LEENDERT W.N. VAN DER TORRE AND YAO-HUA TAN

THE MANY FACES OF DEFEASIBILITY IN DEFEASIBLE DEONTIC LOGIC

1. INTRODUCTION

Deontic logic is the logic of obligations, i.e. reasoning about what should be the case. Defeasible logic is the logic of default assumptions, i.e. reasoning about what normally is the case. In defeasible deontic logic these two are combined. An example of this combination is the sentence 'normally, you should do p'. Now the problem is what to conclude about somebody who does not do p? Is this an exception to the normality claim, or is it a violation of the obligation to do p? This confusion arises because there is a substantial overlap between deontic and defeasibility aspects. In this article we analyze this overlap, and we also show that this confusion can be avoided if one makes the proper distinctions between different types of defeasibility. Furthermore, we also show that these distinctions are essential for an adequate analysis of notorious contrary-to-duty paradoxes such as the Chisholm and Forrester paradoxes.

The main claim of this article is that the defeasible aspect of defeasible deontic logic is different from the defeasible aspect of, for example, Reiter's default rules (Reiter, 1980). Different types of defeasibility in a logic of defeasible reasoning formalize a single notion, whereas defeasible deontic logics formalize two notions. Consider first the logics of defeasible reasoning and the famous Tweety example. In the case of factual defeasibility, we say that the 'birds fly' default is *cancelled* by the fact $\neg f$, and in the case of overridden defeasibility by the 'penguins do not fly' default. By cancellation we mean, for example, that if $\neg f$ is true, then the default assumption that f is true is null and void. The truth of $\neg f$ implies that the default assumption about f is contradicted.

The fundamental difference between deontic logic and logics for defeasible reasoning is that $\neg p \wedge \bigcirc p$ is *not* inconsistent. That is the reason why the deontic operator $\bigcirc$ had to be represented as a modal operator with a possible worlds semantics, to make sure that *both* the obligation and its violation could be true at the same time. Although the obligation $\bigcirc p$ is violated by the fact $\neg p$, the obligation still has its force, so to say. This still being in force of an obligation is reflected, for example, by the fact that someone has to pay a fine even if she does $\neg p$. Even if you drive too fast, you should not drive too fast. But if penguins cannot fly, it makes no sense to state that normally they can fly. We will refer to this relation between the obligation and its violation as

overshadowing to distinguish it from *cancellation* in the case of defeasible logics. By the overshadowing of an obligation we mean that it is still in force, but it is no longer to be acted upon.

The conceptual difference between cancelling and overshadowing is analogous to the distinction between 'defeasibility' and 'violability' made by Smith (1993) and by Prakken and Sergot (1994). An essential difference between those articles and this one is that in this article we argue that violability has to be considered as a type of defeasibility too, because it also induces a constraint on strengthening of the antecedent. The main advantage of the violability-as-defeasibility perspective is that it explains the distinctions *as well as the similarities* between cancelling and overshadowing. Moreover, it can be used to analyze complicated phenomena like prima facie obligations, which have cancelling as well as overriding aspects.

In this article we give a general analysis of different types of defeasibility in defeasible deontic logics. We argue that (at least) three types of defeasibility must be distinguished in a defeasible deontic logic. First, we make a distinction between *factual defeasibility*, that formalizes overshadowing of an obligation by a violating fact, and *overridden defeasibility*, that formalizes cancelling of an obligation by other conditional obligations. Second, we show that overridden defeasibility can be further divided into *strong overridden defeasibility*, that formalizes specificity, and *weak overridden defeasibility*, that formalizes the overriding of prima facie obligations. Our general analysis can be applied to any defeasible deontic logic. Moreover, we illustrate the intuitions behind the various distinctions with preference-based semantics. We also show that these distinctions are essential for an adequate analysis of notorious contrary-to-duty paradoxes such as the Chisholm and Forrester paradoxes in a defeasible deontic logic.

1.1. *Defeasible deontic logic*

In this article we only discuss a dyadic version of deontic logic. Dyadic modal logics were introduced to formalize deontic reasoning about contrary-to-duty obligations in, for example, the Chisholm paradox that we will discuss later. See (Lewis, 1974) for an overview of several dyadic deontic logics. An example of a conditional obligation in a dyadic modal logic is $\bigcirc(h \mid r)$, which expresses that "you ought to be helped (h) when you are robbed (r)". Similarly, $\bigcirc(\neg r \mid \top)$ expresses that "you ought not to be robbed", where $\top$ stands for any tautology. If both $\bigcirc(\neg r \mid \top)$ and r are true, then we say that the obligation is *violated* by the fact r. In recent years it was argued by several authors that these dyadic obligations can be formalized in non-monotonic logics (McCarty, 1994; Horty, 1993; Ryu and Lee, 1993).

In this article we argue that contrary-to-duty obligations do have a defeasible aspect, but a different one than is usually thought. The first part of

this claim follows directly from Alchourrón's (1994) definition of a defeasible conditional as a conditional that lacks strengthening of the antecedent, represented by the inference pattern

$$\text{SA}: \frac{\bigcirc(\alpha|\beta_1)}{\bigcirc(\alpha|\beta_1 \wedge \beta_2)}$$

Alchourrón's definition is based on the idea that lack of strengthening of the antecedent is a kind of implicit non-monotonicity. The relation between strengthening of the antecedent and non-monotonicity can be made explicit with the following inference pattern *Exact Factual Detachment* EFD.[1] Exact factual detachment can be represented by the inference pattern

$$\text{EFD}: \frac{\bigcirc(\alpha|\beta), \mathcal{A}\beta}{\bigcirc(\alpha)}$$

in which $\bigcirc\alpha$ is a new, monadic modal operator, and $\mathcal{A}$ is an all-that-is-known operator (Levesque, 1990): $\mathcal{A}\phi$ is true if and only if (iff) ϕ is logically equivalent with all factual premises given. The inference pattern EFD is based on the intuition that the antecedent of a dyadic obligation restricts the focus to possible situations in which the antecedent is *assumed* to be factually true, and the consequent represent what is obligatory, given that *only* these facts are assumed. If the facts are equivalent to the antecedent, then the consequent can be considered as an absolute obligation. From the properties of $\mathcal{A}$ follows immediately that EFD is monotonic iff the dyadic obligations have strengthening of the antecedent. Dyadic deontic logics that can represent contrary-to-duty reasoning are defeasible deontic logics, because the dyadic obligations typically lack strengthening of the antecedent.[2] In this sense, contrary-to-duty obligations do have a defeasible aspect.

However, we argue that this defeasible aspect of contrary-to-duty obligations is a different one than is usually proposed. In this article, we analyze defeasibility in defeasible deontic logic by analyzing different conditions on

[1] A related idea was proposed by Boutilier (1994): 'to determine preferences based on certain actual facts, we consider only the *most ideal* worlds satisfying those facts, rather than *all* worlds satisfying those facts'. In Boutilier's logic, this means that the antecedent of his conditional is logically equivalent with the premises, i.e. he considers $\vdash \bigcirc(\alpha|KB)$, where KB is the set of premises. Von Wright (1968) proposed two ways to represent monadic obligations $\bigcirc\alpha$ in a dyadic logic: by $\bigcirc(\alpha|\top)$ and by $\bigcirc(\alpha|S)$, where S stands for the actual circumstance. Alchourrón (1994) observes that the former has been followed unanimously by all deontic logicians, although it is wrong (which follows from the semantics). In Alchourrón's words, this misrepresentation is 'the ghost of categorical norms'.

[2] The dyadic obligations can be contrasted to conditional obligations that do validate factual detachment and strengthening of the antecedent, and are typically (see e.g. (Chellas, 1974; Alchourrón, 1994)) represented by a strict implication '$>$' and a monadic operator such that $\bigcirc(\alpha|\beta) =_{\text{def}} \beta > \bigcirc\alpha$.

strengthening of the antecedent. In particular, we analyze the inference relation of defeasible deontic logics with inference patterns, in a similar way as in (Kraus, *et al*., 1990) logics of defeasible reasoning are analyzed. Moreover, we give preference-based semantic intuitions for the inference patterns. Some of the dyadic modal logics that can represent contrary-to-duty obligations have a preference-based semantics (Hansson, 1971; Makinson, 1993). The advantage of our analysis is that (1) it is applicable to any defeasible deontic logic, because of the generality of the inference patterns, and (2) it gives also a semantic explanation of the intuitions behind the inference patterns by the preference semantics.

1.2. *Different types of defeasibility*

In defeasible reasoning one can distinguish at least three types of defeasibility, based on different semantic intuitions. To illustrate the difference between the different types we discuss the penguin example in Geffner and Pearl's assumption-based default theories (Geffner and Pearl, 1992). In such theories, the 'birds fly' default rule is expressed by a factual sentence $\delta_1 \rightarrow f$ and a default sentence $\top \Rightarrow \delta_1$, and the 'penguins do not fly' default by $p \wedge \delta_2 \rightarrow \neg f$ and $p \Rightarrow \delta_2$. Here, '$\rightarrow$' is the classical material implication and '$\Rightarrow$' a kind of default implication. The δ_i constants are the so-called assumptions; for each default in the set of premises a distinct constant is introduced. Geffner and Pearl's so-called conditional entailment maximizes these assumptions, given certain constraints. In conditional entailment, the 'birds fly' default can be defeated by the fact $\neg f$, or it can be overridden by the more specific 'penguins do not fly' default. The first follows directly from $\neg f \rightarrow \neg\delta_1$, i.e. the contraposition of the factual sentence $\delta_1 \rightarrow f$, and the second follows from the fact that $p \rightarrow \neg\delta_1$ can be derived from the constraints of conditional entailment (we do not give the complicated proof; see (Geffner and Pearl, 1992) for these details). We call the first case *factual defeasibility* and the last case *overridden defeasibility*. The distinction between factual and overridden defeasibility is only the start of a classification of different types of defeasibility. To illustrate the further distinction between different types of overridden defeasibility, we consider the adapted 'penguins do not fly and live on the southern hemisphere' default $p \wedge \delta_2 \rightarrow (\neg f \wedge s)$. In some logics of defeasible reasoning, the 'birds fly' default is overridden whenever p is true. In other logics it is overridden when p is true *but only as long as* s *is not false*. If s is false, then the penguin default is no longer applicable. In the first logics the 'birds fly' default is not reinstated, whereas in the second logics it is, because it was only suspended. In other words, in the latter case the penguin default overrides the bird default only when it is applicable itself. We call the first case *strong overridden defeasibility* and the second case *weak overridden defeasibility*. The different types of overridden defeasibility

are based on different semantic intuitions. Strong overridden defeasibility is usually based on a probabilistic interpretation of defaults (most birds fly, but penguins are exceptional), like in Pearl's ϵ-semantics (Pearl, 1988). Weak overridden defeasibility is usually based on an argument-based conflict resolution interpretation (there is a conflict between the two rules, and the second one has highest priority). Examples are conditional entailment, prioritized default logic (Brewka, 1994) and several argument systems (Vreeswijk, 1993; Dung, 1993; Prakken and Sartor, 1995).

The distinction between different types of defeasibility is crucial in logics that formalize reasoning about obligations which can be overridden by other obligations. Overridden defeasibility becomes relevant when there is a (potential) conflict between two obligations. For example, there is a conflict between $\bigcirc(\alpha_1|\beta_1)$ and $\bigcirc(\alpha_2|\beta_2)$ when α_1 and α_2 are contradictory, and β_1 and β_2 are factually true. There are several different approaches to deal with deontic conflicts. In von Wright's so-called standard deontic logic SDL (Von Wright, 1951) a deontic conflict is inconsistent. In weaker deontic logics, like minimal deontic logic MDL (Chellas, 1974), a conflict is consistent and called a 'deontic dilemma'. In a defeasible deontic logic a conflict can be *resolved*, because one of the obligations overrides the other one. For example, overridden structures can be based on a notion of specificity, like in Horty's well-known example that 'you should not eat with your fingers', but 'if you are served asparagus, then you should eat with your fingers' (Horty, 1993). In such cases, we say that an obligation is *cancelled* when it is overridden, because it is analogous to cancelling in logics of defeasible reasoning. The obligation not to eat with your fingers is cancelled by the exceptional circumstances that you are served asparagus. A different kind of overridden structures have been proposed by Ross (1930) and formalized, for example, by Morreau (1996). In Ross' ethical theory, an obligation which is overridden has not become a 'proper' or actual duty, but it remains in force as a prima facie obligation. For example, the obligation not to break a promise may be overridden to prevent a disaster, but even when it is overridden it remains in force as a prima facie obligation. As actual obligation the overridden obligation is cancelled, but as prima facie obligation it is only overshadowed. Because of this difference between cancellation and overshadowing, it becomes essential not to confuse the different types of defeasibility in analyzing the deontic paradoxes. We show that if they are confused, counterintuitive conclusions follow for the Chisholm and Forrester paradoxes. In the table below the three different types of defeasible deontic logic are represented with their corresponding character (cancelling or overshadowing).

In non-deontic defeasible logic the different types of defeasibility, factual and overridden, all have a cancelling character.

	overshadowing	cancelling
Factual defeasibility	X	
Strong overridden defeasibility		X
Weak overridden defeasibility	X	X

This article is organized as follows. In Section 2 we give a detailed comparison of factual and overridden defeasibility in deontic reasoning, and we show that the Chisholm paradox can be analyzed as a case of factual defeasibility rather than overridden defeasibility. In Section 3 we focus on the overshadowing aspect of factual defeasibility as well as the cancellation aspect of overridden defeasibility by analyzing specificity, and we show that in an adequate analysis of an extension of the Forrester paradox both these aspects have to be combined. In Section 4 we focus on the cancelling aspect and the overshadowing aspect of overridden defeasibility by analyzing prima facie obligations.

2. OVERRIDDEN VERSUS FACTUAL DEFEASIBILITY

In this section we analyze the fundamental difference between overridden and factual defeasibility in a defeasible deontic logic by formalizing contrary-to-duty reasoning as a kind of overridden defeasibility as well as a kind of factual defeasibility. Moreover, we show that contrary-to-duty reasoning is best formalized by the latter one.

2.1. *Contrary-To-Duty paradoxes*

Deontic logic is hampered by many paradoxes, intuitively consistent sentences which are formally inconsistent, or from which counterintuitive sentences can be derived. The most notorious paradoxes are caused by so-called *Contrary-To-Duty* (CTD) obligations, obligations that refer to sub-ideal situations. For example, Lewis describes the following example of the CTD obligation that you should be helped when you are robbed.

Example 1 (Good Samaritan paradox) "It ought not to be that you are robbed. *A fortiori*, it ought not to be that you are robbed and then helped. But you ought to be helped, given that you have been robbed. This robbing excludes the best possibilities that might otherwise have been actualized, and the helping is needed in order to actualize the best of those that remain. Among the best possible worlds marred by the robbing, the best of the bad lot are some of those where the robbing is followed by helping." *(*Lewis, 1974*)*

In the early seventies, several dyadic modal systems were introduced to formalize CTD obligations, see (Lewis, 1974) for an overview. Unfortunately, several technical problems related to CTD reasoning persisted in the dyadic

logics, see (Tomberlin, 1981). A dyadic obligation $\bigcirc(\alpha|\beta)$ can be read as 'if β (the antecedent) is the case then α (the consequent) should be the case'. A CTD obligation is a dyadic obligation of which the antecedent contradicts the consequent of another obligation. For example, if we have $\bigcirc(\alpha_1|\top)$ and $\bigcirc(\alpha_2|\neg\alpha_1)$ then the last one is a CTD (or *secondary*) obligation and the first one is called its *primary* obligation. CTD obligations refer to optimal sub-ideal situations. In the sub-ideal situation that $\bigcirc(\alpha_1|\top)$ is violated by $\neg\alpha_1$, the best thing to do is α_2. Recently, it was observed that this aspect of violations can be formalized in non-monotonic logics (McCarty, 1994; Horty, 1993), theories of diagnosis (Tan and Van der Torre, 1994a; Tan and Van der Torre, 1994b) or qualitative decision theories (Boutilier, 1994) (see also (Powers, 1967; Jennings, 1974; Pearl, 1993; Thomason and Horty, 1996)).

Since the late seventies, several temporal deontic logics and deontic action logics were introduced, which formalize satisfactorily a special type of CTD obligations, see for example (Thomason, 1981; Van Eck, 1982; Loewer and Belzer, 1983; Makinson, 1993; Alchourrón, 1994). Temporal deontic logics formalize conditional obligations in which the consequent occurs later than the antecedent. In this temporal approach, the underlying principle of the formalization of CTD obligations is that facts of the past are not in the 'context of deliberation' (Thomason, 1981). Hence, they can formalize the Good Samaritan paradox in Example 1. However, they cannot formalize the variant of the paradox described by Forrester (see Example 4) and the following Chisholm paradox, because in these paradoxes there are CTD obligations of which the consequent occurs at the same time or even before its antecedent.

The following example describes the notorious Chisholm paradox, also called the CTD paradox, or the paradox of deontic detachment (Chisholm, 1963). The original paradox was given in a monadic modal logic. Here we give the obvious formalization in a non-defeasible dyadic logic. See (Tomberlin, 1981) for a discussion of the Chisholm paradox in several conditional deontic logics. To make our analysis as general as possible, we assume as little as possible about the deontic logic we use. The analyses given in this article in terms of inference patterns are, in principle, applicable to any deontic logic.

Example 2.1 (Chisholm paradox) Assume a dyadic deontic logic that validates at least substitution of logical equivalents and the following inference patterns (unrestricted) *Strengthening of the Antecedent* SA, *Weakening of the Consequent* WC and a version of *Deontic Detachment* DD′.[3]

[3] We do not use the 'standard' names of conditional logic (Chellas, 1980), like for example RCM for weakening of the consequent, to emphasize that our inference patterns are analysis tools at the level of inference relations. See for example the inference patterns RSA$_O$ and RSA$_V$ later in this article, which contain conditions C_O and C_V.

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

Notice that the following inference pattern *Deontic Detachment* (or transitivity) DD can be derived from WC and DD′.

$$\text{DD}: \frac{\bigcirc(\alpha|\beta), \bigcirc(\beta|\gamma)}{\bigcirc(\alpha|\gamma)}$$

Furthermore, assume the premises $\bigcirc(a \mid \top)$, $\bigcirc(t \mid a)$ and $\bigcirc(\neg t \mid \neg a)$, where $\top$ stands for any tautology, a can be read as the fact that a certain man goes to the assistance of his neighbors and t as the fact that he tells them he is coming. The premise $\bigcirc(\neg t \mid \neg a)$ is a CTD obligation of the (primary) obligation $\bigcirc(a|\top)$, because its antecedent is inconsistent with the consequent of the latter. Notice that t occurs before a in this interpretation of the propositional atoms. Hence, the example cannot be represented in a temporal deontic logic.

The paradoxical derivation of $\bigcirc(t \mid \neg a)$ from the Chisholm paradox is represented in Figure 1. The intuitive obligation $\bigcirc(a \wedge t|\top)$ can be derived by DD′ from the first two obligations. It seems intuitive, because in the ideal situation the man goes to the assistance of his neighbors and he tells them he is coming. The obligation $\bigcirc(t|\top)$ can be derived from $\bigcirc(a \wedge t|\top)$ by WC (or from the premises by DD). The obligation $\bigcirc(t|\top)$ expresses that if the man does not tell his neighbors, then the ideal situation is no longer reachable. However, from $\bigcirc(t|\top)$ the counterintuitive $\bigcirc(t|\neg a)$ can be derived by SA. This is counterintuitive, because there is no reason to tell the neighbors he is coming when the man does not go. In contrast, in this violation context the man should do the opposite! Moreover, in several deontic logics the set of obligations $\{\bigcirc(\neg t|\neg a), \bigcirc(t|\neg a)\}$ is inconsistent.

$$\cfrac{\cfrac{\cfrac{\bigcirc(t|a) \quad \bigcirc(a|\top)}{\bigcirc(a \wedge t|\top)}\ \text{DD}'}{\bigcirc(t|\top)}\ \text{WC}}{\bigcirc(t|\neg a)}\ \text{SA}$$

Fig. 1. Chisholm paradox

In this example the Chisholm paradox is presented in a normal dyadic deontic logic, to show its paradoxical character. In the next section, we analyze the paradox in a defeasible deontic logic that has only overridden defeasibility. This analysis solves the paradox, but for the wrong reasons. Finally, in

Section 2.3 we give an analysis of the Chisholm paradox in terms of factual defeasibility, which is more satisfactory. In Section 2.4 we analyze factual defeasibility with a preference semantics.

2.2. *Overridden defeasibility*

In recent years several authors have proposed to solve the Chisholm paradox by analyzing its problematic CTD obligation as a type of overridden defeasibility (see e.g. (McCarty, 1994; Ryu and Lee, 1993)).[4] The underlying idea is that a CTD obligation can be considered as a conflicting obligation that overrides a primary obligation. Although this idea seems to be very intuitive at first sight, we claim that the perspective of CTD obligations as a kind of overridden defeasibility is misleading. It is misleading, because although this perspective yields most (but not all!) of the correct conclusions for the Chisholm paradox, it does so for the wrong reasons. We show that it is more appropriate to consider the CTD obligation as a kind of factual defeasibility. This does not mean that there is no place for overridden defeasibility in deontic logic. By a careful analysis of an extended version of another notorious paradox of deontic logic, the Forrester paradox, we show that sometimes combinations of factual and overridden defeasibility are needed to represent defeasible deontic reasoning. But first we give our analysis of the Chisholm paradox. The following example shows that the counterintuitive obligation of Example 2.1 cannot be derived in a defeasible deontic logic with overridden defeasibility. For our argument we use a notion of overridden based on specificity.

Example 2.2 (Chisholm paradox, continued) Assume that SA is replaced by the following Restricted Strengthening of the Antecedent rule RSA_O. RSA_O contains the so-called non-overridden condition C_O, which requires that $\bigcirc(\alpha|\beta_1)$ is not overridden for $\beta_1 \wedge \beta_2$ by some more specific $\bigcirc(\alpha'|\beta')$.[5]

$$\text{RSA}_O : \frac{\bigcirc(\alpha|\beta_1), C_O}{\bigcirc(\alpha|\beta_1 \wedge \beta_2)}$$

where condition C_O is defined as follows:

> C_O: there is no premise $\bigcirc(\alpha' \mid \beta')$ such that $\beta_1 \wedge \beta_2$ logically implies β', β' logically implies β_1 and not vice versa and α and α' are contradictory.

The 'solution' for the paradox is represented in Figure 2. This figure should be read as follows. The horizontal lines represent *possible* derivation steps.

[4] McCarty (1994) does not analyze the Chisholm paradox but the so-called Reykjavic paradox, which he considers to contain 'two instances of the Chisholm paradox, each one interacting with the other'.

[5] The overridden condition C_O is based on a simplified notion of specificity, because background knowledge is not taken into account and an obligation cannot be overridden by

Blocked derivation steps are represented by dashed lines. For example, the last derivation step is blocked, and the cause of the blocking is represented by the obligation $\bigcirc(\neg t|\neg a)$ above the blocked inference rule. We compare the blocked derivation in Figure 2 with the derivation in Figure 1. The intuitive obligation $\bigcirc(t|\top)$ can still be derived by DD (hence, by DD′ and WC) from the first two obligations. From $\bigcirc(t|\top)$ the counterintuitive $\bigcirc(t|\neg a)$ cannot be derived by RSA_O, because $\bigcirc(t\,|\,\top)$ is overridden for $\neg a$ by the CTD obligation $\bigcirc(\neg t|\neg a)$, i.e. C_O is false. Hence, the counterintuitive obligation is cancelled by the exceptional circumstances that the man does not go to the assistance.

$$\dfrac{\dfrac{\dfrac{\bigcirc(t|a) \quad \bigcirc(a|\top)}{\bigcirc(a \wedge t|\top)}\ \text{DD}'}{\bigcirc(t|\top)}\ \text{WC} \qquad \begin{array}{c}\bigcirc(\neg t|\neg a)\\ \downarrow \\ (\text{RSA}_O)\end{array}}{\bigcirc(t|\neg a)}$$

Fig. 2. Chisholm paradox solved by overridden defeasibility

Overridden defeasibility yields intuitive results from the Chisholm paradox, but for the wrong reasons. A simple counterargument against the solution of the paradox in Example 2.2 is that overriding based on specificity does not solve the paradox anymore when the premise $\bigcirc(a|\top)$ is replaced by another premise with a non-tautological antecedent. For example, if it is replaced by $\bigcirc(a|i)$, where i can be read as the fact that the man is personally invited to assist. Another counterargument against the solution of the paradox for *any* definition of overridden is that the derivation of $\bigcirc(t|\neg a)$ is also counterintuitive when the set of premises contains only the first two obligations, as is the case in the following example.

Example 2.3 (Chisholm paradox, continued) Assume only the premises $\bigcirc(a|\top)$ and $\bigcirc(t|a)$. Again the intuitive obligation $\bigcirc(t|\top)$ can be derived by DD. From this derived obligation the counterintuitive $\bigcirc(t\,|\,\neg a)$ can be derived by RSA_O, because there is no CTD obligation which cancels the counterintuitive obligation.

In (Tan and Van der Torre, 1994b) we dubbed the intuition that the inference of the obligation $\bigcirc(t|\top)$ is intuitive but not the inference of the obligation

more than one obligation. A more sophisticated definition of overridden can be found in the literature of logics of defeasible reasoning. For our purposes this simple definition is enough, because it is a weak definition (most definitions of specificity are extensions of this definition). For a discussion on the distinction between background and factual knowledge, see (Van der Torre, 1994).

$$\dfrac{\dfrac{\dfrac{\bigcirc(t|a) \quad \bigcirc(a|\top)}{\bigcirc(a \wedge t|\top)}\ \text{DD}'}{\bigcirc(t|\top)}\ \text{WC}}{\bigcirc(t|\neg a)}\ \text{RSA}_O$$

Fig. 3. Chisholm paradox, continued

$\bigcirc(t|\neg a)$ as 'deontic detachment as a defeasible rule'. Unrestricted strengthening of the antecedent cannot be applied to the obligation $\bigcirc(t|\top)$, derived by DD. This restriction is the characteristic property of defeasible conditionals, see the discussion in Section 1.1. The underlying intuition is that the inference of the obligation of the man to tell his neighbors that he is coming is made *on the assumption that he goes to their assistance*. If he does not go, then this assumption is violated and the obligation based on this assumption is factually defeated. We say that the man should tell his neighbors, unless he does not go to their assistance.

The problematic character of DD is well-known from the Chisholm paradox. A popular 'solution' of the paradox is not to accept DD′ for a deontic logic. However, this rejection of DD′ causes serious semantic problems for these logics. For example, (Tomberlin, 1981) showed that there are semantic problems related to the rejection of DD′ for Mott's solution of the Chisholm paradox (Mott, 1973). Moreover, the following so-called apples-and-pears problem (Tan and Van der Torre, 1996) shows that similar problems occur when RSA_O, WC and the *Conjunction* inference pattern AND are accepted. This last rule is accepted by many deontic logics. For examples of deontic logics *not* satisfying the AND rule, see Chellas' CKD (Chellas, 1974; Chellas, 1980), which is a nonnormal modal deontic logic, or the minimizing logic $\bigcirc_\exists(\alpha \mid \beta)$ in (Tan and Van der Torre, 1996). For examples not validating the WC rule, see S.O. Hansson's Preference-based Deontic Logic (PDL) (Hansson, 1990), Brown and Mantha's logic (Brown and Mantha, 1991) and the ordering logic $\bigcirc(\alpha|\beta)$ in (Tan and Van der Torre, 1996).[6]

Example 3 (Apples-and-Pears problem) Assume a dyadic deontic logic that validates at least substitution of logical equivalents and the inference

[6]An alphabetic variant of Example 3 is the following version of the Chisholm paradox, in which the conditional obligation is represented as an absolute obligation. However, it is usually argued that the premise $\bigcirc(a \rightarrow t|\top)$ does not represent the conditional obligation correctly.

Example 2.4 (Chisholm paradox continued) Consider the premises $\bigcirc(a \mid \top)$ and $\bigcirc(a \rightarrow t|\top)$. The intuitive obligation $\bigcirc(t \mid \top)$ is derived from the two premises by CC (see Example 3). However, from this derived obligation the counterintuitive $\bigcirc(t|\neg a)$ can be derived by SA or RSA_O.

patterns RSA$_O$, WC and the following conjunction rule AND.

$$\text{AND}: \frac{\bigcirc(\alpha_1|\beta), \bigcirc(\alpha_2|\beta)}{\bigcirc(\alpha_1 \wedge \alpha_2|\beta)}$$

Notice that the following inference pattern *Consequential Closure* (CC) can be derived from WC and AND.

$$\text{CC}: \frac{\bigcirc(\alpha_1|\beta), \bigcirc(\alpha_1 \rightarrow \alpha_2|\beta)}{\bigcirc(\alpha_2|\beta)}$$

Furthermore, assume as premise sets $S = \{\bigcirc(a \vee p|\top), \bigcirc(\neg a|\top)\}$ and $S' = \{\bigcirc(a \vee p|\top), \bigcirc(\neg a|\top), \bigcirc(\neg p|a)\}$, where a can be read as 'buying apples' and p as 'buying pears'. A derivation of the counterintuitive obligation $\bigcirc(p|a)$ from S is represented in Figure 4. This obligation is considered to be counterintuitive, because it is not grounded in the premises. If a is true, then the first premise $\bigcirc(a \vee p|\top)$ is fulfilled and the second premise $\bigcirc(\neg a|\top)$ is violated. Since the first premise is already fulfilled, there is intuitively no reason why p should be obliged given the fact that a. The intuitive obligation $\bigcirc(\neg a \wedge p|\top)$ can be derived by AND. From this obligation, the obligation $\bigcirc(p|\top)$ is derived by WC (hence, from the premise set by CC). From this derived obligation, the counterintuitive obligation $\bigcirc(p \mid a)$ can be derived by RSA$_O$. The counterintuitive derivation is not derivable from S' by RSA$_O$, because the CTD obligation $\bigcirc(\neg p|a)$ overrides the obligation $\bigcirc(p|\top)$ for a. However, this solution for S' does not suffice for S, just like the solution in Example 2.2 does not suffice for Example 2.3.

$$\cfrac{\cfrac{\cfrac{\bigcirc(a \vee p|\top) \quad \bigcirc(\neg a|\top)}{\bigcirc(\neg a \wedge p|\top)}\ \text{AND}}{\bigcirc(p|\top)}\ \text{WC}}{\bigcirc(p|a)}\ \text{RSA}_O$$

Fig. 4. Apples-and-pears problem with overridden defeasibility

The examples show that CTD reasoning (i.e., reasoning about sub-ideal behavior) cannot be formalized satisfactorily in a defeasible deontic logic with only overridden defeasibility.

2.3. *Factual defeasibility*

As an illustrative example of a formalization of factual defeasibility, we introduce a deontic version of a labeled deductive system as it was introduced by Gabbay (1991), which is closely related to the proof theoretic approach of

the inference patterns. Assume a finite propositional base logic $\mathcal{L}$ and labeled dyadic conditional obligations $\bigcirc(\alpha|\beta)_L$, with α and β sentences of $\mathcal{L}$ and L a set of sentences of $\mathcal{L}$. Roughly speaking, the label L is a record of the consequents of all the premises that are used in the derivation of $\bigcirc(\alpha|\beta)$. The use of the label can be illustrated by the distinction between explicit and implicit obligations. An explicit obligation is an obligation that has been uttered explicitly (an imperative), and an implicit obligation is an obligation that follows from explicit obligations. The distinction between explicit and implicit obligations is analogous to the distinction between explicit and implicit belief, introduced by Levesque to solve the logical omniscience problem (Levesque, 1984). The consequent of a labeled obligation represents an implicit obligation and its label represents the explicit obligations from which the implicit obligation is derived.

Labeled deontic logic works as follows. Each formula occurring as a premise in the derivation has its own consequent in its label. We assume that the antecedent and the label of an obligation are always consistent. The label of an obligation derived by an inference rule is the union of the labels of the premises used in this inference rule. The labels formalize the assumptions on which an obligation is derived, and the consistency check C_V checks that the assumptions are not violated. Hence, the premises used in the derivation tree are not violated by the antecedent of the derived obligation, or, alternatively, the derived obligation is not a CTD obligation of these premises.[7] Below are some labeled versions of inference schemes.

$$\text{RSA}_V : \frac{\bigcirc(\alpha|\beta_1)_L, C_V}{\bigcirc(\alpha|\beta_1 \wedge \beta_2)_L}, C_V : L \cup \{\beta_1 \wedge \beta_2\} \text{ is consistent}$$

$$\text{WC}_V : \frac{\bigcirc(\alpha_1|\beta)_L}{\bigcirc(\alpha_1 \vee \alpha_2|\beta)_L}$$

$$\text{DD}'_V : \frac{\bigcirc(\alpha|\beta)_{L_1}, \bigcirc(\beta|\gamma)_{L_2}, C_V}{\bigcirc(\alpha \wedge \beta|\gamma)_{L_1 \cup L_2}}, C_V : L_1 \cup L_2 \cup \{\gamma\} \text{ is consistent}$$

$$\text{AND}_V : \frac{\bigcirc(\alpha_1|\beta)_{L_1}, \bigcirc(\alpha_2|\beta)_{L_2}, C_V}{\bigcirc(\alpha_1 \wedge \alpha_2|\beta)_{L_1 \cup L_2}}, C_V : L_1 \cup L_2 \cup \{\beta\} \text{ is consistent}$$

The following example illustrates that RSA_V is better than RSA_O for modeling the Chisholm paradox, because RSA_V yields all of the intended conclusions of the Examples 2.1-2.4, but none of the counterintuitive conclusions produced by RSA_O.

[7] Notice that only the premises are checked from which the obligation is derived. If all premises are checked, then we have some variant of a defeasible reasoning scheme known as System Z (Pearl, 1990; Boutilier, 1994), which has the drawback that it does not validate $\{\bigcirc(p|\top), \bigcirc(q|\top)\} \vdash \bigcirc(p|\neg q)$.

Example 2.5 (Chisholm paradox, continued) Assume a labeled deductive system that validates at least substitution of logical equivalents and the inference patterns RSA_V, WC_V and DD'_V. Furthermore, assume the premises $\bigcirc(a|\top)_{\{a\}}$ and $\bigcirc(t|a)_{\{t\}}$. Figure 5 shows how factual defeasibility blocks the counterintuitive derivation of Figure 1. The obligation $\bigcirc(t \mid \neg a)_{\{a,t\}}$ cannot be derived from $\bigcirc(t \mid \top)_{\{a,t\}}$, because $C_V : \{a,t\} \cup \{\neg a\}$ is not consistent. It does not use a CTD obligation like the blocked derivation in Figure 2, thus it also blocks the counterintuitive derivation in Figure 3.

$$\dfrac{\dfrac{\dfrac{\bigcirc(t|a)_{\{t\}} \quad \bigcirc(a|\top)_{\{a\}}}{\bigcirc(a \wedge t|\top)_{\{a,t\}}}\ \text{DD}'}{\bigcirc(t|\top)_{\{a,t\}}}\ \text{WC}}{\bigcirc(t|\neg a)_{\{a,t\}}}\ (\text{RSA}_V)$$

Fig. 5. Chisholm paradox solved by factual defeasibility

It can easily be checked that the counterintuitive derivation of $\bigcirc(p|a)$ by RSA_O in Example 3 is blocked by RSA_V too. The examples show that CTD structures sometimes look like overridden defeasible reasoning structures, but a careful analysis shows that they are actually cases of factual defeasibility. There is no difference between the overridden and factual defeasibility analyses of CTD structures in Example 2.2 and 2.5, respectively, because in these examples the two restrictions C_O and C_V coincide for strengthening of the antecedent.

The reader might wonder why we consider condition C_V to be a type of factual defeasibility. In this article, we only discuss conditional obligations, and how these can be derived from each other. Facts do not seem to come into the picture here. However, a closer analysis reveals that factual defeasibility is indeed the underlying mechanism. The antecedent of a dyadic obligation restricts the focus to possibilities in which the antecedent is *assumed* to be factually true, and the consequent represents what is obligatory, given that these facts are assumed. Hence, the consequent refers to 'the best of the bad lot'. As we discussed in the introduction, these facts can be made explicit with a kind of factual detachment, for example with EFD. From the Chisholm paradox $\bigcirc(a \mid \top), \bigcirc(t|a), \bigcirc(\neg t \mid \neg a)$ and $\mathcal{A}\top$, we can derive $\bigcirc t$ by EFD, and from $\mathcal{A}\neg a$ we can derive $\bigcirc\neg t$, but not $\bigcirc t$. Hence, by adding a fact $(\neg a)$ we loose a deontic conclusion $(\bigcirc t)$.

Moreover, a comparison with, for example, prioritized default logic (Brewka, 1994) illustrates that C_V is a kind of factual defeasibility. Consider the classical example of non-transitivity of default rules, which consists

of the default rules that 'normally, students are adults' ($\frac{s:a}{a}$) and that 'normally, adults are employed' ($\frac{a:e}{e}$). Given that we know that somebody is a student, we can defeat the default conclusion that this person is employed in two ways. Either, it is defeated by the more specific default rule that students are normally unemployed ($\frac{s:\neg e}{\neg e}$), which is a case of overridden defeasibility, or it is defeated by the defeating fact ($\neg a$) that the particular student is known to be no adult. This latter case of defeasibility is the type of factual defeasibility that is analogous to the defeasibility in the Chisholm paradox.

This analogy with default logic also illustrates what we mean by deontic detachment as a defeasible rule. The transitivity of the two default rules above can be blocked either by overridden or factual defeasibility. If neither of the two are the case, then the transitivity holds. In this sense one could say that in default logic transitivity holds as a defeasible rule. Analogously, we say that deontic detachment holds as a defeasible rule. If we only know $\bigcirc(t \,|\, a)_{\{t\}}$ and $\bigcirc(a \,|\, \top)_{\{a\}}$, then we can apply deontic detachment, which results in $\bigcirc(t|\top)_{\{a,t\}}$. But this detachment is defeated if we assume in the antecedent of this conclusion that $\neg a$ is true.

2.4. *Preference semantics*

In this section we formalize the Chisholm paradox in so-called contextual deontic logic CDL (Van der Torre and Tan, 1996). To illustrate the notion of 'context' of our contextual deontic logic, we consider the following distinction between what we call 'contextual' and 'conditional' obligations for dyadic deontic logics. Technically, the distinction means that a conditional obligation is valid in all cases in which its antecedent is true. It validates strengthening of the antecedent, whereas this is not necessarily the case for contextual obligations. These may be only true in some of these cases.[8] However, in a *dyadic* deontic logic this notion of context is quite restrictive. It only means that in *exactly* the case β the obligation is valid, because any $\beta \wedge \beta'$ can be outside the context. In our contextual deontic logic CDL, dyadic obligations are generalized with an 'unless γ' condition. A contextual obligation is written as $\bigcirc(\alpha|\beta\backslash\gamma)$. The context of a contextual obligation is all cases β except the cases γ. $\bigcirc(\alpha|\beta\backslash\gamma)$ can be compared with the Reiter default rule $\frac{\beta:\neg\gamma}{\alpha}$, where $\neg\gamma$ is the justification of the default rule (Reiter, 1980). For an axiomatization of CDL in Boutilier's modal preference logic CT4O, see (Van der Torre and Tan, 1996).

[8] Loewer and Belzer (1983) make another distinction between dyadic deontic logics that validate deontic detachment DD and factual detachment FD. Our reading is related to the reading of so-called 'contextual' obligations by Prakken and Sergot (1996). They call a dyadic obligation $\bigcirc(\alpha|\beta)$ a contextual obligation if its antecedent (called the context) β stands for 'a constellation of acts or situations that agents regard as being settled in determining what they should do'. See also the discussion on circumstances in (Hansson, 1971).

The unless clause formalizes a kind of factual defeasibility, because it blocks strengthening of the antecedent (thus it is defeasibility) and it does not refer to any other obligation for this blocking (thus it is factual). The crucial observation of the Chisholm paradox below is that if the premises are valid in all cases (i.e. have a context 'unless $\bot$', where $\bot$ is a contradiction), then the derived obligations may still be only valid in a restricted context. The context encodes in such a case the assumptions from which an obligation is derived, i.e. when the obligation is factually defeated. The contextual obligations are in a sense similar to labeled obligations, which shows that the labels of the labeled obligations formalize the context in which an obligation is valid.

Moreover, the formalization of the Chisholm paradox in contextual deontic logic gives an intuitive semantic interpretation of factual defeasibility. The preference semantics represent the notion of deontic choice. A preference of α_1 over α_2 means that if an agent can choose between α_1 and α_2, she should choose α_1 (see e.g. (Jennings, 1974)). An obligation for α is formalized by a preference of α over $\neg\alpha$. Thus, if the agent can choose between α and $\neg\alpha$, then she should choose α. Similarly, a conditional obligation for α if β is formalized by a preference of $\alpha\wedge\beta$ over $\neg\alpha\wedge\beta$. This preference is formalized by condition (3) of Definition 1 below. The other conditions (1) and (2) of Definition 1 formalize the condition that in order to choose between α and $\neg\alpha$, these opportunities must be logically possible (called the contingency clause by von Wright). Notice that condition (2) is a difference with labeled obligations, because we did not impose the condition that it is possible to violate a labeled obligation, although we trivially could have done so.

Definition 1 (Contextual obligation) Let $M = \langle W, \leq, V\rangle$ be a Kripke model that consists of W, a set of worlds, $\leq$, a binary reflexive and transitive relation on W, and V, a valuation of the propositions in the worlds. Moreover, let α, β and γ be propositional sentences. The model M satisfies the obligation 'α should be the case if β is the case unless γ is the case', written as $M \models \bigcirc(\alpha|\beta\backslash\gamma)$, iff

1. $W_1 = \{w \in W \mid M, w \models \alpha \wedge \beta \wedge \neg\gamma\}$ is nonempty, and
2. $W_2 = \{w \in W \mid M, w \models \neg\alpha \wedge \beta\}$ is nonempty, and
3. for all $w_1 \in W_1$ and $w_2 \in W_2$, we have $w_2 \not\leq w_1$.

At first sight, it might seem more intuitive to say '$w_1 < w_2$' in condition 3 of Definition 1. However, it is well-known from preference logics (Von Wright, 1963) that such a condition is much too strong. For example, consider this strong definition, two obligations $\bigcirc(p|\top\backslash\bot)$, and $\bigcirc(q|\top\backslash\bot)$ and a model with $p \wedge \neg q$ and $\neg p \wedge q$ worlds. The obligation $\bigcirc(p\,|\,\top\backslash\bot)$ says that the first world is strictly preferred over the second one, whereas the obligation $\bigcirc(q|\top\backslash\bot)$ implies the opposite. With other words, a model of the obligations cannot contain $p \wedge \neg q$ and $\neg p \wedge q$ worlds. For a further discussion on this topic, see (Tan and Van der Torre, 1996).

To illustrate the properties of CDL, we compare it with Bengt Hansson's dyadic deontic logic. First we recall some well-known definitions and properties of this logic. In Bengt Hansson's classical preference semantics (Hansson, 1971), as studied by (Lewis, 1974),a dyadic obligation, which we denote by $\bigcirc_{HL}(\alpha|\beta)$, is true in a model iff 'the minimal (or preferred) β worlds satisfy α'. A weaker version of this definition, which allows for moral dilemmas, is that $\bigcirc^{w}_{HL}(\alpha|\beta)$ is true in a model iff there is an equivalence class of minimal β worlds that satisfy α, or there is an infinite descending chain in which α is true in all β worlds below a certain β world.

Definition 2 (Minimizing) Let $M = \langle W, \leq, V\rangle$ be a Kripke model and $|\alpha|$ be the set of all worlds of W that satisfy α. M satisfies the weak Hansson-Lewis obligation 'α should be the case if β is the case', written as $M \models \bigcirc^{w}_{HL}(\alpha|\beta)$, iff there is a world $w_1 \in|\alpha \wedge \beta|$ such that for all $w_2 \in|\neg\alpha \wedge \beta|$ we have $w_2 \not\leq w_1$.

The following proposition shows that the expression $\bigcirc^{w}_{HL}(\alpha\,|\,\beta)$ corresponds to a weak Hansson-Lewis minimizing obligation. For simplicity, we assume that there are no infinite descending chains.

Proposition 1 Let $M = \langle W, \leq, V\rangle$ be a Kripke model like in Definition 1, such that there are no infinite descending chains. As usual, we write $w_1 < w_2$ for $w_1 \leq w_2$ and not $w_2 \leq w_1$, and $w_1 \sim w_2$ for $w_1 \leq w_2$ and $w_2 \leq w_1$. A world w is a minimal β-world, written as $M, w \models_{\leq} \beta$, iff $M, w \models \beta$ and for all $w' < w$ holds $M, w' \not\models \beta$. A set of worlds is an equivalence class of minimal β-worlds, written as E_β, iff there is a w such that $M, w \models_{\leq} \beta$ and $E_\beta = \{w' \mid M, w' \models \beta \text{ and } w \sim w'\}$. We have $M \models \bigcirc^{w}_{HL}(\alpha|\beta)$ iff there is an E_β such that $E_\beta \subseteq|\alpha|$.

Proof $\Leftarrow$ Follows directly from the definitions. Assume there is a w such that $M, w \models_{\leq} \beta$ and $E_\beta = \{w' \mid M, w' \models \beta \text{ and } w \sim w'\}$ and $E_\beta \subseteq|\alpha|$. For all $w_2 \in|\neg\alpha \wedge \beta|$ we have $w_2 \not\leq w$.

$\Rightarrow$ Assume that there is a world $w_1 \in|\alpha \wedge \beta|$ such that for all $w_2 \in|\neg\alpha \wedge \beta|$ we have $w_2 \not\leq w_1$. Let w be a minimal β-world such that $M, w \models_{\leq} \beta$ and $w \leq w_1$ (that exists because there are no infinite descending chains), and let $E_\beta = \{w' \mid M, w' \models \beta \text{ and } w \sim w'\}$.

Now we are ready to compare our contextual deontic logic with Bengt Hansson's dyadic deontic logic. The following proposition shows that under a certain condition, the contextual obligation $\bigcirc(\alpha|\beta\backslash\gamma)$ is true in a model if a set of the weak Hansson-Lewis minimizing obligations $\bigcirc^{w}_{HL}(\alpha|\beta')$ is true in the model.

Proposition 2 Let $M = \langle W, \leq, V\rangle$ be a Kripke model such as in Definition 1, that has no worlds that satisfy the same propositional sentences. Hence, we identify the set of worlds with a set of propositional interpretations, such that

there are no duplicate worlds. As usual, for propositional α we say $M \models \alpha$ iff for all $w \in W$ we have $M, w \models \alpha$. $M \models \bigcirc(\alpha|\beta\backslash\gamma)$ iff there are $\alpha \wedge \beta \wedge \neg\gamma$ and $\neg\alpha \wedge \beta$ worlds, and for all propositional β' such that $M \models \beta' \rightarrow \beta$ and $M \not\models \beta' \rightarrow \gamma$, we have $M \models \bigcirc^{w}_{HL}(\alpha|\beta')$.

Proof $\Rightarrow$ Follows directly from the semantic definitions. $\Leftarrow$ Every world is characterized by a unique propositional sentence. Let $\overline{w}$ denote the sentence that uniquely characterizes world w. Proof by contraposition. If we have $M \not\models \bigcirc(\alpha \,|\, \beta \setminus \gamma)$, then there are w_1, w_2 such that $M, w_1 \models \alpha \wedge \beta \wedge \neg\gamma$ and $M, w_2 \models \neg\alpha \wedge \beta$ and $w_2 \leq w_1$. Choose $\beta' = \overline{w_1} \vee \overline{w_2}$. The world w_2 is an element of the preferred β' worlds, because there are no duplicate worlds. (If duplicate worlds are allowed, then there could be a β' world w_3 which is a duplicate of w_1, and which is strictly preferred to w_1 and w_2.) We have $M, w_2 \not\models \alpha$ and therefore $M \not\models \bigcirc^{w}_{HL}(\alpha|\beta')$,

In the beginning of this section, we discussed the distinction between conditional and contextual dyadic obligations. In this terminology, the obligations $\bigcirc(\alpha \,|\, \beta \setminus \bot)$ are conditional obligations and we write $\bigcirc(\alpha \,|\, \beta)$[9] for $\bigcirc(\alpha \,|\, \beta \setminus \bot)$. The following corollary for conditional obligations follows directly from Proposition 2.

Corollary 1 Let $M = \langle W, \leq, V \rangle$ be a Kripke model like in Definition 1, that has no worlds that satisfy the same propositional sentences. We have $M \models \bigcirc(\alpha|\beta\backslash\bot)$ iff there are $\alpha\wedge\beta$ and $\neg\alpha\wedge\beta$ worlds, and for all propositional β' such that $M \models \beta' \rightarrow \beta$ and $M \not\models \neg\beta'$, we have $M \models \bigcirc^{w}_{HL}(\alpha|\beta')$.

The following proposition shows several properties of contextual obligations. It shows that strengthening of the antecedent is blocked by γ (besides by the check that the choice alternatives $\alpha \wedge \beta_1 \wedge \beta_2$ and $\neg\alpha \wedge \beta_1 \wedge \beta_2$ are logically possible).

Proposition 3 Contextual deontic logic validates the following inference patterns.[10]

$$\text{RSA}_V : \frac{\bigcirc(\alpha|\beta_1\setminus\gamma), C_V}{\bigcirc(\alpha|\beta_1 \wedge \beta_2\setminus\gamma)}, \quad C_V : \begin{array}{l} \alpha \wedge \beta_1 \wedge \beta_2 \wedge \neg\gamma \text{ is consistent, } and \\ \neg\alpha_1 \wedge \beta_1 \wedge \beta_2 \text{ is consistent} \end{array}$$

$$\text{WC}_V : \frac{\bigcirc(\alpha_1 \wedge \alpha_2|\beta\setminus\gamma), C_V}{\bigcirc(\alpha_1|\beta\setminus\gamma \vee \neg\alpha_2)}, C_V : \neg\alpha_1 \wedge \beta \text{ is consistent}$$

[9] These so-called *ordering* obligations $\bigcirc(\alpha|\beta)$ lack weakening of the consequent, see (Tan and Van der Torre, 1996) and Proposition 3.

[10] The consistency checks of C_V can also be expressed in the language if we enrich the logic with a modal consistency operator, see (Tan and Van der Torre, 1996; Van der Torre and Tan, 1996).

$$\text{DD}'_V : \frac{\bigcirc(\alpha|\beta\backslash\theta), \bigcirc(\beta|\gamma\backslash\theta), C_V}{\bigcirc(\alpha\wedge\beta|\gamma\backslash\theta)}, C_V : \alpha\wedge\beta\wedge\gamma\wedge\neg\theta \text{ is consistent}$$

Proof The inference patterns can easily be checked in the preference semantics. Consider the inference pattern WC_V. Assume a model M such that $M \models \bigcirc(\alpha_1 \wedge \alpha_2 | \beta \setminus \gamma)$. Let $W_1 = \{w \mid M, w \models \alpha_1 \wedge \alpha_2 \wedge \beta \wedge \neg\gamma\}$ and $W_2 = \{w \mid M, w \models \neg(\alpha_1 \wedge \alpha_2) \wedge \beta\}$. Definition 1 says that W_1 and W_2 are non-empty, and $w_2 \not\leq w_1$ for every $w_1 \in W_1$ and $w_2 \in W_2$. Moreover, let $W'_1 = \{w \mid M, w \models \alpha_1 \wedge \beta \wedge \neg(\gamma \vee \neg\alpha_2)\}$ and $W'_2 = \{w \mid M, w \models \neg\alpha_1 \wedge \beta\}$. We have $W'_1 = W_1$ and $W'_2 \subseteq W_2$, and therefore $w_2 \not\leq w_1$ for all $w_1 \in W'_1$ and $w_2 \in W'_2$. Moreover, W'_1 is non-empty, because W_1 is non-empty. Hence, if W'_2 is non-empty (condition C_V), then $M \models \bigcirc(\alpha_1 \mid \beta \setminus \gamma \vee \neg\alpha_2)$. The proofs of the other inference patterns are analogous and left to the reader.

The following example illustrates that now the Chisholm paradox can be analyzed in contextual deontic logic. In the Chisholm paradox, the *premises* do not have exceptions. Hence, the premises are conditional obligations, i.e. contextual obligations with context 'unless $\bot$'. Moreover, the example shows that factual defeasibility of the Chisholm paradox is caused by contextual reasoning, because the *premises* do not have exceptions, only derived obligations have exceptions. Thus, this aspect of factual defeasibility is quite different from defeasibility related to exceptional circumstances or abnormality formalized in logics of defeasible reasoning, because in that case the premises are subject to exceptions.

Example 2.6 (Chisholm paradox, continued) Consider the set of obligations $S = \{\bigcirc(a|\top\backslash\bot), \bigcirc(t|a\backslash\bot)\}$. The solution of the counterintuitive derivation of the Chisholm paradox in Example 2.3 is represented in Figure 6. The obligation $\bigcirc(t \mid \top \setminus \neg a)$ represents that the man should tell his neighbors, unless he does not go to their assistance.

$$\dfrac{\dfrac{\dfrac{\bigcirc(t|a\backslash\bot) \quad \bigcirc(a|\top\backslash\bot)}{\bigcirc(a\wedge t|\top\backslash\bot)}\ \text{DD}'_V}{\bigcirc(t|\top\backslash\neg a)}\ \text{WC}_V}{\bigcirc(t|\neg a\backslash\neg a)}\ (\text{RSA}_V)$$

Fig. 6. Chisholm paradox solved by factual defeasibility

The following example explains the factual defeasibility of the Chisholm paradox by preference semantics.

Example 2.7 (Chisholm paradox, continued) Consider the set of obligations $S = \{\bigcirc(a|\top\backslash\bot), \bigcirc(t|a\backslash\bot), \bigcirc(\neg t|\neg a\backslash\bot)\}$. A typical model M of S is

given in Figure 7. This figure should be read as follows. The circles represent non-empty sets of worlds, that satisfy the propositions contained in them. Each circle represents an equivalence class of the partial pre-ordering $\leq$ of the model (the ordering partitions the worlds of the model into a set of equivalence classes). The arrows represent strict preferences for all worlds in the equivalence classes. For example, we have $M \models \bigcirc(\neg t|\neg a \setminus \bot)$, because for all $w_1 \in|\neg t \wedge \neg a|$ and $w_2 \in|t \wedge \neg a|$ we have $w_2 \not\leq w_1$. The condition $\neg a$ corresponds to the semantic concept of zooming in on the ordering. In the figure, this zooming in on the ordering is represented by a dashed box. For the evaluation of $M \models \bigcirc(\neg t|\neg a \setminus \bot)$, only the ordering within the dashed box is considered. As we observed in the analyses of the Chisholm paradox given above, the most important thing is that $\bigcirc(t|\neg a \setminus \gamma)$ does not follow from the premises for any γ. This is true for contextual deontic logic. The crucial observation is that we have $M \not\models \bigcirc(t|\neg a \setminus \gamma)$ for any γ such that $M \not\models t \wedge \neg a \wedge \neg\gamma$, because for all $w_1 \in|t \wedge \neg a \wedge \neg\gamma|$ for any γ, and for all $w_2 \in|\neg t \wedge \neg a|$, we have $w_2 \leq w_1$ (and even $w_2 < w_1$). Furthermore, we have $M \not\models \bigcirc^{w}_{HL}(t|\neg a \wedge \neg\gamma)$ for any γ such that there exists a $\neg t \wedge \neg a \wedge \neg\gamma$ world. In other words, t is not true in an equivalence class of most preferred $\neg a \wedge \neg\gamma$ worlds.

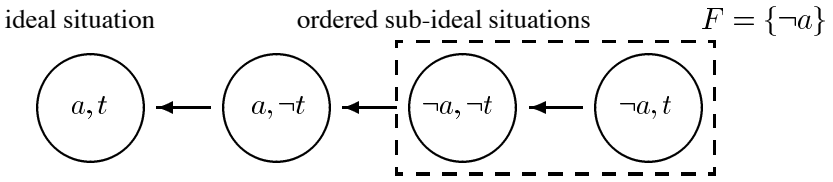

Fig. 7. Preference relation of the Chisholm paradox

Our discussion of the Chisholm paradox showed the fundamental distinction between overridden and factual defeasibility. Contrary-to-duty reasoning can be formalized as a kind of overridden defeasibility as well as a kind of factual defeasibility, and we showed that it is best formalized by the latter. The preference-based semantics illustrates where this type of factual defeasibility comes from. Semantically, the antecedent zooms in on the context of the preference ordering. The inference pattern WC corresponds semantically to introducing exceptions of this context. In the Chisholm paradox, the derivation of $\bigcirc(t|\top \setminus \neg a)$ from $\bigcirc(a \wedge t|\top \setminus \bot)$ says that the preference for t is not valid within the context $\neg a$. As shown in Figure 7, in this violation context the preferences can be the other way around.

Finally, we compare our contextual deontic logic with dyadic deontic logics. First, the Hansson-Lewis minimizing obligations (Hansson, 1971; Lewis, 1974) have too much factual defeasibility, because they do not have any strengthening of the antecedent. This is a result of the fact that every obligation

can itself be derived by weakening of the consequent. Thus, it is never safe to apply strengthening of the antecedent, because any strengthening can result in an exceptional context. Second, Chellas-type of dyadic obligations consisting of a strict implication and a monadic operator $\bigcirc(\alpha|\beta) =_{\text{def}} \beta > \bigcirc\alpha$ (Chellas, 1974; Chellas, 1980; Alchourrón, 1994) have too little factual defeasibility, because they have unrestricted strengthening of the antecedent (and factual detachment). Thus they cannot represent contrary-to-duty obligations, because they suffer from the paradoxes.

3. OVERRIDDEN AND FACTUAL DEFEASIBILITY

In this section, we focus on the cancelling aspect of overridden defeasibility and the overshadowing aspect of factual defeasibility. Overridden defeasibility becomes relevant when there is a (potential) conflict between two obligations, i.e. when there are two contradictory obligations. For example, there is a conflict between $\bigcirc(\alpha_1|\beta_1)$ and $\bigcirc(\alpha_2|\beta_2)$ when α_1 and α_2 are contradictory, and β_1 and β_2 are factually true. In a defeasible deontic logic, such a conflict is resolved when one of the obligations overrides the other one. In the language of dyadic deontic logic, the overriding of $\bigcirc(\alpha_1|\beta_1)$ by $\bigcirc(\alpha_2|\beta_2)$ is formalized by the non-derivability of $\bigcirc(\alpha_1 | \beta_1 \wedge \beta_2)$. An unresolvable conflict is usually called a 'deontic dilemma', in this case represented by the formula $\bigcirc(\alpha_1|\beta_1 \wedge \beta_2) \wedge \bigcirc(\alpha_2|\beta_1 \wedge \beta_2)$.

In particular, we analyze *violated obligations* in a deontic logic that formalizes reasoning about obligations which can be overridden by other obligations. In the language of dyadic deontic logic, an obligation with a contradictory antecedent and consequent like $\bigcirc(\neg\alpha|\alpha)$ represents 'if α is the case, then it is a violation of the obligation that $\neg\alpha$ should be the case'.[11] This representation of violations is related to the more standard representation $\alpha \wedge \bigcirc\neg\alpha$ in SDL as follows. The standard representation of violations is a combination of monadic obligations and factual detachment, see (Van der Torre and Tan, 1995). With the inference pattern EFD discussed in the introduction the obligation $\bigcirc\neg\alpha$ can be derived from $\mathcal{A}\alpha$ and $\bigcirc(\neg\alpha|\alpha)$. Hence, $\bigcirc(\neg\alpha|\alpha)$ can be read as 'if only α is known, then $\bigcirc\neg\alpha$ can be derived' and $\alpha \wedge \bigcirc\neg\alpha$ represents a violation. The contextual obligations we defined in Section 2.4 do not represent violated obligations, but in Section 3.4 we show how the definition of $\bigcirc(\alpha|\beta\setminus\gamma)$ can be adapted to $\bigcirc^r(\alpha|\beta\setminus\gamma)$ to derive violated (i.e. overshadowed) contextual obligations. To keep our analysis as general as possible, in this section we only accept the inference pattern RSA_O. Because RSA_O is the only inference pattern we assume, we do not have to formalize contrary-to-duty reasoning and its related problems which we discussed in

[11] Alternatively, such an obligation could represent the obligation to update the present state of affairs. For example, the obligation 'if you smoke in a no-smoking area, then you should not smoke in a no-smoking area' (Hansson, 1971) can be read as the obligation to quit smoking.

the previous section. Thus, the analyses in this section are independent from our analysis and our solution of the Chisholm paradox.

3.1. *The Fence example*

The following so-called Fence example was introduced in (Prakken and Sergot, 1994) to illustrate the distinction between contrary-to-duty reasoning and defeasible reasoning (based on exceptional circumstances). It is an extended version of the Forrester (or gentle murderer) paradox: you should not kill, but if you kill, then you should do it gently (Forrester, 1984). In (Van der Torre, 1994) we discussed this Fence example in Horty's defeasible deontic logic. The following example is an alphabetic variant of the original example, because we replaced s, to be read as 'the cottage is by the sea', by d, to be read as 'there is a dog'. The distinction between 'the cottage is by the sea' and 'there is a dog' is that the latter proposition is controllable, whereas the former is not. This important distinction between controllable and uncontrollable propositions has to be formalized in a deontic (or action) logic, if only because for any uncontrollable α the obligation $\bigcirc(\alpha|\top)$ does not make sense, see (Boutilier, 1994) for a discussion. For example, it does not make sense to oblige someone to make the sun rise. In this article, we abstract from this problem and we assume that all propositions are controllable.

Example 4.1 (Fence example) Assume a dyadic deontic logic that validates at least substitution of logical equivalents and the inference pattern RSA_O. Furthermore, assume the obligations

$$S = \{\bigcirc(\neg f|\top), \bigcirc(w \wedge f|f), \bigcirc(w \wedge f|d)\},$$

where f can be read as 'there is a fence around your house', $w \wedge f$ as 'there is a white fence around your house' and d as 'you have a dog'. Notice that $\bigcirc(w \wedge f \,|\, f)$ is a CTD obligation of $\bigcirc(\neg f \,|\, \top)$ and $\bigcirc(w \wedge f \,|\, d)$ is not. If there is a fence and a dog ($\mathcal{A}(f \wedge d)$), then the first premise of S is intuitively overridden, and therefore it cannot be violated. Hence, $\bigcirc(\neg f|f \wedge d)$ should *not* be derivable. However, if there is a fence without a dog ($\mathcal{A}f$), then the first premise is intuitively not overridden, and therefore it is violated. Hence, the obligation $\bigcirc(\neg f \,|\, f)$ should be derivable. Moreover, this is exactly the difference between cancellation and overshadowing that we discussed in the introduction of this article. Overriding of $\bigcirc(\neg f|\top)$ by $f \wedge d$ and $\bigcirc(w \wedge f|d)$ means that the obligation to have no fence is cancelled and has no force anymore, hence $\bigcirc(\neg f|f \wedge d)$ should not be derivable. Violation of $\bigcirc(\neg f|\top)$ by f means that the obligation to have no fence has still its force, it is only overshadowed and not cancelled, hence $\bigcirc(\neg f|f)$ should be derivable. The possible derivations of $\bigcirc(\neg f \mid f \wedge d)$ and $\bigcirc(\neg f \mid f)$ are represented in Figure 8. In the first derivation, the counterintuitive obligation $\bigcirc(\neg f|f \wedge d)$

is not derived from $\bigcirc(\neg f \mid \top)$ by RSA$_O$, because the latter obligation is overridden by $\bigcirc(w \wedge f \mid d)$ for $f \wedge d$. However, in the second derivation the intuitive obligation $\bigcirc(\neg f \mid f)$ is not derived either from $\bigcirc(\neg f \mid \top)$ by RSA$_O$, because it is overridden by $\bigcirc(w \wedge f \mid f)$ for f, according to C_O.

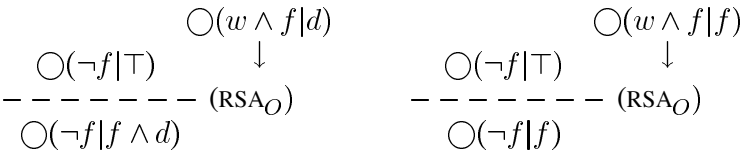

Fig. 8. Fence example with C_O

The problem in this example is that both $\bigcirc(w \wedge f \mid f)$ and $\bigcirc(w \wedge f \mid d)$ are treated as more specific obligations that override the obligation $\bigcirc(\neg f \mid \top)$, i.e. both are treated as cases of overridden defeasibility. However, this is not correct for $\bigcirc(w \wedge f \mid f)$. This last obligation should be treated as a CTD obligation, i.e. as a case of factual defeasibility. This interference of specificity and CTD is represented in Figure 9. This figure should be read as follows. Each arrow is a condition: a two-headed arrow is a consistency check, and a single-headed arrow is a logical implication. For example, the condition C_O formalizes that an obligation $\bigcirc(\alpha \mid \beta)$ is overridden by $\bigcirc(\alpha' \mid \beta')$ if the conclusions are contradictory (a consistency check, the double-headed arrow) and the condition of the overriding obligation is more specific (β' logically implies β). Case (a) represents criteria for overridden defeasibility, and case (b) represents criteria for CTD. Case (c) shows that the pair of obligations $\bigcirc(\neg f \mid \top)$ and $\bigcirc(w \wedge f \mid f)$ can be viewed as overridden defeasibility as well as CTD.

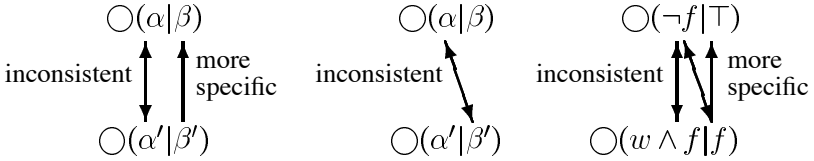

Fig. 9. Specificity and CTD

What is most striking about the Fence example is the observation that when the premise $\bigcirc(\neg f \mid \top)$ is violated by f, then the obligation for $\neg f$ should be derivable, but not when $\bigcirc(\neg f \mid \top)$ is overridden by $f \wedge d$. This means that the CTD or overriding interpretations of $\bigcirc(\neg f \mid \top)$ are quite different in the

sense that they have different consequences. This overriding can be viewed as a type of overridden defeasibility and the violation in the CTD as a type of factual defeasibility. Hence, also the Fence example shows that factual and overridden defeasibility lead to different conclusions. This is a kind of factual defeasibility which differs from its counterpart in default logic in the sense that it is overshadowing factual defeasibility rather than cancelling factual defeasibility.

3.2. *Overridden defeasibility*

One obvious analysis of the problem mentioned in Example 4.1 is to observe that condition C_O is too strong. In (Van der Torre, 1994) we gave an ad hoc solution of the problem by weakening the definition of specificity in C_O to C_O^* with an additional condition which represents that a CTD obligation cannot override its primary obligations. The specificity condition C_O^* has three conditions: the two conditions of C_O and the additional condition that the overriding obligation $\bigcirc(\alpha' \mid \beta')$ is not a CTD of $\bigcirc(\alpha \mid \beta)$, i.e. $\beta' \wedge \alpha$ must be consistent. Due to this extra condition the overriding interpretation in case (c) in Figure 9 is no longer valid. The following example shows that the definition of specificity C_O^* gives the intuitive conclusions and avoids the counterintuitive ones.

Example 4.2 (Fence example, continued) Assume that RSA_O is replaced by the following RSA_O^*.

$$\text{RSA}_O^* : \frac{\bigcirc(\alpha|\beta_1), C_O^*}{\bigcirc(\alpha|\beta_1 \wedge \beta_2)}$$

C_O^*: there is no premise $\bigcirc(\alpha' \mid \beta')$ such that $\beta_1 \wedge \beta_2$ logically implies β', β' logically implies β_1 and not vice versa, α and α' are contradictory and $\alpha \wedge \beta'$ is consistent. (Van der Torre, 1994)

The derivations from S with RSA_O^* are represented in Figure 10. RSA_O^* does not derive the counterintuitive $\bigcirc(\neg f|f \wedge d)$, just like RSA_O in Figure 8. However, RSA_O^* does derive the intuitive $\bigcirc(\neg f|f)$ from $\bigcirc(\neg f|\top)$, in contrast to RSA_O. RSA_O^* solves the problem of Example 4.1, because it does not derive the counterintuitive obligation, but it does derive the intuitive obligation.

This solution of the Fence example is ad hoc, because there is no *a priori* reason to prefer C_O^* and RSA_O^* (the violability interpretation) to C_O and RSA_O (the overridden interpretation). The informal reason given in (Van der Torre, 1994) to prefer the former inference pattern is that with RSA_O, the obligation $\bigcirc(\neg f|\top)$ can never be violated, which is a highly counterintuitive property of an obligation. In the following subsection, we give a formal analysis of the Fence example, based on the essential property of obligations that they can be violated.

$$\dfrac{\bigcirc(\neg f|\top) \qquad \bigcirc(w \wedge f|d) \downarrow}{\bigcirc(\neg f|f \wedge d)}\ (\mathrm{RSA}^*_O) \qquad \dfrac{\bigcirc(\neg f|\top)}{\bigcirc(\neg f|f)}\ \mathrm{RSA}^*_O$$

Fig. 10. Fence example with C^*_O

3.3. *Factual defeasibility*

Instead of analyzing the problem of Example 4.1 by examining specificity condition C_O (overridden defeasibility), we can also look at properties of violability (factual defeasibility). The following inference patterns *Contrary-to-Duty* (CD) and *According-to-Duty* (AD) formalize the intuitions that an obligation cannot be defeated by only violating or fulfilling it. The CD rule models the intuition that after violation the obligation to do α is still in force (i.e. overshadowing). Even if you drive too fast, you are still obliged to obey the speed limit.[12]

$$\mathrm{CD}: \frac{\bigcirc(\alpha|\beta)}{\bigcirc(\alpha|\beta \wedge \neg\alpha)} \qquad \mathrm{AD}: \frac{\bigcirc(\alpha|\beta)}{\bigcirc(\alpha|\beta \wedge \alpha)}$$

We reconsider the Fence example and we show that CD with RSA_O derives exactly the intuitive conclusions, just like RSA^*_O.

Example 4.3 (Fence example, continued) Assume the inference patterns RSA_O and CD. Figure 11 represents the same two situations as Figure 8. First consider the situation when there is a fence and a dog ($f \wedge d$). The counterintuitive obligation $\bigcirc(\neg f \mid f \wedge d)$ cannot be derived, because the derivation via $\bigcirc(\neg f \,|\, d)$ from $\bigcirc(\neg f \,|\, \top)$ is blocked by C_O. Now consider the situation when there is a fence but not a dog (f). The intuitive obligation $\bigcirc(\neg f|f)$ can be derived from $\bigcirc(\neg f|\top)$ by CD.

Example 4.2 and 4.3 illustrate that the problem of RSA_O is that it does not imply CD (because its specificity condition C_O is too strong). In other words,

[12]The inference patterns CD and AD should not be confused with the following inverses of CD and AD, which seem to say that violations or fulfilled obligations do not come out of the blue.

$$\mathrm{CD}^{-}: \frac{\bigcirc(\alpha|\beta \wedge \neg\alpha)}{\bigcirc(\alpha|\beta)} \qquad \mathrm{AD}^{-}: \frac{\bigcirc(\alpha|\beta \wedge \alpha)}{\bigcirc(\alpha|\beta)}$$

Although these inference patterns seem intuitive at first sight, they are highly counterintuitive on further inspection. Reconsider the Fence example. There should be a white fence, if there is a fence $\bigcirc(w \wedge f \,|\, f)$. Hence, by AD, there should be a white fence, if there is a white fence $\bigcirc(w \wedge f|w \wedge f)$ (a fulfilled obligation). However, this does not mean that there is an unconditional obligation that there should be a white fence $\bigcirc(w \wedge f|\top)$. Hence, the inference pattern AD^{-} is not valid. A similar argument can be given for CD^{-}.

$$\dfrac{\dfrac{\bigcirc(w \wedge f|d)}{\bigcirc(\neg f|\top)} \quad \downarrow \text{(RSA}_O\text{)}}{\dfrac{\bigcirc(\neg f|d)}{\bigcirc(\neg f|f \wedge d)}\ \text{CD}} \qquad \dfrac{\bigcirc(\neg f|\top)}{\bigcirc(\neg f|f)}\ \text{CD}$$

Fig. 11. Fence example with CD

the problem of RSA_O is that there can be obligations, like $\bigcirc(\neg f \mid \top)$, that can never be violated. In Example 4.3, CD and RSA_O yield exactly the same intuitive conclusions as RSA_O^* in Example 4.2. An advantage of CD is that the inference pattern is very intuitive and not an ad hoc like solution of the problem like the adaptation of C_O. Moreover, AD also formalizes an intuitive notion of fulfilled obligations, because it deals with fulfilled obligations in exactly the same way as CD with violated obligations. We illustrate the applicability of our approach by the analysis of the following Reykjavic Scenario, introduced by Belzer (1986).

Example 5.1 (Reykjavic Scenario) Consider the premise set of obligations $S = \{\bigcirc(\neg r \mid \top), \bigcirc(\neg g \mid \top), \bigcirc(r \mid g), \bigcirc(g \mid r)\}$, where r can be read as 'the agent tells the secret to Reagan' and g as 'the agent tells the secret to Gorbatsjov'. Figure 12 illustrates that the Reykjavic Scenario is a more complex instance of the Fence example, illustrated in Figure 9. In the Fence example, the obligation $\bigcirc(w \wedge f \mid f)$ can be interpreted as a more specific overriding obligation, and it can be interpreted as a CTD obligation. In the Reykjavic Scenario, the latter two obligations of S can be considered as more specific obligations overriding the former two, and they can be considered as CTD obligations.

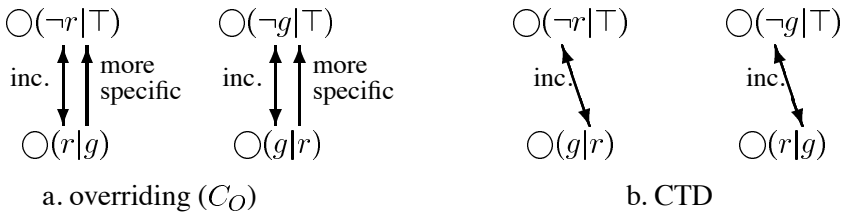

Fig. 12. Specificity and CTD in the Reykjavic Scenario

The Reykjavic Scenario is a highly ambiguous paradox, as a result of the fact that the latter two obligations can be considered as overriding as well

as CTD obligations. In (Van der Torre, 1994), we gave the following two interpretations of this paradox.

1. **Overridden interpretation.** In this interpretation, the third sentence of S is an exception to the first sentence, and the fourth sentence is an exception to the second sentence (see Figure 12.a). The agent's primary obligation is not to tell Reagan or Gorbatsjov. When he tells Reagan, he should not tell Reagan but he should tell Gorbatsjov. It is a case of overridden defeasibility, because $\bigcirc(\neg g \,|\, r)$ cannot be derived from $\bigcirc(\neg g|\top)$ due to the premise $\bigcirc(g|r)$. When he tells both, he does not violate any obligations because r and g are considered as exceptions.[13]
2. **Violability interpretation.** In this interpretation the two dyadic obligations $\bigcirc(\neg r|r \wedge g)$ and $\bigcirc(\neg g|r \wedge g)$ are both derivable from S. Hence, when the agent tells both, he should have told neither of them, $\bigcirc(\neg r|r \wedge g)$ and $\bigcirc(\neg g|r \wedge g)$, a case of violability. The third sentence of S is a CTD obligation of the second sentence and the fourth sentence is a CTD obligation of the first sentence (see Figure 12.b).

In our view the violability interpretation is to be preferred to the overridden interpretation, The following example illustrates that the overridden interpretation conflicts with CD.

Example 5.2 (Reykjavic Scenario, continued) Assume a dyadic deontic logic that validates at least substitution of logical equivalents and the inference patterns AND, RSA_O, CD and the following disjunction rule OR.

$$\text{OR}: \frac{\bigcirc(\alpha_1|\beta), \bigcirc(\alpha_2|\beta)}{\bigcirc(\alpha_1 \vee \alpha_2|\beta)}$$

Moreover, assume the set of obligations S of Example 5.1. According to the overridden interpretation, there is no violation when the agent tells both Reagan and Gorbatsjov. We cannot use RSA_O to derive a violation from S, because the premises are overridden as represented in Figure 12.a. However, we can use CD to derive the violation $\bigcirc(\neg r \vee \neg g \,|\, r \wedge g)$, as represented in Figure 13. Hence, if we accept CD then we have to reject the overridden interpretation. Since we gave a general motivation for CD that is independent from particular examples, we reject the overridden interpretation.

The examples show that the inference patterns CD and AD are adequate tools to analyze conflicts between overridden and contrary-to-duty interpretations. However, they cannot discriminate between the following two violability

[13] According to the overridden interpretation, it might be argued that the paradox is not modeled correctly by the set of obligations S. When the last two conditional obligations should be interpreted as CTD obligations when the agent tells both, the first two obligations should be represented by one conditional obligation $\bigcirc(\neg r \wedge \neg g|\top)$. In that case, the last two sentences are interpreted as CTD obligations by C^*_O.

$$\dfrac{\dfrac{\bigcirc(\neg r|\top) \quad \bigcirc(\neg g|\top)}{\bigcirc(\neg r \vee \neg g|\top)}\,\text{OR}}{\bigcirc(\neg r \vee \neg g|r \wedge g)}\,\text{CD} \qquad \dfrac{\dfrac{\bigcirc(\neg r|\top) \quad \bigcirc(\neg g|\top)}{\bigcirc(\neg r \wedge \neg g|\top)}\,\text{AND}}{\bigcirc(\neg r \wedge \neg g|r \vee g)}\,\text{CD}$$

Fig. 13. Reykjavic scenario with CD

interpretations of the Reykjavic Scenario. McCarty (1994) argues for the first violability interpretation.

2.1 **Violability-1 interpretation** When he tells only Reagan, then one could interpret this as an overridden case, i.e. a case of defeasibility. In this interpretation $\bigcirc(\neg g|\top)$ is overridden by $\bigcirc(g|r)$ and the fact r. Hence, in this interpretation $\bigcirc(\neg g|r)$ is not derivable from the premises. The remarkable thing about this interpretation is that $r \wedge g$ is treated as a violability case, whereas r in isolation is treated as an overridden case.

2.2 **Violability-2 interpretation** If we accept the reasonable principle that if an obligation is overriden for some situation, that it is then also overridden for a more specific situation, then the obligation $\bigcirc(\neg g|\top)$ cannot be overridden by r only, because it is in the violability interpretation not overridden by the more specific situation $r \wedge g$.[14] According to this interpretation, when the agent tells only Reagan, then he still has the obligation $\bigcirc(g|r)$ to tell Gorbatsjov, but also he has the derivable obligation not to tell Gorbatsjov $\bigcirc(\neg g|r)$. The remarkable thing about this interpretation is that if we accept a reasonable principle, then the Reykjavic Scenario becomes a deontic dilemma.

This again illustrates the fact that this scenario is highly ambiguous, and additional principles have to be accepted if we want to decide between these two interpretations 2.1. and 2.2..

3.4. *Preferential semantics:* CD *and* AD

Before we can examine the conflicts between specificity and contrary-to-duty in the semantics, there are two ways in which we have to adapt the definition of contextual obligations. First, in this section we adapt the definition of $\bigcirc(\alpha|\beta \setminus \gamma)$ to $\bigcirc^r(\alpha|\beta \setminus \gamma)$. The logic of $\bigcirc^r(\alpha|\beta \setminus \gamma)$ represents fulfilled and violated obligations, because it validates CD and AD. Second, we have to introduce a semantic notion to model specificity, which is done in Section 3.5 when we introduce obligations $\bigcirc^{re}(\alpha|\beta \setminus \gamma)$.

[14] This principle certainly holds for defeasible logics. For example, if the 'birds fly' default is overridden by the more specific 'penguins do not fly default, then this latter default also holds for the subset super-penguins of penguins, unless it is explicitly stated that by default 'super-penguins do fly'.

The contextual obligations $\bigcirc(\alpha \mid \beta \setminus \gamma)$ do not represent violated and fulfilled obligations, because the first two conditions of Definition 1 say that $\bigcirc(\alpha|\beta\setminus\gamma)$ is false if either $\alpha \wedge \beta \wedge \neg\gamma$ or $\neg\alpha \wedge \beta$ is inconsistent. Obviously, we have to relax these two conditions. We allow the set of worlds W_1' and W_2' of $\bigcirc^r(\alpha|\beta\setminus\gamma)$ to be supersets of W_1 and W_2 from $\bigcirc(\alpha|\beta\setminus\gamma)$. If W_1 and W_2 of Definition 1 are nonempty, then the definition of $\bigcirc$ is equivalent to the definition of $\bigcirc^r$. However, if the set W_1 or W_2 is empty, then we have $M \not\models \bigcirc(\alpha|\beta\setminus\gamma)$, whereas $M \models \bigcirc^r(\alpha|\beta\setminus\gamma)$ if there is any $M \models \bigcirc(\alpha|\beta'\setminus\gamma)$ where β logically implies β' (see Proposition 4 and 5).

Definition 3 (Contextual obligation, with violations) Let $M = \langle W, \leq, V\rangle$ be a Kripke model that consists of W, a set of worlds, $\leq$, a binary reflexive and transitive relation on W, and V, a valuation of the propositions in the worlds. The model M satisfies the obligation 'α should be the case if β is the case unless γ is the case', written as $M \models \bigcirc^r(\alpha|\beta\setminus\gamma)$, iff

1. there is a nonempty $W_1 \subset W$ such that
   - for all $w \in W_1$, we have $M, w \models \alpha \wedge \neg\gamma$, and
   - for all w such that $M, w \models \alpha \wedge \beta \wedge \neg\gamma$, we have $w \in W_1$, and
2. there is a nonempty $W_2 \subset W$ such that
   - for all $w \in W_2$, we have $M, w \models \neg\alpha$, and
   - for all w such that $M, w \models \neg\alpha \wedge \beta$, we have $w \in W_2$, and
3. for all $w_1 \in W_1$ and $w_2 \in W_2$, we have $w_2 \not\leq w_1$.

To give an intuition for the previous formalization of contextual obligations we give the following metaphor, based on a parallel with belief revision. Let $W_1 = \{w \mid M, w \models \alpha \wedge \beta \wedge \neg\gamma\}$ and $W_2 = \{w \mid M, w \models \neg\alpha \wedge \beta\}$ be the choice alternatives of $\bigcirc(\alpha|\beta\setminus\gamma)$. Definition 1 in Section 2.4 says that W_1 and W_2 are non-empty, and $w_2 \not\leq w_1$ for every $w_1 \in W_1$ and $w_2 \in W_2$. Thus, we evaluated $\bigcirc(\alpha|\beta)$ by a choice between $\alpha \wedge \beta$ and $\neg\alpha \wedge \beta$, which can be considered as the AGM expansions of β by α and $\neg\alpha$.[15] Now, we evaluate $\bigcirc^r(\alpha|\beta)$ by a choice between the AGM-style revisions of β by α or $\neg\alpha$, which explains our notation $\bigcirc^r$.[16] Condition (1) and (2) formalize that revision must be possible. The following proposition shows that contextual

[15] For details on expansion and etraction, see (Gärdenfors, 1988).

[16] A similar idea is present in a proposal of Tan and Pearl (1994), where a conditional desire $D(l|n \wedge \neg l)$ is interpreted as $D(l|n)$, representing that 'I desire the light to be on if it is night and the light is off' compares night-worlds in which the light is on with those in which the light is off. However, their formalization is problematic, as is shown in (Boutilier, 1994). Moreover, in our case it is violation detection and revision (it refers to deontic alternatives in the past), in their case it is world improvement and update (it refers to alternatives in the future).

Revision can be considered as a combination of retraction and expansion, known as the Levi identity. In (Van der Torre and Tan, 1995), we interpreted the essential mechanism to represent violations in terms of a so-called retraction test. Boutilier and Becher (1995) use a similar

obligations validate strengthening of the antecedent.[17] Hence, the logic also validates CD and AD, because CD and AD follow from SA.

Proposition 4 The logic validates unrestricted strengthening of the antecedent.

$$\text{SA}: \frac{\bigcirc^r(\alpha|\beta_1 \setminus \gamma)}{\bigcirc^r(\alpha|\beta_1 \wedge \beta_2 \setminus \gamma)}$$

Proof Assume $M \models \bigcirc^r(\alpha \mid \beta_1 \setminus \gamma)$. There are W_1 and W_2 such that the conditions of Definition 3 are fulfilled. The same W_1 and W_2 also fulfill the conditions for $M \models \bigcirc^r(\alpha|\beta_1 \wedge \beta_2 \setminus \gamma)$.

The following proposition shows the relation between expansion-based contextual obligations in Section 2.4 (Definition 1) and the revision-based contextual obligations (Definition 3).

Proposition 5 The logic validates the following inference pattern.

$$\frac{\bigcirc(\alpha|\beta \setminus \gamma)}{\bigcirc^r(\alpha|\beta \setminus \gamma)}$$

Proof Assume a model M such that $M \models \bigcirc(\alpha|\beta \setminus \gamma)$. Let $W_1 = \{w \mid M, w \models \alpha \wedge \beta \wedge \neg\gamma\}$ and $W_2 = \{w \mid M, w \models \neg\alpha \wedge \beta\}$ be the choice alternatives of Definition 1. Then $M \models \bigcirc^r(\alpha|\beta \setminus \gamma)$, because W_1 and W_2 fulfill the conditions of Definition 3.

3.5. *Multi preference semantics*

In this section we adapt the definition of contextual obligations to model specificity, i.e. overridden defeasibility. Overridden defeasibility can be formalized by introducing a normality ordering in the semantics. Hence, the logic has a

kind of retraction to model predictive explanations: 'In order to evaluate the predictive force of factual explanations, we require that the agent (hypothetically) give up its belief in β and then find some α that would (in this new belief state) restore β. In other words, we contract K by β and evaluate the conditional $\alpha \Rightarrow \beta$ with respect to this contracted belief state: $\beta \in (K_\beta^-)^*_\alpha$. Thus, when we hypothetically suspend belief in β, if α is sufficient to restore this belief then α counts as a valid explanation. The contracted belief set K_β^- might fruitfully be thought of as the belief set held by the agent before it came to accept the observation β'.

[17] For example, we can derive $\bigcirc^r(t|\neg a \setminus \neg a)$ from $\bigcirc^r(t|a \setminus \bot)$ and $\bigcirc^r(a|\top \setminus \bot)$ in the Chisholm paradox (see Figure 6). There are two ways to view this derived obligation. The first is to say it is meaningless, because the antecedent $\neg a$ implies the unless clause $\neg a$. The second way is to say that it is counterintuitive, because it looks like the counterintuitive dyadic obligation $\bigcirc(t \mid \neg a)$. We can add a fourth condition to Definition 3 if we consider SA too strong, which states that there are worlds $\beta \wedge \neg\gamma$. In that case, there is a condition C_V on SA and $\bigcirc^r(t|\neg a \setminus \neg a)$ is not derivable from the Chisholm paradox.

multi preference semantics: an *ideality ordering* ($\leq_I$) to model contrary-to-duty structures (factual defeasibility) and a *normality ordering* ($\leq_N$) to model exceptional circumstances (overridden defeasibility), see (Tan and Van der Torre, 1995) for details. To facilitate the comparison with the definitions of $\bigcirc(\alpha|\beta\backslash\gamma)$ and $\bigcirc^r(\alpha|\beta\backslash\gamma)$, we assume that the preferential orderings are bounded.[18]

Definition 4 (Contextual obligation, with violations and overriding) Let $M = \langle W, \leq_I, \leq_N, V\rangle$ be a Kripke model that consists of W, a set of worlds, $\leq_I$ and $\leq_N$, two binary reflexive and transitive relations on W, and V, a valuation of the propositions in the worlds, such that there are no infinite descending chains. The model M satisfies the obligation 'α should be the case if β is the case unless γ is the case', written as $M \models \bigcirc^{re}(\alpha|\beta\backslash\gamma)$, iff

1. there is a nonempty $W_1 \subset W$ such that
   - for all $w \in W_1$, we have $M, w \models \alpha \wedge \neg\gamma$, and
   - for all w such that $M, w \models_{\leq_N} \alpha \wedge \beta \wedge \neg\gamma$, we have $w \in W_1$, and
2. there is a nonempty $W_2 \subset W$ such that
   - for all $w \in W_2$, we have $M, w \models \neg\alpha$, and
   - for all w such that $M, w \models_{\leq_N} \neg\alpha \wedge \beta$, we have $w \in W_2$, and
3. for all $w_1 \in W_1$ and $w_2 \in W_2$, we have $w_2 \not\leq_I w_1$.

The following example illustrates the multi preference semantics of the Fence example.

Example 4.4 (Fence example, continued) Consider the set of obligations $S = \{\bigcirc^{re}(\neg f|\top\backslash\bot), \bigcirc^{re}(w \wedge f|d\backslash\bot)\}$. The typical[19] multi preference model of S is given in Figure 14 and can be read as follows. The circles denote equivalence classes of worlds that satisfy the literals inside the circles and the 'horizontal' arrows denote the deontic preference ordering. The boxes denote equivalence classes in the normality ordering and the 'vertical' arrow the normality preference ordering. S constructs two preference orderings on the worlds: one ordering for ideality (like before) and one for normality. The idea of the preference ordering on normality is that the worlds with exceptional circumstances (where you have a dog) are semantically separated from the normal situation (where you do not have a dog). The upper box represents

[18] The fact that $\leq_N$ is bounded, ensures that the set of w such that $w \in W_1$ and $M, w \models_{\leq_N} \alpha \wedge \beta \wedge \neg\gamma$ is well-defined. The more general definition for unbounded orderings is: for all w such that $M, w \models \alpha \wedge \beta \wedge \neg\gamma$, there is a world $w' \leq_N w$ such that $M, w' \models \alpha \wedge \beta \wedge \neg\gamma$, and for all w'' such that $M, w'' \models \alpha \wedge \beta \wedge \neg\gamma$ and $w'' \leq_N w'$, we have $w'' \in W_1$. See also Definition 2 and Proposition 1.

[19] Computing these typical models in general is difficult, see (Tan and Van der Torre, 1995). For example, it seems more difficult than defeasible reasoning schemes to complete a single ordering like 'maximally connected' (Tan and Van der Torre, 1996) or System Z (Pearl, 1990).

the 'normal' worlds, which is determined by the fact that d is false, i.e. you do not have a dog. Deontically, the $\neg d$ worlds are ordered according to the obligation that, normally, there should be no fence. The lower box contains the worlds where d is true and which are therefore exceptional. These worlds are deontically ordered by the obligation that in this situation, there should be a white fence. Because of the exceptional circumstances, the worlds are not subject to the obligation that normally, there should not be a fence. In the ideality ordering, the normal $\neg d \wedge \neg f$ worlds and the exceptional $d \wedge w \wedge f$ worlds are equivalent.

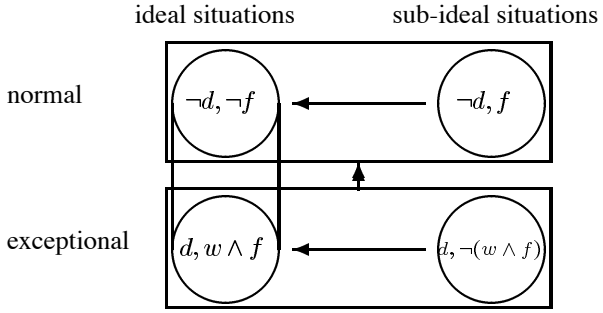

Fig. 14. Multi-preference relation of the Fence example

For example, we have $M \models \bigcirc^{re}(\neg f | \top \setminus \bot)$, because for all $w_1 \in |\neg f \wedge \neg d|$ (the most normal $\neg f$ worlds) and for all $w_2 \in |f \wedge \neg d|$ (the most normal f worlds) we have $w_2 \not\leq_I w_1$. Moreover, we have $M \models \bigcirc^{re}(w \wedge f | d \setminus \bot)$, because we zoom in on the d worlds, and $w \wedge f \wedge d$ worlds are preferred over $\neg(w \wedge f) \wedge d$ worlds.

Notice that we first minimize in the normality ordering when we evaluate the obligation $\bigcirc^{re}(\neg f | \top \setminus \bot)$ in Example 4.4, because we first determine the sets $W_1 = |\neg f \wedge \neg d|$ and $W_2 = |f \wedge \neg d|$, and subsequently we compare the sets W_1 and W_2 in the ideality ordering. We compare the best most normal worlds and we do not compare the most normal best sets $W_1' = |\neg f \wedge \neg d|$ and $W_2' = |w \wedge f \wedge d|$. This is based on the heuristic rule that if an option (like f) can be a violation (like W_2) or an exception (like W_2'), then it is assumed to be a violation. The motivation of this rule is that a criminal should have as little opportunities as possible to excuse herself by claiming that her behavior was exceptional rather than criminal. If an agent has a fence, then it is assumed to be a violation and she cannot excuse herself by claiming that it is an exceptional case (unless, of course, there is a dog).[20] The following proposition shows that the obligations validate CD and AD.

[20] However, our approach is quite different from lexicographic minimizing (minimize first $\leq_N$ and then $\leq_I$) like in (Makinson, 1993), because our second step is not minimizing. In fact,

Proposition 6 The logic of the obligations $\bigcirc^{re}$ does not validate SA, but it validates CD and AD.

Proof First, consider the invalidity of SA. The contextual obligation $\bigcirc^{re}(\alpha \,|\, \beta_1 \wedge \beta_2 \setminus \bot)$ cannot be derived from $\bigcirc^{re}(\alpha \,|\, \beta_1 \setminus \bot)$, because the most normal worlds $\beta_1 \wedge \beta_2$ can contain worlds not among the most normal β_1 worlds. Thus the logic does not validate SA. Secondly, consider CD and AD. Assume $M \models \bigcirc^{re}(\alpha|\beta\setminus\gamma)$. Hence, there are W_1 and W_2 such that the conditions of Definition 4 are fulfilled. The same W_1 and W_2 also satisfy the conditions for $M \models \bigcirc^{re}(\alpha|\beta \wedge \neg\alpha\setminus\gamma)$ and $M \models \bigcirc^{re}(\alpha|\beta \wedge \alpha\setminus\gamma)$.

The following example illustrates the conflict between overridden and CTD.

Example 4.5 (Fence example, continued) Consider the set of obligations $S' = \{\bigcirc^{re}(\neg f|\top\setminus\bot), \bigcirc^{re}(w \wedge f|d\setminus\bot), \bigcirc^{re}(w \wedge f|f\setminus\bot)\}$. The typical multi preference model M' of S' is given in Figure 15. The normal worlds have deontically been specified more precisely, compared to the model M in Figure 14 of the set of obligations S in Example 4.4. We have $M' \models \bigcirc^{re}(\neg f \,|\, \top \setminus \bot)$, for similar reasons as $M \models \bigcirc^{re}(\neg f \,|\, \top \setminus \bot)$ in Example 4.4. We also have $M' \models \bigcirc^{re}(\neg f \,|\, f \setminus \bot)$, which can be shown as follows. Semantically, the sets W_1 and W_2 must contain the most normal $\neg f \wedge f$ and $f \wedge f$ worlds, respectively. Hence, W_1 can be any subset of $|\neg f|$, and W_2 is a subset of $|f|$ that contains at least $|f \wedge \neg d|$. We can choose W_1 and W_2 as $|\neg f \wedge \neg d|$ and $|f \wedge \neg d|$, and we have $w_2 \not\leq w_1$ for all $w_1 \in W_1$ and $w_2 \in W_2$. However, we do not have $M' \models \bigcirc^{re}(\neg f \,|\, f \wedge d \setminus \bot)$, as can be verified as follows. The sets W_1 and W_2 must contain the most normal $\neg f \wedge f \wedge d$ and $f \wedge f \wedge d$ worlds, respectively. Hence, W_1 can be any subset of $|\neg f|$, and W_2 is a subset of $|f|$ that contains at least $|f \wedge d|$. Any world $w_2 \in |w \wedge f \wedge d|$ is deontically preferred, hence there cannot be a world $w_1 \in W_1$ such that $w_2 \not\leq w_1$, thus the first condition cannot be fulfilled. This illustrates that the logic does not validate SA, because it does not strengthen $\bigcirc^{re}(\neg f \,|\, \top \setminus \bot)$ to $\bigcirc^{re}(\neg f \,|\, f \wedge d \setminus \bot)$ (although it does strengthen to $\bigcirc^{re}(\neg f \,|\, f \setminus \bot)$). These are precisely the intuitive conclusions that one would draw from S'. If one only knows that there is a fence, then one concludes that the first obligation from S' still holds, hence one derives $\bigcirc^{re}(\neg f \,|\, f \setminus \bot)$. However, if one knows that there is a dog as well as a fence, then the first obligation is overridden by the second one, and hence one does not derive $\bigcirc(\neg f \,|\, f \wedge d \setminus \bot)$.

In this section, we focussed on the cancelling aspect of overridden defeasibility and the overshadowing aspect of factual defeasibility. We argued that the distinction should be reflected by two distinct preference orderings

under certain assumptions lexicographic minimizing is equivalent to minimizing in a single preference ordering (the lexicographic ordering of $\leq_N$ and $\leq_I$).

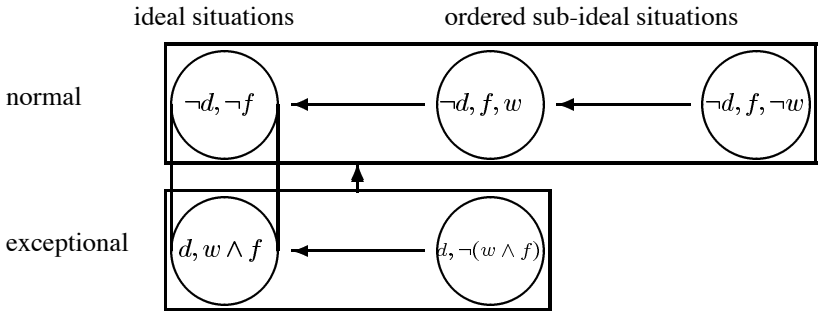

Fig. 15. Extended multi-preference relation of the Fence example

in the semantics: one normality ordering for the cancelling aspect of overridden defeasibility, and one ideality ordering for the overshadowing aspect of factual defeasibility. This is a major distinction between defeasible deontic logics and logics of defeasible reasoning, because in the latter both kinds of defeasibility are cancelling, and they can be modeled by a single preference ordering (see e.g. (Makinson, 1993; Geffner and Pearl, 1992)).

4. STRONG VERSUS WEAK OVERRIDDEN DEFEASIBILITY

In this section we focus on the cancelling aspect and the overriding aspect of overridden defeasibility by formalizing prima facie obligations. First, we show that the overridden defeasibility related to multi preference semantics cannot be used for prima facie obligations. Secondly, we introduce a new kind of preference semantics, based on priorities, to model prima facie obligations.

We call the overridden defeasibility related to multi-preference semantics *strong overridden defeasibility*, and the overridden defeasibility based on priorities *weak overridden defeasibility*. The distinction between the different types of overridden defeasibility is shown by three inference patterns which are not valid for the first type, but which are valid for the second type: forbidden conflict and two versions of reinstatement. To distinguish the two types of defeasibility we will use the deontic operator $\bigcirc$ to represent the logic of the first type and $\bigcirc_{pf}$ for the latter one. One of the inferential differences between weak and strong overridden defeasibility is the inference pattern

$$\frac{\bigcirc(\neg f \mid \top), \bigcirc(w \wedge f \mid d)}{\bigcirc(\neg d \mid \top)}$$

which is not valid in strong overridden defeasibility, whereas

$$\frac{\bigcirc_{pf}(k \mid \top), \bigcirc_{pf}(p \wedge \neg k \mid d)}{\bigcirc_{pf}(\neg d \mid \top)}$$

is valid in weak overridden defeasibility. This might look strange, because the premises in both inference schemes have the same syntactic form (obviously the substitution of $\neg k$ for f does not make any difference). However, it simply means that the $\bigcirc$ that represents obligations like 'there should be no fence' is different from the $\bigcirc_{pf}$ that represents prima facie obligations.

4.1. *Prima facie obligations*

Ross (1930) introduced the notion of so-called prima facie obligations. In his own words: 'I suggest '*prima facie* duty' or 'conditional duty' as a brief way of referring to the characteristic (quite distinct from that of being a duty proper) which an act has, in virtue of being of a certain kind (e.g. the keeping of a promise), of being an act which would be a duty proper if it were not at the same time of another kind which is morally significant' (Ross, 1930, p.19). A prima facie duty is a duty proper when it is not overridden by another prima facie duty. When a prima facie obligation is overridden, it is not a proper duty but it is still in force: 'When we think ourselves justified in breaking, and indeed morally obliged to break, a promise [...] we do not for the moment cease to recognize a prima facie duty to keep our promise' (Ross, 1930, p.28). See (Morreau, 1996) for a formalization of Ross' theory in a deontic logic. The following example describes the typical kind of defeasibility involved in reasoning about prima facie obligations.

Example 6.1 (Promises) Assume the inference pattern RSA_O and the premises $\bigcirc_{pf}(k|\top)$ and $\bigcirc_{pf}(p \wedge \neg k|d)$, where k can be read as 'keeping a promise', p as 'preventing a disaster' and d as 'a disaster will occur if nothing is done to prevent it'. There is a potential conflict between the two obligations, because when the facts imply d then the first obligation says that you should keep your promise and the second one implies that you should not. Assuming that the second obligation is stronger than the first one, the first obligation is overridden by the second one. Hence, the inference

$$\frac{\bigcirc_{pf}(k|\top), \bigcirc_{pf}(p \wedge \neg k|d)}{\bigcirc_{pf}(k|d)}$$

is *not* valid. Important here is that this priority does not depend on specificity. In this example the priority is compatible with specificity, but the converse priority could also have been chosen. You do not have an absolute (alias proper) obligation to keep your promise, but you still have the prima facie obligation. The situation is not ideal anymore. All situations where k is false, i.e. where the prima facie obligation for k is violated, are sub-ideal. This can be verified as follows. Consider a person having the obligation to keep a promise to show up at a birthday party, but she does not want to. So, she does something which might result in a disaster later on (leaving the coffee

machine on, for instance) and at the moment of the party, she rushes home to turn off the coffee machine. She has the actual obligation to go home and turn off the machine, but leaving the machine on (on purpose) was a violation already. Hence, the inference

$$\frac{\bigcirc_{pf}(k|\top), \bigcirc_{pf}(p \wedge \neg k|d)}{\bigcirc_{pf}(\neg d|\top)}$$

is valid. It says that it is not permitted to do something that might result in a disaster (remember that all propositions are assumed to be controllable). Finally, assume that there may be a disaster but you do not prevent it. Hence, the second obligation has been violated. In this situation, the proper obligation is not fulfilled, but we can still fulfill the prima facie obligation. Violating one obligation is better than violating both. Hence, the inference

$$\frac{\bigcirc_{pf}(k|\top), \bigcirc_{pf}(p \wedge \neg k|d)}{\bigcirc_{pf}(k|d \wedge \neg p)}$$

is valid.

The following inference pattern is called *Forbidden Conflict* (FC). If the inference pattern is accepted, then it is not allowed to bring about a conflict, because a conflict is sub-ideal, even when it can be resolved.

$$\text{FC}: \frac{\bigcirc_{pf}(\alpha_1|\beta_1), \bigcirc_{pf}(\neg\alpha_1 \wedge \alpha_2|\beta_1 \wedge \beta_2)}{\bigcirc_{pf}(\neg\beta_2|\beta_1)}$$

The situation considered in the following inference pattern *Reinstatement* (RI) is whether an obligation can be overridden by an overriding obligation that itself is factually defeated. The obligation $\bigcirc_{pf}(\alpha_1|\beta_1)$ is overridden by $\bigcirc_{pf}(\neg\alpha_1 \wedge \alpha_2|\beta_1 \wedge \beta_2)$ for $\beta_1 \wedge \beta_2$, but is it also overridden for $\beta_1 \wedge \beta_2 \wedge \neg\alpha_2$? If the last conclusion is not accepted, then the first obligation α_1 should be in force again. Hence, the original obligation is reinstated.

$$\text{RI}: \frac{\bigcirc_{pf}(\alpha_1|\beta_1), \bigcirc_{pf}(\neg\alpha_1 \wedge \alpha_2|\beta_1 \wedge \beta_2)}{\bigcirc_{pf}(\alpha_1|\beta_1 \wedge \beta_2 \wedge \neg\alpha_2)}$$

The following inference pattern RIO is a variant of the previous inference pattern RI, in which the overriding obligation is not factually defeated but overridden. $\bigcirc_{pf}(\alpha_1|\beta_1)$ is overridden by $\bigcirc_{pf}(\neg\alpha_1 \wedge \alpha_2|\beta_1 \wedge \beta_2)$ for $\beta_1 \wedge \beta_2$, and the latter is overridden by $\bigcirc_{pf}(\neg\alpha_2 \,|\, \beta_1 \wedge \beta_2 \wedge \beta_3)$ for $\beta_1 \wedge \beta_2 \wedge \beta_3$. The inference pattern RIO says that an obligation cannot be overridden by an obligation that is itself overridden. Hence, an overridden obligation becomes reinstated when its overriding obligation is itself overridden.

$$\text{RIO}: \frac{\bigcirc_{pf}(\alpha_1|\beta_1), \bigcirc_{pf}(\neg\alpha_1 \wedge \alpha_2|\beta_1 \wedge \beta_2), \bigcirc_{pf}(\neg\alpha_2|\beta_1 \wedge \beta_2 \wedge \beta_3)}{\bigcirc_{pf}(\alpha_1|\beta_1 \wedge \beta_2 \wedge \beta_3)}$$

Example 6.1 illustrates that the kind of overridden defeasibility related to Ross' notion of 'prima facie' obligations validates the inference patterns FC, RI and RIO.[21] In the next section, we show that the type of overridden defeasibility we used to model specificity in the Fence example does not validate these inference patterns. Hence, there are two different types of overridden defeasibility. We call the type related to prima facie obligations *weak overridden defeasibility* in contrast to *strong overridden defeasibility*. In Section 4.3, we illustrate this new type of defeasibility by a preference ordering with priorities, instead of the multi preference semantics of strong overridden defeasibility in Section 3.5.

4.2. *Strong overridden defeasibility*

In the following example, we reconsider the Fence example and we argue that it should not validate inference patterns similar to FC, RI and RIO. Since this example is based on strong overridden defeasibility, it also shows that these inference patterns are not valid for this type of defeasibility.

Example 4.6 (Fence example, continued) Reconsider the two obligations $\bigcirc(\neg f \mid \top)$ and $\bigcirc(w \wedge f \mid d)$ of Example 4.1. There is a potential conflict between the two obligations. When the facts imply d, then there is a conflict, because the first obligation says that there should not be a fence, and the second obligation implies that there should be a fence. However, the first obligation is overridden by the second one, because the second one is more specific. Hence, the conflict is resolved and there should be a white fence. The inference

$$\frac{\bigcirc(\neg f \mid \top), \bigcirc(w \wedge f \mid d)}{\bigcirc(\neg f \mid d)}$$

is *not* valid. The first sentence can be read as: 'normally, there should not be a fence around your house'. Hence, in most situations there should not be a fence, but in exceptional circumstances a fence is allowed. Similarly, the second sentence can be read as 'normally there should be a white fence, when you have a dog'. Hence, the situation when you have a dog is one of the exceptional situations in which the first obligation is not in force. The

[21] Alchourrón (1994) criticizes B. Hansson's logic (Hansson, 1971) for being a logic of prima facie obligations instead of a logic of CTD obligations. Hansson's logic validates FC when the antecedent β_1 is $\top$ (establishing a conflict is sub-ideal) but not RI (reinstatement). Actually, there is no strengthening of the antecedent at all in the logic of B. Hansson.

situation is not sub-ideal yet, it is only exceptional. Hence, the inference

$$\frac{\bigcirc(\neg f \mid \top), \bigcirc(w \wedge f \mid d)}{\bigcirc(\neg d \mid \top)}$$

is *not* valid. Finally, assume that there is a dog but there cannot be a white fence (e.g. there might be a black fence or no fence at all). Hence, the second obligation has been violated. In this situation, which is even more specific than the situation where there is a dog (d), nothing is said whether no fence is preferred over a non-white fence. Hence, the inference

$$\frac{\bigcirc(\neg f \mid \top), \bigcirc(w \wedge f \mid d)}{\bigcirc(\neg f \mid d \wedge \neg w)}$$

is *not* valid.

The following example illustrates that the invalidity of the inference patterns FC, RI and RIO can be explained by the multi preference semantics in Section 3.5.

Example 4.7 (Fence example, continued) Reconsider the multi preference model M in Figure 14 of the defeasible contextual obligations $\bigcirc^{re}(\neg f|\top\backslash\bot)$ and $\bigcirc^{re}(w \wedge f|d\backslash\bot)$ in Example 4.4. Figure 14 shows why the two inference patterns FC and RI are not valid. First of all, the obligation not to establish a conflict is not valid, $M \not\models \bigcirc^{re}(\neg d|\top \setminus \bot)$, because the $\neg d$ worlds (the most normal $\neg d$ worlds) are no better than the $d \wedge w \wedge f$ worlds (the optimal most normal d worlds). Secondly, the inference pattern reinstatement is not valid, $M \not\models \bigcirc^{re}(\neg f|d \wedge \neg w\backslash\bot)$, because all $d \wedge \neg w$ worlds are equivalent. Hence, if we zoom in on these worlds, there is no preference for f or $\neg f$.

The invalidity of inference patterns similar to FC, RI and RIO shows that strong overridden defeasibility is not sufficient to model reasoning about prima facie obligations. In other words, the obligations that model the Fence example are a different type of obligations than the obligations that model prima facie obligations.

4.3. *Weak overridden defeasibility*

The notion of weak overridden defeasibility can be formalized in a prioritized system. We do not give the formal definitions of a prioritized system, because they can be found in many articles on defeasible reasoning (see e.g. (Brewka, 1994; Geffner and Pearl, 1992; Vreeswijk, 1993)), but we illustrate the idea of a prioritized system by our promises example.

Example 6.2 (Promises, continued) In a prioritized system, a single preference ordering (an ideality ordering) is constructed for the obligations

$\bigcirc_{pf}(k|\top)$ and $\bigcirc_{pf}(p\wedge\neg k|d)$. To construct the ordering, a naming mechanism is used, similar to the one in conditional entailment (Geffner and Pearl, 1992). When the ordering is constructed, the prioritization of (the violations of) the obligations is taken into account. A typical prioritized preference ordering of Example 6.1 in Section 4.1 is given in Figure 16. The important relations in this preference model are $w_1 < w_2$ for all $w_1 \in |\neg k \wedge p \wedge d| \cup |\neg k \wedge \neg d|$ and $w_2 \in |k \wedge \neg p \wedge d|$, which state that violating the second obligation is worse than violating the first obligation. Without the prioritization, these worlds would be incomparable. Figure 16 shows why the inference patterns FC and RI are valid. First of all, forbidden conflict FC is valid, because $M \models \bigcirc_{pf}(\neg d|\top\backslash\bot)$. This follows from the fact that all d worlds are sub-ideal. Secondly, reinstatement is valid because $M \models \bigcirc_{pf}(k|d \wedge \neg p \backslash \bot)$. The $d \wedge \neg p$ worlds are not equivalent. Hence, if we zoom in on these worlds, as represented by a dashed box, there is an obligation for k.

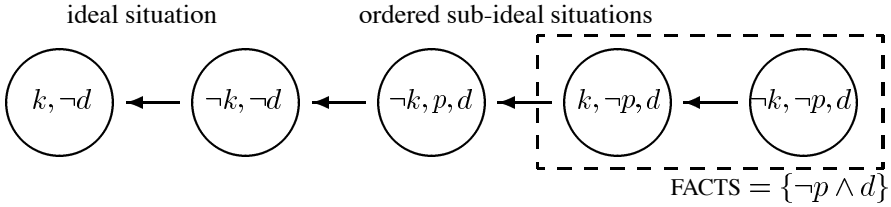

Fig. 16. Prioritized Preference Relation

Weak overridden defeasibility is quite close to overshadowing, but these notions are not identical. The typical case of overshadowing is that an obligation $\bigcirc(p|\top)$ is violated by the fact $\neg p$. We can introduce the notion of an absolute obligation $\bigcirc p$ to express that, in spite of the factual violation, the obligation is still in force. In the typical case of weak overridden defeasibility there are two conflicting obligations, say $\bigcirc_{pf}(p \mid \top)$ and $\bigcirc_{pf}(\neg p \mid q)$ and the fact q, with a preference ordering. To illustrate the difference with overshadowing, let us assume that the second obligation is preferred over the first one. We could generalize the logic of absolute obligations to take preference orderings into account, and then these two obligations would imply the actual obligation $\bigcirc\neg p$, but not $\bigcirc p$. This obligation expresses the duty proper, the obligation that should be acted upon. But these obligations would also imply both prima facie obligations $\bigcirc_{pf}\neg p$ and $\bigcirc_{pf} p$, which express that both obligations are still in force. These prima facie obligations resemble the absolute obligations of overshadowing. Hence, overshadowing and weak overridden defeasibility are equivalent from the point of view of 'cue for action': once an obligation is violated, it is still fully in force, but no longer a cue for action. Once an obligation is weakly overridden, it is no longer fully in force, but it is still in force as a prima facie obligation.

5. CONCLUSIONS

In this article we analyzed different types of defeasibility in defeasible deontic logics. We discriminated between two concepts, i.e. overshadowing and cancelling, and three types of defeasibility, i.e. factual defeasibility, strong overridden defeasibility and weak overridden defeasibility. We argued that factual defeasibility should be used to model violability, that strong overridden defeasibility should be used to model specificity and that weak overridden defeasibility should be used to model prima facie obligations. We also showed that the distinction between different types of defeasibility is essential for a better understanding of some of the notorious paradoxes of deontic logic, namely the Chisholm and Forrester paradoxes. Moreover, we introduced several preference-based semantics for deontic logics to analyze the different types of defeasibility.

ACKNOWLEDGEMENTS

This research was partially supported by the ESPRIT III Basic Research Project No.6156 DRUMS II and the ESPRIT III Basic Research Working Group No.8319 MODELAGE. Thanks to Patrick van der Laag, John-Jules Meyer, Henry Prakken and Marek Sergot for useful comments on earlier versions of this article.

Leendert W.N. van der Torre
EURIDIS, Tinbergen Institute and Department of Computer Science
Erasmus University
Rotterdam, The Netherlands

Yao-Hua Tan
EURIDIS
Erasmus University
Rotterdam, The Netherlands

REFERENCES

Alchourrón, C. E. (1994). Philosophical foundations of deontic logic and the logic of defeasible conditionals. In Meyer and Wieringa (eds.), *Deontic Logic in Computer Science: Normative System Specification,* John Wiley & Sons, pages 43–84.

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

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

Boutilier, C. and Becher, V. (1995). Abduction as belief revision. *Artificial Intelligence* 77:43–94.

Brewka, G. (1994). Adding specificity and priorities to default logic. In *Proceedings of Logics in artificial intelligence : European workshop,* Springer-Verlag.

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

Chellas, B. F. (1974). Conditional obligation. In Stunland, S. (ed.), *Logical Theory and Semantical Analysis: Essays dedicated to Stig Kauger,* D. Reidel Publishing Company, Dordrecht, Holland, pages 23–33.

Chellas, B. F. (1980). In Stunland S. (ed.), *Modal Logic: An Introduction* Cambridge University Press.

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

Dung, P. M. (1993). An argumentation semantics for logic programming with explicit negation. In *Proceedings of the Tenth Logic Programming Conference,* MIT Press, pages 616–630.

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

Gabbay, D. (1991). Labelled deductive systems. *Technical report, Centrum fur Informations und Sprachverarbeitung,* Universität Munchen.

Gärdenfors, P. D. (1988). *Knowledge in Flux*. MIT Press, Cambridge.

Geffner, H. and Pearl, J. (1992). Conditional entailment: bridging two approaches to default reasoning. *Artificial Intelligence* 53:209–244.

Hansson, B. (1971). An analysis of some deontic logics. In Hiplinen (ed.), *Deontic Logic: Introductory and Systematic Readings,* D. Reidel Publishing Company, Dordrecht, Holland, pages 121–147.

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

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

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

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

Levesque, H. J. (1984). A logic of implicit and explicit belief. In *Proceedings of the National Conference on Artificial Intelligence,* pages 198–202.

Levesque, H. J. (1990). All I know: A study in autoepistemic logic. *Artificial Intelligence* 42:263–309.

Lewis, D. (1974). Semantic analysis for dyadic deontic logic. In Stunland (ed.), *Logical Theory and Semantical Analysis* D. Reidel Publishing Company, Dordrecht, Holland, pages 1–14.

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

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

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

Morreau, M. (1996). Prima facie and seeming duties. *Studia Logica* 57:47–71.

Mott, P. L. (1973). On Chisholm's paradox. *Journal of Philosophical Logic* 2.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, Los Altos, CA.

Pearl, J. (1990). System Z: A natural ordering of defaults with tractable applications to default reasoning. In M. Vardi (ed.), *Proceedings of Theoretical Aspects of Reasoning about Knowledge,* San Mateo, Morgan Kaufmann, pages 121–135.

Pearl, J. (1993). A logic of pragmatic obligation. In *Proceedings of Uncertainty in Artificial Intelligence,*.

Powers, L. (1967). Some deontic logicians. *Noûs* 1:381–400.
Prakken, H. and Sartor, G. (1995). On the relation between legal language and legal argument: assumptions, applicability and dynamic properties. In *Proceedings of the Fifth International Conference on Artificial Intelligence and Law,* ACM Press, pages 1–9.
Prakken, H. and Sergot, M. J. (1996). Contrary-to-duty obligations. *Studia Logica* 57:91–115.
Prakken, H. and Sergot, M. J. (1997). Dyadic deontic logic and contrary-to-duty obligations. In Nute, D. (ed.), *This volume*.
Reiter, R. (1980). A logic for default reasoning. *Artificial Intelligence* 13:81–132.
Ross, D. (1930). *The Right and the Good*. Oxford University Press.
Ryu, Y. U. and Lee, R. M. (1993). Defeasible deontic reasoning: A logic programming model. In Meyer and Wieringa (eds.), *Deontic Logic in Computer Science: Normative System Specification,* John Wiley & Sons, pages 225–241.
Smith, T. (1993). Violation of norms. In *Proceedings of the Fourth International Conference on AI and Law,* ACM, New York, pages 60–65.
Tan, S.-W. and Pearl, J. (1994). Specification and evaluation of preferences under uncertainty. In *Proceedings of the Fourth International Conference on Principles of Knowledge Representation and Reasoning,* pages 530–539.
Tan, Y.-H. and van der Torre, L. W. N. (1994). DIODE: Deontic logic based on diagnosis from first principles. In *Proceedings of the Workshop 'Artificial normative reasoning' of the Eleventh European Conference on Artificial Intelligence,* Amsterdam.
Tan, Y.-H. and van der Torre, L. W. N. (1994). Representing deontic reasoning in a diagnostic framework. In *Proceedings of the Workshop on Legal Applications of Logic Programming of the Eleventh International Conference on Logic Programming,*.
Tan, Y.-H. and van der Torre, L. W. N. (1995). Why defeasible deontic logic needs a multi preference semantics. In *Proceedings of the Symbolic and Quantitative Approaches to Reasoning and Uncertainty,* Springer Verlag, pages 412–419.
Tan, Y.-H. and van der Torre, L. W. N. (1996). How to combine ordering and minimizing in a deontic logic based on preferences. In *Deontic Logic, Agency and Normative Systems. Proceedings of the Third Workshop on Deontic Logic in Computer Science,* Springer Verlag, pages 216–232.
Thomason, R. (1981). Deontic logic as founded on tense logic. In R. Hilpinen (ed.), *New Studies in Deontic Logic,* D. Reidel, pages 165–176.
Thomason, R. and Horty, R. (1996). Nondeterministic action and dominance: foundations for planning and qualitative decision. In *Proceedings of the Sixth Conference on Theoretical Aspects of Rationality and Knowledge,* Morgan Kaufmann, pages 229–250.
Tomberlin, J. E. (1981). Contrary-to-duty imperatives and conditional obligation. *Noûs* 16:357–375.
van der Torre, L. W. N. (1994). Violated obligations in a defeasible deontic logic. In *Proceedings of the Eleventh European Conference on Artificial Intelligence,* John Wiley & Sons, pages 371–375.
van der Torre, L. W. N. and Tan, Y.-H. (1995). Cancelling and overshadowing: two types of defeasibility in defeasible deontic logic. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence,* Morgan Kaufman.
van der Torre, L. W. N. and Tan, Y.-H. (1996). Contextual obligations. *Technical Report 96-04-01,* EURIDIS, Erasmus University Rotterdam.
van Eck, J. (1982). A system of temporally relative modal and deontic predicate logic and its philosophical applications. *Logique et Analyse* 25:249–290 and 339–381.
von Wright, G. H. (1951). Deontic logic. *Mind* 60:1–15.
von Wright, G. H. (1963). *The logic of preference*. Edinburgh University Press.
von Wright, G. H. (1968). *An Essay on Deontic Logic and the General Theory of Action*. North-Holland Publishing Company, Amsterdam.

Vreeswijk, G. (1993). *Studies in Defeasible Argumentation*. PhD thesis, Free University Amsterdam.