1 \alpha_2 \rightarrow (\alpha_1 \wedge \beta_1) \succ_1 (\alpha_2 \wedge \beta_2) \\
gen1 & (\alpha \succ \beta_1 \wedge \alpha \succ \beta_2) \rightarrow \alpha \succ_1 (\beta_1 \vee \beta_2) \\
gen2 & (\alpha_1 \succ_1 \beta) \wedge (\alpha_2 \succ_1 \beta) \rightarrow (\alpha_1 \vee \alpha_2) \succ_1 \beta
\end{array}$$

Which are equivalent to the following deontic formulas.

$$\begin{array}{ll}
 & \bigcirc(\alpha \vee \beta|\beta) \\
 & \bigcirc(\alpha \wedge \neg\beta|\beta) \\
 & \bigcirc(\alpha|\beta_1 \wedge \beta_2) \leftrightarrow \bigcirc(\alpha \wedge \beta_1|\beta_1 \wedge \beta_2) \\
spec & \bigcirc(\alpha|\beta_1) \rightarrow \bigcirc(\alpha|\beta_1 \wedge \beta_2) \\
gen1 & \bigcirc(\alpha|\alpha \vee \beta_1) \wedge \bigcirc(\alpha|\alpha \vee \beta_2) \rightarrow \bigcirc(\alpha|\alpha \vee \beta_1 \vee \beta_2) \\
gen2 & \bigcirc(\alpha_1|\alpha_1 \vee \beta) \wedge \bigcirc(\alpha_2|\alpha_2 \vee \beta) \rightarrow \bigcirc(\alpha_1 \vee \alpha_2|\alpha_1 \vee \alpha_2 \vee \beta)
\end{array}$$

The deontic betterness relation $\succ_1$ is quite weak. For example, it is not anti-symmetric (we cannot derive $\neg(\alpha_2 \succ_1 \alpha_1)$ from $\alpha_1 \succ_1 \alpha_2$) and it is not transitive (we cannot derive $\alpha_1 \succ_1 \alpha_3$ from $\alpha_1 \succ_1 \alpha_2$ and

$\alpha_2 \succ_1 \alpha_3$). The lack of these properties is the result of the fact that we do not have totally connected orderings. In this section we do not further discuss the properties of $\succ_1$, but we focus on the properties of the dyadic ordering obligations. Intuitively, an obligation $①(\alpha \mid \beta)$ expresses a strict deontic preference of all $\alpha \wedge \beta$ over $\neg\alpha \wedge \beta$. An $\alpha \wedge \beta$ world is preferred to an $\neg\alpha \wedge \beta$ world, or they are incomparable (in case of conflicting obligations). The following proposition shows that this preference is equivalent to the 'negative' condition that $\neg\alpha \wedge \beta$ worlds are not as preferable as $\alpha \wedge \beta$ worlds.[24]

Proposition 2. Let $M = \langle W, \leq, V\rangle$ be a 2DL model, $|\,\alpha\,|$ be the set of worlds of W that satisfy α, and $|\,\alpha_1\,| \not\leq |\,\alpha_2\,|$ denote that $\forall w_1 \in |\,\alpha_1\,|$ and $\forall w_2 \in |\,\alpha_2\,|$, we have $w_1 \not\leq w_2$. For a world $w \in W$, we have $M, w \models ①(\alpha|\beta)$ iff ($M \models ①(\alpha|\beta)$ iff) $|\neg\alpha \wedge \beta| \not\leq |\alpha \wedge \beta|$ and $|\alpha \wedge \beta|$ is nonempty.

Proof $\Rightarrow$ By contraposition. If $|\alpha \wedge \beta|$ is empty then the proof is trivial. Assume a model $M = \langle W, \leq, V\rangle$ with worlds $w_1, w_2 \in W$ such that $M, w_1 \models \neg\alpha \wedge \beta$, $M, w_2 \models \alpha \wedge \beta$ and $w_1 \leq w_2$. We have $M, w_2 \not\models (\alpha \wedge \beta) \to \Box(\beta \to \alpha))$. $M, w \models \overset{\leftrightarrow}{\Box} \alpha$ for a world $w \in W$ iff for *all* worlds $w' \in W$ we have $M, w' \models \alpha$. Hence, $M, w \not\models ①(\alpha|\beta)$.

$\Leftarrow$ By contraposition. Assume $M, w \not\models ①(\alpha|\beta)$ for some world w. Hence, there is no $\alpha \wedge \beta$ world (trivial) or there is a world $w_2 \in W$ such that $M, w_2 \not\models (\alpha \wedge \beta) \to \Box(\beta \to \alpha))$. In the latter case, it follows that $M, w_2 \models \alpha \wedge \beta$ and $M, w_2 \not\models \Box(\beta \to \alpha)$. Hence, there is a world $w_1 \in W$ such that $M, w_1 \models \neg\alpha \wedge \beta$ and $w_1 \leq w_2$.

The following proposition shows that ordering obligations can be used as phase-1 obligations, because they validate variants of SA and DD, but they do not validate WC and ORA.[25]

Proposition 3. The logic 2DL has the following theorems.

$$\mathbf{RSA}_1\text{:} \quad (①(\alpha|\beta_1) \wedge \overset{\leftrightarrow}{\Diamond}(\alpha \wedge \beta_1 \wedge \beta_2)) \to ①(\alpha|\beta_1 \wedge \beta_2)$$
$$\mathbf{DD}_1\text{:} \quad (①(\alpha|\beta \wedge \gamma) \wedge ①(\beta|\gamma) \to ①(\alpha \wedge \beta|\gamma)$$

The logic 2DL does *not* have the following theorems.

$$\mathbf{WC}_1\text{:} \quad ①(\alpha_1|\beta) \to ①(\alpha_1 \vee \alpha_2|\beta)$$
$$\mathbf{ORA}_1\text{:} \quad (①(\alpha|\beta_1) \wedge ①(\alpha|\beta_2)) \to ①(\alpha|\beta_1 \vee \beta_2)$$

Proof The (non)theorems can be proven by proving (un)satisfiability in the preference-based semantics. First, consider the validity of strengthening of the antecedent $\mathbf{RSA}_1$. The validity of strengthening $①(\alpha|\beta_1)$

to ①$(\alpha|\beta_1 \wedge \beta_2)$ follows directly from the fact that a strong preference of $\alpha \wedge \beta_1$ over $\neg\alpha \wedge \beta_1$ implies a strong preference of $\alpha \wedge \beta_1 \wedge \beta_2$ over $\neg\alpha \wedge \beta_1 \wedge \beta_2$. Secondly, consider the non-theorem $\mathbf{WC}_1$. ①$(\alpha_1 \,|\, \beta)$ is not weakened to ①$(\alpha_1 \vee \alpha_2 \,|\, \beta)$, because ①$(\alpha_1 \,|\, \beta)$ expresses a preference of *every* $\alpha_1 \wedge \beta$ worlds over any $\neg\alpha_1 \wedge \beta$ world, and from such a preference does not follow that every $(\alpha_1 \vee \alpha_2) \wedge \beta$ world is preferred to any $\neg\alpha_1 \wedge \neg\alpha_2 \wedge \beta$ world. For a counterexample, consider the preference-based model M in Figure 17.c. We have $M \models$ ①$(a|\top)$ and $M \not\models$ ①$(a \vee t \,|\, \top)$, because $|\neg a \wedge \neg t| \leq |\neg a \wedge t|$. Hence, the ordering obligations do not have weakening of the consequent. Verification of the other (non)theorems is left to the reader. Alternatively, the theorems can be proven in the logic 2DL.

In this section we introduced a logic of ordering obligations. It is a weak variant of Prohairetic Deontic Logic, because dilemmas like ①$(p \,|\, \top) \wedge$ ①$(\neg p \,|\, \top)$ are consistent, whereas they are inconsistent in Prohairetic Deontic Logic. In the next section we discuss 'standard' minimizing obligations and we compare them with the ordering obligations.

3.2.3. *Minimizing: a weak variant of Hansson-Lewis dyadic deontic logic*

In the dyadic deontic logic of Bengt Hansson, an obligation $\bigcirc(\alpha|\beta)$ is true if and only if α is true in all minimal (preferred) β worlds (Hansson, 1971). We therefore say that his logic is based on minimizing. Boutilier (1994b) gave a reconstruction of B. Hansson's logic in a modal preference structure. In this section we give a related but weaker logic, in which an obligation ②$(\alpha|\beta)$ is true if and only if α is true in an *equivalence class* of minimal (preferred) β worlds.[26] To discriminate between the two types of minimizing conditionals we call the Hansson-Lewis type universal-minimizing.

The minimizing obligation is defined in a weak deontic betterness relation, written as $\alpha_1 \succ_2 \alpha_2$. As we discussed in the Section 3.1, we say that a preference ordering $\succ$ formalizes weak preferences when a preference $\alpha_1 \succ \alpha_2$ does not logically imply a preference for $\alpha_1' \succ \alpha_2'$ when α_1' implies α_1 and α_2' implies α_2. We say that α_1 is weakly preferred to α_2 if and only if there is an α_1 world such that there is no α_2 world which is as preferable. That is, there is a preferred α_1 world such that for all preferred α_2 worlds we have either that the α_1 world is preferred to the α_2 world, or that the two worlds are incomparable.

Definition 9 (Dyadic minimizing obligation). The dyadic minimizing obligation 'α should be the case if β is the case', written as ②$(\alpha|\beta)$ and ②$^c(\alpha \,|\, \beta)$, is defined as a weak preference of $\alpha \wedge \beta$ over $\neg\alpha \wedge \beta$.

A weak preference of α_1 over α_2, written as $\alpha_1 \succ_2 \alpha_2$, is defined as follows.

$$
\begin{array}{lll}
\alpha_1 \succ_2 \alpha_2 & =_{\text{def}} & \overset{\leftrightarrow}{\Diamond}(\alpha_1 \wedge \neg\alpha_2 \wedge \Box\neg(\alpha_2 \wedge \neg\alpha_1)) \\
②(\alpha|\beta) & =_{\text{def}} & (\alpha \wedge \beta) \succ_2 (\neg\alpha \wedge \beta) \\
 & = & \overset{\leftrightarrow}{\Diamond}((\alpha \wedge \beta) \wedge \Box\neg(\neg\alpha \wedge \beta)) \\
 & \leftrightarrow & \overset{\leftrightarrow}{\Diamond}(\beta \wedge \Box(\beta \rightarrow \alpha)) \\
②^c(\alpha|\beta) & =_{\text{def}} & (\alpha \wedge \beta) \succ_2 (\neg\alpha \wedge \beta) \wedge \overset{\leftrightarrow}{\Diamond}(\neg\alpha \wedge \beta)
\end{array}
$$

The following proposition shows that the obligation $②(\alpha|\beta)$ refers to the optimal β worlds, and that $②(\alpha|\top)$ refers to the ideal worlds.

Proposition 4. Let $M = \langle W, \leq, V\rangle$ be a 2DL model and let $\mid \alpha \mid$ be the set of worlds that satisfy α. For a world $w \in W$, we have $M, w \models ②(\alpha \mid \beta)$ if and only if there is a world $w_2 \in\mid \alpha \wedge \beta \mid$ such that for all worlds $w_1 \in\mid \neg\alpha \wedge \beta \mid$ it is true that $w_1 \not\leq w_2$. Hence, we have $M, w \models ②(\alpha|\beta)$ if and only if

1. α is true in an equivalence class of most preferred β worlds of M, or
2. there is an infinite descending chain in which there is a β world w_2 such that α is true in all β worlds w_1 with $w_1 \leq w_2$.

Proof Analogous to the proof of Proposition 2 (see also (Boutilier, 1994a)). $\Rightarrow$ By contraposition. Assume a model $M = \langle W, \leq, V\rangle$ such that for all worlds $w_2 \in W$ such that $M, w_2 \models \alpha \wedge \beta$ there is a world $w_1 \in W$ such that $M, w_1 \models \neg\alpha \wedge \beta$ and $w_1 \leq w_2$. We have $M, w_2 \not\models (\alpha \wedge \beta) \wedge \Box(\beta \rightarrow \alpha))$. $M, w \models \overset{\leftrightarrow}{\Diamond} \alpha$ for a world $w \in W$ if and only if there is a world $w' \in W$ such that $M, w' \models \alpha$. Hence, $M, w \not\models ②(\alpha|\beta)$.

$\Leftarrow$ By contraposition. Assume $M, w \not\models ②(\alpha|\beta)$ for some world w. Hence, for all worlds $w_2 \in W$ we have $M, w_2 \not\models \beta \wedge \Box(\beta \rightarrow \alpha)$. It follows that for all worlds w_2 such that $M, w_2 \models \alpha \wedge \beta$ we have $M, w_2 \not\models \Box(\beta \rightarrow \alpha)$. Hence, there is a world $w_1 \in W$ such that $M, w_1 \models \neg\alpha \wedge \beta$ and $w_1 \leq w_2$.

The following proposition shows that minimizing obligations can be used as phase-2 obligations, because they do not validate SA, ANDC or DD, but they do validate WC and ORA.[27]

Proposition 5. The logic 2DL has the following theorems.

WC$_2$ $\textcircled{2}(\alpha_1|\beta) \rightarrow \textcircled{2}(\alpha_1 \vee \alpha_2|\beta)$
ORA$_2$: $(\textcircled{2}(\alpha|\beta_1) \wedge \textcircled{2}(\alpha|\beta_2)) \rightarrow \textcircled{2}(\alpha|\beta_1 \vee \beta_2)$

The logic 2DL does not have the following theorems.

SA$_2$: $\textcircled{2}(\alpha|\beta_1) \rightarrow \textcircled{2}(\alpha|\beta_1 \wedge \beta_2)$
AND$_2$: $\textcircled{2}(\alpha_1|\beta) \wedge \textcircled{2}(\alpha_2|\beta) \rightarrow \textcircled{2}(\alpha_1 \wedge \alpha_2|\beta)$
DD$_2$: $\textcircled{2}(\alpha|\beta) \wedge \textcircled{2}(\beta|\gamma) \rightarrow \textcircled{2}(\alpha|\gamma)$

Proof The (non)theorems can be proven by proving (un)satisfiability in the preference-based semantics. Consider first the validity of weakening of the consequent **WC**$_2$. The logic has weakening of the consequent of $\textcircled{2}(\alpha_1 \,|\, \beta)$ to $\textcircled{2}(\alpha_1 \vee \alpha_2 \,|\, \beta)$, because the most preferred β worlds that satisfy α_1 also satisfy $\alpha_1 \vee \alpha_2$. Secondly, consider strengthening of the antecedent **SA**$_2$. The logic does not have strengthening of the antecedent of $\textcircled{2}(\alpha\,|\,\beta_1)$ to $\textcircled{2}(\alpha\,|\,\beta_1 \wedge \beta_2)$, because the preferred β_1 worlds may be different from the preferred $\beta_1 \wedge \beta_2$ worlds. For a counterexample, consider the Kripke model M in Figure 17.c. We have $M \models \textcircled{2}(t \,|\, \top)$ and $M \not\models \textcircled{2}(t \,|\, \neg a)$. We do not have $M \models \textcircled{2}(t \,|\, \neg a)$, because the preferred $\neg a$ worlds are the $\neg a \wedge \neg t$ worlds. Hence, $\textcircled{2}$ does not have strengthening of the antecedent. Verification of the other (non)theorems is left to the reader.

3.2.4. *Relation between ordering and minimizing obligations*

In this section we discuss several relations between ordering and minimizing. First, from Proposition 3 and 5 follows that ordering and minimizing obligations are duals when we consider the inference patterns strengthening of the antecedent and weakening of the consequent, because the former only validates the first inference pattern whereas the latter only validates the second one. Moreover, they are also duals when we consider the conjunction rule for the consequent and the disjunction rule for the antecedent, although universal-minimizing logics (e.g. Hansson-Lewis logics) combine these two inference patterns. The second relation is given by the following proposition.

Proposition 6. The logic 2DL has the following theorems.

Rel: $\textcircled{1}(\alpha|\beta) \rightarrow \textcircled{2}(\alpha|\beta)$
Rel^c: $\textcircled{1}^c(\alpha|\beta) \rightarrow \textcircled{2}^c(\alpha|\beta)$

Proof The theorems can easily be proven by proving satisfiability in the preference-based semantics. For example, consider the theorem **Rel**. $\textcircled{1}(\alpha|\beta)$ is true in a model if and only if we have $|\neg\alpha \wedge \beta| \not\leq |\alpha \wedge \beta|$ and

$|\alpha \wedge \beta|$ is non-empty. Then any world $w \in |\alpha \wedge \beta|$ is part of a preferred β equivalence class (or infinite descending chain) or they can see one. Hence, there is at least one preferred β equivalence class (or infinite descending chain) of which the worlds satisfy $\alpha \wedge \beta$. The other theorems follow directly from this result and the definitions of the obligations. Alternatively, the theorems can be proven by proving validity in 2DL. For example, **Rel** is equivalent with the following theorem of 2DL.

$$\textbf{Rel} \quad (\overleftrightarrow{\Box}(\beta \wedge \alpha \to \Box(\beta \to \alpha)) \wedge \overleftrightarrow{\Diamond}(\alpha \wedge \beta)) \to \overleftrightarrow{\Diamond}(\beta \wedge \Box(\beta \to \alpha))$$

Finally, the theorem is easier to read as an instance of the following formula that relates the preference orderings $(\alpha_1 \succ_1 \alpha_2) \to (\alpha_1 \succ_2 \alpha_2)$.

$$\textbf{Rel} \quad (\overleftrightarrow{\Box}(\alpha_1 \to \Box\neg\alpha_2) \wedge \overleftrightarrow{\Diamond}\alpha_1) \to \overleftrightarrow{\Diamond}(\alpha_1 \wedge \Box\neg\alpha_2)$$

The following proposition gives another relation between ordering and minimizing obligations. It shows that an ordering obligation is equivalent to a set of existential-minimizing obligations, when we impose a constraint on the models.

Proposition 7. Let M be a 2DL model such that M does not contain duplicate worlds, i.e. for all $w_1, w_2 \in W$ such that $w_1 \neq w_2$, there is a propositional α such that $M, w_1 \models \alpha$ and $M, w_2 \not\models \alpha$. We have

$M, w \models \textcircled{1}(\alpha \mid \beta)$ iff for all β' such that $M, w \models \overleftrightarrow{\Diamond}(\alpha \wedge \beta')$ and $M, w \models \overleftrightarrow{\Box}(\beta' \to \beta)$, we have $M, w \models \textcircled{2}(\alpha|\beta')$.

$M, w \models \textcircled{1}^c(\alpha|\beta)$ iff for all β' such that $M, w \models \overleftrightarrow{\Diamond}(\neg\alpha \wedge \beta')$, $M, w \models \overleftrightarrow{\Diamond}(\alpha \wedge \beta')$ and $M, w \models \overleftrightarrow{\Box}(\beta' \to \beta)$, we have $M, w \models \textcircled{2}(\alpha|\beta')$.

Proof We only give the proof for $\textcircled{1}$ and $\textcircled{2}$; the other case is analogous. $\Rightarrow$ Follows directly from $\mathbf{RSA}_1$ and **Rel**. $\Leftarrow$ Every world is characterized by a unique propositional sentence. Let $\overline{w}$ denote this sentence that characterizes world w. Proof by contraposition. If $M, w \not\models \textcircled{1}(\alpha|\beta)$, then there is no $\alpha \wedge \beta$ world (in which case the proof is trivial), or there are w_1, w_2 such that $M, w_1 \models \alpha \wedge \beta$, $M, w_2 \models \neg\alpha \wedge \beta$ and $w_2 \leq w_1$. Choose $\beta' = \overline{w_1} \vee \overline{w_2}$. w_2 is one of the preferred β' worlds, because there are no duplicate worlds. (If duplicate worlds are allowed, then there could be a β' world w_3 which is a duplicate of w_1, and which is strictly preferred to w_1 and w_2.) We have $M, w_2 \not\models \alpha$ and therefore $M, w \not\models \textcircled{2}(\alpha|\beta')$.

In the next section we show how minimizing and ordering can be combined in a two-phase deontic logic. The two-phase approach combines strengthening of the antecedent and the conjunction rule for the consequent with weakening of the consequent and the disjunction rule of the antecedent.

3.2.5. *Combining ordering and minimizing*
In this section we analyze the examples discussed in Section 1 in 2DL. The two phases in a deontic logic correspond to the two different kinds of obligations ① and ② (or $①^c$ and $②^c$). From a proof-theoretic point of view, the first phase corresponds to applying valid inferences of ① like RSA, RAND and DD, and the second phase corresponds to applying valid inferences of ② like WC and ORA. We have to use phase-1 obligations ① as premises, link phase-1 obligations to phase-2 obligations with REL, and end with phase-2 obligations. Hence, Proposition 6 is crucial.

We start with reasoning about dilemmas. In the analysis of Van Fraassen's paradox in Section 1.1 we showed that the problem of the paradox is the combination of the restricted conjunction rule and weakening. In fact, we showed that the two inference patterns can only be combined in a two-phase deontic logic. Now we will show that it is also a sufficient condition to block the counterintuitive derivation. This is proven by showing that the model in Figure 17.a is a counter-model of the derivation. Thus far, we have discussed the ordering logic ① that has conjunction rule but not weakening, and the minimizing logic ② that has weakening but not conjunction. The following example shows that the counterintuitive obligation cannot be derived in 2DL. We can combine restricted conjunction and weakening only if there are two phases. The first phase does not have weakening but it has restricted conjunction, and the second phase vice versa.[28]

Example 8 (Van Fraassen's paradox, continued). Consider the set of obligations $S = \{①(p \mid \top), ①(\neg p \mid \top)\}$ that represents a dilemma, because the consequents of the obligations are contradictory. The set S is consistent, and a typical model M of S is given in Figure 17.a. We have $M \models ①(p \mid \top)$ and $M \models ①(\neg p \mid \top)$, because $|\neg p| \not\leq |p|$ and $|p| \not\leq |\neg p|$, respectively, and such p and $\neg p$ worlds exist. The model illustrates that the dilemma S corresponds semantically to incomparable p and $\neg p$ worlds. The model illustrates that the dilemma S corresponds semantically to incomparable p and $\neg p$ worlds. The first phase of the ordering creates this ordering that consists of two states. The optimal states consists of p or $\neg p$ worlds, both containing a violation. The second phase of 2DL uses this ordering for minimization.

We now analyze contrary-to-duty reasoning. In Section 1.2 we showed that the problem of the paradoxes is the combination of strengthening of the antecedent and weakening of the consequent. We now show that it is a sufficient condition in 2DL. The following example illustrates how the two-phase approach analyzes the CTD paradoxes.

Example 9 (Contrary-to-duty paradoxes, continued). Consider the set of premises $S = \{\textcircled{1}(\neg k \mid \top), \textcircled{1}(g \mid k), k, \overset{\leftrightarrow}{\Box}(g \rightarrow k)\}$ of Forrester's paradox in Example 3. The crucial observation is that $\textcircled{2}(\neg g \mid k)$ is not entailed by S. A typical counter-model M is represented in Figure 17.b. We have $M \models \textcircled{1}(\neg k|\top)$ and $M \models \textcircled{1}(g|k)$, because $|k| \not\leq |\neg k|$ and $|\, k \wedge \neg g \,| \not\leq |\, k \wedge g \,|$ respectively. We have $M \not\models \textcircled{2}(\neg g \,|\, k)$, because $|k \wedge g| \leq |k \wedge \neg g|$.

Consider the set $S = \{\textcircled{1}(a|\top), \textcircled{1}(t|a), \textcircled{1}(\neg t|\neg a), a\}$ of Chisholm's paradox in Example 4. The crucial observation is that $\textcircled{2}(t|\neg a)$ is not entailed by S. A typical counter-model M is represented in Figure 17.c. We have $M \models \textcircled{1}(a \,|\, \top)$, $M \models \textcircled{1}(t \,|\, a)$ and $M \models \textcircled{1}(\neg t \,|\, \neg a)$, because $|\neg a| \not\leq |a|$, $|a \wedge \neg t| \not\leq |a \wedge t|$ and $|\neg a \wedge t| \not\leq |\neg a \wedge \neg t|$ respectively. We have $M \not\models \textcircled{2}(t|\neg a)$, because $|\neg a \wedge \neg t| \leq |\neg a \wedge t|$.

Finally we analyze reasoning by cases. In Section 1.3 we showed that the problem of reasoning by cases is the combination of strengthening of the antecedent and the disjunction rule for the antecedent. We now show that the sequencing of derivations is a sufficient condition to block the counterintuitive derivation in 2DL.

Example 10 (Disarmament paradox, continued). Consider the set of premises $S = \{\textcircled{1}(d \mid w), \textcircled{1}(d \mid \neg w), \textcircled{1}(\neg d \mid d \leftrightarrow w)\}$. The set S is consistent, and the crucial observation is that $\textcircled{2}(d \mid d \leftrightarrow w)$ is not entailed by S. A typical counter-model M is represented in Figure 17.d. We have $M \models \textcircled{1}(d \,|\, w)$, $M \models \textcircled{1}(d \,|\, \neg w)$, and $M \models \textcircled{1}(\neg d \,|\, d \leftrightarrow w)$, because $|\neg d \wedge w| \not\leq |d \wedge w|$, $|\neg d \wedge w| \not\leq |d \wedge \neg w|$ and $|\neg d \wedge \neg w| \not\leq |d \wedge w|$ respectively. We have $M \not\models \textcircled{2}(d|d \leftrightarrow w)$, because $|\neg d \wedge \neg w| \leq |d \wedge w|$.

4. Summary

In this paper we have introduced the notion of a *two-phase deontic logic* and we have shown how such a logic analyzes, and escapes, a number of paradoxes and problems. We began with a concern to allow for deontic dilemmas, which requires limiting the conjunction rule $(\bigcirc\alpha \wedge \bigcirc\beta) \rightarrow \bigcirc(\alpha \wedge \beta)$ (AND) of standard deontic logic (SDL) so as to disallow $(\bigcirc p \wedge \bigcirc \neg p) \rightarrow \bigcirc q$, which would make deontic dilemmas formally inconsistent. Yet, following Van Fraassen and others, we do not want to eliminate the conjunction rule altogether, but to allow for such inferences when α and β are at least logically consistent. We called such a limited rule restricted conjunction rule (RAND). Yet, merely to limit AND in that way is not enough to avoid paradox, if the logic also contains a rule of 'weakening' $\bigcirc\alpha \to \bigcirc(\alpha \vee \beta)$ (W), which is derivable from a monotonicity rule $\frac{\vdash \alpha \to \beta}{\vdash \bigcirc\alpha \to \bigcirc\beta}$, which seems hard to give up. This, in effect, is van Fraassen's paradox, which is not much discussed in the literature. In order to preserve both RAND and W, we have presented our two-phase deontic logic. Both rules are, individually, intuitive and acceptable, but they cannot be used together. That is, RAND is barred from application following W, and vice versa. This allows the inferences one wants while blocking those one does not want. This may be analyzed either proof-theoretically, through sequencing the applications of inference rules, or by introducing two sorts of obligation operators. For the first phase operator ① RAND but not W is valid, and for the second phase operator ② W but not RAND is valid. The two are linked by a rule ①$\alpha \to$ ②α. Sections 2 and 3 develop these two approaches in formal detail. The first section, however, continues to show how the idea of a two-phase deontic logic also applies to the more familiar Contrary-To-Duty paradoxes, which extends the idea of the two-phase structure to conditional obligation. Then it examines problems that arise from Reasoning By Cases, the 'disarmament paradox', where once again problematic inferences are blocked by the two phase structure. The standard approach to these kinds of paradox is to reject some of the inference patterns of SDL, considering that logic too strong. The present approach, however, diagnoses the problems rather as resulting from, in effect, SDL being too weak, insofar as it does not distinguish between the two phases. Once that distinction is drawn, all the familiar, and presumably intuitive, patterns of inference may be preserved. Through examining the variety of deontic puzzles and applying the idea of the two-phase structure to them, we have demonstrated the power of our proposal.

Section 2 presents a 'phase labelled deontic logic (PLDL) in the manner of Gabbay's labeled deductive systems. Formulas take the form $\bigcirc(\alpha|\beta)_L$, where L is a label that, in effect, tracks the steps in a derivation. These are the premises from which they are derived and their phases, which allows rules to be restricted in order to preserve consistency. We observe that this sort of logic is severely limited. It does not admit a possible world semantics, and its language does not allow for factual statements. Hence, although it can be applied to Van Fraassen's paradox and the problems of reasoning by cases, it does not extend to the contrary-to-duty paradoxes. Nevertheless, the introduction of the phase labelled deontic logics formally demonstrates the adequacy of the idea of restricting rule applications that is inherent in the two-phase deontic logic.

Section 3 presents the 'real' two-phase deontic logic, 2DL, chiefly through a preference-based possible worlds semantics. Within this semantics two operations of preference, ordering, $\succ_1$, and minimizing, $\succ_2$, can be distinguished. They allow for the definition of the two dyadic deontic operators ①$(\alpha|\beta)$ and ②$(\alpha|\beta)$ that will obey distinct patterns of inference in the way described in Section 1 for the two phases (as extended to dyadic conditional obligation). It is interesting that the two 'better' operations, and hence the two 'ought' operations, are *definable* in a normal bimodal logic with a S5-like necessity operator $\overset{\leftrightarrow}{\Box}$ and an S4-like operator $\Box$. Since these are well-behaved and familiar, it makes the development of the deontic concepts particularly simple, and soundness and completeness come cheap and easy. The section further motivates the logic by showing how the preference-based models for 2DL can be used to analyze all three types of problem raised earlier, Van Fraassen's, the CTD paradoxes, and reasoning by cases.

An open problem is the conceptual interpretation of the distinction between the two phases. One direction to search such an interpretation is the decision-theoretic account of normative statements (van der Torre & Tan, 1999b; van der Torre & Weydert, 2001; Lang, van der Torre, & Weydert, 2002), an idea which goes back at least to Powers' comparison of deontic logic with the calculus of a pay-back machine (Powers, 1967). The preference-based definition $\bigcirc(\alpha|\beta) =_{\text{def}} (\alpha \wedge \beta) \succ (\neg\alpha \wedge \beta)$ can be read as: 'if the agent *chooses* between $\alpha \wedge \beta$ and $\neg\alpha \wedge \beta$, then she ought to choose $\alpha \wedge \beta$.' The consistency conditions of von Wright contingency principle can also be explained with the concept of choice: if it is not possible to violate or fulfill the obligation, then there is no possibility to choose. One could explain the intuition behind the distinction between phase-1 and phase-2 reasoning with the following metaphor. The moral agent has to make up her mind before she can take decisions; she has to think before she acts.

Metaphor of the moral agent. Phase-1 reasoning is what a person does when she 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. Phase-2 reasoning is what the person does when she also tries to realize the best states. Distinctions between varying degrees of bad states are irrelevant. A moral agent does both, because first she interprets the legal message and then she also tries to realize the best worlds.

Notes

[1] In this paper we do not consider nested modal operators. See (Weydert, 1992) for an interpretation and formalization of nested modal operators. Our logics can be extended along the lines proposed there.

[2] For a long time, Van Fraassen's argument for a consistent representation of dilemmas has been ignored. Since the eighties, there has been some discussion on dilemmas (Conee, 1982; Prakken, 1996). In application-oriented research it is often been argued that dilemmas exist in practical situations, and they should therefore be consistent, see e.g. (Brown, Mantha, & Wakayama, 1993).

[3] SDL is a modal system of type KD according to the Chellas classification (Chellas, 1980). It is the smallest set that contains the propositional theorems and the axioms $K : \bigcirc(\beta \to \alpha) \to (\bigcirc\beta \to \bigcirc\alpha)$ and $D : \neg(\bigcirc\alpha \wedge \bigcirc\neg\alpha)$, and that is closed under the inference rules modus ponens and necessitation.

[4] A drawback of this solution (Loewer & Belzer, 1986; Goble, 1991; Goble, 1996) is that not every adverb of action is amenable to treatment as a predicate. For example, Goble (1991) 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. As another example, Prakken and Sergot (1996) 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 (Prakken & Sergot, 1996): from 'you should keep your promise' $\bigcirc p$, 'if you do not keep you promise, then you should apologize' $\neg p \to \bigcirc a$ and the fact 'you do not keep your promise' $\neg p$ we can derive the counterintuitive 'you should keep your promise and apologize for not keeping it' $\bigcirc(p \wedge a)$.

[5] The most convincing argument that weakening is invalid is the paradox of the knower (Åqvist, 1967), represented by the sentence 'if you ought to know p, then p ought to be (done)' $\bigcirc Kp \to \bigcirc p$.

[6] A drawback of this solution is that only in a *few* cases it *seems* that $\bigcirc\alpha \wedge \bigcirc(\neg\alpha \wedge \beta)$ is not a dilemma and should therefore be consistent. This solution seems like overkill.

[7] 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 so-called 'pragmatic oddities' discussed by Prakken and Sergot (1996).

[8] In 3D a dyadic obligation $\bigcirc(\alpha|\beta)$ 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 (1986) also discuss the relation between their solution and Castañeda's approach to the contrary-to-duty paradoxes (Castañeda, 1981).

[9] 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 3, see also the discussion in (Prakken & Sergot, 1996; Yu, 1995).

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

[11] See e.g. (Prakken & Sergot, 1996; Prakken & Sergot, 1997; van der Torre & Tan, 2000) for a discussion on the relation between deontic contrary-to-duty reasoning and contextual reasoning.

[12] Reasoning by cases is related to Savage's (1988) sure-thing principle used in the foundations of decision theory. In general it is considered to be a desirable property of conditionals, although several people have raised criticism against it in decision theory (see e.g. (McClennen, 1988)). Another criticism against reasoning by cases is that there is often an implicit modal operator in the scope of the antecedent. E.g. from 'α if you believe β' and 'α if you believe $\neg\beta$' we can derive by reasoning by cases that 'α if you believe β or you believe $\neg\beta$' but not 'α if you believe $\beta \vee \neg\beta$.' Example 5 is based on the following classic illustration of Jeffrey (1983), see also (Thomason & Horty, 1996).

> Either there will be a nuclear war or there will not. If there will not be a nuclear war, then it is better for us to disarm because armament is expensive and pointless. If there will be a nuclear war, then we will be dead whether or not we arm, so we are better of saving money in the short term by disarming. So, we should 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. Moreover, the example illustrates that we should make a distinction between controllable and uncontrollable propositions (Boutilier, 1994b), because we cannot control whether there is a nuclear war or not (although we can influence it!). Using our terminology, Jeffrey's complaint about the disarmament paradox is the derivability of $\bigcirc(d \mid \top)$ from $\bigcirc(d|w)$ and $\bigcirc(d|\neg w)$. In two-phase deontic logic we can derive ②$(d \mid \top)$ but not ①$(d|\top)$ from ①$(d|w)$ and ①$(d|\neg w)$. This can be explained as follows. To make decisions, probabilities have to be taken into account. In that case, not only the most ideal state but all states are relevant, because the ideal state may be highly unlikely. Consequently ② cannot be used, but ① should be used.

[13] In other words, in this derivation the obligation $\bigcirc(d|d \leftrightarrow w)$ is considered to be counterintuitive, because it is not grounded in the premises. If $d \leftrightarrow w$ and w (the antecedent of the first premise) are true then d is trivially true, and if $d \leftrightarrow w$ and $\neg w$ (the antecedent of the second premise) are true then d is trivially false. With other words, if $d \leftrightarrow w$ then the first premise cannot be violated and the second premise cannot be fulfilled. Hence, the two premises do not ground the conclusion that for arbitrary $d \leftrightarrow w$ we have that $\neg d$ is a violation.

The example is difficult to interpret, because it makes use of a bi-implication. An alternative set of premises, also based on bi-implications, with analogous counterintuitive conclusions is $\{\bigcirc(d|d \leftrightarrow w),\ \bigcirc(d|\neg d \leftrightarrow w),\ \bigcirc(\neg d|w)\}$.

[14] Dilemmas are 'inconsistent' in PLDL if the restricted conjunction rule is replaced by the following unrestricted conjunction rule AND$'$, and the deontic axiom 'ought implies can' $\neg\bigcirc(\bot|\alpha)$ is accepted.

$$\text{AND}' : \frac{\bigcirc(\alpha_1 \mid \beta)_{(F_1,p_1)}, \bigcirc(\alpha_2 \mid \beta)_{(F_2,p_2)}}{\bigcirc(\alpha_1 \wedge \alpha_2 \mid \beta)_{(F_1 \times F_2, \max(p_1,p_2))}}$$

[15] This example has been discussed in the context of the Reykjavik Scenario (Belzer, 1986; McCarty, 1994; van der Torre, 1994; Makinson, 1999), in which $S_1 = \{\bigcirc(\neg r|\top), \bigcirc(\neg g|\top)\}$, $S_2 = \{\bigcirc(\neg r \wedge \neg g|\top)\}$ and r is read as telling the secret to Reagan and g as telling it to Gorbatchov. The example suggests that premises not only encode obligations but also independence (or irrelevance) assumptions. In S_1 the obligations for p and q are independent, but in S_2 they are not.

[16] Note that the reverse replacements are possible for WC and for SA and ORA, as follows from adaptations of the first, second and fourth replacement in Figure 16. However, it is not possible to inverse the replacement of DD and ORA. A counterex-

ample is the following derivation.

$$\dfrac{\dfrac{\bigcirc(p|q \wedge r) \quad \bigcirc(q|r)}{\bigcirc(p \wedge q|r)}\,\text{DD} \qquad \bigcirc(p \wedge q|\neg r)}{\bigcirc(p \wedge q|\top)}\,\text{ORA}$$

[17] For example, assume that the preference ordering is not transitive and consider three worlds w_1, w_2 and w_3 such that $w_1 \leq w_2$, $w_2 \leq w_3$ but not $w_1 \leq w_3$. Is w_1 a preferred world or not? There is no world preferred to it, which seems to indicate that it is a preferred world. However, it is not as preferred as world w_3, which seems to indicate that it is not a preferred world.

[18] Moreover, preferences referring to the normal circumstances can be used to formalize defeasible obligations. Ceteris paribus preferences, referring to similar circumstances, are popular to formalize desires in (qualitative) decision theory (Doyle & Wellman, 1991; Doyle, Shoham, & Wellman, 1991).

[19] However, some minimizing logics have the unrestricted conjunction rule for the consequent, as is explained later in this paper.

[20] The first deontic logic based on a preference ordering was introduced by B. Hansson (Hansson, 1971). It is a dyadic logic and it belongs to the second category, because it is based on minimizing. B. Hansson's logic has been criticized because it lacks strengthening of the antecedent. For example, Alchourrón (1993) argues that lack of strengthening of the antecedent is acceptable for logics of defeasible reasoning or logics of defeasible obligations (sometimes called prima facie obligations), but not for non-defeasible obligations. Moreover, the semantic concept of minimization is unexplained: whereas in a defeasible logic 'normally p' might refer to the most normal worlds only, 'obligatory p' does *not* seem to refer to the most ideal worlds only. Recently, several authors (Jackson, 1985; Goble, 1990b; Hansson, 1990; Brown, Mantha, & Wakayama, 1993; Huang & Masuch, 1997) introduced a preference ordering in a monadic deontic logic. These logics belong to the first category of preference-based deontic logics, because the truth of $\bigcirc\alpha$ depends on the whole ordering. This approach can be traced through a long history of research in preference logics, see e.g. (von Wright, 1963; Rescher, 1967; Jennings, 1974). For our framework we use variants of the dyadic deontic logic Prohairetic Deontic Logic proposed in (van der Torre & Tan, 1999a). An obligation $\bigcirc(\alpha|\beta)$ is true if for all $\alpha \wedge \beta$ and $\neg\alpha \wedge \beta$ worlds, we have that the $\alpha \wedge \beta$ world is preferred to the $\neg\alpha \wedge \beta$ world, or the two worlds are incomparable.

[21] The logic PDL and Hansson's minimizing logic can be defined in 2DL as follows (see the definition of 2DL and $\succ_1$ later in this paper, and (Boutilier, 1994b; van der Torre & Tan, 1999a)).

$$\begin{array}{lll}
\alpha_1 \succ_{2'} \alpha_2 & =_{\text{def}} & \overset{\leftrightarrow}{\Box}(\alpha_2 \wedge \neg\alpha_1 \rightarrow \overset{\leftrightarrow}{\Diamond}(\alpha_1 \wedge \neg\alpha_2 \rightarrow \Box\neg(\alpha_2 \wedge \neg\alpha_1))) \\
①_{\text{PDL}}(\alpha|\beta) & =_{\text{def}} & (\alpha \wedge \beta) \succ_1 (\neg\alpha \wedge \beta) \wedge (\alpha \wedge \beta) \succ_{2'} (\neg\alpha \wedge \beta) \wedge \overset{\leftrightarrow}{\Diamond}(\alpha \wedge \beta) \\
②_{\text{HL}}(\alpha|\beta) & =_{\text{def}} & (\alpha \wedge \beta) \succ_{2'} (\neg\alpha \wedge \beta) \wedge \overset{\leftrightarrow}{\Diamond}(\alpha \wedge \beta)
\end{array}$$

Boutilier (1994b) uses a more complex system based on Humberstone's logic of inaccessible worlds. We do not do this in this paper, because it is quite complex and the crucial issue is not the axiomatization of the underlying monadic modal logic, which is indeed trivial now we use $\overset{\leftrightarrow}{\Box}\alpha \rightarrow \Box\alpha$ as the axiomatization, but the definition of the dyadic operator in the monadic logic.

[22] We cannot derive $①(\alpha|\beta)$ from $①(\alpha|\top)$ by $\mathbf{RSA}_1$, unless we have the consistency expression $\overset{\leftrightarrow}{\Diamond}(\alpha \wedge \beta)$ as another premise. Instead of explicitly writing down these consistency expressions in every example, we can consider only models in which all propositionally satisfiable formulas α are true in *some* world. This can be 'axiomatized' with Boutilier's axiom scheme **LP**, see (Boutilier, 1994a) for a discussion. The axiom scheme **LP** states that every formula α without any occurrences of modal

operators, which is propositionally satisfiable, is true in some world. The logic 2DL* is 2DL extended with the following axiom scheme **LP**.

LP: $\overleftrightarrow{\Diamond}\alpha$ for all satisfiable propositional α

Let $\mathcal{P}$ be the set of propositional atoms of the propositional base language $\mathcal{L}$. A 2DL*-model is a 2DL-model $M = \langle W, \leq, V\rangle$ that satisfies the following condition:

$$\{f \mid f \text{ maps } \mathcal{P} \text{ into} \{0,1\}\} \subseteq \{V(w) \mid w \in W\}$$

We write $\models^*$ for logical entailment in 2DL*.

The logic 2DL* is illustrated by the following example. Consider the set $S = \{\bigcirc(p_1|\top), \bigcirc(p_2|\top)\}$. Semantically, the axiom **LP** ensures that the $p_1 \wedge p_2$ worlds exist in all 2DL* models. Hence, we have $\models^* \overleftrightarrow{\Diamond}(p_1 \wedge p_2)$ whereas we have $\not\models \overleftrightarrow{\Diamond}(p_1 \wedge p_2)$. Proof-theoretically, in 2DL we can derive $\bigcirc(p_1 \wedge p_2 \mid \top)$ from S and the premise $\overleftrightarrow{\Diamond}(p_1 \wedge p_2)$ by **RAND**$_1$. In 2DL* the consistency expression $\overleftrightarrow{\Diamond}(p_1 \wedge p_2)$ can be derived from **LP**, and hence $\bigcirc(p_1 \wedge p_2|\top)$ can be derived from S. This shows that we do not have to write the consistency expressions explicitly in the logic 2DL*.

[23] The conditions only check logical possibility. In an agent environment, the alternatives are to consider stronger conditions which refer to the agent's opportunities or to her abilities. The logical conditions are already stronger than necessary to invalidate the counterintuitive theorems, because the consistency conditions $\overleftrightarrow{\Diamond}\alpha$ and $\overleftrightarrow{\Diamond}\neg\alpha$ would (in principle) also do the trick.

[24] Alternatively, we can take the dyadic ordering obligation as primitive, defined by the semantic definition in Proposition 2, or we can take the preference ordering $\succ_1$ as primitive. In the latter case, we can define the monadic operator $\Box$ in terms of $\succ_1$ by $\Box\alpha =_{\text{def}} \neg\alpha \succ_1 \top$. Analogous definitions of unary modalities in terms of minimizing conditionals $\beta \Rightarrow \alpha$ by $\Box\alpha =_{\text{def}} \neg\alpha \Rightarrow \alpha$ are well-known, see e.g. (Stalnaker, 1981; Lewis, 1973), and an analogous grounding of the logic CT4 (hence S4) in a minimizing conditional can be found in (Boutilier, 1992, p.89). Note that we cannot define a conditional by $\bigcirc(\alpha|\beta) =_{\text{def}} \Box((\alpha \wedge \beta) \rightarrow \Box(\beta \rightarrow \alpha))$ in a monomodal logic, because these obligations cannot be combined with facts.

[25] Moreover, the logic 2DL also has the following theorems.

ORC$_1$:	$(\bigcirc(\alpha_1|\beta) \wedge \bigcirc(\alpha_2|\beta)) \rightarrow \bigcirc(\alpha_1 \vee \alpha_2|\beta)$
DD$-_1$:	$(\bigcirc(\alpha|\beta \wedge \gamma) \wedge \bigcirc(\neg\beta|\gamma)) \rightarrow \bigcirc((\alpha \wedge \beta) \vee \neg\beta|\gamma)$
NC$_1$:	$\neg\bigcirc(\bot|\alpha)$
NC$^c_1$:	$\neg\bigcirc^c(\bot|\alpha)$
Id$_1$:	$\overleftrightarrow{\Diamond}\alpha \rightarrow \bigcirc(\alpha|\alpha)$
NId$^c_1$:	$\neg\bigcirc^c(\alpha|\alpha)$

The logic 2DL does *not* have the following theorem.

DD$_1$:	$(\bigcirc(\alpha|\beta) \wedge \bigcirc(\beta|\gamma)) \rightarrow \bigcirc(\alpha|\gamma)$
DD$\top_1$:	$(\bigcirc(\alpha|\beta) \wedge \bigcirc(\beta|\top)) \rightarrow \bigcirc(\alpha|\top)$
D*$_1$:	$\neg(\bigcirc(\alpha|\beta) \wedge \bigcirc(\neg\alpha|\beta))$

[26] The definition is adapted from a modal formula of Boutilier. The minor distinction is that Boutilier defines $\overleftrightarrow{\Diamond}(\beta \wedge \Box(\beta \rightarrow \alpha)) \vee \overleftrightarrow{\Box}\neg\beta$. We have adapted the definition for our two-phase approach. Boutilier's definition is false if $\overleftrightarrow{\Box}\neg(\beta \wedge \alpha) \wedge \neg\overleftrightarrow{\Box}\neg\beta$, and therefore does not validate Proposition 6.

[27] Moreover, the logic 2DL has the following theorems.

$\mathbf{NC}_2$: $\neg ②(\perp|\alpha)$
$\mathbf{NC}^c_2$: $\neg ②^c(\perp|\alpha)$
$\mathbf{ID}_2$: $\overleftrightarrow{\Diamond}\alpha \rightarrow ②(\alpha|\alpha)$
$\mathbf{NID}^c_2$: $\neg ②^c(\alpha|\alpha)$

The logic does not have the following theorem.

$\mathbf{D^*}_2$: $\neg(②(\alpha|\beta) \wedge ②(\neg\alpha|\beta))$

[28] In Horty's (1993) reconstruction of van Fraassen's theory in Reiter's (1980) default logic the two phases are *not* explicit. In our terminology, the distinct operators ① and ② are represented by the same modal operator ◯, just like in 2LDL. As a consequence, it is very difficult if not impossible to construct a semantics for these logics.

Acknowledgement

Thanks to David Makinson for many suggestions concerning the issues raised in this paper, in particular with respect to the labeled deontic logic in Section 2, and to an annonymous referee of this journal for several helpful comments.

References

Alchourrón, C. 1993. Philosophical foundations of deontic logic and the logic of defeasible conditionals. In Meyer, J.-J., and Wieringa, R., eds., *Deontic Logic in Computer Science: Normative System Specification.* John Wiley & Sons. 43–84.

Åqvist, L. 1967. Good Samaritans, contrary-to-duty imperatives, and epistemic obligations. *Noûs* 1:361–379.

Belzer, M. 1986. A logic of deliberation. In *Proceedings of the Fifth National Conference on Artificial Intelligence (AAAI'86)*, 38–43.

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

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

Boutilier, C. 1994b. Toward a logic for qualitative decision theory. In *Proceedings of the Fourth International Conference on Principles of Knowledge Representation and Reasoning (KR'94)*, 75–86.

Brown, A.; Mantha, S.; and Wakayama, T. 1993. Exploiting the normative aspect of preference: a deontic logic without actions. *Annals of Mathematics and Artificial Intelligence* 9:167–203.

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

Chellas, B. 1980. *Modal Logic: An Introduction.* Cambridge University Press.

Chisholm, R. 1963. Contrary-to-duty imperatives and deontic logic. *Analysis* 24:33–36.

Conee, E. 1982. Against moral dilemmas. *The Philosophical Review* 91:87–97.
Davidson, D. 1967. The logical form of action sentences. In Rescher, N., ed., *The Logic of Decision and Action*. University of Pittsburg Press.
Doyle, J., and Wellman, M. 1991. Preferential semantics for goals. In *Proceedings of the Ninth National Conference on Artificial Intelligence (AAAI-91)*, 698–703.
Doyle, J.; Shoham, Y.; and Wellman, M. 1991. The logic of relative desires. In Ras, Z. W., and Zemankova, M., eds., *Methodologies for Intelligent Systems*, volume 542 of *Lecture Notes in Artificial Intelligence*, 16–31.
Fagin, R.; Halpern, J.; Moses, Y.; and Vardi, M. 1995. *Reasoning About Knowledge*. MIT press.
Forrester, J. 1984. Gentle murder, or the adverbial Samaritan. *Journal of Philosophy* 81:193–197.
Gabbay, D. 1996. *Labelled Deductive Systems*, volume 1. Oxford University Press.
Goble, L. 1989. A logic of *better*. *Logique et Analyse* 32:297–318.
Goble, L. 1990a. A logic of good, would and should, part 1. *Journal of Philosophical Logic* 19:169–199.
Goble, L. 1990b. A logic of good, would and should, part 2. *Journal of Philosophical Logic* 19:253–276.
Goble, L. 1991. Murder most gentle: the paradox deepens. *Philosophical Studies* 64:217–227.
Goble, L. 1993. The logic of obligation, 'better' and 'worse'. *Philosophical Studies* 70:133–163.
Goble, L. 1996. 'ought' and extensionality. *Noûs* 30:330–355.
Hansson, B. 1971. An analysis of some deontic logics. In Hilpinen, R., ed., *Deontic Logic: Introductory and Systematic Readings*. Dordrecht, Holland: D. Reidel Publishing Company. 121–147.
Hansson, S. 1990. Preference-based deontic logic (PDL). *Journal of Philosophical Logic* 19:75–93.
Horty, J. 1993. Deontic logic as founded in nonmonotonic logic. *Annals of Mathematics and Artificial Intelligence* 9:69–91.
Huang, Z., and Masuch, M. 1997. The logic of permission and obligation in the framework of ALX3: how to avoid the paradoxes of deontic logic. *Logique et Analyse* 149.
Hughes, H., and Creswell, M. 1984. *A Companion to Modal Logic*. London: Methuen.
Jackson, F. 1985. On the semantics and logic of obligation. *Mind* 94:177–196.
Jeffrey, R. 1983. *The Logic of Decision*. University of Chicago Press, 2nd edition.
Jennings, R. 1974. A utilitarian semantics for deontic logic. *Journal of Philosophical Logic* 3:445–465.
Keeney, R., and Raiffa, H. 1976. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. New York: Wiley.
Lamarre, P. 1991. S4 as the conditional logic of nonmonotonicity. In *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning (KR'91)*, 357–367.
Lang, J.; van der Torre, L.; and Weydert, E. 2002. Utilitarian desires. *Autonomous Agents and Multi-Agent Systems*.
Lewis, D. 1973. *Counterfactuals*. Oxford: Blackwell.
Lewis, D. 1974. Semantic analysis for dyadic deontic logic. In Stunland, S., ed., *Logical Theory and Semantical Analysis*. Dordrecht, Holland: D. Reidel Publishing Company. 1–14.
Loewer, B., and Belzer, M. 1983. Dyadic deontic detachment. *Synthese* 54:295–318.
Loewer, B., and Belzer, M. 1986. Help for the good Samaritan paradox. *Philosophical Studies* 50:117–127.
Makinson, D., and van der Torre, L. 2000. Input-output logics. *Journal of Philosophical Logic* 29:383–408.

Makinson, D., and van der Torre, L. 2001. Constraints for input-output logics. *Journal of Philosophical Logic* 30(2):155–185.

Makinson, D. 1999. On a fundamental problem of deontic logic. In McNamara, P., and Prakken, H., eds., *Norms, Logics and Information Systems. New Studies on Deontic Logic and Computer Science*, 29–53. IOS Press.

McCarty, L. 1994. Defeasible deontic reasoning. *Fundamenta Informaticae* 21:125–148.

McClennen, E. 1988. Sure-thing doubts. In Gärdenfors, P., and Sahlin, N., eds., *Decision, Probability, and Utility.* Cambridge University Press. 166–182.

Meyer, J.-J. 1988. A different approach to deontic logic: deontic logic viewed as a variant of dynamic logic. *Notre Dame Journal of Formal Logic* 29:109–136.

Nute, D., and Yu, X. 1997. Introduction. In Nute, D., ed., *Defeasible Deontic Logic.* Kluwer. 1–16.

Pearl, J. 1993. From conditional oughts to qualitative decision theory. In *Proceedings of Uncertainty in Artificial Intelligence (UAI'93)*, 12–20.

Powers, L. 1967. Some deontic logicians. *Noûs* 1:380–400.

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

Prakken, H., and Sergot, M. 1997. Dyadic deontic logic and contrary-to-duty obligations. In Nute, D., ed., *Defeasible Deontic Logic.* Kluwer. 223–262.

Prakken, H. 1996. Two approaches to the formalisation of defeasible deontic reasoning. *Studia Logica* 57:73–90.

Reiter, R. 1980. A logic for default reasoning. *Artificial Intelligence* 13:81–132.

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

Ryu, Y., and Lee, R. 1993. Defeasible deontic reasoning: A logic programming model. In Meyer, J.-J., and Wieringa, R., eds., *Deontic Logic in Computer Science: Normative System Specification.* John Wiley & Sons. 225–241.

Savage, L. 1988. The sure-thing principle. In Gärdenfors, P., and Sahlin, N., eds., *Decision, Probability, and Utility.* Cambridge University Press. 80–85.

Sinnot-Armstrong, W. 1985. A solution to Forrester's paradox of gentle murder. *Journal of Philosophy* 82:162–168.

Stalnaker, R. 1981. A theory of conditionals. In Harper, W.; Stalnaker, R.; and Pearce, G., eds., *Ifs.* Dordrecht: D. Reidel. 41–55.

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

Thomason, R., and Horty, R. 1996. Nondeterministic action and dominance: foundations for planning and qualitative decision. In *Proceedings of the Sixth Conference on Theoretical Aspects of Rationality and Knowledge (TARK'96)*, 229–250. Morgan Kaufmann.

Tomberlin, J. 1981. Contrary-to-duty imperatives and conditional obligation. *Noûs* 16:357–375.

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

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

van der Torre, L., and Tan, Y. 1998. The temporal analysis of Chisholm's paradox. In *Proceedings of the Proceedings of 15th National Conference on Artificial Intelligence (AAAI'98)*, 650–655.

van der Torre, L., and Tan, Y. 1999a. Contrary-to-duty reasoning with preference-based dyadic obligations. *Annals of Mathematics and Artificial Intelligence* 27:49–78.

van der Torre, L., and Tan, Y. 1999b. Diagnosis and decision making in normative reasoning. *Artificial Intelligence and Law* 7:51–67.

van der Torre, L., and Tan, Y. 2000. Contextual deontic logic: Violation contexts and factual defeasibility. In *Formal Aspects of Context*, volume 20 of *Applied Logic Series*. Kluwer. 143–166.

van der Torre, L., and Weydert, E. 2001. Parameters for utilitarian desires in a qualitative decision theory. *Applied Intelligence* 14:285–301.

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

van der Torre, L. 1998a. Labeled logics of goals. In *Proceedings of the Thirteenth European Conference on Artificial Intelligence (ECAI'98)*, 368–369.

van der Torre, L. 1998b. Phased labeled logics of goals. In *Proceedings of the 6th European Workshop on Logics in AI (JELIA'98)*, volume 1489 of *Lectore Notes in Computer Science*, 92–106. Springer. Also appeared in: Proceedings of the First International Workshop on Labeled Deduction (LD'98).

van Eck, J. 1982. A system of temporally relative modal and deontic predicate logic and its philosophical application. *Logique et Analyse* 100:249–381.

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

von Neumann, J., and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton University Press.

von Wright, G. 1963. *The logic of preference*. Edinburgh University Press.

von Wright, G. 1971. A new system of deontic logic. In Hilpinen, R., ed., *Deontic Logic: Introductory and Systematic Readings*. Dordrecht, Holland: D. Reidel Publishing company. 105–120. A reprint of 'A New System of Deontic Logic,' *Danish Yearbook of Philosophy* 1:173–182, 1964, and 'A Correction to a New System of Deontic Logic', *Danish Yearbook of Philosophy* 2:103–107, 1965.

von Wright, G. 1981. On the logic of norms and actions. In Hilpinen, R., ed., *New studies in Deontic Logic: Norms, Actions and the Foundations of Ethics*. D.Reidel Publishing company. 3–35.

Weydert, E. 1992. Hyperrational conditional logic. In *Proceedings of the ECAI'92 Workshop on Theoretical Foundations of Knowledge Representation and Reasoning*. Springer Verlag.

Yu, X. 1995. *Deontic Logic with Defeasible Detachment*. Ph.D. Dissertation, University of Georgia.

Address for correspondence:
L.W.N. van der Torre
Department of Artificial Intelligence
Vrije Universiteit Amsterdam
De Boelelaan 1081a
1081 HV Amsterdam
The Netherlands