DYADIC DEONTIC LOGIC AND CONTRARY-TO-DUTY OBLIGATIONS

ABSTRACT.

This paper investigates to what extent contrary-to-duty obligations can be represented in dyadic deontic logics of the Hansson-Lewis family, which interpret obligations in terms of a preference ordering on worlds. The Hansson-Lewis systems are extended in two ways. First a notion of alethic necessity is added, which sheds light on the difference between what we have earlier called 'contextual' obligations and conditional obligations (whether defeasible or not) as ordinarily understood. This extension also facilitates a comparison with temporal deontic logics, including the critical observation that the commonly accepted treatment of temporal contrary-to-duty structures neglects some important problems. The second extension is a set of conditions on the preference orderings intended to ensure that non-ideal worlds are ranked according to how well they resemble or measure up to more ideal worlds. The aim here is to establish that the Hansson-Lewis account of obligation must be extended to capture even the most basic features of contrary-to-duty reasoning, that these extensions cannot be undertaken using standard model-theoretic devices, but that there are nevertheless some promising avenues to explore.

1. INTRODUCTION

One of the main issues in the discussion on standard deontic logic (SDL) is the representation of contrary-to-duty (CTD) obligations. A well-known example is Forrester's (1984) paradox of the gentle murderer: it is forbidden to kill, but if one kills, one ought to kill gently. Intuitively, one would feel that these sentences are consistent, but in SDL no (obvious) consistent formalisation is available: assuming that $kill\text{-}gently$ logically implies $kill$, the formalisation

(1) $\mathrm{O}\neg kill$
(2) $\mathrm{O}\ kill\text{-}gently$

is inconsistent, since SDL contains the inference rule of consequential closure:

$$\text{ROM.} \qquad \frac{A \rightarrow B}{\mathrm{O}A \rightarrow \mathrm{O}B}$$

and the valid scheme[1]

$$\text{D.} \qquad \neg(\mathrm{O}A \wedge \mathrm{O}\neg A)$$

The reason why this paradox is so challenging is that some well-studied approaches that work for other paradoxes fail here. It does not help to distinguish between the times of violating an obligation and fulfilling its associated contrary-to-duty obligation (cf. (Åqvist and Hoepelman, 1981;

[1]The names of the schemes in this article are based on those of (Chellas, 1980).

van Eck, 1982)), since here these times are equal. Neither do solutions apply where the condition of a conditional obligation is regarded as a state of affairs and the content of an obligation as an act ((Castañeda, 1981; Meyer, 1988)). Clearly, in the gentle murderer both the condition and the content of the CTD obligation are acts. Moreover, there are variants of the example where both are states of affairs; a holiday cottage regulation could say: there must be no fence around the cottage, but if there is a fence, it must be a white fence (Prakken and Sergot, 1994, 1996).

Another option is to reject the rule ROM. Reasons for giving up this principle have been put forward independently of CTD reasoning, notably in connection with Ross's paradox 'You ought to mail this letter, so you ought to mail it or burn it'. However, we feel that this solution is not adequate. As we remarked earlier in (Prakken and Sergot, 1994, 1996), there are also strong reasons to believe that ROM should be retained, at least in some restricted form: someone who is told not to kill must surely be able to infer that he or she ought not to kill by strangling, say.

Yet another option is to reject the D scheme, a move which has also been suggested for other reasons, viz. as a way to represent moral dilemmas in a meaningful way (cf. e.g. (Horty, 1994)). However, as one of us has defended in (Prakken, 1996), we feel that that aim is better served by embedding a deontic logic validating the D scheme in some suitable non-monotonic logic. In such a logic contradictions do not necessarily trivialise the premises, and thus they provide a way to unify a realistic view on moral dilemmas with a rationality requirement for normative systems that obligations do not conflict.

Contextual obligations

In accord with e.g. (Lewis, 1974), (Jones and Pörn, 1985) and (Tan and van der Torre, 1994) we feel that the cause of SDL's failure to deal with the gentle murderer is different. SDL cannot distinguish between various grades or levels of non-ideality; in the semantics of SDL worlds are either ideal or non-ideal. Yet the expression 'if you kill, kill gently' says that some non-ideal worlds are more ideal than other non-ideal worlds; it says: presupposing that one kills, then in those non-ideal worlds that best measure up to the deontically perfect worlds, one kills gently. In formalising CTD reasoning the key problem is formalisation of what is meant by 'best measure up'.

In (Prakken and Sergot, 1994) and its extended version (Prakken and Sergot, 1996) we gave a first formalisation of our intuitions. The key idea was to interpret obligations as being relative to a *context*. For instance, the obligation to kill gently should be taken to pertain to the context where the killing is taking place. We formalised this by making the deontic modalities dyadic: $\mathrm{O}_B\, A$ says that A is obligatory given, or pre-supposing, or from the point of view of, the context B. $\mathrm{O}_B\, A$ and $\mathrm{O}_C\, \neg A$ can both hold at the same time,

since they pertain to different contexts, or points of view, and one context can be more or less close to ideal than another.

As formulated in (Hilpinen, 1993, p. 96), a context stands for a constellation of acts or situations that agents regard as being settled in determining what they should do. In deciding how to kill, a person takes it for granted that he or she kills. Normgivers, in stating contrary-to-duty obligations, anticipate the choices of context that agents can make. However, it is important to see that the settledness of contexts is subjective, since a normgiver is in no way required to respect a person's choice of context; anything within a context can be designated as forbidden, and anything outside the context as obligatory. This is why a moral code can consistently say: you should not kill, but if you kill, kill gently.

Although we were and still are convinced that the introduction of contexts is the right way to analyse CTD reasoning, the system presented in our earlier paper contained a flaw, and we concluded that further research is needed. The present paper reports on an aspect of that further research. The following section outlines the general idea, after recapturing the basics of our earlier proposal.

2. GENERAL CONSIDERATIONS

In this paper we employ the following notational conventions: capitals A, B and C are metavariables for arbitrary formulas, lower case letters w, v, ..., possibly subscripted, represent worlds, capitals P, Q, R, W stand for propositions in the sense of sets of worlds, and X, Y, Z stand for sets generally. $\|A\|^{\mathcal{M}}$ denotes the set of worlds (of a model $\mathcal{M}$) in which A is true. We leave the model $\mathcal{M}$ implicit where it is obvious from context. Finally, we assume the basic definitions of SDL to be known.

As for terminology, in examples of CTD structures we will often call an obligation and its associated CTD obligation, the 'primary' and 'secondary' obligation, respectively.

Our earlier approach

In making obligations relative to contexts, the main idea of (Prakken and Sergot, 1994, 1996) was to represent a context as a proposition, i.e. as a set of possible worlds, and then to pick out the ideal worlds not only relative to a world but also relative to a set of worlds. To this end we augmented the language of SDL with, for every formula B, a modal operator O_B, standing for 'obligatory from the point of view of the (sub-ideal) context B'. To capture the semantics we defined a function dc of contextual deontic ideality: for any world w and set of worlds Q, $dc(Q, w)$ picks out those worlds that are the best alternatives of w as assessed from the context Q. The truth conditions of

$\mathrm{O}_B\, A$ were defined as:

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

$\mathrm{P}_B\, A$ was defined as $\neg \mathrm{O}_B \neg A$ and $\mathrm{O}A$ as $\mathrm{O}_\top\, A$.

Thus the gentle murderer can be expressed consistently as follows.

(1) $\mathrm{O}\neg kill$
(2) $\mathrm{O}_{kill}\, kill\text{-}gently$

Apart from finding a consistent representation of such examples, another main concern was to state conditions on dc that would make the B-best worlds resemble the best worlds as closely as possible, given that B. Firstly, we wanted to prevent that a CTD context could contain new obligations for no reason: if a CTD context introduces a new obligation it should be to regulate the violation of a 'higher' obligation. The model conditions that imposed bounds on dc validated the following 'Up' principle

Up. $\qquad \mathrm{P}_B\, C \rightarrow (\mathrm{O}_{(B \wedge C)}\, A \rightarrow \mathrm{O}_B\, A)$

More importantly, we also wanted to formulate conditions on dc under which conflicting primary and secondary obligations are consistent. Here it becomes important how exactly we read a contextual CTD obligation, such as 'Given that you kill, offer a cigarette first'. If we read this as saying that in the best of worlds in which you kill, you offer a cigarette first, then intuitively this seems consistent with a primary obligation not to offer cigarettes. The worlds in which you kill are, after all, already non-ideal. But in certain other readings this seems wrong. If the CTD obligation is read as saying 'of the worlds where you kill, in those that resemble as closely as possible the ideal worlds, you offer a cigarette first', this conflicts with a primary obligation not to offer cigarettes. On this reading, a legislator who wants to regulate violation of the obligation not to kill must take account of the primary obligation not to offer cigarettes, since this is also intended to regulate killing contexts. Regulation of norm-violation must still respect other norms that are in force. On the other hand, not all primary obligations have to be taken into account in this way: the primary obligation not to kill can be ignored by the legislator, since the CTD context where you kill already covers its violation. In other words, the context where you kill is *related* to the obligation not to kill.

Accordingly, we let the logic of our earlier paper validate a scheme according to which primary obligations are 'downwards inherited' by unrelated contexts.

Down. $\qquad \varepsilon \rightarrow (\mathrm{O}_B\, A \rightarrow \mathrm{O}_{(B \wedge C)}\, A)$

The actual form of ε is shown later, in section 6.1. For now, the point is that it captured, or so we thought, the notion of relatedness of an obligation to a context. However, we ended our paper with the observation that our system

validated some undesirable inferences. And although we were able to pinpoint a number of ways these problems could be removed, we could not see how finding the right version of Down could be solved by adjusting the conditions on the function dc without further guidance. We need some way of building it up from more basic, simpler, components.

We therefore suggested a different semantical perspective, viz. that of preference orderings on worlds. The idea here is to define an ordering on the set of possible worlds in such a way that, roughly speaking, the more obligations a world satisfies, the better it is. Then the truth of a contextual obligation $\mathrm{O}_B\, A$ can be defined as: A holds in the best of the worlds where B is the case. Our hope was that this view would reflect our intuitions on down inheritance. In this paper we develop this approach, and we will investigate whether our hope was justified.

The relation with dyadic deontic logic

The reader will have noticed the formal similarities between our approaches to contextual obligation and well-known systems of dyadic deontic logic. In particular, 'best of the worlds where B is the case' is the basis for the logics developed by David Lewis (1974), which in turn resemble and generalise the system of Bengt Hansson (1969). We therefore want to develop our new account as a variant of the Hansson-Lewis systems. In doing so, we want to address two main points. The first is of a philosophical nature, viz. an examination of dyadic deontic logics of the Hansson-Lewis type as systems of contextual rather than conditional obligation. This will also help to clarify what we mean by 'context' and the difference between what we are calling 'contextual obligations' and conditional obligations as ordinarily understood. The second point is to make a technical contribution, by investigating how dyadic deontic logics of this type can be augmented to capture our intuitions on upward and downward inheritance of contextual obligations.

The relation with defeasible deontic logic

We should also motivate why this paper appears in a volume on *defeasible* deontic logics. This is for three reasons.

First, some authors, e.g. (Loewer and Belzer, 1983; Alchourrón, 1993), have interpreted Hansson-Lewis systems as candidates for logics of *prima facie*, or defeasible conditional obligations. They have done so since these logics exhibit some of the kinds of properties one would expect of defeasible conditionals. They invalidate, in particular, the principle of strengthening of the antecedent for the dyadic operators: from 'promises ought to be kept' it does not follow that 'promises to do immoral things ought to be kept'. Thus, so it is argued, they seem to capture the idea that *prima facie* obligations

can be defeated in exceptional circumstances. We will argue that this view on these systems, although understandable, is mistaken; it is not *prima facie* obligations that these logics represent.

The second issue concerns a proposal of some to formalise timeless CTD structures using non-monotonic techniques (e.g. (McCarty, 1994; Ryu and Lee, 1996)). In these proposals, in circumstances where a primary obligation is violated, consistency is maintained by regarding the derivation of the primary obligation as somehow blocked by the derivation of the secondary obligation that comes into force.

We here briefly summarise our arguments in (Prakken and Sergot, 1994, 1996) as to why we think this view is incorrect. What it fails to capture is that when the secondary obligation, say to kill gently, is being fulfilled, at the same time the primary obligation not to kill is being violated: violating an obligation in a situation does not make it inapplicable to that situation.

In (Prakken and Sergot, 1994, 1996) we illustrated this with the following example.

(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.

(2) is intended as a CTD obligation of (1) and (3) as an exception to (1).

A person who has a cottage by the sea with a fence does not violate (1), since (1) is defeated by (3): (1) does not apply when the cottage is by the sea. Someone whose cottage is not by the sea and who has a white fence complies with (2) but still violates (1): any fine imposed for violating (1) will have to be paid. A logic that in these circumstances regards (1) as being defeated by (2) cannot express this.

The third connection with defeasibility is that in later sections of the paper we shall argue, not only that the Hansson-Lewis systems must be extended if they are to deal with contrary-to-duty reasoning, but that these extensions apparently cannot be undertaken using standard model-theoretic devices. We shall sketch the outline of a solution which yields a non-monotonic consequence relation, of a kind not unlike those studied in the field of defeasible reasoning, but respecting that secondary obligations do not defeat primary ones.

The structure of this paper

We will develop the discussion as follows. To make the link with the Hansson-Lewis logics, section 3 will present a representative system, a modified version which includes an additional operator for alethic necessity. Consideration of how to interpret this logic leads to a discussion in section 4 of various notions of obligation that have appeared in the literature. In section 5 we re-assess the system in the light of the preceding discussion, with the aim of

providing a more detailed conceptual analysis of contextual obligations and CTD reasoning. Section 6 presents a possible extension of the Hansson-Lewis logic intended to provide some form of up and down 'inheritance'; section 7 identifies its shortcomings and sketches a solution. In section 8 we will assess what we have achieved.

3. HANSSON-LEWIS CONDITIONALS

The basic idea in the Hansson-Lewis account of obligation is that expressions 'Given that B, it ought to be that A' are interpreted as saying that A holds in a chosen subset of the (accessible) B-worlds: these are the 'best' (accessible) B-worlds, as determined by an ordering on worlds representing preferences or the relative 'goodness' of worlds. The idea originates in (Hansson, 1969) and was subsequently developed by several authors, notably Lewis. (Lewis, 1974) presents several different value structures, in addition to orderings on worlds, and also provides a useful comparison with other proposals, including Hansson's.

In this section we present a representative system of the Hansson-Lewis family. Most of the details can be found in (Lewis, 1974), although our version will also be different in several respects. We focus only on preference orderings, and not on the other kinds of value structures considered by Lewis; we make a notational change designed to make the intended reading of the deontic operators more perspicuous; and we isolate and discard some assumptions that are made by Lewis about preference orderings. We also add a new component, viz. an alethic accessibility relation. This is not present in the systems studied in (Lewis, 1974) but it is a standard feature in counterfactual conditionals, which formally are constructed in exactly the same way.

The language is that of propositional logic, augmented with two dyadic deontic operators $\mathrm{O}[B]A$ and $\mathrm{P}[B]A$, meant to be interdefinable as usual: $\mathrm{P}[B]A =_{def} \neg\mathrm{O}[B]\neg A$. (Lewis's notation is $\mathrm{O}(A/B)$ and $\mathrm{P}(A/B)$.) The intended reading of $\mathrm{O}[B]A$ is 'Given that B, it ought to be that A' *in the sense that* A holds in all of the best (accessible) B-worlds. Notice that $\mathrm{O}[\top]A$ ($\top$ any tautology) then says that A holds in all of the best of all worlds, a reading which coincides with that of Standard Deontic Logic (SDL). Accordingly, the expression $\mathrm{O}A$ is used as an abbreviation for $\mathrm{O}[\top]A$. As in (Prakken and Sergot, 1994, 1996), we add to the language two more operators $\Box$ and $\Diamond$, standing for 'necessary' and 'possible' respectively.

Models are structures

$$\mathcal{M} = \langle W, f, \geq^W, V \rangle$$

W is a set of possible worlds and V is a valuation function for atomic sentences in each of the possible worlds. f is a function from W into $\mathrm{Pow}(W)$ representing the alethic accessibility relation: $f(w)$ is the set of worlds in W

accessible from w. The relevant truth conditions are

$$\mathcal{M}, w \models \Box A \text{ iff } f(w) \subseteq \|A\|$$

$\Diamond A$ is defined as $\neg\Box\neg A$. We make no further assumptions about the nature of f at this stage.

A formula A is *true in a model* $\mathcal{M}$ iff $\mathcal{M}, w \models A$ for all w. And A is *valid* iff A is true in all models. Furthermore, for any set Γ of sentences $\mathcal{M}, w \models \Gamma$ iff $\mathcal{M}, w \models B$ for all $B \in \Gamma$. Finally, we will use the following notion of entailment: A set Γ of sentences *entails* a sentence A iff $\mathcal{M}, w \models A$ for all models $\mathcal{M}$ and worlds w such that $\mathcal{M}, w \models \Gamma$.

The main semantical device is $\geq^W$, which is a *preference frame* over W: for each w in W it assigns an ordering $\langle K_w, \geq_w \rangle$ where K_w is a (possibly empty) subset of W and $\geq_w$ is a pre-order (a reflexive and transitive relation) on K_w. In (Lewis, 1974) it is further assumed that each $\geq_w$ is a total ('strongly connected') pre-order, i.e. that, for all w_1 and w_2 in K_w, either $w_1 \geq_w w_2$ or $w_2 \geq_w w_1$. We do not make this assumption. We comment on its significance, and on other features of Lewis's systems, in later discussions. The K_w component provides an extra degree of flexibility but for present purposes it can be discarded; it is sufficient to restrict attention to the case where all accessible worlds are evaluable, i.e. to frames in which $f(w) \subseteq K_w$ for all worlds w in W.

The intended reading of $w_1 \geq_w w_2$ is that world w_1 is *at least as good as* w_2 according to some valuation of worlds, as measured from w. The orderings are indexed by worlds w to allow for the possibility that preferences or measures of goodness may differ from one world to another.

As already indicated, the idea now is that $\mathrm{O}[B]A$ will hold at a world w just in case A holds at all the $\geq_w$-best B-worlds. However, there are alternative ways of formalising this idea, depending on the level of generality required.

Terminology and notation For any pre-order $\geq_w$ on K_w, $>_w$ is the associated strict (irreflexive and transitive) ordering. w_m will be said to be *maximal* under $\geq_w$ in a subset X of K_w iff it is maximal in X under the associated ordering $>_w$, i.e., iff $w_m \in X$ and there is no $w' \in X$ such that $w' >_w w_m$. We use the notation $\max_w(X)$ for the set of elements of X ($X \subseteq K_w$) that are maximal under $\geq_w$.

The truth conditions for $\mathrm{O}[B]A$ are required to capture the idea that the best of the (accessible) B-worlds in which A holds are strictly better than the best of the (accessible) B-worlds in which A does not hold. One way of formalising this is as follows:

(obl) $w \models \mathrm{O}[B]A$ iff there exists some world $w_m \in f(w) \cap \|B \wedge A\|$ such that $w_m >_w w'$ for all $w' \in f(w) \cap \|B \wedge \neg A\|$

(An alternative, that $\mathrm{O}[B]A$ should hold at world w iff *all* worlds in $f(w) \cap \|B \wedge A\|$ are strictly preferable ($>_w$) to all worlds in $f(w) \cap \|B \wedge \neg A\|$ is much too strong a requirement to be useful.)

The truth definition (obl) caters for the possibility that there are infinite sequences of better and better and better worlds. If attention is restricted to frames satisfying the *limit assumption*, in conditional logics also referred to as *stoppering* — frames in which there are no infinite sequences of better and better worlds — or to the slightly more restrictive class of *well-founded* orderings — frames in which $\max_w(X)$ is non-empty for every non-empty subset X of K_w — then the truth conditions may be stated equivalently as follows:

$(\mathrm{obl}_{\max})$ $\quad w \models \mathrm{O}[B]A$ iff $\max_w(f(w) \cap \|B\|) \neq \emptyset$ and $\max_w(f(w) \cap \|B\|) \subseteq \|A\|$

The condition $\max_w(f(w) \cap \|B\|) \neq \emptyset$ appears here, as in (Lewis, 1974), because then the truth definitions (obl) and $(\mathrm{obl}_{\max})$ coincide under the limit/well-foundedness assumption. This makes $\mathrm{O}[B]A$ false for the degenerate case where B is inconsistent ($\|B\|$ is empty) or not 'possible' ($f(w) \cap \|B\|$ is empty).

In what follows we shall tend to refer to the truth definition $(\mathrm{obl}_{\max})$, but this is just to simplify the presentation. The logic of $\mathrm{O}[B]A$ does not change if the limit/well- foundedness assumption is removed, as long as the truth conditions are then stated in the form (obl) to compensate.

With these truth conditions the logic of each $\mathrm{O}[B]$ (for consistent, 'possible' B) is that of SDL. More precisely, $\mathrm{O}[B]A$ is (almost) a 'normal conditional logic' in the terminology of (Chellas, 1980, Ch10). It contains the following rules:

RCOEA. $\quad \dfrac{B \leftrightarrow B'}{\mathrm{O}[B]A \leftrightarrow \mathrm{O}[B']A}$

RCOK. $\quad \dfrac{A_1 \wedge \ldots \wedge A_n \rightarrow A}{\mathrm{O}[B]A_1 \wedge \ldots \wedge \mathrm{O}[B]A_n \rightarrow \mathrm{O}[B]A} \qquad (n > 0)$

Because B can be inconsistent/'impossible' the rule RCOK does not hold for $n = 0$; it holds in a restricted form:

RCON. $\quad \dfrac{A}{\Diamond B \rightarrow \mathrm{O}[B]A}$ $\qquad$ i.e. $\quad$ ON. $\quad \Diamond B \rightarrow \mathrm{O}[B]\top$

Validity of the scheme

OD. $\quad \mathrm{O}[B]A \rightarrow \mathrm{P}[B]A$

follows from the evaluation rule for $\mathrm{O}[B]A$ (without any assumptions about f). Seriality of f (i.e. $f(w)$ is non-empty for all w in W) validates:

P. $\quad \mathrm{O}[\top]\top \qquad$ i.e. $\Diamond\top$

(This may seem a little surprising at first sight but it is really a consequence of the way the truth conditions are defined. A variant of SDL could be constructed

in similar style. Define $w \models \mathrm{O}A$ iff $d(w) \neq \emptyset$ and $d(w) \subseteq \|A\|$ where d is the usual deontic accessibility relation of SDL. Validity of $\mathrm{O}A \rightarrow \mathrm{P}A$ follows without any assumption about seriality of d. But then $\mathrm{O}\top$ is not valid: it is validated by adding the assumption that d is serial.)

It can be seen that this Hansson-Lewis system is a generalisation of SDL. Semantically, notice that the deontic accessibility relation $d(w)$ of SDL can be defined as $d(w) =_{def} \max_w(f(w))$; then $d(w) \subseteq f(w)$ for all w in W, and d will be serial if f is serial. (Furthermore, every deontic accessibility relation can be so characterised.)

Some further properties of $\mathrm{O}[B]A$ are forthcoming without any further assumptions about the orderings $\geq_w$ or the nature of f. For instance, the following, named as in (Chellas, 1980), is valid in all models

ODIL. $\quad \mathrm{O}[B]A \ \wedge \ \mathrm{O}[C]A. \rightarrow \mathrm{O}[B \vee C]A$

ODIL will be referred to later in connection with 'Up inheritance' principles.

Since $w \models \Box(B \rightarrow C)$ iff $f(w) \cap \|B\| \subseteq f(w) \cap \|C\|$, it is easy to verify that the logic contains the valid scheme:

O□A. $\quad \Box(B \rightarrow C) \rightarrow (\mathrm{O}[B]A \rightarrow \mathrm{O}[B \wedge C]A)$

and also:

□ON. $\quad \Box A \rightarrow (\Diamond B \rightarrow \mathrm{O}[B]A)$

which implies, for instance:

□OCK. $\quad \Box(A \rightarrow C) \rightarrow (\mathrm{O}[B]A \rightarrow \mathrm{O}[B]C)$

□ON together with OD gives:

$$\mathrm{O}[B]A \rightarrow \Diamond(B \wedge A)$$

which we shall refer to as the 'ought implies can' property.

Of particular importance for the interpretation of the Hansson-Lewis family of dyadic deontic logics are the following properties.

Since $\max_w(f(w) \cap \|B\|) \subseteq \|B\|$, $\mathrm{O}[B]B$ holds for every 'possible' B, i.e. the following scheme is valid:

OI. $\quad \Diamond B \rightarrow \mathrm{O}[B]B$

and also, more generally:

□OI. $\quad \Box(B \rightarrow A) \rightarrow (\Diamond B \rightarrow \mathrm{O}[B]A)$

And since $f(w) \subseteq \|B\|$ implies that $f(w) \cap \|B\| \cap \|C\| = f(w) \cap \|C\|$, we get:

SFD. $\quad \Box B \rightarrow (\mathrm{O}[B \wedge C]A \rightarrow \mathrm{O}[C]A)$

of which a special case is:

$$\Box B \rightarrow (\mathrm{O}[B]A \rightarrow \mathrm{O}A)$$

The significance of these two properties for the interpretation of $\mathrm{O}[B]A$ will be discussed separately, in section 5.

Further properties

The reader familiar with the logics presented in (Lewis, 1974) will recall that all the systems presented there contain three valid schemes which are of interest in that they already begin to resemble the 'Up' and 'Down' inheritance principles we are seeking. The schemes are (with the numbering of (Lewis, 1974), but employing the $\Box$ and $\Diamond$ operators):

A6. $\quad \mathrm{O}[B]A \ \wedge \ \mathrm{O}[C]A. \rightarrow \mathrm{O}[B \vee C]A$

A7. $\quad \neg\Diamond C \ \wedge \ \mathrm{O}[B \vee C]A. \rightarrow \mathrm{O}[B]A$

A8. $\quad \mathrm{P}[B \vee C]B \ \wedge \ \mathrm{O}[B \vee C]A. \rightarrow \mathrm{O}[B]A$

A6 can be regarded as a form of 'upward inheritance'. It is the scheme we referred to as ODIL earlier, valid in all models without any further assumptions about the preference orderings. It will be discussed in connection with deontic detachment presently.

A7 and A8 are special cases of 'downward inheritance'. A7 is equivalent to the scheme O$\Box$A above, valid in all models. Since $\Box(B \rightarrow C) \rightarrow \Box(B \leftrightarrow (B \wedge C))$, A7/O$\Box$A just says that a contextual obligation is inherited by necessarily equivalent contexts. A8 is validated by the further assumption that preference orderings on worlds are strongly connected (i.e. 'total' or 'linear'). A8 can be written equivalently as:

A8′. $\quad \mathrm{P}[B]C \rightarrow (\mathrm{O}[B]A \rightarrow \mathrm{O}[B \wedge C]A)$

of which the following is a special case:

$$\mathrm{P}B \rightarrow (\mathrm{O}A \rightarrow \mathrm{O}[B]A)$$

Both of A7/O$\Box$A and A8′ provide a kind of 'Down', except that, of course, they do not cover CTD contexts: A7/O$\Box$A is the boundary case in which contextual obligations are inherited by the same context, and in A8′ the more specific context $B \wedge C$ does not violate any obligation of the more general context B.[2]

This last observation is the reason why we want to go beyond the Hansson-Lewis systems. For our purposes O$[B]A$ as defined so far is too weak. A8 depends on the very strong, and in our view inappropriate, assumption that preference orderings are linear. But in any case it is too weak. It states logical relations between obligations for different contexts but does not cover the case where one context is a CTD context of the other. In the Hansson-Lewis framework, such obligations are mutually consistent, regardless of whether they are conflicting. We, by contrast, want to analyse what kind of consistency relations hold between obligations pertaining to contexts which stand in a CTD relation to one another.

[2]Where there is no danger of confusion we often refer to a context by a formula B rather than set of worlds $\|B\|$.

Upward inheritance and deontic detachment

Let us now investigate what forms of Up principle are available. A version analogous to that of Up in (Prakken and Sergot, 1994, 1996) would take the form:

$$\mathrm{P}[B \vee C]B \rightarrow (\mathrm{O}[B]A \rightarrow \mathrm{O}[B \vee C]A)$$

It is not valid. The following, weaker version of the original Up is valid:

$$\mathrm{P}[B \vee C]B \rightarrow (\mathrm{O}[B]A \rightarrow \mathrm{P}[B \vee C]A)$$

However, we can already derive a stronger property, a generalised form of deontic detachment, derivable from ODIL (which is not exclusive to the Hansson-Lewis family) and OI ($\Diamond A \rightarrow \mathrm{O}[A]A$) which is characteristic of Hansson-Lewis.

The derivation is as follows. First observe that $\mathrm{O}[B]A \rightarrow \mathrm{O}[B](B \rightarrow A)$. And from $\Diamond\neg B \rightarrow \mathrm{O}[\neg B]\neg B$ (OI), we have also $\Diamond\neg B \rightarrow \mathrm{O}[\neg B](B \rightarrow A)$. Putting these together, using ODIL, we obtain $\Diamond\neg B \rightarrow (\mathrm{O}[B]A \rightarrow \mathrm{O}[B](B \rightarrow A))$. Note finally that $\neg\Diamond\neg B \rightarrow (\mathrm{O}[B]A \rightarrow \mathrm{O}[B](B \rightarrow A))$ is a special case of the valid scheme SFD. So then we have:

DK. $\quad \mathrm{O}[B]A \rightarrow \mathrm{O}(B \rightarrow A)$

Deontic detachment (DD) follows from DK and $\mathrm{O}(B \rightarrow A) \rightarrow (\mathrm{O}B \rightarrow \mathrm{O}A)$:

DD. $\quad \mathrm{O}[B]A \rightarrow (\mathrm{O}B \rightarrow \mathrm{O}A)$

Note that from DD and OD we obtain a weak form of 'Up':

WeakUp′. $\quad \mathrm{P}B \rightarrow (\mathrm{O}[B]A \rightarrow \mathrm{P}A)$

This in turn can be re-written in equivalent form as:

Ctd′. $\quad \mathrm{O}[B]A \wedge \mathrm{O}\neg A. \rightarrow \mathrm{O}\neg B$

The derivation of DK and DD may be generalised easily. We obtain:

GDK. $\quad \mathrm{O}[B \wedge C]A \rightarrow \mathrm{O}[C](B \rightarrow A)$

From GDK follows the generalised form of deontic detachment:

GDD. $\quad \mathrm{O}[B \wedge C]A \rightarrow (\mathrm{O}[C]B \rightarrow \mathrm{O}[C]A)$

and weak 'Up':

WeakUp. $\quad \mathrm{P}[C]B \rightarrow (\mathrm{O}[B \wedge C]A \rightarrow \mathrm{P}[C]A)$

and its equivalent form:

Ctd. $\quad \mathrm{O}[B \wedge C]A \wedge \mathrm{O}[B]\neg A. \rightarrow \mathrm{O}[C]\neg B$

Other dyadic deontic logics

We conclude this presentation by remarking that there are of course other families of dyadic deontic logics besides the Hansson-Lewis kind. (See e.g. the discussion in (Loewer and Belzer, 1983).) Many of these logics are candidates for representing (defeasible) conditional obligations, and perhaps even

contextual obligations. Sorting out their various claims can be difficult because, given the way they are typically constructed, the resulting systems are inevitably almost identical. A detailed discussion is well beyond the scope of this paper, but it is instructive to refer again to the dyadic deontic logic $\mathrm{O}_B\, A$ in our earlier work (Prakken and Sergot, 1994, 1996).

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

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

The form of these truth conditionals is very common: it is just that of a normal conditional logic. And if there is another (alethic) accessibility relation f, then imposing $dc(Q, w) \subseteq f(w)$ as a model condition yields almost all of the rules and valid schemes identified earlier in this section. What is critical is what further conditions are imposed on the function dc.

The characteristic feature of Hansson-Lewis systems is that $dc(Q, w) \subseteq Q$, which here would validate $\mathrm{O}_A\, A$. From this flows deontic detachment. But this is not a feature of our earlier (Prakken and Sergot, 1994, 1996) system, not because of an oversight but because we were there not thinking of $dc(Q, w)$ as necessarily picking out some 'best' subset of the Q worlds. In the present paper we have chosen the Hansson- Lewis framework as a starting point, because it fits the kind of semantic structure we want to investigate in later sections. But there are also other possiblities we want to explore; they will not be discussed further in this paper.

4. SOME NOTIONS CONCERNING OBLIGATION

We have said that dyadic deontic logics of the Hansson-Lewis type are good starting points for the analysis of CTD structures. This seems to agree with the positions of Hansson and Lewis themselves. Hansson says explicitly that "... dyadic obligations are secondary, reparational obligations, telling someone what he should do if he has violated (...) a primary obligation" (Hansson, 1969, p.142). Although (Lewis, 1974) does not discuss this issue at length (the paper is mainly concerned with technical aspects of the logics), the only informal example is of a CTD structure, viz. the well-known Good Samaritan: it ought to be that you are not robbed, but given that you have been robbed you ought to be helped. This again fits the suggestion to regard Lewis's system as a candidate logic for CTD structures.

However other authors have interpreted the Hansson-Lewis systems in other ways. For instance, Loewer and Belzer (1983) have argued that these systems should be interpreted as logics for *prima facie*, or *ideal* obligation, as opposed to a logic for *all-things-considered*, or *actual* obligation. Sometimes they even seem to use the terms 'conditional' and 'unconditional' obligations to denote this distinction. And as mentioned above, others have interpreted

the Hansson-Lewis family as candidate logics for defeasible conditional obligations.

So what are we analysing in this paper? What is the relation between CTD structures and the various distinctions that have appeared in the literature? This section attempts to answer these questions. A secondary aim is to compare some of the different senses in which the various notions are used in the literature, with the aim of preventing terminological confusion.

Conditional vs. unconditional obligations

Since we propose to formalise CTD obligations with dyadic modalities, the question naturally arises whether we regard CTD obligations as a kind of conditional obligation. In particular, does the obligatoriness of A in the context B mean that A is obligatory on the condition that B holds? Our discussion in section 2 of the gentle murderer and the white fence is intended to indicate that this is not the case. Recall that we have described CTD obligations as obligations that are relative to a certain context, or certain circumstances. They may give cues for action for persons who regard the context as settled, but, and this is critical, regarding something *subjectively* as settled does not make obligations that hold outside the context go away. Even if I regard it as settled that I kill, the obligation not to kill is still binding upon me. The key to a consistent representation of timeless CTD structures is that the context is an essential part of the obligation: the obligation to kill gently pertains to the context where you kill: placing yourself outside the context, even if it is the case that you kill, makes the obligation cease to be a cue for action. This is one reason why we regard Hansson-Lewis systems as a basis for contextual rather than conditional obligations: they fail to satisfy factual detachment: $\mathrm{O}[B]A, B \not\models \mathrm{O}A$.

But perhaps this inference holds in a weaker sense? Perhaps it is only defeasibly valid, since contextual obligations are *prima facie* obligations? To answer this, we have to discuss what could be meant by the term *prima facie* obligation.

Prima facie vs. all-things-considered obligation

The term *prima facie* obligation originates from (Ross, 1930). According to Ross an act is *prima facie* obligatory if it has a characteristic that makes the act (by virtue of an underlying moral principle) *tend* to be a 'duty proper'. Fulfilling a promise is a *prima facie* duty because it is the fulfilment of a promise, i.e. because of the moral principle that you should do what you have promised. But the act may also have other characteristics which make the act tend to be forbidden. For instance, if I have promised a friend to visit him for a cup of tea, and then my mother suddenly falls ill, then I also have a

prima facie duty to do my mother's shopping, based, say, on the principle that we ought to help our parents when they need it. To find out what one's duty proper is, one should 'consider all things', i.e. compare all *prima facie* duties that can be based on any aspect of the factual circumstances and find which one is 'more incumbent' than any conflicting one.

To see how *prima facie* obligations might be formalised, we have to discuss the notion of 'defeasible' obligation.

Defeasible vs. non-defeasible obligations

Following Ross (1930), Loewer and Belzer (1983) say that *prima facie* duties are defeasible, or subject to exceptions. What can they mean by this? As we have just seen, *prima facie* duties tend to be duties by virtue of an underlying moral principle, that stresses only some of the characteristics of an act. Now *normally* such a principle can be applied to a situation without conflicting with other principles, but there can be exceptional circumstances in which conflicting principles also apply and perhaps even prevail. What happens in such a situation is that an argument, or inference, using such a moral principle is overridden, or defeated by another argument, using a stronger principle. To go back to the example, if in the circumstances I regard the principle concerning helping one's parents as more incumbent than the principle concerning keeping promises, then the argument for the obligation to help my mother defeats the argument for the obligation to have a cup of tea with my friend.

Arguments for conclusions that can be overridden are usually called 'defeasible'. More precisely, an argument is called defeasible if, although valid on the basis of a certain set of premises, it might be invalidated if new premises are added.

Note that defined in this way defeasibility is a property of an argument and not of a conditional. This is because one can have strengthening of the antecedent with or without the validity of modus ponens, and modus ponens with or without strengthening of the antecedent, and all these inferences can be both defeasible and deductive: so strengthening properties of a conditional have no bearing on the nature of arguments that use the conditional (see also (Makinson, 1993)). The often-used phrase 'defeasible conditional' is on this account elliptical for 'a conditional which, when used in an argument, makes the argument defeasible'.

The study of defeasible arguments is the field of so-called non-monotonic logic, a subfield of Artificial Intelligence. This, then, is the area where tools can be found for formalising reasoning about *prima facie* obligations. Reasoning about such obligations is reasoning with defeasible moral principles. Such principles can be formalised as defeasible conditionals; the antecedents of such conditionals stand for the aspect of a situation on which the *prima facie* obligation is based. More specifically, an obligation is *prima facie* if

it is the conclusion of an argument that is (non-monotonically) valid under a subset of the actual circumstances, although under the totality of the circumstances it may be invalidated. Only if the latter is not the case, is the *prima facie* duty also an all-things-considered duty (although still defeasibly derived, since we might come to know even more about the situation: 'all things considered' means 'all things considered that are known').

As a terminological matter it should be noted that in our terms (as in (Morreau, 1994)), it is not a *prima facie* obligation that is defeasible, but the argument from something being a *prima facie* obligation to its being an all-things-considered obligation. Defeat of such an argument does not mean that the conclusion ceases to be a *prima facie* obligation; a reason to act remains a reason to act, even if in the circumstances other reasons prevail.

It should be noted that earlier discussions of *prima facie* obligations, of e.g. (Hintikka, 1971; Loewer and Belzer, 1983; Jones and Pörn, 1985), do not use non-monotonic techniques. However, this is perhaps because at the time non-monotonic logics were not yet (widely) available. Since this has changed, the view that reasoning with *prima facie* obligations is non-monotonic has become increasingly popular; see e.g. (Horty, 1994; Morreau, 1994; Prakken, 1996).

Are contextual obligations prima facie obligations?

So then, are contextual obligations *prima facie* obligations? From our discussion it follows that if they are, then they should satisfy some form of defeasible factual detachment: from it ought to be that A given context B, and the truth of B, it should follow defeasibly that it ought to be that A. But this cannot be, since then in the gentle murderer and the white fence examples we would end up with a normative conflict between the primary and secondary obligation, which runs counter to our intuitions about these examples. As we explained in section 2, the primary obligation not to kill and the secondary one to kill gently need not in any sense be weighed to see which one should prevail in the situation: they both apply to the situation. There is no need for conflict resolution: if someone complies with the CTD obligation to kill gently, the sanction for killing can still be applied. As we argued at length in (Prakken and Sergot, 1994, 1996) and have repeated above in section 2, this is the key difference between contrary-to-duty and *prima facie* obligations.

Of course, a different matter is that primary or CTD obligations can be based on *prima facie* principles, just as any other type of obligation can. But what is essential is that defeasibly derived contrary-to-duty obligations still have a context attached. To give an example, imagine I have a friend who has some kind of personal problem. In conversations where he is present I should not mention this problem; in the context where I do mention it, there can be reasons why I should apologise for mentioning it and reasons why I

should not apologise and let the matter rest. Weighing these two *prima-facie* CTD-obligations might result in an all-things-considered, but still contextual, obligation to apologise or not to apologise.

Hansson-Lewis systems and prima facie obligations

Let us now return to the interpretation of Hansson-Lewis dyadic deontic logics as logics for *prima facie* obligations. If this interpretation is correct, then what we have just said implies that we cannot use such systems for representing contextual obligations. At first sight, it would seem that the way these systems define dyadic obligations indeed fits our description of *prima facie* obligations: being dyadic, the obligations depend only on certain aspects of a situation; moreover, conflicting obligations with different antecedents are consistent, which seems to capture that *prima facie* obligations remain a reason to act, even if in a given situation they do not become all-things-considered obligations.

Yet this interpretation is not appropriate. Recall that if the aspect on which a *prima facie* obligation is based is present in a situation, the *prima facie* obligation defeasibly implies an all-things-considered obligation. Now if Hansson-Lewis logics are regarded as logics for *prima facie* obligations, then such inferences should be captured in a non-monotonic extension of these systems: $B \wedge \mathrm{O}[B]A$ should defeasibly imply $\mathrm{O}A$. But how can this inference be formalised given the semantic interpretation of these logics? The only way seems to be to formalise an assumption to the effect that any world is as good as is consistent with the premises. However, apart from the question whether this assumption is realistic, it does not work: suppose that in our example it warrants that the actual world is among the best B worlds; all we can then derive is that in the actual world A is the case, not that in the actual world A is obligatory. Thus it seems that the attempt to interpret dyadic deontic logics of the Hansson-Lewis type as logics for *prima facie* obligations is fundamentally flawed.

Yet these attempts are understandable. As pointed out by Makinson (1993), $\mathrm{O}[B]A$ as interpreted in a Hansson-Lewis semantics can very well be read as 'A holds in all the most normal B worlds'. The O then stands for 'normally', which makes the dyadic formula express a defeasible conditional. However, the point is that it then expresses a conditional fact, not a conditional obligation.

Ideal vs. actual obligations

In discussing Lewis's logic, Loewer and Belzer (1983) sometimes use the terms 'ideal' and 'actual' obligation. Others have also linked these terms with CTD structures, e.g. (Jones and Pörn, 1985). How do these terms relate to

the notion of contextual obligation? An answer is not straightforward, since it seems that in the literature this distinction has been used in several different ways. It seems useful to point out these differences.

A common element in all analyses is that 'ideally it ought to be that A' at least implies that in a world where nothing has gone wrong A is the case. Now there are several ways in which things can go wrong. Sometimes the undesirable event is something unusual, motivating an exception to an obligation, as the obligation for soldiers to kill in war is an exception to the prohibition to kill. If used to describe such situations, as Loewer and Belzer (1983) seem to do, the distinction ideal/actual seems to stand for *prima facie*/all-things-considered.

Another usage of the terms is that of Jones and Pörn (1985), who also discuss CTD structures, but of a different form. We can illustrate it by reference to a timeless version of the Chisholm (1963) scenario, as in (Prakken and Sergot, 1994, 1996). Suppose that: there must be no dog around the house, and if there is no dog, there must be no warning sign, but if there is a dog, there must be a warning sign. Obviously, if there is a dog, the conditional obligation that there must be no sign does not become unconditional, since its condition is not fulfilled. On the other hand, it can also be inferred that if no obligations are violated, there will be no sign (modulo exceptions, of course). Jones and Pörn (1985) have argued that this is an inference of an ideal but not actual obligation not to have a sign, and they formalise it as a deontic detachment inference, maintaining consistency by introducing distinct modal operators for ideal and actual obligation.

Finally, the terms ideal/actual are sometimes used in a different sense again, especially in connection with CTD structures like the gentle murderer and the white fence. Here the intended point seems to be that, while several conflicting contextual obligations can apply to a situation, the job of actual obligations is to tell us what in the end we must do. And on this reading there can be at most one actual obligation: either don't kill, or kill gently; either tear down the fence, or paint it white. So further, if one regards the violation of the primary obligation as settled, the actual obligation is the secondary one, but if one regards the violation as still avoidable, the actual obligation is the primary one. This seems to be the sense motivating a recent proposal by Carmo and Jones (1996).

Can we describe primary obligations as ideal obligations and contrary-to-duty obligations as actual ones? If the distinction ideal/actual stands for *prima facie*/all-things-considered then, as explained, it is independent of primary/CTD. The sense in which Jones and Pörn (1985) make the distinction is meaningful, but it does not apply to the CTD examples we are studying. In the dog example where there is a dog, having a sign does not violate an obligation that applies to the situation: no fine is due for having a warning sign, only for having a dog. By contrast, the primary obligations not to kill, or

not to have a fence, do apply to the situation where a killing is taking place, or where there is a fence. The sanctions for violating them can be executed.

Finally, what of the last interpretation? An answer to this question requires a study of the detachment properties of contextual obligations. We now examine the properties of the Hansson-Lewis systems from that point of view.

5. HANSSON-LEWIS CONDITIONALS AS CONTEXTUAL OBLIGATIONS

In section 3 we presented a Hansson-Lewis logic extended with an operator for alethic necessity. Technically this is a simple addition, but it is a significant one. It enables us to clarify the sense in which contextual obligations are conditional, and to comment on the link between this system and the commonly accepted approach to temporal CTD structures in deontic logics with time.

5.1. *Detachment properties of contextual obligations*

We have said that contextual obligations are not to be confused with conditional obligations, that is, obligations which apply in certain circumstances but not in others. Yet there is a relationship between contextual and conditional obligations. Although contextual obligations do not satisfy modus ponens or any form of (possibly defeasible) factual detachment, they do satisfy another form of detachment which makes them conditional obligations of a sort. The point is that they are conditional in a special sense. The valid formula SFD of section 3 implies that contextual obligations, at least those of the Hansson-Lewis kind, satisfy a kind of 'strong-factual' detachment:

$$\models (\mathrm{O}[B]A \wedge \Box B) \rightarrow \mathrm{O}A$$

Contextual obligations are conditional, not upon the mere truth of the context, but upon the fact that the context is necessarily true, or 'objectively settled' as we shall also say.

For CTD obligations this form of strong factual detachment seems very appropriate, but it must be read with extreme care. As long as it is possible to avoid violation of a primary obligation $\mathrm{O}\neg B$ a CTD obligation $\mathrm{O}[B]A$ remains restricted to the context; it is only if the violation of $\mathrm{O}\neg B$ is unavoidable, if $\Box B$ holds, that the CTD obligation comes into full effect, pertains to the context $\top$. But it is important to note that the kind of necessity expressed by the $\Box$ operator is objective necessity, rather than some kind of subjective necessity, such as when an agent decides to regard it as settled *for him* that there will be a fence. It may be that a given agent is determined or becomes convinced that there is going to be a fence, come what may, but this does not make the obligation to have no fence go away. By contrast, 'it is *objectively* settled that B', $\Box B$, is much stronger: $\Box B$ implies $\mathrm{O}B$ and is inconsistent

with $\mathrm{O}\neg B$. By SFD, $\mathrm{O}[B]A$ implies $\Box B \rightarrow \mathrm{O}A$, but it is not equivalent to the latter.

5.2. *Temporal necessity*

In (Prakken and Sergot, 1994, 1996), we presented a series of examples to draw parallels between temporal and timeless CTD structures. In temporal CTD structures, that is to say, in CTD structures where there is a difference between the times of the primary and secondary (CTD) obligations, the objective necessity will often be of a temporal kind, i.e. of the kind whereby it is settled now that yesterday I violated an obligation to keep a certain promise. Whatever course the world will take from now on, the past cannot be undone. If I have an obligation to apologize for not keeping my promise, it pertains to a context that has been settled: I cannot undo the not keeping of my promise; all I can do now is apologize. This is the kind of inference that is captured by strong factual detachment. Perhaps we should call it 'contextual detachment'?

The temporal effect is a little awkward to demonstrate using the timeless system of section 3 since that logic is expressively restricted. We are forced to choose whether to represent the situation before the violation, at the time of the violation, or after the violation. Before the violation:

(1) $\mathrm{O}\,keep$
(2) $\mathrm{O}[\neg keep\,]apologize$

At the moment of breaking the promise the following can be added:

(3) $\neg keep$

To represent the situation after the violation, (3) can be strengthened. We then have:

(2) $\mathrm{O}[\neg keep\,]apologize$
(3*) $\Box\neg keep$

From (3*) and (2) follows the obligation $\mathrm{O}\,apologize$. The sentence (1) cannot be included, however, because it is inconsistent with (3*): once it is (objectively) settled that the promise is broken, there can be no obligation to keep it.

The effect is most clearly illustrated in temporal deontic logics that allow time to be expressed in the object language. Then it is not necessary to choose which of the three situations to represent because obligations pertaining to different times can be conjoined. For instance, in the system of van Eck (1982), we can say consistently:

(1′) $\mathrm{O}_{t_1} keep_{t_1}$
(2′) $\mathrm{O}_{t_2}[\neg keep_{t_1}]\, apologize_{t_3}$
(3′) $\neg keep_{t_1}$

Here time t_1 is before time t_2 is before time t_3. Expression (1′) says that at time t_1 it is obligatory that the promise is kept at t_1, and so on. The use of

our notation for contextual obligations at (2′) is justifiable because van Eck (1982) also employs a dyadic deontic operator (but relativised to time points) and adopts the Hansson-Lewis 'best of accessible worlds' interpretation of obligation; indeed the logic of each temporal obligation operator $\mathrm{O}_t[B]A$ is that of $\mathrm{O}[B]A$ in section 3.

It is beyond the scope of this paper to give a detailed account of the temporal semantics but a general sketch is in order to illustrate the extent of the similarities with the timeless case. Typically (see e.g. (van Eck, 1982; Åqvist and Hoepelman, 1981)) the set of possible worlds has the form of a tree. Each branch represents a possible world, or better: a possible course of events through time, and each node in a branch is a point in time. The notion of temporal necessity is captured in the following way. At any node in the tree, i.e. at any given point of time t in a given world w, the worlds accessible for w are those worlds that up to but not including t have the same past as w, i.e. that so far are indistinguishable from w. Those worlds that branched off in the past have become inaccessible: if yesterday I did not keep my promise, worlds in which I did keep it are today inaccessible. The deontic modalites are interpreted as follows: some of the possible futures of w at t will be marked as the ideal ones; what holds in all of them are the obligations at t, what holds in some of them is permissible at t.

Now it follows immediately from this that, for t_1 before t_2, (3′) $\neg keep_{t_1}$ implies

(3*′) $\quad \Box_{t_2} \neg keep_{t_1}$

When t_2 is after t_1, it is settled at t_2 that the promise was not kept at t_1. (3*′) and (2′) imply $\mathrm{O}_{t_2} apologize_{t_3}$ by strong factual detachment, exactly as in the timeless case. To complete the comparison with the timeless representation, note that (3*′) implies $\mathrm{O}_{t_2} \neg keep_{t_1}$, but this does not contradict (1′). Violation of (1′) can be expressed by conjoining it with (3′).

The reader may be wondering why the contextual obligation at (2′) could not have been represented instead as a conditional obligation, of the form:

(2*′) $\quad \neg keep_{t_1} \rightarrow \mathrm{O}_{t_2}\, apologize_{t_3}$

The answer is that in van Eck's system, (2′) and (2*′) are equivalent; where t_1 is before t_2, the following is valid (van Eck, 1982, Thm. 17):

$$\mathrm{O}_{t_2}[B_{t_1}]A \leftrightarrow (B_{t_1} \rightarrow \mathrm{O}_{t_2} A)$$

van Eck does not identify what we have been calling 'strong factual detachment'; his Thm. 17 serves much the same purpose.

One can see a corresponding feature in the timeless logic of section 3. In that system the following is valid:

$$\Box B \rightarrow (\mathrm{O}[B]A \leftrightarrow (B \rightarrow \mathrm{O}A))$$

One half is an instance of the 'strong factual detachment' property SFD. The other half, $\Box B \rightarrow ((B \rightarrow \mathrm{O}A) \rightarrow \mathrm{O}[B]A)$, is a special case of the scheme O□A (A7 in Lewis's numbering) that relates necessarily equivalent contexts. Here again one can see the essential point we are striving to make: contextual obligations and conditional obligations are indistinguishable when the context is necessarily true, but not otherwise.

In (Prakken and Sergot, 1994, 1996) we gave temporal CTD examples no detailed discussion because we thought they were not problematic; we thought that where fulfilling the secondary obligation comes after violating the primary one, the problems of the timeless examples do not occur, and that use of existing temporal deontic logics gives an adequate representation. We see now that this is not so. Suppose that at time t_1 I have the following two obligations, one to visit my neighbour's birthday party at t_2, and one to pay my taxes at t_4. Suppose that I do not visit my neighbour's party. Then at t_3, after t_2 but before t_4, what were the best futures of w at t_1 are not accessible any more; at t_3 there will be a completely different set of ideal futures. But how should these new ideal futures of w at t_3 be picked out? Surely in such a way that in all of them I still pay my taxes at t_4; the obligation to pay my taxes cannot disappear simply because I have violated another obligation. So the ideal futures at t_3, although disjoint from those at t_1, should still measure up to the old t_1-ideals as much as possible. Here we have a problem intimately related to the problem we want to study in a timeless setting in the present paper; the relevance of our discussions is not restricted to timeless CTD structures. However, since temporal notions introduce a host of further complications of their own, we confine ourselves here to the simpler, timeless case.

5.3. *The conditional nature of timeless CTD obligations*

For timeless CTD structures, such as the gentle murderer and the white fence, the use of temporal deontic logics does not help, since all statements pertain to the same point of time. Our earlier formulation of the gentle murderer re-written in the system of van Eck would look like this:

(1) $\mathrm{O}_t \neg kill_t$
(2) $\mathrm{O}_t[kill_t]kill\text{-}gently_t$

Adding

(3) $kill_t$

changes nothing. At the time of the killing this event is not yet unavoidable: in van Eck's system $kill_t$ does not imply $\Box_t kill_t$. This is how it should be, otherwise there could be no obligation at time t not to kill.

This is the key to the consistent representation of the gentle murderer, the white fence, and similar timeless examples. Since contextual CTD obligations

do not satisfy (ordinary) factual detachment, we never derive two conflicting obligations that pertain to the same context; and although we have strong factual detachment, violation of the primary obligation cannot be (objectively) settled, since this makes the primary obligation inconsistent. Of course in timeless CTD structures we usually do not judge the situation from the point of view where violation of the primary obligation is (temporally) settled, since then it will also be settled either that the secondary obligation is fulfilled, or that it is violated.

We feel that we should remark also on the validity in the Hansson-Lewis family of the formula

$$\mathrm{O}[A]A$$

(for consistent, 'possible' A), since this is often cited as a fundamental flaw of these systems. When interpreted as expressing a standard conditional obligation to the effect that 'if A then it ought to be the case that A' then of course the criticisms cannot be disputed. But this is not the reading that is ascribed to the O operator. $\mathrm{O}[A]A$ says only that A holds in all of the best accessible A worlds, which is no more (or less) unacceptable than the validity of $\mathrm{O}\top$ in standard deontic logic. Nor is there anything problematic about the reading of $\mathrm{O}[A]A$ as a special kind of conditional: if the context A is objectively settled, the truth of A is unalterable; again there seems nothing particularly odd about saying that what is unalterably true is also obligatory. Notice finally that to violate $\mathrm{O}[A]A$, a world would have to satisfy $A \wedge \neg A$. Our conclusion is that the Hansson-Lewis systems are not philosophically flawed, as long as they are not interpreted as systems for 'ordinary' conditional obligations. Their characteristic feature is the validity of $\mathrm{O}[A]A$, from which stems deontic detachment.

Finally, we are able to comment on one use of the distinction between ideal/actual obligations, viz. the use in which just one of the conflicting primary and secondary obligations in CTD examples is taken to be the 'actual' obligation, depending on circumstances. For temporal CTD examples what is actual is clear-cut (though superfluous) since what is obligatory changes with time and the primary and secondary obligations do not hold simultaneously. For timeless CTD examples, we cannot see that the ideal/actual distinction is useful. It is only meaningful to consider the case where violation of the primary obligation is not settled (for otherwise violation or fulfilment of the secondary obligation is also settled). In that case both primary and secondary obligations apply to the situation. Why should we single out one of them as 'actual' and the other as merely ideal? The situation is different if $\Box$ is interpreted as some other kind of settledness, such as the *subjective* notion whereby an agent is determined that violation of the primary obligation will take place (cf. (Carmo and Jones, 1996)), in which case it can be that violation of the primary obligation is 'settled' while violation/fulfillment of

the secondary obligation is not. This is not the kind of 'settledness' that we have been discussing.

5.4. *A timeless Chisholm scenario*

Let us now look at a timeless version of the Chisholm scenario, and ask how it can be represented in the modified Hansson-Lewis system of section 3. The example is taken from our earlier work:

(1) There must be no dog.
(2) If there is no dog there must be no sign.
(3) If there is a dog, there must be a sign.

Let us consider the following partial representation, where the proper formalisation of (2) is left open for now:

(1′) $\mathrm{O}\neg dog$
(2′)
(3′) $\mathrm{O}[dog]sign$

Clearly (3), being a CTD obligation, should be formalised as a contextual obligation. But how must (2) be formalised? The view is sometimes expressed that it should have the same conditional form as (3), i.e.

(2″) $\mathrm{O}[\neg dog]\neg sign$

But is this really what the statement (2) says? This is, of course, a matter of interpretation, but we think that on a very plausible reading of (2) the obligation not to have a sign is conditional upon the mere fact that there is no dog, not on the stronger condition that non-violation of (1) has been settled. And in this reading, (2) is not adequately formalised by (2″).

In our reading, (2) just says that if a world is such that there happens to be no dog, then there must be no sign. And on this reading, if we combine (2) with (1), they surely do not imply that in this world there *is* an obligation to have no sign; it just depends on how good this world is. Accordingly, it seems natural to regard (2) not as a CTD obligation but as a primary obligation conditional upon the (mere) fact that there is no dog. This reading is captured by the following representation:

(2′) $\neg dog \Rightarrow \mathrm{O}\neg sign$

where $\Rightarrow$ is any suitable conditional satisfying factual detachment. (We put to one side questions of defeasibility and the possibility of implicit exceptions.)

Let us now examine how well this representation satisfies the usual requirements for formalisations of the Chisholm set. As just observed, the conditional statements (2) and (3) have received different representations, but we think that because (2) is a primary and (3) is a secondary obligation, this is how it should be: primary obligations satisfy weak factual detachment but CTD obligations, by their very nature, only satisfy strong factual detachment.

Interestingly, in the temporal deontic logics of (Åqvist and Hoepelman, 1981) and (van Eck, 1982) these two statements of the Chisholm set also receive different formalisations (although naturally there a temporal variant of $(2')$ is chosen).

Another requirement is that from (2) and the (mere) fact that there is no dog, it should follow unconditionally that there must be no sign. $(2')$ satisfies this condition; $(2'')$ does not.

Chisholm's own requirements are consistency and logical independence of the statements (1)–(3) and the further assertion that (4) there is a dog. Logical independence is determined by the choice of the conditional $\Rightarrow$ in $(2')$. As regards consistency, the sentences $(1')$–$(3')$ are consistent. Adding

$(4')$ dog

does not make the formalisation inconsistent. But then another commonly stated requirement is not satisfied, viz. that $(3')$ and $(4')$ allow detachment of an obligation O$sign$ that there should be a sign. For us this is not problematic. We have just discussed at length why factual detachment is not valid for contextual obligations. However, there are other problems with the formalisation, to which we now turn.

5.5. *Inadequacies*

Consider now the following fragment of the previous example, taken, let us suppose, from regulations governing the use of holiday cottages.

(1) There must be no dog.
(2) If there is a dog, there must be a sign.

Suppose that to these requirements is added a further regulation:

(3) There must be no sign.

And suppose that the relevant authorities have explicitly declared that these three statements are not to be understood defeasibly: there are no exceptions. Are regulations (1)–(3) consistent? Is it logically possible, given (1)–(3), that there is a dog?

Much of our earlier work has been motivated by the very strong intuition that these regulations, and other examples discussed in (Prakken and Sergot, 1994, 1996), are inconsistent when given a particular, rather natural reading.[3] The Hansson-Lewis framework does not capture this reading. On the Hansson-Lewis account of obligation, (1) says that in the best of all worlds there is no dog, (2) that in the best of dog worlds there is a sign, and (3) that

[3] When we say that a set of regulations or set of (contextual) obligations such as (1)–(3) is inconsistent, we mean by this that the set is logically inconsistent with some further assumptions, in this case that it is logically possible there is a dog. In presenting examples we tend to leave these further assumptions unstated. Wherever the point of the example depends on it, we will state the assumptions explicitly.

in the best of all worlds there is no sign. There is nothing contradictory about that: (2) and (3) say that there is no dog in any ideal world, which is also what (1) says. Although the Hansson-Lewis systems do include principles that relate what is best in B-worlds with what is best in C-worlds for certain contexts B and C, these principles do not cover the case, essential for contrary-to-duty reasoning, where one context B is sub-ideal with respect to, contains a violation of, what is obligatory in context C. What these systems do not capture is that sub-ideal worlds should still measure up to the ideal worlds as much as possible.

In (Prakken and Sergot, 1994, 1996) our aim was to capture this aspect of CTD structures. At the very least we wanted to derive consistency requirements which would detect examples such as (1)–(3), an instance of what we called 'the considerate assassin', as inconsistent. Preferably we would like to obtain stronger 'down inheritance' principles, i.e. 'strengthening of the antecedent' principles in the terminology of conditional logics, allowing the inference from $\mathrm{O}[B]A$ to $\mathrm{O}[B \wedge C]A$, for certain combinations of A, B and C. So in the example, it seems to us that in at least one reading, we should be able to infer from (3) that even in dog worlds there should be no sign: in dog worlds that are as close as possible to ideal there is no sign. This inference would contradict (2), making the regulations inconsistent, which is what we want to infer in this example.

There is nothing wrong or incoherent about the Hansson-Lewis reading of (1)–(3). It is just that it does not capture adequately the notion of 'obligation' which makes us think that (1)–(3) are contradictory. What we want to investigate next is whether that notion of obligation can be captured without abandoning the Hansson-Lewis framework altogether.

Before moving on to that question, there is one further remark to make. If we have 'down inheritance' of contextual obligations, *no matter how it is obtained*, then the formalisation of the Chisholm set discussed in the previous section must be adjusted. The reason is that, in the case where there is no dog, the sentences

(1′) $\mathrm{O}\neg dog$
(2′) $\neg dog \Rightarrow \mathrm{O}\neg sign$
(3′) $\mathrm{O}[dog]sign$

are inconsistent: (2′) and factual detachment gives $\mathrm{O}\neg sign$, 'down inheritance' of this obligation gives $\mathrm{O}[dog]\neg sign$, and this contradicts (3′). Since we want down inheritance, it seems that we must adopt the formalisation used in (Prakken and Sergot, 1994, 1996) and replace (3′) by a conditional contextual obligation of the form:

(3″) $dog \Rightarrow \mathrm{O}[dog]sign$

The alternative is to find a 'down inheritance' principle which blocks this instance of inheritance but not others, a task that is not straightforward.

6. ON FORMALISING DOWN INHERITANCE

6.1. *Our earlier attempts*

As just explained, in the Hansson-Lewis framework obligations from different contexts are logically related only if none of the obligations are violated. It seems to us that for the analysis of contrary-to-duty structures it is essential that some logical relations hold also between obligations in contexts where one is a CTD context of another. In the present section we will investigate how this can be achieved.

What we were seeking in (Prakken and Sergot, 1994, 1996) was a principle of 'downward inheritance' of obligations, of the following form:

Down. $\varepsilon \rightarrow (\mathrm{O}[B]A \rightarrow \mathrm{O}[B \wedge C]A)$

ε was intended to capture the notion of relatedness of an obligation to a context: it had the form

$$\Diamond(A \wedge B \wedge C) \wedge \neg\Box((B \wedge \neg A) \rightarrow C)$$

The first conjunct says that the context $B \wedge C$ leaves compliance with the B-context obligation open, i.e. the context $B \wedge C$ does not imply a violation, $\neg A$, of the B-context obligation that A. The second conjunct states that violation of the B-context obligation does not necessarily put us into the $B \wedge C$ context, i.e. the context $B \wedge C$ does not already cover violation of the B-context obligation. The intended effect of these conditions is perhaps most easily illustrated with some examples.

The 'gentle murderer':

(1) $\mathrm{O}\neg kill$
(2) $\mathrm{O}[kill]kill\text{-}gently$

should be consistent with the further assumptions that $\Box(kill\text{-}gently \rightarrow kill)$ and $\Diamond(kill \wedge \neg kill\text{-}gently)$. Since the first of these assumptions contradicts the first conjunct of ε, $\mathrm{O}\neg kill$ is not 'downwards inherited' to $\mathrm{O}[kill]\neg kill$. Furthermore, the derived primary obligation $\mathrm{O}\neg kill\text{-}gently$ is not 'downwards inherited' to $\mathrm{O}[kill]\neg kill\text{-}gently$ since the second assumption contradicts the second conjunct of ε.

On the other hand, the 'considerate assassin' (Prakken and Sergot, 1994, 1996):

(1) $\mathrm{O}\neg kill$
(2) $\mathrm{O}[kill]offer\text{-}cigarettes$
(3) $\mathrm{O}\neg offer\text{-}cigarettes$

should come out inconsistent with the assumptions that $\Diamond(kill \wedge \neg offer\text{-}cigarettes)$ and $\Diamond(offer\text{-}cigarettes \wedge \neg kill)$. These assumptions satisfy the ε conditions, and so the primary obligation $\mathrm{O}\neg offer\text{-}cigarettes$ is 'down inherited' to $\mathrm{O}[kill]\neg offer\text{-}cigarettes$, which is inconsistent with (2).

However, as pointed out by Leon van der Torre and further discussed in (Prakken and Sergot, 1994, 1996), there are examples in which these ε conditions give unacceptable results. Rather than repeat that discussion here, we now give a general argument why a down inheritance principle of the form Down cannot be acceptable in any Hansson-Lewis system. For simplicity we shall just show the details for the special case where one context is $\top$. The ε conditions are then $\Diamond(A \wedge C) \wedge \neg\Box(\neg A \rightarrow C)$, i.e. $\Diamond(A \wedge C) \wedge \Diamond(\neg A \wedge \neg C)$.

Suppose that the following is valid in some class of Hansson-Lewis models:

Down′. $\quad \Diamond(A \wedge C) \ \wedge \ \Diamond(\neg A \wedge \neg C). \rightarrow (\mathrm{O}A \rightarrow \mathrm{O}[C]A)$

Now consider any two (unrelated) obligations $\mathrm{O}A$ and $\mathrm{O}B$, and the context $C = (A \wedge \neg B) \vee (\neg A \wedge B)$. It is easy to check that the conditions for Down′ inheritance of obligation $\mathrm{O}A$ to the context C are $\Diamond(A \wedge \neg B) \wedge \Diamond(\neg A \wedge \neg B)$. Similarly, the conditions for Down′ inheritance of obligation $\mathrm{O}B$ to the context C are $\Diamond(B \wedge \neg A) \wedge \Diamond(\neg B \wedge \neg A)$. So, if all three conditions $\Diamond(A \wedge \neg B)$, $\Diamond(\neg A \wedge B)$, and $\Diamond(\neg A \wedge \neg B)$ hold, we have both $\mathrm{O}A \rightarrow \mathrm{O}[C]A$ and $\mathrm{O}B \rightarrow \mathrm{O}[C]B$.

Suppose $\Diamond(A \wedge \neg B)$, $\Diamond(\neg A \wedge B)$, and $\Diamond(\neg A \wedge \neg B)$, and $\mathrm{O}A$ and $\mathrm{O}B$. Then both $\mathrm{O}[C]A$ and $\mathrm{O}[C]B$, from which $\mathrm{O}[C](A \wedge B)$ follows by RCOK. But for $C = (A \wedge \neg B) \vee (\neg A \wedge B)$, $\mathrm{O}[C](A \wedge B)$ must be false in any Hansson-Lewis system, by 'ought implies can': $\mathrm{O}[C](A \wedge B) \rightarrow \Diamond(C \ \wedge \ (A \wedge B))$, but $\Diamond(C \ \wedge \ (A \wedge B)) = \Diamond(((A \wedge \neg B) \vee (\neg A \wedge B)) \ \wedge \ (A \wedge B))$, which is logically equivalent to $\Diamond\bot$, which is false.

We must conclude that if $\mathrm{O}A$ and $\mathrm{O}B$, then either $\neg\Diamond(A \wedge \neg B)$, or $\neg\Diamond(\neg A \wedge B)$, or $\neg\Diamond(\neg A \wedge \neg B)$. If Down′ is valid then the following is valid also:

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

Proposition 6.1 If Down is valid in any class of Hansson-Lewis models then, for any A, B and C, the following is valid also:

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

Proof Generalise the derivation shown above. □

It follows that in any Hansson-Lewis system in which Down is valid, there can be no logically independent obligations.

This is a general argument. It does not depend on any particular orderings. It does not even depend on the exact form of ε. Since it uses only 'ought implies can' and the scheme

OC. $\quad \mathrm{O}[C]A \ \wedge \ \mathrm{O}[C]B. \rightarrow \mathrm{O}[C](A \wedge B)$

proposition 6.1 can easily be generalised to argue that ε conditions for any strong down inheritance principle of the form Down cannot be axiomatised.

But this does not imply that there are no acceptable down inheritance principles at all. Does the present setting, where contextual obligations are interpreted in terms of a preference relation on worlds, enable us to find a suitable version of down inheritance? This is what we now want to investigate. The essence of the problem is that the Hansson-Lewis framework allows for the possibility of sub-ideal worlds but has very little to say about what they are like and nothing to say about how they compare with ideal worlds. Somehow we have to find a way of relating, for any pair of contexts $Q \subset Q'$, the best of Q-worlds with the best of Q'-worlds. The idea is that this can be done by putting more structure on the preference orderings $\geq_w$ to reflect the requirement that sub-ideal contexts should still measure up to the standards of more ideal contexts as much as possible. After formalising this idea, we return to our intuitions concerning downwards (and upwards) inheritance, and check to what extent these intuitions are captured by this approach.

6.2. *The general idea*

The general idea can be explained by simple diagrams, of a kind not uncommon in both non-monotonic logic and some recent presentations of deontic logic.

For simplicity consider a language with just two atomic propositions p and q. Suppose that Op holds. In a model structure with four distinct worlds, one would feel intuitively that they are ordered according to their relative goodness as in figure 1. (Worlds within circles are equally good, and if two circles are connected by a line, those on the left are strictly better than those on the right.)

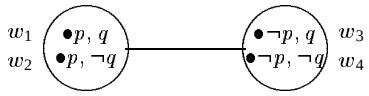

Fig. 1. A model for Op

In terms of Hansson-Lewis models, the picture can be seen as depicting a fragment of a model where the worlds w_1–w_4 are those that are accessible from some world w and the ordering shown is $\geq_w$.

Adding an extra obligation Oq would give the ordering shown in figure 2. It is easy to see that in the best $\neg p$ worlds q is true, for which reason we like to say that Oq is inherited by, or transported down, to context $\neg p$.

A further element of the idea, the key feature, is that contextual obligations refine the ordering within the context to which they pertain. Thus, to construct a model for Op, Pq, P$\neg q$ and O$[\neg p]q$ we first order the worlds as to how well they fulfil the primary obligation Op, and then refine the ordering within the

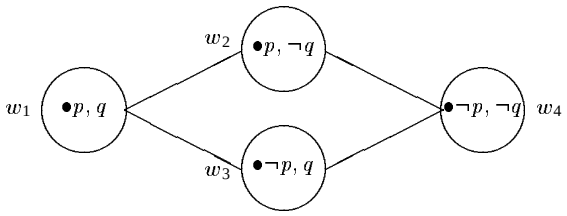

Fig. 2. A model for Op, Oq

set $||\neg p||$ with respect to O$[\neg p]q$. This is illustrated in figure 3, where the box focusses on $||\neg p||$.

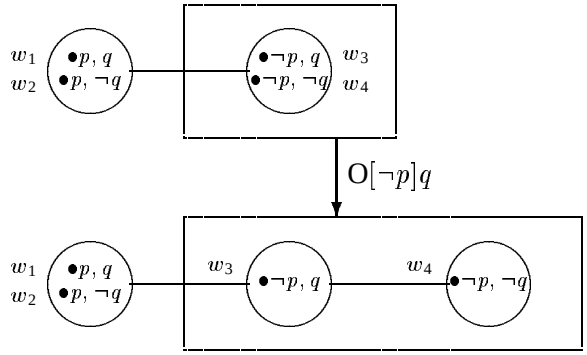

Fig. 3. A model for Op, O$[\neg p]q$

6.3. *The ordering on the set of worlds*

We now investigate a way to formalise the ideas behind these pictures. We will do this by adapting a technique used by Ryan in the study of non-monotonic reasoning. (Ryan, 1992) investigates the problem of ordering models of sets of defaults according to how well each model satisfies the defaults, while taking into account a further priority ordering on defaults. We will use his technique to order, not models, but worlds within models. Worlds will be ordered according to how well they satisfy various obligations, but allowing for the fact that not all obligations are equally important. The key idea, illustrated by figure 3, is that it is better to fulfil an obligation from a more ideal context and violate one from a less ideal context than the other way around. Obligations from 'better' contexts dominate over obligations from worse contexts. If our intuitions on down inheritance are validated at all, it will be because of this idea.

The general strategy, adapted from (Ryan, 1992), is to employ a generalised form of lexicographic ordering constructed from a number of intermediate orderings. First, we want to determine the relative goodness of worlds by comparing which of the obligations in force they fulfil and which they violate. But what are the obligations 'in force' whose violation or fulfilment is to be checked? As a first shot we will say that they are all the contextual obligations that are true in the model. We shall have reason to change this decision later, in section 7.

We define for each $v \in W$ the *violation set* of v relative to any context C, i.e. the set of all C-obligations violated by v.

Definition 6.2 (Violation sets) *For all $w, v \in W$ and $Q \subseteq W$ the violation set $V_{w,Q}(v)$ is defined as $\{P \subseteq W \mid \max_w(f(w) \cap Q) \subseteq P \;\&\; v \notin P\}$.*

Violation sets can be used to determine, for each context Q, how well worlds w_1 and w_2 fulfil the obligations that pertain to context Q: let this be represented by an ordering $\sqsupseteq_w^Q$ (all orderings are also relativised to worlds w, as usual). Specifically, we define the intermediate orderings $w_1 \sqsupseteq_w^Q w_2$ iff $V_{w,Q}(w_1) \subseteq V_{w,Q}(w_2)$.

Now, the main technical problem is to combine these intermediate orderings into an ordering $\sqsupseteq_w$ to reflect the relative 'ideality' or importance of contexts. Given the truth conditions for $\mathrm{O}[B]A$, it is natural to say that a context Q_1 is more 'ideal', more important, than a context Q_2 when $Q_1 \supset Q_2$. So, given two intermediate orderings $\sqsupseteq_w^{Q_1}$ and $\sqsupseteq_w^{Q_2}$ such that $Q_1 \supset Q_2$, their combined effect can be captured by the standard lexicographic construction in which $\sqsupseteq_w^{Q_1}$ takes precedence and $\sqsupseteq_w^{Q_2}$ just refines the ordering of the $\sqsupseteq_w^{Q_1}$-equivalent worlds. This is the basic step. It has to be generalised to combine the effects of orderings $\sqsupseteq_w^Q$ for all contexts $\emptyset \subset Q \subseteq W$, not just two of them. This can be done by employing a generalisation of the lexicographic ordering, as used in (Ryan, 1992) and further studied in (Ryan and Schobbens, 1993). We state the definition directly in terms of violation sets, without explicit reference to intermediate orderings.

Definition 6.3 ('Layered ordering') *For any triple of worlds w, w_1 and w_2 in W it holds that*

- *$w_1 \sqsupseteq_w w_2$ iff $\forall Q \subseteq W$ such that $Q \neq \emptyset$:*
  1. *$V_{w,Q}(w_1) \subseteq V_{w,Q}(w_2)$ or*
  2. *$\exists Q' \subseteq W.\, (Q' \supset Q$ and $V_{w,Q'}(w_1) \subset V_{w,Q'}(w_2))$*

The effect of the definition is to generalise the lexicographic construction to a set of orderings which is itself partially ordered. Results of (Ryan and Schobbens, 1993) ensure that $\sqsupseteq_w$ is a pre-order when W is finite. (The restriction to finite W can be removed by complicating the truth conditions for $\mathrm{O}[B]A$ but we will not discuss that here.)

The definition says that w_1 is as good as w_2 ($w_1 \sqsupseteq_w w_2$) if w_1 is as good as w_2 at all contexts, except possibly those at which there is a higher context at which w_1 is strictly better than w_2. And $w_1 \sqsupset_w w_2$ means that all (maximal) chains $Q_1 \subset Q_2 \subset \ldots \subset W$ end with first a context Q_i such that $V_{w,Q_i}(w_1) \subset V_{w,Q_i}(w_2)$ and then zero or more contexts $Q_j \supset Q_i$ such that $V_{w,Q_j}(w_1) \subseteq V_{w,Q_j}(w_2)$. Note that the last element of every such chain is W. In terms of obligations, $w \models \mathrm{O}[B]A$ means that all (maximal) chains $\|B\| \subset Q_1 \subset \ldots \subset W$ end with first a Q_i such that $\max_w(f(w) \cap Q_i) \subseteq \|A\|$, and then a sequence of zero or more $Q_j \supset Q_i$ such that $\|A\| \cap \max_w(f(w) \cap Q_j) \neq \emptyset$.

So in summary: the (contextual) obligations that are true at a world w in a model $\mathcal{M}$ are determined by the orderings $\sqsupseteq_w$; these orderings are defined in terms of violation sets; violation sets are determined by the (contextual) obligations that are true at w in model $\mathcal{M}$. This construction generates a set of constraints on $\sqsupseteq_w$: any model whose orderings $\sqsupseteq_w$ satisfy these constraints will be said to be a 'Layered ordering' model. We now want to know what additional formulas are valid in the class of such models.

6.4. *Valid and invalid formulas*

We now investigate to what extent these semantic ideas capture our intuitions concerning up and down inheritance. First, we observe that the analogue of the strong 'Up' principle in (Prakken and Sergot, 1994, 1996)

StrongUp. $\quad \mathrm{P}[B]C \rightarrow (\mathrm{O}[B \wedge C]A \rightarrow \mathrm{O}[B]A)$

is invalid. Figure 4 shows a counterexample, already for the special case where $B = \top$. It depicts a model of $\mathrm{O}[p]q$ (and $\mathrm{O}[\neg p]\neg q$) which does not satisfy $\mathrm{O}q$ (and $\mathrm{O}\neg q$).

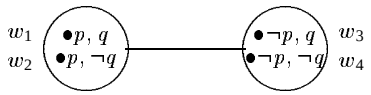

Fig. 4. A counterexample to Up

We still have the weaker 'Up' principle valid in all Hansson-Lewis systems:

WeakUp. $\quad \mathrm{P}[B]C \rightarrow (\mathrm{O}[B \wedge C]A \rightarrow \mathrm{P}[B]A)$

As shown in section 3 this follows from deontic detachment (DD) and the scheme OD, and does not depend on any particular way of obtaining the orderings $\sqsupseteq_w$. The following, equivalent formulation clearly shows that WeakUp formulates a consistency condition on conflicting obligations from different contexts.

$$\mathrm{O}[B]A \ \wedge \ \mathrm{O}[B \wedge C]\neg A. \rightarrow \mathrm{O}[B]\neg C$$

We next turn to our main concern in this paper, down inheritance. First we check that this semantics does not validate the strong Down principle we were originally seeking in (Prakken and Sergot, 1994, 1996). By Proposition 6.1 we simply need to construct a model for any two obligations $\mathrm{O}A$ and $\mathrm{O}B$ where $\Diamond(A \wedge \neg B)$ and $\Diamond(B \wedge \neg A)$ and $\Diamond(\neg A \wedge \neg B)$, such that the requirements of the 'Layered ordering' (definition 6.3) are satisfied. This is easily done. (Figure 2 already shows an example.)

We can also see semantically why there are no prospects for finding a sensible condition ε under which an obligation in a certain context transports down to a more specific context. The reason is that this condition will have to depend on all other obligations that possibly transport downwards. For every obligation a model can always be constructed with yet another interfering candidate. We cannot block all of them: ε cannot be axiomatised.

We turn now to the more modest goal of finding consistency conditions for obligations from different contexts. On what conditions are the formulas $\mathrm{O}[B]A$ and $\mathrm{O}[B \wedge C]\neg A$ inconsistent? For which formulas φ, if any, is the following scheme valid?

$$\text{WeakDown.} \quad \varphi \rightarrow (\mathrm{O}[B]A \rightarrow \neg\mathrm{O}[B \wedge C]\neg A)$$

We have been able to find such a valid scheme for the special case where $B = \top$.

Proposition 6.4 The following scheme is valid in the class of 'Layered ordering' models, for any A and C:

$$\text{WeakDown}'. \quad \Diamond(A \wedge C) \rightarrow (\mathrm{O}A \rightarrow \neg\mathrm{O}[C]\neg A)$$

Proof The most concise statement uses the more general form (obl) of truth conditions for $\mathrm{O}[B]A$ (see section 3). The proof works just as well with conditions $(\mathrm{obl}_{\max})$, as is easily checked.

Consider any w satisfying (i) $\Diamond(A \wedge C)$, (ii) $\mathrm{O}A$ and (iii) $\mathrm{O}[C]\neg A$. By (i) there exists a w_1 in $f(w) \cap \|A \wedge C\|$. By (iii) there is a w_2 in $f(w) \cap \|C \wedge \neg A\|$ which is strictly better than any world in $f(w) \cap \|C \wedge A\|$: so $w_2 \sqsupset_w w_1$.

Now, by (ii), $w \models \mathrm{O}A$. Since $w_1 \models A$ and $w_2 \not\models A$, $V_{w,W}(w_2) \not\subseteq V_{w,W}(w_1)$. So condition (1) of definition 6.3 is not satisfied. Then, to have even $w_2 \sqsupseteq_w w_1$, we must find a $Q \supset W$ such that $V_{w,Q}(w_2) \subset V_{w,Q}(w_1)$. Clearly such a Q does not exist. □

This result is surprising. The antecedent of this WeakDown′ seems too weak. It contains only the first conjunct of the ε condition of Down in section 6.1, while there we saw that for consistent representation of the gentle murderer the second conjunct of ε is needed also. The validity of WeakDown′ causes us to question some of our fundamental assumptions, as we discuss next.

7. A CASE FOR EXPLICIT OBLIGATIONS

As we have just seen, the 'Layered ordering' of section 6 does not fully capture our intuitions. Yet still it seems that the general idea of the previous section is a reasonable one; in this section we discuss how it can be modified to yield something more promising.

Consider first the following example, which is the one used in section 5.5 to motivate the need for 'down inheritance' (strengthening of the antecedent) for contextual obligations.

(1) There must be no dog.
(2) If there is a dog, there must be a sign.
(3) There must be no sign.

As we indicated earlier, our view is that on a particular, rather natural reading, these statements are inconsistent. The basic reason is this: CTD obligations are intended to regulate norm violation but they cannot just ignore all other norms. Here, the CTD obligation (2) regulating the violation of (1) does not respect another primary obligation, (3). We want to say that for this reason, (2) and (3) are inconsistent.

Consider now the following variant, which is a version of the gentle murderer:

(1a) There must be no dog.
(2a) If there is a dog, it must be a poodle.

Since poodles are dogs, one can see that this is indeed a version of the gentle murderer. Intuitively this should be consistent. But suppose now that we add another primary obligation, to the effect that:

(3a) There must be no poodle.

What we want to suggest is that (1a)–(3a) are inconsistent in precisely the same way that (1)–(3) are. (2a) regulates the violation of (1a) but it does not respect another primary obligation, (3a): (2a) and (3a) are inconsistent. The new feature of the poodle example, of course, is that (3a) is implied by (1a). Thus, we want to say that (1a)–(3a) are inconsistent, *even though* (1a)–(2a) are consistent and (3a) is implied by (1a).[4]

How can we formalise these intuitions? At first sight it would seem that we must abandon consequential closure (i.e. $\Box$OCK of section 3). This is the route taken by e.g. Tan and van der Torre (1996) who have also discussed the validity of what we call 'down inheritance' in connection with one of their dyadic deontic logics. We do not think this is the right solution. As we have said earlier, if a normgiver forbids having dogs he surely also implicitly forbids having any particular kind of dog. Our view is that consequential

[4]Or consider: (1b) the door must be painted red; (2b) if the door is not painted red, it must be left unpainted; (3b) the door must not be left unpainted. Again we want to say that (1b)-(3b) are inconsistent, even though (1b)-(2b) are consistent, and (1b) implies (3b).

closure should not be disregarded when determining what is *obligatory* but only when determining how good sub-ideal worlds are.

If we look more closely at these examples, the difference between what appears consistent and inconsistent seems to depend critically on what is stated explicitly. In (1a)–(2a) the obligation not to have a poodle is only implicit whereas (3a) states this obligation explicitly. Suppose we take this difference seriously. We now sketch the development of an entailment relation $\Gamma \,|\!|\!\sim A$ in which designated *explicit obligations* in premises Γ will be given special status. (Some traces of this idea can also be found in (Tan and van der Torre, 1994) though the details are different. Also the 'Type 2 obligation' of (Brown, 1996) seems to be related.) Given a set S of designated explicit (contextual) obligations, we restrict attention to models where the orderings $\sqsupseteq_w$ are obtained as in definition 6.3 but where the violations sets are determined only by the *explicit* obligations S. We thereby obtain an entailment relation parametrised by a set of explicit obligations, as designated in the premises.

In the poodle example it will be obvious what the designated explicit obligations are, but in general this will not be so, since the premises can be of any form. Therefore, we will assume a function D that assigns to each set Γ of sentences the set of explicit obligations that are designated by Γ. We leave it to future research to investigate in general the properties of the function D; in this paper we will employ a simple notational device to specify what is in $D(\Gamma)$. We write $\widehat{\mathrm{O}}[B]A$ to indicate when $\mathrm{O}[B]A$ is one of the explicit obligations of a set of premises: $\widehat{\mathrm{O}}[B]A$ in premises Γ signifies that $\mathrm{O}[B]A \in D(\Gamma)$; moreover, no other obligations are designated as members of $D(\Gamma)$ beyond those written as $\widehat{\mathrm{O}}$–obligations.

Next we define the entailment relation, in terms of the standard notion of validity in a class of models (cf. e.g. (Chellas, 1980, p. 36)), which is that A is valid in a class C of models ($\models_{\mathsf{C}} A$) iff A is true in all models $\mathcal{M}$ in C. Then as usual, for any set Γ of sentences, $\Gamma \models_{\mathsf{C}} A$ iff $\mathcal{M}, w \models A$ for all models $\mathcal{M} \in \mathsf{C}$ and worlds w such that $\mathcal{M}, w \models \Gamma$.

For any set Γ of sentences, let the set $\mathsf{C}(\Gamma)$ of Γ–*ordered models* be the set of all 'Layered ordering' models where the violation sets inducing the ordering $\sqsupseteq_w$ are determined by the explicit obligations $D(\Gamma)$. Then the entailment relation $|\!|\!\sim$ is defined as follows:

$$\Gamma \,|\!|\!\sim A \quad \text{iff} \quad \Gamma \models_{\mathsf{C}(\Gamma)} A$$

What does this framework have to say about the logic of explicit obligations? By imposing some (modest) restrictions in the definition of violation sets it is easy to arrange that, for any Γ, if $\widehat{\mathrm{O}}[B]A$ is an explicit obligation of Γ then $\models_{\mathsf{C}(\Gamma)} \mathrm{O}[B]A$, i.e.

$$\widehat{\mathrm{O}}[B]A,\ \Delta \,|\!|\!\sim \mathrm{O}[B]A$$

for any set of additional premises Δ. This in turn induces some constraints on sets of premises Γ: in particular any set of premises Γ designating both $\widehat{\mathrm{O}}[B]A$ and $\widehat{\mathrm{O}}[B]\neg A$ is inconsistent, in the sense that the set of Γ-ordered models is empty.

In the example, the following set of premises Γ_1:

(1′) $\widehat{\mathrm{O}}\neg dog$
(2′) $\widehat{\mathrm{O}}[dog]poodle$

has, e.g., $\Gamma_1 \mid\!\!\sim \mathrm{O}\neg dog$ and $\Gamma_1 \mid\!\!\sim \Box(poodle \rightarrow dog) \rightarrow \mathrm{O}\neg poodle$. It is important to note that what is obligatory, represented by expressions of the form $\mathrm{O}[B]A$, is closed under consequence ($\Box$OCK of section 3 is valid in the class of all Γ–ordered models). Explicit obligations, by contrast, are not closed under consequence: premises Γ_1 do not entail the *explicit* obligation $\widehat{\mathrm{O}}\neg poodle$.

Another property of sets of explicit obligations can be obtained from the derivation of the WeakDown′ principle of Proposition 6.4. When violation sets are defined in terms of designated explicit obligations only, WeakDown′ does not hold for $\mathrm{O}[B]A$ but for $\widehat{\mathrm{O}}[B]A$; more precisely, the same derivation as for Proposition 6.4 now yields a weak down principle according to which, if $\widehat{\mathrm{O}}A$ and $\widehat{\mathrm{O}}[B]\neg A$ are both in Γ, then $\models_{\mathrm{C}(\Gamma)} \neg\Diamond(A \wedge B)$, i.e.

$$\widehat{\mathrm{O}}A,\ \widehat{\mathrm{O}}[B]\neg A,\ \Delta \mid\!\!\sim \neg\Diamond(A \wedge B)$$

We shall refer to this principle as WeakDown*.

In the example, with the further assumptions that

(4′) $\Box(poodle \rightarrow dog)$
(5′) $\Diamond(dog \wedge \neg poodle)$

the difference is between the set of premises Γ_1 which satisfies WeakDown*, and the following set of premises Γ_2

(1′) $\widehat{\mathrm{O}}\neg dog$
(2′) $\widehat{\mathrm{O}}[dog]poodle$
(3′) $\widehat{\mathrm{O}}\neg poodle$

which is inconsistent, in the sense that, by WeakDown*, there are no Γ_2-ordered models of (4′) and (5′).

What then of strong ‘down’? Our earlier argument for the unacceptability of strong ‘down’ depended on the observation that all other obligations that are true have to be considered; and since for every obligation a model can always be constructed with yet another interfering candidate, we concluded that down inheritance cannot be axiomatised. However, this argument does not hold if it is only explicitly stated obligations that are taken into account when determining violation sets. Assume that we have as premises

(1) $\widehat{\mathrm{O}}A$

(2) $\widehat{\mathrm{O}}B$

In verifying down inheritance we need to consider models where the ordering is determined by (1) and (2) only; other obligations, whether implied by (1) and (2) or not, can be ignored since they do not appear explicitly in the premises. In all such models any $\neg A \wedge \neg B$ worlds will be worse than the best $A \wedge \neg B$ worlds; so then in all those models $\mathrm{O}A$ is downwards inherited by the context $\neg B$ (assuming of course that there are $A \wedge \neg B$ worlds). We get:

$$\widehat{\mathrm{O}}A,\ \widehat{\mathrm{O}}B \ \|\!\sim\ \Diamond(A \wedge \neg B) \rightarrow \mathrm{O}[\neg B]A$$

Suppose that we add as another premise

(3) $\widehat{\mathrm{O}}C$

Now (3) is also relevant to construction of the ordering, and if $\Diamond(C \wedge \neg B)$, then not all best $\neg B$ worlds will contain A; some may contain $C \wedge \neg A$. So:

$$\widehat{\mathrm{O}}A,\ \widehat{\mathrm{O}}B,\ \widehat{\mathrm{O}}C \ \ \|\!\not\sim \Diamond(A \wedge \neg B) \rightarrow \mathrm{O}[\neg B]A$$

So the consequence relation $\|\!\sim$ is non-monotonic (although each individual $\models_{\mathrm{C}(\Gamma)}$ is not).

Clearly, much work remains to be done to refine these ideas. The main aim of this part of the paper has been to show that this work is needed; that the Hansson-Lewis account of obligation must be extended to capture even the most basic features of contrary-to-duty reasoning; that these extensions cannot be undertaken using standard model-theoretic devices; but that there are nevertheless promising avenues to explore.

We do not expect that finalising these details will be easy. Similar reasoning patterns have been studied in non-monotonic reasoning, where they have proved notoriously hard to formalise. In particular, down inheritance of contextual obligations, even with explicitly designated obligations, raises similar problems to the problem of irrelevance in possible-worlds accounts of defeasible conditionals: from 'birds fly', 'penguins do not fly' and 'birds are small' we want to infer for Frank the penguin that he does not fly but that he is small; if we know that birds fly and we know that Gloria is a black bird, we should be able to conclude that Gloria can fly given that she is black.

Another area where similar problems have arisen is temporal reasoning. A main problem here is the frame problem: how to formalise the persistence of facts through time and the ramifications of change. Likewise, as we discussed in section 5, temporal deontic logics must account for the persistence of obligations through time. In systems such as those of (van Eck, 1982; Åqvist and Hoepelman, 1981) the best futures at t in w must be related somehow to the best futures at $t+1$ in w; otherwise we can lose obligations in the future simply by violating an unrelated obligation now. It is not that persistence is the same problem as 'down inheritance', but that the same problems, of

irrelevance, have to be confronted. How close these resemblances actually are is also a topic of our current investigations. One point we might make, however, is this: if, as we now believe, CTD reasoning depends on what obligations are stated explicitly as premises, then the temporal persistence problem might not be too difficult; perhaps it is sufficient to account for how explicitly designated obligations persist through time, without having to worry about all the obligations that can be derived from them as consequences. This remains to be seen.

8. CONCLUSION

In this paper we have tried to improve on our earlier analysis of CTD reasoning by interpreting obligations in terms of preference structures, that is, orderings on the relative 'goodness' of worlds. We presented our attempts as a modification of the well-known systems of dyadic deontic logics of (Hansson, 1969) and (Lewis, 1974). Let us recapitulate.

First of all, we have argued that dyadic deontic logics validating the principle $O[A]A$ are not flawed, as long as they are regarded as candidates for representing contextual rather than 'ordinary' (defeasible or non-defeasible) conditional obligations. To argue for this, it was necessary to clarify several distinctions of kinds of obligations that have appeared in the literature.

A main ingredient of our argument was the addition of a notion of alethic necessity to the systems of Hansson and Lewis. Thus we were able to formalise the idea that contextual obligations do not satisfy factual detachment, as ordinary conditionals do, but only a stronger form, whereby the obligation becomes unrestricted to context when its antecedent is necessarily true.

The introduction of alethic necessity has also clarified the link between contextual obligations and generally accepted treatments of temporal CTD structures, to which we were previously able to allude only implicitly. The problems we have studied in a timeless setting do not disappear when time is introduced. This is not only because there are examples of CTD structures where all obligations pertain to the same point of time, so that the greater expressive power of temporal deontic logics remains unused; it is also because there are some outstanding problems concerning the persistence of obligations in time that seem to have been overlooked in the literature on temporal CTD structures. We have suggested that there are close parallels between the formalisation of down inheritance (strengthening of the antecedent) in contextual obligations and persistence in temporal deontic reasoning. It remains to be seen if this is borne out by future investigations.

The last part of the paper was concerned with an investigation of how the Hansson-Lewis framework could be augmented if it is to deal with the analysis of CTD structures. According to our diagnosis, the weakness of the Hansson-Lewis account of obligation is that, although it allows for the possibility of

non-ideal worlds, it has nothing to say about them. We therefore focussed on a class of models in which the orderings on the relative goodness of worlds are given additional structure: we adapted a technique from the study of default reasoning to construct orderings which rank non-ideal worlds according to how well they measure up to the ideal ones.

Perhaps the most significant result of these technical investigations is the emergence of CTD examples whose consistency seems to depend critically on whether an obligation is stated explicitly or is simply implied by other statements. This has led us to construct a non-monotonic consequence relation parametrised by a set of explicitly designated obligations. We obtain thereby a logic of explicit obligations, which is not closed under logical consequence, and a separate logic of what is obligatory, which is closed under consequence. Although the construction is quite natural, it is not something we undertake lightly. We have avoided it for as long as possible; we now feel the evidence is irresistible.

ACKNOWLEDGEMENTS

The authors wish to thank Andrew Jones, Donald Nute, Mark Ryan and Leon van der Torre for valuable discussions on the topic of this paper.

REFERENCES

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

M.A. Brown. Doing as we ought: towards a logic of simply dischargeable obligations. In M.A. Brown and J. Carmo (eds.): *Deontic Logic, Agency and Normative Systems*. Workshops in Computing, Springer, London, 1996, 50–65.

J. Carmo and A.J.I. Jones. A new approach to contrary-to-duty obligations. *This volume*.

H.-N. Castañeda. The paradoxes of deontic logic: the solution to all of them in one fell swoop. In R. Hilpinen (ed.): *New Studies in Deontic Logic*. Reidel, Dordrecht, 1981, 37–85.

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

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

J.A. van Eck. A system of temporally relative modal and deontic predicate logic and its philosophical applications. *Logique ét Analyse* 100:249–381 (1982).

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

B. Hansson. An analysis of some deontic logics. *Nôus*, 3:373–398 (1969). Reprinted in R. Hilpinen (ed.): *Deontic Logic: Introductory and Systematic Readings*. Reidel, Dordrecht, 1971, 121–147.

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.

J. Hintikka. Some main problems of deontic logic. In R. Hilpinen (ed.): *Deontic Logic: Introductory and Systematic Readings*. Reidel, Dordrecht, 1971, 59–104.

J.F. Horty. Moral dilemmas and non-monotonic logic. *Journal of Philosophical Logic* 23:35–65 (1994).

A.J.I. Jones and I. Pörn. Ideality, sub-ideality and deontic logic. *Synthese* 65:275–290 (1985).

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

B. Loewer and M. Belzer. Dyadic deontic detachment. *Synthese* 54:295–318 (1983).

L.T. McCarty. Defeasible deontic reasoning. *Fundamenta Informaticae* 21:125–148 (1994).

D. Makinson. Five faces of minimality. *Studia Logica* 52(3):339–379 (1993).

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).

M. Morreau. Prima facie and Seeming Duties. In A.J.I. Jones and M.J. Sergot (eds.): *Proc. Second International Workshop on Deontic Logic and Computer Science (DEON'94)*, Oslo, January 1994. Complex 1/94, Tano Publishers, Norway, 221–251.

H. Prakken Two approaches to the formalisation of defeasible deontic reasoning. *Studia Logica* 57(1):73–90 (1996).

H. Prakken and M.J. Sergot. Contrary-to-duty imperatives, defeasibility and violability. In A.J.I. Jones and M.J. Sergot (eds.): *Proc. Second International Workshop on Deontic Logic in Computer Science (DEON'94)*, Oslo, January 1994. Complex 1/94, Tano Publishers, Norway, 296–318.

H. Prakken and M.J. Sergot. Contrary-to-duty obligations. *Studia Logica* 57(1):91–115 (1996).

W.D. Ross. *The Right and the Good*. Oxford University Press, 1930.

M.D. Ryan. Representing defaults as sentences with reduced priority. *Proc. Third International Conference on Principles of Knowledge Representation and Reasoning (KR'92)*, Morgan Kaufman, 1992.

M.D. Ryan and P.-Y. Schobbens. The lexicographic combination of preferences. Working notes from the the Dutch/German workshop on non-monotonic reasoning techniques and their applications, Aachen 1993.

Y.H. Ryu and R.M. Lee. Deontic logic viewed as defeasible reasoning. *This volume*.

Y.-H. Tan and L.W.N. van der Torre. Multi preference semantics for a defeasible deontic logic. In H. Prakken, A.J. Muntjewerff, A. Soeteman (eds.): *Legal knowledge based systems. The relation with legal theory*. Koninklijke Vermande BV, Lelystad, 1994, 115–126.

Y.-H. Tan and L.W.N. van der Torre. How to combine ordering and minimizing in a deontic logic based on preferences. In M.A. Brown and J. Carmo (eds.): *Deontic Logic, Agency and Normative Systems*. Workshops in Computing, Springer, London, 1996, 216–232.

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.