Contrary-to-duty obligations

Henry Prakken
Computer/Law Institute
Free University Amsterdam
De Boelelaan 1105
1081 HV Amsterdam
henry@rechten.vu.nl

Marek Sergot
Department of Computing
Imperial College of Science,
Technology and Medicine
180 Queen's Gate, London SW7 2BZ
mjs@doc.ic.ac.uk

September 1994; January 1995*

Abstract

We investigate under what conditions contrary-to-duty (CTD) structures lacking temporal and action elements can be given a coherent reading. We argue, contrary to some recent proposals, that CTD is not an instance of defeasible reasoning, and that methods of nonmonotonic logics are inadequate since they are unable to distinguish between defeasibility and violation of primary obligations. We propose a semantic framework based on the idea that primary and CTD obligations are obligations of different kinds: a CTD obligation pertains to, or pre-supposes, a certain context in which a primary obligation is already violated. This framework is presented initially as an extension of Standard Deontic Logic (SDL), a normal modal logic of type KD, and is illustrated by application to a series of examples. The concluding section is concerned with some resemblances between CTD and defeasible reasoning. We show first that the SDL-based framework contains a flaw and must be adjusted. A discussion of possible adjustments, including an alternative treatment in terms of a preference-based semantics, reveals difficulties that are reminiscent of problems in defeasible reasoning and intensional accounts of defeasible conditionals.

1 Introduction

A well-known problem in the study of deontic logic is the proper representation of contrary-to-duty structures, situations in which there is a primary obligation and what we might call a secondary obligation, which comes into effect when the primary obligation is violated. The problem is that for many intuitively coherent examples of contrary-to-duty (CTD) structures it is hard to find a consistent and also otherwise acceptable formalization. Here is a simple example:

[*] An earlier version of this paper appeared in *Proc. Second International Workshop on Deontic Logic in Computer Science (DEON'94)*, Oslo, January 1994. Complex 1/94, Tano Publishers, Norway, pp 296–318.

Example 1

1. There must be no fence.
2. If there is a fence then it must be a white fence.
3. There is a fence.

This example is quite genuine. It comes from a set of formal and informal regulations governing the appearance and use of holiday cottages. As a matter of fact both of the regulations (1) and (2) are intended to hold at one and the same time, and (2) is intended as a CTD-rule for (1) rather than as some kind of exception.

The example has two important features. The first is that it concerns states of affairs rather than actions. This rules out proposals based on action logics (e.g. [Meyer 88]), or more generally on the distinction betweeen 'ought-to-be' (*Seinsollen*) and 'ought-to-do' (*Tunsollen*) statements. The second feature is that all three statements pertain to the same point in time, thus making proposals based on temporal logics (e.g. [Åqvist & Hoepelman 81]) inapplicable also. To some eyes timeless 'ought-to-be' examples of this kind already seem to be inconsistent in their natural language formulation. In section 3, however, we will argue that it is worthwhile investigating under what conditions such CTD structures can be given consistent readings; this investigation will be the main topic of the paper.

Recently, it has been suggested that CTD reasoning is just an instance of defeasible reasoning, to which, accordingly, familiar techniques of nonmonotonic logics can be applied rather directly (e.g. [Ryu & Lee 91, McCarty 94]). In section 4 we will argue that these proposals are mistaken, although understandable because of the ambiguity of many natural-language examples. Instead we will follow suggestions of, among others, [Jones & Pörn 85], that CTD-structures can be represented consistently if a distinction is made between different senses of 'ought'. A semantic framework formalising this idea will be presented in section 5 and used to analyse some further examples in section 6. Although the framework exhibits many attractive features, the version presented in section 5 contains a flaw and requires some adjustment. Section 7 is concerned with identifying the source of the problem and options for removing it. From this discussion it emerges that, although CTD reasoning is not a special kind of defeasible reasoning, there are nevertheless similarities between *certain aspects* of contrary-to-duty reasoning and problems encountered in defeasible reasoning.

The framework of section 5 is defined initially as an extension of Standard Deontic Logic (SDL), a normal logic of type KD in the Chellas classification [Chellas 80]. We do this for convenience; it is possible to construct similar extensions of other systems of deontic logic as well, and we shall comment further on this in section 7. In the following section we rehearse the basic features of SDL, for future reference and as a way of introducing notation.

2 Standard Deontic Logic

We will employ the following notational conventions: capitals A, B and C are metavariables for arbitrary formulas, lower case letters w, v, $\dots$ represent worlds, capitals P, Q, R stand for propositions in the sense of sets of worlds, and $\|A\|^{\mathcal{M}}$ denotes the set of worlds (of a model $\mathcal{M}$) in which A is true. Now the language of SDL is that of propositional logic, augmented with two operators O and P, standing for 'obligatory' and 'permitted'. Models of SDL are structures $\mathcal{M} =< W, d, V >$ where W is a set of possible worlds and V is a valuation function for atomic sentences in each of the possible worlds. d corresponds to the deontic accessibility relation, for which we will use a functional notation: $d(w)$ is the set of worlds in W that are the (deontic) alternatives to w. The truth conditions for the O-operator are

$$\models_w \mathrm{O}A \text{ iff } d(w) \subseteq \|A\|$$

(Throughout the paper we omit reference to the model $\mathcal{M}$ where it is obvious from context.)

PA is defined as $\neg\mathrm{O}\neg A$.

Models of this kind validate the rule of consequential closure, i.e.

$$\frac{A_1 \wedge \cdots \wedge A_n \rightarrow B}{\mathrm{O}A_1 \wedge \cdots \wedge \mathrm{O}A_n \rightarrow \mathrm{O}B} \qquad (n \geq 0)$$

For SDL we furthermore require that the accessibility relation is serial, i.e. for all $w \in W$ it holds that $d(w) \neq \emptyset$. This validates the characteristic scheme

$$\mathrm{O}A \rightarrow \neg\mathrm{O}\neg A$$

We trivially extend the language of SDL with the two operators $\Box$ and $\Diamond$ for 'necessary' and 'possible', and its model structure to $< W, f, d, V >$. f is also a function from W into Pow(W): it assigns to each world a set of accessible worlds. The relevant truth conditions are

$$\models_w \Box A \text{ iff } f(w) \subseteq \|A\|$$

with $\Diamond A$ defined as $\neg\Box\neg A$, as usual. As for the relation between d and f we will require only that for any $w \in W$, $d(w) \subseteq f(w)$. Together with seriality of d this makes f also serial and validates a form of 'ought implies can' principle:

$$\mathrm{O}A \rightarrow \Diamond A$$

Except where indicated otherwise, our discussion will not rely on any further assumptions on the nature of f.

3 Contrary-to-duty structures

In this section we examine in more detail what kinds of elementary CTD structures there are and to what problems they give rise. We will leave the discussion of two benchmark examples of CTD structures, Chisholm's paradox

[Chisholm 63] and Belzer's Gorbachov-Reagan example [Belzer 87], until section 6, since they both exhibit other features besides CTD, which can distract attention from what we take to be the essence of CTD reasoning.

In our examples we shall need to represent conditional obligations. Since opinions vary as to what kind of conditional is appropriate for this task, and our discussion is independent of these considerations, we employ the symbol $\Rightarrow$ to stand for any suitable conditional. The only property we assume is the validity of unrestricted factual detachment, i.e.

$$\frac{A, A \Rightarrow \mathrm{O}B}{\mathrm{O}B}$$

Readers may replace $\Rightarrow$ by any conditional, for example a strict or counterfactual implication, as long as it validates unrestricted factual detachment.[1]

We also need to point out that, although the paper is concerned with 'ought-to-be' structures, some of the examples will contain 'ought-to-do' statements. We do this merely to preserve continuity with well-known examples in the literature. 'Ought-to-be' variants are easily constructed and will be provided for all the main examples.

3.1 Temporal examples

Example 2

1. Keep your promise.
2. If you haven't kept your promise, apologise.
3. You haven't kept your promise.

This seems to be a common kind of CTD structure, but already it presents a problem for SDL, as the following formalisation shows.

1. $\mathrm{O}k$
2. $\neg k \Rightarrow \mathrm{O}a$
3. $\neg k$

Now both $\mathrm{O}k$ and $\mathrm{O}a$ are true but it is a bit odd to say that in all ideal versions of this world you keep your promise and you apologise for not keeping it. This oddity — we might call it a 'pragmatic oddity' — seems to be absent from the natural language version, which means that the SDL representation is not fully adequate.

For examples containing a temporal element a natural solution is available, in the form of temporal deontic logics as mentioned above. Details vary, but generally speaking, in such logics the obligations and permissions are determined by what holds in the deontic alternatives which are accessible, as in SDL, but

[1] Concretely, a suggestion in the spirit of [Jones & Pörn 85] would be to identify $\Box$ with their notion of 'deontic necessity' – a modality of type KT – and then read expressions of the form $A \Rightarrow \mathrm{O}B$ as standing for $\Box(A \rightarrow \mathrm{O}B)$.

what is new is that this accessibility is defined in temporal terms: at any given point in time only those worlds are accessible which have the same past as the actual worlds. This makes the past necessary, in the sense that it cannot be changed. A natural effect of this is that obligations pertaining to a particular point in time cease to hold after they have been violated, since the violation makes every world in which the obligation has been fulfilled inaccessible. To apply this to our example, let t stand for the moment the promise is broken: then $\mathrm{O}k$ holds until and including t, while from t on $\mathrm{O}k$ is false and $\mathrm{O}a$ is true. After the violation there is no accessible ideal world in which you keep your promise but still apologise for not keeping it.

The need for a temporal representation is even more urgent if, instead of a pragmatic oddity, it is a case of conflicting obligations which has to be avoided. This is illustrated by the next example, containing instructions of a party host to his personnel.

Example 3

1. Woody and Mia should not meet.
2. If Woody and Mia meet, then they should be forced to embrace.
3. Woody and Mia meet.
4. Woody and Mia cannot be forced to embrace if they do not meet.

Assume we translate this in SDL as

1. $\mathrm{O}\neg m$
2. $m \Rightarrow \mathrm{O}e$
3. m
4. $\neg\Diamond(e \wedge \neg m)$

Because of (4) the obligations are conflicting, and are logically inconsistent since O is of type KD. However, on at least one plausible reading the natural language example seems to be consistent, in such a way that a temporal analysis can again remove the inconsistency: $\mathrm{O}\neg m$ holds *until* Woody and Mia have met each other, while $\mathrm{O}e$ holds *after* they have met.

3.2 Non-temporal examples

Unfortunately the temporal solution is not always available, since sometimes the primary and CTD rule pertain to the same point in time. A timeless example with a pragmatic oddity is a set of holiday cottage regulations on keeping dogs, stating that there ought to be no dog, but that if there is a dog, there ought to be a warning sign. Surely, it is strange to say that in all ideal worlds there is no dog and also a sign warning that there is a dog. Like example 2, this example has a stronger version, with the pragmatic oddity replaced by conflict between obligations: example 1 is of this form.

Now it might be felt that there is after all something suspicious about the holiday cottage regulations. Whether or not that is the case, here are two sentences which would be regarded as incontrovertibly true by any parent of small children living in London:

Example 4

1. The children ought not to be cycling on the street.
2. If the children are cycling on the street, then
   they ought to be cycling on the left hand side of the street.

Now we add:

3. The children are cycling on the street.
4. The children cannot be cycling on the left hand side of the street
   if they are not cycling on the street.

With the temporal dimension the way of avoiding the inconsistency between the primary and the CTD obligation has also disappeared. Thus the example is, like example 1, an 'ought-to-be' variant of the paradox of the 'gentle murderer' [Forrester 84]: do not kill, but if you kill, kill gently.

3.3 Discussion

Do examples 1 and 4 provide a problem for SDL or are the natural-language versions themselves inconsistent? It might be suggested that the inconsistency arises simply because insufficient attention has been given to the proper logical form of the examples. For instance, it can be argued that example 1 can be formalized more accurately (assuming some appropriate treatment of quantifiers) as:

1. $\mathrm{O}\neg\exists \mathrm{x} F\mathrm{x}$
2. $\forall \mathrm{x}(F\mathrm{x} \rightarrow \mathrm{O}W\mathrm{x})$
3. Fa

(3) says that a is an object which is a fence. From this set of sentences we can derive $\mathrm{O}\neg Fa \wedge \mathrm{O}Wa$, but this is not inconsistent. A similar treatment might be devised for example 4, with more intricate devices perhaps.

However, this does not solve the problems. Note first that in this translation we still have the pragmatic oddity that in all ideal worlds a is not a fence but is painted white. Moreover, the example can easily be modified into a case of contradicting obligations — it is sufficient to add a second level of contrary-to-duty requirements stating that if a fence is not white then it should be black. Here we have an example resembling the examples 1 and 4 but where it is not at all obvious how attention to logical form will circumvent the problem.

Must we now accept that the natural-language examples are inconsistent? In our opinion the key issue here is whether the 'oughts' in statements (1) and

(2) of these examples really are the same: if they are, then the natural language version itself is inconsistent, but not otherwise. Now, whatever one may think of the holiday cottage regulations, there is in our view nothing strange or contrived about the sentences (1) and (2) of example 4. There is at least one plausible interpretation in which the oughts are of different types: the first is a 'normal' ought, stating what holds in ideal circumstances, but the second 'ought' seems to presuppose a certain context or point of view, viz. the one in which the children are on the road in spite of the fact that they should not be there and the question is what should hold given this state of affairs. (A similar view is expressed, although not formalised, by Hilpinen [Hilpinen 93, p96]. This way of looking at CTD obligations is also suggested by Lewis [Lewis 74]. He uses it to motivate a semantic framework for a class of dyadic deontic logics but does not develop its application to the analysis of CTD structures.) The point we want to stress is that when the children are on the road *both* kinds of obligations hold. In section 5 we shall propose a formal account of this way of consistently reading timeless 'ought-to-be' CTD structures. Before that we will investigate the adequacy of another approach, proposed recently.

4 Are nonmonotonic methods sufficient?

Are contrary-to-duty structures just a special case of the kinds of defeasible reasoning structures examined extensively in the past 15 years? This question arises naturally from the above discussion, since the main problem was how to deal with conflicting primary and secondary obligations, while one of the virtues of nonmonotonic logics is that they are intended to cope with conflicting information. Indeed at least two suggestions along these lines have already appeared in the literature, [Ryu & Lee 91] and [McCarty 94]. Let us look again at example 1:

1. There must be no fence.
2. If there is a fence, it must be a white fence.

One plausible reading is that (1) has been formulated as a defeasible rule, understood to be subject to exceptions, and that (2) takes effect in these exceptional circumstances: there is no conflict, because the two rules do not apply in the same circumstances. Another possible reading — which can be regarded as a special case of the first — is that (2) itself expresses an exception to (1): on this reading the problem of inconsistency is resolved by regarding the exceptional rule (2) as *defeating* the general rule (1) in the circumstances in which they both apply. And a natural way of formalizing this reading is to adopt or adapt some suitable formalism for nonmonotonic reasoning.

Is this, then, a proper way of dealing with the problem of consistency of CTD structures? In our view it is not. A crucial effect of letting the primary obligation be defeated by the secondary obligation is that the primary one cannot be violated, since it is simply not applicable to the situation. Clearly this is not what is intended if (2) is meant as a CTD rule of (1). Consider the following extension of example 1.

1. There must be no fence.
2. If there is a fence, it must be a white fence.
3. If the cottage is by the sea, there may be a fence.

Assume that (2) is intended as a CTD obligation of (1) and (3) as an exception to (1), and consider first the case of someone who has a fence because the cottage is by the sea. Then (1) has not been violated since the rule does not apply in these circumstances, while whether (2) has been violated depends on whether this rule is intended to take effect only if (1) has been violated or also in other circumstances in which there is a fence. Now what about someone whose cottage is not by the sea and who has a fence to keep children out? Then the conclusion is different: certainly (1) has been violated and if the fence is not white then (2) has also been violated. One of the key elements of our proposal in the next section will be to retain the possibility of expressing the distinction between these various situations.

The confusion of CTD and defeasible structures is nevertheless quite understandable, since many examples can be read either way. For instance, example 1 might also be read as 'There must be no fence, unless it is a white fence'. Similarly the natural-language version of the CTD rule in example 4 is ambiguous also.

5 A semantic framework for CTD modalities

We now formalise the suggestion of section 3 to view CTD obligations as obligations pertaining to, or pre-supposing, a certain context, in which a primary obligation is already violated. The main requirement concerns the relation between the various 'levels' of obligations. As already indicated, we regard conflicting primary and secondary obligations as consistent because they are obligations of different kinds. However, conflicting primary and secondary obligations are consistent only if they 'belong together' in some sense; secondary obligations should still be inconsistent with conflicting 'unrelated' obligations. In example 4, for instance, the CTD obligation to keep left when cycling in the street should surely be inconsistent with a traffic regulation making it obligatory to keep to the right. Most of the detail will be concerned with capturing the required notion of 'relatedness'.

Our proposal will be constructed initially as an extension of SDL, described in section 2. We augment the language by adding, for every formula B, a modal operator O_B, standing for 'obligatory pre-supposing the (sub-ideal) context B': the expression $\mathrm{O}_B A$ is intended to be read as 'there is a secondary obligation that A given that, or pre-supposing, the sub-ideal context B', or 'given that B, which is a violation of some primary obligation, there is a secondary, compromise obligation that A'.

To capture the semantics we define a new function $dc : \mathrm{Pow}(W) \times W \to \mathrm{Pow}(W)$ of contextual deontic ideality: $dc(\|B\|, w)$ picks out those worlds which are the ideal (perhaps: best) alternatives of w given the sub-ideal context $\|B\|$.

The truth conditions of $\mathrm{O}_B A$ are defined as:

$$\models_w \mathrm{O}_B A \text{ iff } dc(\|B\|, w) \subseteq \|A\|$$

$\mathrm{P}_B A$ is defined as $\neg\mathrm{O}_B \neg A$.

In order to deal with multiple levels of CTD rules we redefine $\mathrm{O}A$ as a boundary case:

$$\mathrm{O}A =_{df} \mathrm{O}_\top A$$

Defined in this way, $\mathrm{O}_B A$ is a form of relative necessity, and its logic is an instance of what Chellas calls a 'normal conditional logic' [Chellas 80, Ch10]. However, we want to stress that the expression $\mathrm{O}_B A$ is not to be read as a conditional (primary) obligation. In this paper a conditional obligation 'if B then it ought to be the case that A' is represented by an expression of the form $B \Rightarrow \mathrm{O}A$. The expression $\mathrm{O}_B A$, in contrast, represents a particular kind of obligation. There is no meaningful sense, in particular, in which the obligation $\mathrm{O}A$ can be detached from the expression $\mathrm{O}_B A$. We should add that when representing CTD structures in examples, we will always have formulas of the form $B \Rightarrow \mathrm{O}_B A$; this expresses that the secondary obligation $\mathrm{O}_B A$ holds in circumstances B. One might ask why, if the context B always appears in the antecedent of a conditional (secondary) obligation, we do not compress $B \Rightarrow \mathrm{O}_B A$ into a single conditional form. The answer is that we would then want to detach a secondary obligation from this conditional when B holds: $\mathrm{O}_B A$ is that secondary obligation.

Note also that an expression of the form $\neg B \Rightarrow \mathrm{O}_B A$ has no intuitive meaning in this framework. We have considered strengthening the truth conditions for $\mathrm{O}_B A$ as follows

$$\models_w \mathrm{O}_B A \text{ iff } w \in \|B\| \text{ and } dc(\|B\|, w) \subseteq \|A\|$$

which would validate

$$\mathrm{O}_B A \rightarrow B$$

We have not done so because we can see no clear advantage to this complication for present purposes.

We turn now to the requirements stated above for relating the various 'levels' of obligations. These are addressed by imposing conditions on dc. We adopt the following, for any w, P, Q and R ($-P$ stands for $W - P$).

(i) $dc(Q, w) \subseteq f(w)$

(ii) $dc(Q, w) \neq \emptyset$ if $Q \neq \emptyset$

(iii) $dc(Q, w) \not\subseteq dc(Q \cap R, w)$ only if $dc(Q, w) \cap R = \emptyset$

(iv) If $dc(Q, w) \subseteq P$ and $P \cap (Q \cap R) \cap f(w) \neq \emptyset$
and $(Q \cap -P) \cap f(w) \not\subseteq R$ then $dc(Q \cap R, w) \subseteq P$

Condition (i) gives $\mathrm{O}_{B}\, A \rightarrow \Diamond A$ and (ii) makes the logic for every individual operator O_{B}, where B is consistent, of type KD. The possibility that B is inconsistent ($\|B\|$ is empty) can be eliminated but we leave it as an uninteresting boundary case because it can cause no difficulties in practice. More interesting are the conditions (iii) and (iv), each of which captures part of the idea that the contextually ideal worlds should resemble the primarily ideal worlds as closely as possible; or, in the case of multiple CTD levels, that a more specific version of a context should still measure up to the standards in the original context as far as this is possible.

Condition (iii) imposes for any context $Q \cap R$ a lower bound on the set of contextually ideal worlds. We need to prevent the introduction of new CTD obligations for a context $Q \cap R$ simply by ignoring some of the worlds that are ideal in context Q. We want to say that a world which is ideal in context Q is still ideal in a more specific context $Q \cap R$ unless this more specific context itself implies a violation in the original context, i.e. $dc(Q, w) \subseteq dc(Q \cap R, w)$ unless $dc(Q, w) \cap R = \emptyset$. Another way of reading this condition is to think of $Q \cap R$ as a more specific version of the *same* context Q when $dc(Q, w) \cap R \neq \emptyset$ and as a more specific but *different* (CTD) context when $dc(Q, w) \cap R = \emptyset$. Condition (iii) results from making making this latter case the only exception to the general requirement that $dc(Q, w) \subseteq dc(Q \cap R, w)$.

Condition (iv) is more complicated, for which reason we motivate it in stages. We need to impose also an upper bound on the contextually ideal worlds, to prevent obligations disappearing simply by regarding some worlds as ideal for no reason. The basic form of the condition will be

$$dc(Q \cap R, w) \subseteq dc(Q, w)$$

or equivalently, the requirement that, for all P:

$$\text{if } dc(Q, w) \subseteq P \text{ then } dc(Q \cap R, w) \subseteq P$$

This states that all primary obligations are also secondary, which of course is not what we want. We need to qualify this basic form in two ways.

The first qualification reflects the intuitive motivation for introducing contexts: a more specific context $Q \cap R$ should still measure up to the standards of context Q, as long as compliance with any particular obligation $\mathrm{O}P$ remains possible in $Q \cap R$, i.e. on condition that $P \cap (Q \cap R) \cap f(w) \neq \emptyset$. The second qualification captures part of what it means for primary and secondary obligations to 'belong together': it is possible that the context $Q \cap R$ already covers the case of non-compliance with a standard P of context Q; this is the case when the context $Q \cap -P$ is already a special case of context $Q \cap R$, i.e. when

$$(Q \cap -P) \cap f(w) \subseteq (Q \cap R) \cap f(w)$$

Putting these conditions together and re-writing in equivalent form yields the condition (iv) as presented above.

We now investigate some of the valid formulas. For simplicity we start with the case of only one level of CTD rules, i.e. the special case where we have

$Q = W$ in conditions (iii) and (iv). To see the validity of formulas depending on (iii) it is helpful to note that this condition is equivalent to

$$\text{If } dc(Q, w) \cap P \neq \emptyset \text{ and } dc(Q, w) \cap R \neq \emptyset \text{ then } dc(Q \cap R, w) \cap P \neq \emptyset$$

Let us now see on what conditions deontic expressions can be transported upward and downward between the primary and the secondary level.

From (iii) we have

$$(\mathrm{P}A \wedge \mathrm{P}B) \rightarrow \mathrm{P}_B A$$

or equivalently

Up: $\quad \mathrm{P}B \rightarrow (\mathrm{O}_B A \rightarrow \mathrm{O}A)$

This formula says that secondary obligations are also primary ones on the condition that the context of the secondary obligation is itself permitted. It is perhaps helpful to look at an equivalent formulation:

Ctd: $\quad \neg\mathrm{O}A \rightarrow (\mathrm{O}_B A \rightarrow \mathrm{O}\neg B)$

Here the expression $(\mathrm{O}_B A \rightarrow \mathrm{O}\neg B)$ is the core of the intuitive reading suggested for the O_B operator: if A is obligatory in a sub-ideal context B then there must be a (primary) obligation $\mathrm{O}\neg B$ that makes B sub-ideal. But this holds only if the obligation is not itself primary, which is the reason for the condition $\neg\mathrm{O}A$. Without this condition the system would collapse into absurdity: for example, the following would be valid

$$\mathrm{O}_\top A \rightarrow \mathrm{O}\neg\top$$

which yields $\neg\mathrm{O}_\top A$, i.e. $\neg\mathrm{O}A$ for any A.

Since Up without its antecedent does not hold, the formula

$$(\mathrm{O}A \wedge \mathrm{O}_{\neg A} B) \rightarrow \mathrm{O}B$$

is *not* valid, not even if $\Diamond(A \wedge B)$ holds. This prevents pragmatic oddities, as will be illustrated in section 6.

We turn now to condition (iv), which yields

Down: $\quad (\Diamond(A \wedge B) \wedge \neg\Box(\neg A \rightarrow B)) \rightarrow (\mathrm{O}A \rightarrow \mathrm{O}_B A)$

This formula should be read with care. What it says is that a primary obligation is also a secondary one if the context B leaves compliance with the primary obligation open, i.e. if $(A \wedge B)$ is possible, and if violation of the primary obligation does not necessarily put us into the context, i.e. if $(\neg A \wedge \neg B)$ is possible. As will be illustrated by examples in the next section, this ensures that, in the case of violation, obligations not related to the violation still hold. To see the significance of Down, notice that it implies the formula

$$(\mathrm{O}A \wedge \mathrm{O}_B \neg A) \rightarrow (\Box(B \rightarrow \neg A) \vee \Box(\neg A \rightarrow B))$$

This says that if in a certain context a primary and a secondary obligation conflict, then these obligations must 'belong together' or be logically related, in the following sense: either the context itself necessarily implies a violation of the primary obligation, or violation of the primary obligation already necessarily puts us in the context.

The logic also supports a very limited form of reasoning with contexts: the semantics validates the following inference rule

$$\frac{A \leftrightarrow B}{\mathrm{O}_{A} C \leftrightarrow \mathrm{O}_{B} C}$$

Finally, we record the relation between arbitrary levels of CTD rules. Still holding are the multiple-level versions of Up (Ctd) and Down:

$\mathrm{Up_m}$: $\mathrm{P}_{B} C \rightarrow (\mathrm{O}_{(B \wedge C)} A \rightarrow \mathrm{O}_{B} A)$

$\mathrm{Ctd_m}$: $\neg \mathrm{O}_{B} A \rightarrow (\mathrm{O}_{(B \wedge C)} A \rightarrow \mathrm{O}_{B} \neg C)$

$\mathrm{Down_m}$: $(\Diamond(A \wedge B \wedge C) \wedge \neg\Box((B \wedge \neg A) \rightarrow C)) \rightarrow (\mathrm{O}_{B} A \rightarrow \mathrm{O}_{(B \wedge C)} A)$

And the following valid formula identifies the general conditions in which an obligation in a certain context must be 'related' to a conflicting obligation in a more specific context:

$$(\mathrm{O}_{B} A \wedge \mathrm{O}_{(B \wedge C)} \neg A) \rightarrow (\Box((B \wedge C) \rightarrow \neg A) \vee \neg\Box((B \wedge \neg A) \rightarrow C))$$

6 Examples

We now apply the formalism to some examples. Each is intended to illustrate a particular CTD structure, and they are presented more or less in order of increasing complexity. We begin by showing how multiple levels of CTD obligations are dealt with. We add to example 4 one further rule, saying that 'If the children are cycling on the right of the street, they ought to be cycling on the extreme right'. The example becomes

Example 5 *multiple CTD levels*

1. $\mathrm{O}\neg c$
2. $c \Rightarrow \mathrm{O}_{c} l$
3. $(c \wedge \neg l) \Rightarrow \mathrm{O}_{(c \wedge \neg l)} e$
4. $\Box(l \rightarrow c) \wedge \Box(e \rightarrow c) \wedge \Box(e \rightarrow \neg l)$
5. $c \wedge \neg l$

From (2) and (5) it follows that $\mathrm{O}_{c} l$, which with (4) gives $\mathrm{O}_{c} c$. Then, since the 'blocking condition' $\Box(c \rightarrow c)$ is logically valid, (1) is not transported downwards. Furthermore, from (3) and (5) we have $\mathrm{O}_{(c \wedge \neg l)} e$, which with (4) implies $\mathrm{O}_{(c \wedge \neg l)} \neg l$. Since the blocking condition $\Box((c \wedge \neg l) \rightarrow \neg l)$ is logically valid, $\mathrm{O}_c l$ is also not transported downwards. Hence the set of sentences is not inconsistent, as desired. Finally, observe that $\mathrm{O}_c \neg(c \wedge \neg l)$ holds—it follows from $\mathrm{O}_c l$ by consequential closure—but $\mathrm{O}_{(c \wedge \neg l)} \neg c$ does not; in other words, in the context c the context $(c \wedge \neg l)$ is forbidden but not vice versa, which shows that $(c \wedge \neg l)$ is a CTD context of c and not the other way around.

Example 6

This example shows how pragmatic oddities are eliminated, and illustrates the transportation of unrelated primary obligations. Assume we have

1. There must be no dog.
2. If there is a dog, there must be a warning sign.
3. There must be no vicious animals.

A Rotweiler is both a dog and a vicious animal, not all dogs are vicious animals, and not all vicious animals are dogs. For convenience of formalising the example only, we assume the validity of

$$\Box A \rightarrow A$$

We formalise the example as

1. $\mathrm{O}\neg d$
2. $d \Rightarrow \mathrm{O}_d s$
3. $\mathrm{O}\neg v$
4. $\Box(r \rightarrow d) \wedge \Box(r \rightarrow v)$
5. $\neg\Box(d \rightarrow v) \wedge \neg\Box(v \rightarrow d)$
6. r

From (6) and (4) we have d and hence $\mathrm{O}_d s$ from (2). But the transportation up $\mathrm{O}_d s \rightarrow \mathrm{O}s$ is blocked by (1), which is equivalent to $\neg\mathrm{P}d$. Therefore we cannot derive the pragmatic oddity that in all ideal worlds there is both no dog and a sign warning that there is a dog.

Furthermore, $\mathrm{O}\neg d$ does not transport down to $\mathrm{O}_d \neg d$ because the blocking condition $\neg\Diamond(d \wedge \neg d)$ is logically valid. Also, since (4) implies $\Box(\neg d \rightarrow \neg r)$, $\mathrm{O}\neg d$ implies $\mathrm{O}\neg r$; but that obligation does not transport down to $\mathrm{O}_d \neg r$, since (4) implies $\Box(r \rightarrow d)$. However, $\mathrm{O}\neg v$ does transport to $\mathrm{O}_d \neg v$, since both $\Diamond(\neg v \wedge d)$ and $\neg\Box(v \rightarrow d)$ follow from (5). Finally, from $\mathrm{O}_d \neg v$ we have $\mathrm{O}_d \neg r$ since (4) implies $\Box(\neg v \rightarrow \neg r)$.

In summary, even in the sub-ideal context where there is a dog, there is still a secondary obligation that there should be no Rotweiler, because there is an obligation that there should be no vicious animal. We might say that not only are the right obligations transported down from the primary level, but also that they are transported for the right reasons.

Example 7 *the considerate assassin*

This example concerns a conflict between a secondary obligation and an (unrelated) primary obligation. It will also help in the analysis of Belzer's Gorbachov-Reagan example, to be discussed next. A Mafia assassin is told:

1. You should not kill the witness.
2. If you kill the witness, you should offer him a cigarette.

Furthermore, we know that both killing without offering a cigarette and offering a cigarette without killing are possible, and we have the moral rule

3. You should not offer cigarettes.

Finally we assume that the assassin kills the witness. (We present an 'ought-to-be' variant of this example as part of the next example, example 8.)

If we interpret (2) as a CTD obligation of (1), then a natural formalisation is

1. $\mathrm{O}\neg k$
2. $k \Rightarrow \mathrm{O}_k\, c$
3. $\mathrm{O}\neg c$
4. $\Diamond(k \wedge \neg c) \wedge \Diamond(c \wedge \neg k)$
5. k

From (2) and (5) of the formalisation we have $\mathrm{O}_k\, c$; but since (3) transports down to $\mathrm{O}_k\, \neg c$ that means that this set is inconsistent. Is this acceptable? In our opinion it is precisely what we want: what is crucial here is that (2) is not a CTD rule of (3) but of (1), for which reason $\mathrm{O}_k\, c$ and $\mathrm{O}\neg c$ are unrelated obligations. Now one may ask how this conflict should be resolved and, of course, one plausible option is to regard (2) as an exception to (3) and to formalize this with a suitable nonmonotonic defeat mechanism. But it is important to note that resolution of the conflict is a separate issue, which has nothing to do with the CTD aspects of the example.

Example 8 *the Reykjavik scenario*

The following example, discussed by [Belzer 87] and [McCarty 94], concerns an instruction to officials accompanying Reagan and Gorbachov during their Reykjavik summit, on telling them a certain secret.

1. Don't tell Reagan.
2. Don't tell Gorbachov.
3. If you tell Reagan, tell Gorbachov.
4. If you tell Gorbachov, tell Reagan.

Let us concentrate on the most problematic case, called by McCarty the 'parallel scenario', in which some official has simultaneously told the secret to Reagan and Gorbachov. For an 'ought-to-be' variant, consider, say, the following set of fashion guidelines (the Paris-Milan scenario?):

1. The trousers should not be red.
2. The jacket should not be green.
3. If the trousers are red, the jacket should be green.
4. If the jacket is green, the trousers should be red.

In the case where someone wears both red trousers and a green jacket, which of these fashion rules are violated, if any?

Formalisation is not easy since here we have an example whose natural-language reading is highly ambiguous. One reading is that (3) and (4) are straightforward exceptions to, respectively, (2) and (1). In this reading it is plausible to say that telling them both (or wearing both) does not violate any of the obligations. However, to most eyes it is more natural to interpret (3) and (4) as CTD rules of, respectively, (1) and (2) and in these readings telling both Reagan and Gorbachov (or wearing both red and green) should imply some violation at the primary level.

A formalization in the style of the previous examples is

1. $\mathrm{O}\neg r$
2. $\mathrm{O}\neg g$
3. $r \Rightarrow \mathrm{O}_r\, g$
4. $g \Rightarrow \mathrm{O}_g\, r$
5. $\Diamond(r \wedge \neg g) \wedge \Diamond(g \wedge \neg r)$
6. $r \wedge g$

Interestingly, this contains two instances of the previous example, the 'considerate assassin': to see this, drop either (1) or (2). Now if without (1) or (2) we already have an inconsistency, then surely the same holds with both of them included. This disagrees with the usual intuitive opinion that the natural-language instructions are not only consistent, but consistent with the fact that r and g both hold. However, because of the relation with the considerate assassin example the same analysis applies: we must conclude that in at least one plausible reading the instructions are inconsistent with the fact that someone wears red trousers and a green jacket (or with the fact that Reagan and Gorbachov are both told the secret). Again the question of how to restore consistency is a separate issue, which can be addressed in the same way as in the considerate assassin; only now there is one reasonable extra constraint, viz. that the two conflicts are interpreted symmetrically: for example, if (3) is regarded as an exception to (2) then (4) should be regarded as an exception to (1).

Another plausible interpretation of the example can be illustrated by reference to McCarty's treatment [McCarty 94]. McCarty claims that he can account for a violation at the primary level in his nonmonotonic deontic logic. In his analysis the example gives rise to two mutually exclusive 'justifiable presumptions' (a concept roughly comparable to 'extensions' in default logic): one in which only (1) has been violated and one in which only (2) has been violated. However, in our opinion this outcome is only adequate for a particular interpretation of the example, viz. for a mixed exception/CTD reading 'Don't tell Reagan (Gorbachov) unless *you violate your obligation* not to tell Gorbachov (Reagan)'. In our framework this interpretation can be formalised by changing (3) and (4) into

3′. $(r \wedge \mathrm{O}\neg r) \Rightarrow \mathrm{O}_r\, g$

4′. $(g \wedge \mathrm{O}\neg g) \Rightarrow \mathrm{O}_g\, r$

(It is easy to check that a similar reformulation makes no difference to the analysis of the previous examples discussed.)

This reformulation also gives an inconsistency when r and g hold, which again could be resolved by some nonmonotonic defeat mechanism formalising an exception interpretation; only this time a symmetric interpretation is not available, which accounts for McCarty's outcome in terms of multiple extensions.

After two inconsistent CTD readings it is natural to ask whether our framework can also express consistent CTD readings of the example. The answer is that it can. One natural way of reading the instructions is as: 'Don't tell either; but if you tell one of them, then also tell the other'. We represent this as

1. $\mathrm{O}\neg(r \vee g)$
2. $(r \vee g) \Rightarrow \mathrm{O}_{(r \vee g)}\, (r \wedge g)$
3. $r \wedge g$

We cannot transport (1) down to $\mathrm{O}_{(r \vee g)}\, \neg(r \vee g)$, since the blocking condition $\Box((r \vee g) \rightarrow (r \vee g))$ is logically valid; therefore, no inconsistency arises. Furthermore, if we check the violations, we see that the primary obligation has been violated but the secondary one has been obeyed.

Notice that with this reading, Belzer's Gorbachov-Reagan scenario has exactly the same logical form as the 'gentle murderer'/'white fence' examples: the only ingredient not given explicitly is just

$$\Box((r \wedge g) \rightarrow (r \vee g))$$

Example 9 *the Chisholm paradox*

Since it was Chisholm's paradox which first drew attention to some problems in formalising CTD rules, it is natural to ask which of the problems discussed so far turn up in this paradox. In accordance with the topic of this paper we will discuss a timeless, or 'parallel', 'ought-to-be' version, formalised in SDL.

1. $\mathrm{O}\neg d$ (There must be no dog)
2. $\neg d \Rightarrow \mathrm{O}\neg s$ (If there is no dog, there must be no warning sign)

3. $d \Rightarrow \mathrm{O}s$ (If there is a dog, there must be a warning sign)

4. d (There is a dog)

Generally, the problem of the Chisholm paradox is described as the task of finding a representation of the example which is consistent and in which no premise is logically implied by another premise. In addition, we have identified a third requirement: the formalisation may not contain pragmatic oddities. In our opinion the problem of logical dependence — in this case of (2) and (4) — is not a problem of deontic reasoning specifically but of conditionals in general, and can be dealt with by taking ⇒ to be some suitable strict or counterfactual conditional. This leaves the issues of consistency and pragmatic oddity.

The SDL formalisation just given is consistent; therefore the question arises of what exactly has made this paradox the subject of so many discussions in the literature. If we look at the example more closely, then we see that the consistency problems only arise if we regard 'There ought not to be a warning sign' as derivable from (1) and (2). This derivation is usually called 'deontic detachment'. In SDL it can be captured by

$2'$. $\mathrm{O}(\neg d \rightarrow \neg s)$

but this makes the example inconsistent.

A discussion of the validity of deontic detachment is beyond the scope of this paper. We confine ourselves to the observation that these consistency problems are more problems of deontic detachment than of CTD structures; therefore it might be better to call the Chisholm paradox the 'Deontic detachment paradox' instead of the 'Contrary-to-duty paradox'.

Finally, even without deontic detachment this example contains a problem, since in SDL we still have the pragmatic oddity that in all ideal worlds there is both no dog and a sign warning that there is a dog. But, as already shown in example 6, in our proposal this can be fixed by replacing (3) by

$3'$. $d \Rightarrow \mathrm{O}_d s$

7 Evaluation and further work

One purpose of the previous section is to demonstrate that the semantic framework constructed in section 5 exhibits some very desirable properties and can give useful insights in the analysis of CTD structures. However, it is also the case that the framework as presented contains a flaw, and requires some adjustment.

Whatever its other properties, a framework for CTD reasoning must be able to deal with examples such as the following one, being the gentle murderer:

1. $\mathrm{O}\neg k$
2. $k \Rightarrow \mathrm{O}_k g$
3. $\Box(g \rightarrow k) \wedge \neg\Box(k \rightarrow g)$
4. k

Although we have shown that the framework gives a satisfactory analysis of the apparent conflict between (1) and (2), we now illustrate that there is a problem.[2] It concerns the model condition (iv) and the transportation principle Down which this validates.

From (1) and (3) we can derive $\mathrm{O}\neg g$, which is reasonable. But suppose that there is *another* primary obligation, $\mathrm{O}c$ say, unrelated to $\mathrm{O}\neg k$ in any way. Then, since O is normal, we also have $\mathrm{O}(\neg g \wedge c)$. Transportation of *this* obligation by Down is not blocked, since we have both $\Diamond(k \wedge \neg g \wedge c)$ and $\neg\Box((g \vee \neg c) \rightarrow k)$. Hence we have $\mathrm{O}_k(\neg g \wedge c)$ from which $\mathrm{O}_k \neg g$, which is inconsistent with the secondary obligation $\mathrm{O}_k g$, derivable from (2) and (4). For convenience we shall refer to this as the problem of 'the irrelevant obligation'.

Evidently we need a better way of defining when a primary and secondary obligation are related. Model condition (iv) is intended to capture that notion but, as the example shows, it is inadequate. To see what is required, consider a primary obligation $\mathrm{O}A$ and a CTD obligation $\mathrm{O}_{\neg A}\neg A$. Obviously, these two obligations are related, but the concept of relatedness must also capture two further situations. Firstly, it must take into account that there may be obligations in the premises that have $\mathrm{O}A$ as a consequence. These obligations must be related to the context $\neg A$, since otherwise they would be transported downwards and imply $\mathrm{O}_{\neg A}A$, yielding inconsistency with the CTD obligation $\mathrm{O}_{\neg A}\neg A$. Secondly, some primary obligations will be consequences of $\mathrm{O}A$ and other primary obligations, and these derived obligations must also be related to the context, otherwise problems as in the problem of the 'irrelevant obligation' will occur. However, we must not qualify too many obligations as related to the context, otherwise examples which are intuitively inconsistent become consistent. Some notion of minimality is required.

In this respect CTD reasoning resembles features of theory revision and defeasible reasoning. Specifying which obligations are transported downwards from a context B to a context $(B \wedge C)$ is similar to specifying which formulas A remain in a set of beliefs after contraction by $\neg C$. It is also similar to specifying the conditions under which 'strengthening of the antecendent'

$$\frac{B > A}{(B \wedge C) > A}$$

is a valid rule for a (defeasible) conditional $B > A$.

Correspondences between belief revision, defeasible reasoning and the logic of conditionals are well documented: see in particular Makinson's account in which he identifies five 'faces of minimality' [Makinson 93]. We might say that CTD reasoning presents a sixth 'face of minimality', though how precisely it relates to the forms discussed by Makinson remains a topic for future investigation.

One way of tackling the problem of relatedness within the present framework is to weaken the logical closure properties of the O_B operators, in order to make irrelevant obligations as in our example logically underivable. The literature on deontic logic already contains many independent arguments for the inadequacy of normal systems of modal logic, and the above example could be welcomed as

[2] This is a simplified version of a construction pointed out by Leon van der Torre.

a further illustration of this inadequacy. One obvious possibility is to move to a non-normal system which does not contain the rule ROM:

$$\text{ROM.} \quad \frac{A \to C}{\mathrm{O}_B A \to \mathrm{O}_B C}$$

Abandoning the rule ROM does eliminate the problem of the 'irrelevant obligation'. However, there are also strong reasons to believe that some form of consequential closure should be retained: someone who is told not to kill must surely be able to infer that he or she ought not to kill by strangling, say. A less drastic, and also standard, way of restricting the closure properties of O_B is to retain ROM but reject the scheme OC:

$$\text{OC.} \quad (\mathrm{O}_B A \wedge \mathrm{O}_B C) \to \mathrm{O}_B (A \wedge C)$$

(See for instance the discussion by Chellas of what he calls 'minimal deontic logic' [Chellas 80, pp201–202].) Abandoning the scheme OC weakens the O_B operators sufficiently that the problem of the 'irrelevant obligation' is eliminated, and it retains enough of consequential closure to validate the inference from 'You ought not to kill' to 'You ought not to strangle'.

These adjustments can be made very straightforwardly using standard semantical devices. However, we feel that tackling the problems by imposing restrictions on the properties of O_B under which the problems disappear is not entirely satisfactory. It seems to us that it should be possible to construct a framework for CTD reasoning that is applicable independently of particular properties of the deontic operators. With this general aim in mind we are investigating another way of capturing relatedness, based on the idea that obligations are related when they originate from the same source.[3] And the most direct way of expressing sources, instead of introducing dyadic obligation operators as in this paper, would be to extend the language of SDL by indexed operators O_i where the index i is used in place of what we have called the 'context'. This index can be seen as a name for the obligation or as an identifier for a rule or regulation from which the obligation is derived. The attraction of this proposal is that with the indices we could specify explicitly which obligation is CTD to which other, just as we have tended to do when presenting informally the intended readings of examples in this paper. However, formalisation still needs further research: the main problem concerns how to express consistency requirements across different labels.

Another strategy, which we are currently investigating, is to take a completely different semantical perspective. A view that naturally suggests itself is that of preference relations among worlds: order all possible worlds as to how well they fulfil the primary obligations, and define the CTD obligations of an obligation $\mathrm{O}A$ as those formulas which are true in all the best worlds of those in which $\mathrm{O}A$ has been violated.[4] It might even be felt that approaching the

[3] Note that this might also provide intuitive motivation for rejecting the scheme OC.

[4] McCarty also mentions preference orderings as a natural way of analysing the Gorbachov-Reagan example but then rejects it — in our opinion for the wrong reasons, since his argument seems to shift away from the 'parallel' reading he was discussing.

representation of CTD structures from this direction will avoid altogether the kinds of problems we have been addressing in this paper. Proposals for constructing deontic logics in terms of explicit notions of preference are reported in [Hansson 90] and [Brown et al. 93] but neither of these has addressed contrary-to-duty questions and neither proposal seems to be adaptable for our purposes. A systematic account of how semantic analyses for dyadic deontic logics can be based on value structures is presented in [Lewis 74].

This new perspective leads to a system in which conflicting primary obligations are inconsistent, conflicting primary and secondary obligations are consistent if related and inconsistent if unrelated, and pragmatic oddities are avoided. It therefore seems very attractive. And at first sight it seems easy to formalise: just consider for each world the set of obligatory propositions that it satisfies, compare these sets with respect to set inclusion and rank the worlds accordingly. However, if for the reasons just given we want to retain ROM, then problems again emerge. Assume that the premises are $\mathrm{O}A$ and $\mathrm{O}B$: then intuitively one would feel that a world w_1 satisfying $\neg A \wedge B$ is better than a world w_2 satisfying $\neg A \wedge \neg B$, for which reason $\mathrm{O}_{\neg A} B$ would hold. However, with the rule ROM, $\mathrm{O}A$ implies $\mathrm{O}(A \vee \neg B)$ and this obligation is satisfied by w_2 but not by w_1. In general, with consequential closure of O the preference ordering will collapse: ideal worlds will be preferred to all worlds where there is any violation, and any two worlds that are not ideal will be incomparable. Here, too, an account is needed of when obligations are related.

It is interesting to note that similar problems occur in intensional accounts of defeasible conditionals. Veltman, for instance, constructs a logic for defeasible conditionals by ordering worlds as to how well they satisfy given default rules [Veltman 91]. His conditional is defined by means of a unary 'normality' operator, and he solves the problem by rejecting consequential closure of 'normally'. Morreau [Morreau 94] attaches to every defeasible conditional a 'regard' and lets conditionals derived from a certain conditional 'inherit' its regard. Then he makes sure that if a given conditional is overridden, all conditionals with the same regard are also overridden. In [Morreau 94] reasoning about defaults *across* regards is impossible. It would be interesting to see how this relates to our remarks above on the possible reasons for rejecting the validity of the scheme OC.

We started the paper by arguing that contrary-to-duty reasoning is not a special case of defeasible reasoning; we have ended by showing that there are nevertheless resemblances between formal accounts of the two kinds of reasoning. These are perhaps most apparent in the respective uses of preference relations on worlds (or contexts): CTD reasoning can be captured by ordering worlds (or contexts) as to how close to *ideal* they are, and defeasible reasoning by ordering them as to how *normal* they are. But note that an essential difference remains: while defeasible conclusions are derived on the basis of the retractable assumption that things are normal, primary obligations are not derived on the retractable assumption that things are ideal. Exploiting these formal similarities while being aware of the remaining differences is the direction of our current research.

Acknowledgements

The bulk of this work was completed at Imperial College while Henry Prakken was supported by ESRC/MRC/SERC Joint Council Initiative project G9212036. The work was also supported by the Portuguese research programme STRIDE, FEDER and JNICT under the project JURAD, and by the ESPRIT Working Group 8319 *ModelAge*. The authors wish to thank Leon van der Torre for pointing out the problem identified in section 7 and Andrew Jones for many illuminating discussions and for suggesting the relevance of [Lewis 74].

References

[Belzer 87] M. Belzer. Legal reasoning in 3-D. *Proc. First International Conference on Artificial Intelligence and Law*, Boston. ACM Press. 1987, 155–163.

[Brown et al. 93] A.L. Brown, Jr., S. Mantha and T. Wakayama. Exploiting the normative aspect of preference: a deontic logic without actions. *Annals of Mathematics and Artificial Intelligence* 9:167–204, 1993.

[Chellas 80] B. Chellas. *Modal logic: An introduction.* Cambridge University Press, 1980.

[Chisholm 63] R.M. Chisholm. Contrary-to-duty imperatives and deontic logic. *Analysis* 24:33–36, 1963

[Forrester 84] J.W Forrester. Gentle murder, or the adverbial Samaritan. *Journal of Philosophy* 81:4:193–197, 1984.

[Hansson 90] S.O. Hansson. Preference-based deontic logic (PDL). *Journal of Philosophical Logic* 19:75–93, 1990.

[Hilpinen 93] R. Hilpinen. Actions in Deontic Logic. In J.-J.Ch. Meyer and R.J. Wieringa (eds.): *Deontic Logic in Computer Science: Normative System Specification.* John Wiley & Sons, Chichester, 1993, 85–100.

[Jones & Pörn 85] A.J.I. Jones and I. Pörn. Ideality, sub-ideality and deontic logic. *Synthese* 65:275–290, 1985.

[Lewis 74] D. Lewis. Semantic analyses for dyadic deontic logic. In S. Stenlund (ed.): *Logical Theory and Semantic Analysis.* D. Reidel, Dordrecht, 1974, 1–14.

[McCarty 94] L.T. McCarty. Defeasible deontic reasoning. *Fundamenta Informaticae* 21:125–148, 1994.

[Makinson 93] D. Makinson. Five faces of minimality. *Studia Logica* 52(3), 1993.

[Meyer 88] J.-J.Ch. Meyer. A different approach to deontic logic: deontic logic viewed as a variant of dynamic logic. *Notre Dame Journal of Formal Logic* 29:1:109–136, 1988.

[Morreau 94] M. Morreau. Prima facie and Seeming Duties. In A.J.I. Jones and M.J. Sergot (eds.): *Proceedings of DEON'94: Second International Workshop on Deontic Logic and Computer Science*, Oslo, January 1994. Complex 1/94, Tano Publishers, Norway.

[Ryu & Lee 91] Y.H. Ryu and R.M. Lee. Defeasible deontic reasoning: a logic programming model. *Proceedings of the First International Workshop on Deontic Logic in Computer Science, DEON-91*, Amsterdam, 1991, 347–363.

[Veltman 91] F. Veltman. Defaults in Update Semantics. Report LP-91-02, Institute for Language, Logic and Information, University of Amsterdam, 1991.

[Åqvist & Hoepelman 81] L. Åqvist and J. Hoepelman. Some theorems about a "tree" system of deontic tense logic. In R. Hilpinen (ed.): *New studies in deontic logic*. Reidel, Dordrecht, 1981, 187–221.