Cancelling and Overshadowing Two Types of Defeasibility in Defeasible Deontic Logic*

Leendert W N van der Torre
EURIDIS, Tinbergen Inst and Dept of CS
Erasmus University Rotterdam
P O Box 1738
3000 DR Rotterdam
The Netherlands

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

Abstract

IN this paper we give a general analysis of dyadic, deontir logics that were introduced in the earlv eventies to formalize deontic reasoning about subideal behavior Recently it was observed that they are closely related to non-monotonic logics, theories of diagnosis and dc cision theories In particular, we argue that two types of defeasibihty must be distinguished in a defeasible deontic logic overridden defeasibihty that formalizes cancelling of in obligation b\ other conditional obligations and factual defeasibility that formalizes, overshadowing of an obligation by a violating fact We also show that this, distinction is essential for an adequate analysis of notorious 'paradoxes of deontic logic such as the Chisholm and Forrester Paradoxes

1 Introduction

In recent >ears defeasible deontic logic has become increasingly popular as a tool to model legal reasonmg in expert systems [McCart\, 1992 Mever and Wiennga, 1994, Jones and Sergot 1994] because defeasible re asoning is an important aspect of legal reasoning [Prakken 1993] Deontic logic is a modal logic in which the modal operator *O* is used to express that something is obligatory [1] For example if the proposition r stands for the fact that you are robbed then (O(->7) expresses that you ought not to be robbed Dyadic modal logics were introduced to formalize deontic reasoning about subideal behavior in, for example the Chisholm 'Paradox that we will discuss later An example of a conditional obligation m a dyadic modal logic is *0(h* | r) which expresses that you ought to be helped *(h)* when you are robbed (r) If both *0(-r* | T) and are true then we say that the obli gation is *violated by* the fact r In recent years it was ar gued by several authors that these dyadic obligations can be formalized in non monotonic logics [McCarty 1992, Harty 1993, Rvu and Lee 1994]

In defeasible reasoning one (an distinguish two types of defeasibihty To illustrate the difference between the two we consider the default rule -^ This default can be defeated by the fact -p, or it can be overridden by another more specific default q-f/p, for example in Brcwkds prioritized elefault logic [Brewka, 1994] \Ve call the first case facatual *defeasibihty* and the last one *ovverridden de feasibility* In both the se cases the default p/p is cancelled either by the fact -p or by the default rule respec tivelv By cancelation wc mean, for example, that if *—p* is true then the default assumption that *p* is true is null and void The truth of *-p* implies that the default assumption about *p* is completely" falsified To say that a fact is inconsistent with a default rule makes no sense literallv because a default rule has no truth value However if we consider tbe autoepisttmu translation [honohge 1988] ->l-y> -> *p* of the default ■* (with the an toepisteniu belief operator *L)* then -<*L~^p* A -ip is ine (insistent In other words the default assumption —>£—>p is not consistent with the fact ->/» The fundamental difference between deontic logic and logics for defeasible reasoning is that *0{p* | T) A -<*p* i« *not* inconsistent That is the reason why the deonlic operator *O* 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 ai the sainr time \1 though the obligation *0{p* | T) is violated by the fact -<*p,* the obli gation still has its fe)rce For example even if you are robbed, you should not have been robbed But if a penguin cannot fly, it makes no sense to state that normally he can fly We will refer to this relation between the obligation and its violation as *overghadounng* to distinguish

This research was partially supported b> the ESPRIT III Basic Research Project No 6156 DRUMS II and the ESPRIT III Basic Research Working Group No 8319 MODEL A F

[1]The best known deontic logic is so-called 'standard de ontic logic (SDL) a normal modal system of type RD accord ing to the Chellas classification [Chellas, 1980] It satisfies besides the inference rules modus ponens $\frac{p, p \to q}{q}$ and necessitaion $\frac{\vdash p}{\vdash \mathcal{O}(p)}$, the propositional tautologies and the axioms K $(\mathcal{O}(p \to q) \wedge \mathcal{O}(p)) \to \mathcal{O}(q)$ and D $\neg(\mathcal{O}(p) \wedge \mathcal{O}(\neg p))$

it from cancellation in the ca.se of defeasible logics [2]

In this paper we defend two claims First that a number of notorious 'paradoxes of deontic logic can be solved when they are analyzed as forms of defeasible reasoning This has alread} been defended by other authors before us Second]} and this is a new claim we argue that an analogous distinction between factual dcfcasibilit\ and overridden defcasibility also holds for defeasibli obligations A defeasible obligation $\mathcal{O}(p \mid \top)$ can be violated by a fact $\neg p$ or overridden bv another obligation $\mathcal{O}(\neg p \mid q)$ However the important difference is that in the case of default logics both types of defeasibility are cancellation whereas in the case of deontic logic onl\ overnding leads, to cancellation, while violation leads to overshadowing Because of this difference between can ccllation in the first cast and overshadowing in the second case, it becomes essential not to confuse the two t\pes of defeasibility in analyzing the 'paradoxes' We show that if thev are confused, counterintuitive conclusions follow for the Chisholm and Forrester Paradoxes The distinction between these two kinds of defeasibihty is in our opinion the real paradox of defeasible deontit. logic because the\ have 90% overlap but they are different This distinction will become clear when we analyze the detachment of absolute obligations from the dyadic obligations

This paper is organized as follows In Section 2 we give a detailed comparison of factual and overridden defeasibihtv in deontit reasoning and we show that the Chisholm 'Paradox can better be analyzed as a case of factual defeasibility rather than overridden dcfeasibilitv, as is usually done In Section 3 we focus on the overshadowmg aspect of factual defeasibility and the cancellation aspect of overridden defeasibility, and we show that in an adequate anal} sis of the Forrester 'Paradox both these aspects ha\e to be combined In Section 4 we discuss further research

2 Overridden versus Factual Defeasibihty

In this section wc give a general anal} sis of defeasible deontic logic by analyzing some intuitive inference patterns [1] We show the fundamental difference between factual and overridden defeasibihty

[2] Tbe conceptual difference between cancelling and over shadowing is similar to the distinction between 'defeasibihtv and 'violanility' made in [Smith 1993] and [Prakken and Ser got, 1994] However, the essential difference between these papers and this one is that in this paper we argue that vio lability has to be considered as a type of defeasibihty too

[1] See [Tan and van der Torre, 1994b, 1995] for a semantic analysis of the two types In the multi preference spmantics there are two distinct preferenre orderings one ideality preference ordering which can be used to formalize deontic reasoning about subideal behavior and one normality preference ordering which can be used to formalize a notion of ovf rndden

2 1 Contrary-To-Duty obligations

Deontic logic is plagued by manv 'paradoxes' intuitvely consistent sentences which are formally inconsistent or derive counterintuitive sentences The most notorious 'paradoxes' arc caused by so-called Contrary-To-Dut> (CTD) obligations obligations that refer to subideal situations For example, Lewis describes the CTD obligation that \ou ought to be helped when \ou 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 possibility s that might otherwise have been actualized, and the helping is needed in order to actualize the but of those that remain Among the best possible worlds marred by the robbing tht best of the bad lot are some of thosf where the robbing is followed by helping [Lewis 1 974]

In the early seventies several dvadic modal systems were introduced to formalize CTD obligations sne [LCWJS 1974] for an overview Unfortunately several technical problems related to CTD reasoning persisted in the d\adic logics see [Tomberhn, 1981] A dyadic obligation $0(n \mid ø)$ can be read as 'if ø (the antecedent) is the case then a (the consequent) ought to be the case A CTD obligation is a dyadic obligation of which the antecedent contradicts the conclusion of another obligation For example if wc have $\mathcal{O}(\alpha \mid \beta)$ and $\mathcal{O}(\gamma \mid \neg\alpha)$ then the last one is a CTD (alias secondary obligation and the first one is called its primary obligation CTD obligations refer to optimal subideal situations In the subideal situation that $\mathcal{O}(\alpha \mid \beta)$ is violated $\beta \wedge \neg\alpha$, the best thing to do is γ Recently if was observed that the violation can be formalized in non-monotonic logics [McCarty, 1992 Horty 1993] theories of diagnosis [Tan and van der Torre 1994a, 1994c] and decision theories [Boutilier 1994b] Wc sa\ that dyadic obligations satisfy the Kantian principle 'ought implies ran when ought refers to 'the best of those that remain' This will be explained in more detail in Section 2 3

Since the late seventies, several temporal deontic lopes and deontic action logics were introduced, eg [Loewer and Belzer, 1983, Makinson, 1993 Alchourron, 1994] which formalize sat isfact only a special type of CTD obligations Temporal deontic logics formalize conditional obligations in which the consequent occurs later than the antecedent The underlying principle of the formalization of CTD obligations is that facts of the past are not in the 'context of judgment' [Loewer and Belzer 1983] 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 3) and the Chisholm 'Paradox' (see Example 2), because in these 'paradoxes' there arc CTD obligations of which the consequent occurs at the same time or even before its antecedent

2 2 Overridden defeasibility

A *defeasible* deontic logic can formalize obligations that ran be overridden hy other obligations Overridden structures can be based on a notion of specificity, like in Horty s well-known example that vou should not eat with your fingers, but if you are served asparagus you should cat with your fingers [Hortv 1993] We say that an obligation is cancelled by exceptional circumstances when it is overridden For example the obligation not to cat with your fingers is cancelled bv the exceptional circumstances that vou are served asparagus

In a defeasible deontic logic conditional obligations are defeasible conditionals In recent years several authors have proposed to solve the Chisholm Paradox' bv analyzing the problematic CTD obligation that occurs in it as a tvpe of overridden defeasibilitv (see eg [MrCartv, 1992, Ryu and Lee, 1994]) [4] The underlying idea is that a CTD obligation can be considered as a conflicting obligation that override a primary obligation Although this idea seems to be ver\ intuitive at fust sight, we argue in this paper that this per/spec live of CTD obligations as a kind of overridden defeasibility is misleading It is misleading because although this perspective vields most (but not all') of the correct conclusions for the Chisholm Paradox , it does so for the wrong rca sons We show that it is more appropriate to consider the. CTD obligation as a kind of factual defeasibilitv This docs not mean that the re is no place for overridden defeasibilitv in deontic logic Bv a careful analysis of an extended version of another notorious paradox of d< otitic logie, the Fonester Paradox we show that sometimes combinations of factual and overridden defeasibilitv art needed to represent defeasible deontic reasoning But first we give our analysis of the (hisholm 'Paradox. Fust we present the 'paradox[1] in a normal dvadu deontic logic to show its paradoxical character Subsequently we analyze the paradox in a defeasible deontic logic m which there is only overridden defeasibilitv and discuss the shortcomings of this approach Finally we give an analysis of the Chisholm 'Paradox in terms of factual dcfeasibility To make our analvsis as general as possible, we assume as little as possible about the deontic logie we use The analyses given in this paper m terms of inference patterns are, m principle, applicable to anv defeasible deontic logic

Assume a deontic. logic with a finite piopositional base logic *C* and dyadic modal obligations $O(\alpha \mid \beta)$, where β (the antecedent) and α (the consequent) arc sentences of *C* Assume further the unrestricted strengthening of the antecedent rule SA

$$\text{SA}\quad \frac{O(\alpha \mid \beta)}{O(\alpha \mid \beta \wedge \gamma)} \tag{1}$$

[4][McCarty, 1992] does not analyze the Chisholm 'Paradox but thf so-called Revkjavic 'paradox' which he considers to contain two instances of the Chisholm 'Paradox' each one interacting with the other'

Finally, assume the deontic detachment (alias transitivity) rule DD

$$\text{DD}\quad \frac{O(\alpha \mid \beta), O(\beta \mid \gamma)}{O(\alpha \mid \gamma)} \tag{2}$$

The notorious Chisholm Paradox' [Chisholm, 1963] (alias the CTD 'paradox , alias the 'paradox of deontic detachment) is as follows [5]

Example 2 1 (Chisholm 'Paradox') *Consider the premises* $O(a \mid \top)$, $O(t \mid a)$ *and* $O(\neg t \mid \neg a)$ *where* $\top$ *stands for any tautology a can be read as the fat t that a certain man qoes to the assistance of his ntiqhbors and t* as *the fact that he tells thevi he is (omtng Tht premise* $O(\neg t \mid \neg a)$ *is a CTD obliqeitum of the (primary) obligation* $O(a \mid \top)$ *because its antecedent is inconsistent vnth thf (onsequcnt of the latter*

The intuitive obligation $O(t \mid \top)$ *can be derived by* DD *from the first two obligations ft seems intuitive because in the ideal situation the man qoes to the assistant f of his neiqhbors* and he tells them he *is* coming *Hener if he does not tell them tht ideal situation* is no *longer reathable However from* $O(t \mid \top)$ *the counterintuitive* $O(t \mid \neg a)$ *can he derived by* S A *This* is *counterintuitive because there is no reason to tell them he is coming when the man docs not qo Moreover in many denotes logics* $O(\neg t \mid \neg p)$ *and* $O(t \mid \neg p)$ *are inconsistent*

This counterintuitive obligation rannot be derived in a *defeasible* deontic logic with overridden defeasibility For our argument we use a notion of overridden based on specificity Assume that SA is replaced b\ the following restricted strength' mug of the antecedent rule RSA_O RSA_O contains the so-called non-overridden condition C_O which represents that $O(\alpha \mid \beta)$ is not overridden b\ some $O(\alpha' \mid \beta')$ for $\beta \wedge \gamma$ It is based on a simplihed notion of spcciht it\ because background knowledge is not taken into acrount and an obligation rannot be o\ernddcn b\ more than one obligation

$$\text{RSA}_O\quad \frac{O(\alpha \mid \beta), C_O}{O(\alpha \mid \beta \wedge \gamma)} \tag{3}$$

where condition C_O is defined as follows

> C_O there is no premise $O(\alpha' \mid \beta')$ such that $\beta \wedge \gamma$ logically implies β', β' logically implies β and not vice versa and α and α' are inconsistent

The following solution can now be given for the 'paradox'

Example 2 2 *The intuitive obligation* $O(t \mid \top)$ *can still be derived by* DD *from the first two obligations From*

[5]The original 'paradox[1] was Riven in a monadic modal logit, here we give the obvious formalization in a dvadu logic See [Tomberlin 1981] for a discussion of the Chisholm 'Paradox' in several conditional deontic logics

$O(t \mid \top)$ the counterintuitive $O(t \mid \neg a)$ cannot be derived by RSA_O, because $O(t \mid \top)$ is overridden for $\neg a$ by the CTD obligation $O(\neg t \mid \neg a)$, 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.

Though this yields intuitive results from the set of premises, we think that it does so for the wrong reasons. A simple counterargument against the solution of the 'paradox' above is that overriding based on specificity does not work any more when the first premise is $O(a \mid i)$, 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 trick does not work either when the set of premises contains only the first two obligations, as is the case in the following example.

Example 2.3 *Consider only the premises $O(a \mid \top)$ and $O(t \mid a)$. Again the intuitive obligation $O(t \mid \top)$ can be derived by DD. From this derived obligation the counterintuitive $O(t \mid \neg a)$ can again be derived by RSA_O, because there is no CTD obligation which cancels the counterintuitive obligation.*

In [Tan and van der Torre, 1994c] we dubbed the intuition that the obligation $O(t \mid \top)$ is intuitive and the obligation $O(t \mid \neg a)$ is counterintuitive as *deontic detachment as a defeasible rule*. The obligation $O(t \mid \top)$ that is derived by DD lacks unrestricted strengthening of the antecedent, the characteristic property of defeasible conditionals. The underlying assumption of the 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.

The problematic character of DD is well-known from the Chisholm 'Paradox', and it is therefore usually not accepted for deontic logics. However, the same phenomena occurs when SA (or RSA_O) and the following rule of consequential closure CC is accepted, and this rule is accepted by many deontic logics.[6]

$$\mathrm{CC} \quad \frac{O(\alpha_1 \mid \beta),\ O(\alpha_1 \to \alpha_2 \mid \beta)}{O(\alpha_2 \mid \beta)} \tag{4}$$

This is shown by the following variant of the example, where the conditional obligation is represented as an absolute obligation.[7]

Example 2.4 *Consider the premises $O(a \mid \top)$ and $O(a \to t \mid \top)$. The intuitive obligation $O(t \mid \top)$ is derived from the two premises by CC. However, from this derived obligation the counterintuitive $O(t \mid \neg a)$ can be derived by SA or RSA_O.*

The examples show that CTD reasoning, i.e. reasoning about subideal behavior, cannot be formalized satisfactorily in a defeasible deontic logic with only overridden defeasibility. Hence we cannot accept SA (or RSA_O), represent (a-temporal) CTD obligations and accept DD or CC.

2.3 Factual defeasibility

As an illustrative example of a formalization of factual defeasibility, we discuss the so-called non-violability condition C_V of our deontic logic DIODE, see [Tan and van der Torre, 1994a; 1994c]. DIODE is a diagnostic model for deontic reasoning based on Reiter's theory of diagnosis [Reiter, 1987]. The underlying idea of DIODE is that violated obligations are analogous to faulty components in diagnostic reasoning. In DIODE the assumption-based reasoning discussed in Example 2.3 is related to the assumptions about faulty components made in diagnostic reasoning.

Assume a finite propositional base logic $\mathcal{L}$ and labeled dyadic conditional obligations $O(\alpha \mid \beta)_L$ with α, β and L sentences of $\mathcal{L}$. Roughly speaking, the label L is a record of the consequences of the premises that are used in the derivation of $O(\alpha \mid \beta)$. Each formula occurring as a premise 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 conjunction of the labels of the premises used in this inference rule. The non-violability condition C_V is used to realize the *Kantian principle* that '*ought implies can*'. Informally, 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.

$$\mathrm{RSA}_V \quad \frac{O(\alpha \mid \beta)_L,\ C_V}{O(\alpha \mid \beta \wedge \gamma)_L} \tag{5}$$

C_V: $L \wedge \beta \wedge \gamma$ is consistent

$$\mathrm{RSA}_{OV} \quad \frac{O(\alpha \mid \beta)_L,\ C_O, C_V}{O(\alpha \mid \beta \wedge \gamma)_L} \tag{6}$$

C_V: $L \wedge \beta \wedge \gamma$ is consistent

$$\mathrm{DD}_V \quad \frac{O(\alpha \mid \beta)_{L_1}, O(\beta \mid \gamma)_{L_2}, C_V}{O(\alpha \mid \gamma)_{L_1 \wedge L_2}} \tag{7}$$

C_V: $L_1 \wedge L_2 \wedge \gamma$ is consistent

[6] For examples of deontic logics *not* satisfying the CC rule, see Chellas' CKD [Chellas, 1980] (a nonnormal modal deontic logic) and Hansson's Preference Deontic Logic (PDL) [Hansson, 1990].

[7] It may be argued that the premise $O(a \to t \mid \top)$ does not represent the conditional obligation correctly. However, [Horty, 1993] gave an example (adapted from an example of [van Fraassen, 1973]) in which similar inferences are made from the premises $O(a \vee s \mid \top)$ and $O(\neg a \mid \top)$, where a can be read as the fact that you are in the army and s as the fact that you perform alternative service. Moreover, when DD is not accepted and CC is, then this leads to semantic problems as [Tomberlin, 1981] showed for *Mott's solution of* the Chisholm 'Paradox' [Mott, 1973].

$$\mathrm{CC_V} \quad \frac{O(\alpha_1 \mid \beta)_{L_1}, O(\alpha_1 \rightarrow \alpha_2 \mid \beta)_{L_2}\ C_V}{O(\alpha_2 \mid \beta)_{L_1 \wedge L_2}} \qquad (8)$$

C_V: $L_1 \wedge L_2 \wedge \beta$ is consistent

It can easily be checked that for example $\mathrm{RSA_V}$ is better than $\mathrm{RSA_O}$, because $\mathrm{RSA_V}$ yields *all of the intended conclusions* of the Examples 2.1-2.4, but none of the counterintuitive conclusions produced by $\mathrm{RSA_O}$.

The reader might wonder why we consider condition C_V to be a type of factual defeasibility. In this section we only discuss conditional obligations, and how these can be derived from each other. Facts do not seem to come into the picture. However, a closer analysis reveals *that factual defeasibility is indeed the underlying mechanism*. First of all, the antecedent of a dyadic obligation restricts the focus to possibilities in which the antecedent is *assumed* to be factually true, and the consequent represent what is obligatory given that these facts are assumed. Hence, the consequent refers to 'the best of those possibilities that remain'. The Kantian principle ought implies can, states essentially that what ought to be the case (i.e. the consequent) has to be possible given these *assumed-to-be true* facts (i.e. the antecedent). This is the meaning of 'can' here. An analogy that illustrates that the Kantian Principle induces factual defeasibility is to compare the $\mathrm{DD_V}$ rule with its default logic counterpart. The order of application of default rules in the generation of an extension can be compared to the transitivity of obligations in the $\mathrm{DD_V}$ rule. For example, the default $\frac{\beta:\alpha}{\alpha}$ can be applied after the default $\frac{\gamma:\beta}{\beta}$, because the consequent of the first can be used to obtain the prerequisite of the second. But this chain would be broken if γ would factually defeat β (and assuming there is no other way to obtain β).

The Examples 2.1-2.4 show that CTD structures sometimes look like overridden defeasible reasoning structures, but a careful analysis shows that they are actually cases of factual defeasibility. The difference between the conditions C_O and C_V explains the confusion between CTD structures and overridden structures in Example 2.2, because in this example the two restrictions coincide for strengthening of the antecedent.

We introduced C_V in DIODE to deal with CTD obligations. This condition is analogous to a proposal of Van Fraassen [van Fraassen, 1973], which was formalized by Horty [Horty, 1993] in Reiter's default logic. They introduced it to represent moral dilemmas, deontic inconsistencies like $O(p \mid \top)$ and $O(\neg p \mid \top)$. See [Horty, 1993] for the motivation and [Tan and van der Torre, 1994a] for the comparison with DIODE. For other proposals of factual defeasibility, see Hansson's dyadic deontic logic [Hansson, 1971] (no strengthening of the antecedent) and Boutilier's extension of Hansson's logic [Boutilier, 1994b].

2.4 Facticity

As discussed in Section 2.1, in dyadic deontic logic ought refers to 'the best of those that remain'. Dyadic obligations $O(\alpha \mid \beta)$ can be read as 'α is the case in the best states where β is the case'. The following rule of facticity F is intuitive under this reading of conditional obligations. As has been pointed out many times, see e.g. [Alchourron, 1994], F is counterintuitive with the original reading of the dyadic obligations, because it says that 'if α is the case then α ought to be the case'.

$$\mathrm{F} \quad \frac{}{O(\alpha \mid \alpha)} \qquad (9)$$

The next example shows a possible use of F.

Example 2.5 *Consider the single premise $O(a \rightarrow t \mid \top)$. The obligation $O(a \rightarrow t \mid a)$ is derived from the premise by RSA_V, and the obligation $O(a \mid a)$ is derived by F. From these two obligations the intuitive obligation $O(t \mid a)$ is derived by CC_V.*

3 Cancelling versus Overshadowing

In this section we analyze the derivation of absolute obligations from the dyadic obligations in a defeasible deontic logic. To keep our analysis as general as possible, we only accept the inference pattern $\mathrm{RSA_O}$ for the dyadic obligations. The inference pattern that derives absolute obligations from conditional obligations is called factual detachment. To represent the detached absolute obligations we assume monadic modal obligations $O(\alpha)$. No further properties of the monadic operator are assumed. The simplest definition of factual detachment is the following rule FD, alias deontic modus ponens.

$$\mathrm{FD} \quad \frac{O(\alpha \mid \beta), \beta}{O(\alpha)} \qquad (10)$$

Obviously, FD is not acceptable in a *defeasible* deontic logic, because it detaches overridden obligations. The following exact factual detachment rule EFD does not derive overridden obligations. Here it is formalized with Levesque's All-I-Know (alias only knowing) operator $\mathcal{A}$ (see [Boutilier, 1994a]). $\mathcal{A}(\alpha)$ is true when α is logically equivalent with the conjunction of all factual premises that are given.

$$\mathrm{EFD} \quad \frac{O(\alpha \mid \beta), \mathcal{A}(\beta)}{O(\alpha)} \qquad (11)$$

Note that EFD yields a kind of overridden defeasibility with respect to factual detachment. If we have, for example, as premises $O(\alpha \mid \beta)$ and $O(\neg\alpha \mid \beta \wedge \gamma)$, then EFD derives the conclusion $O(\alpha)$ if we *only* have as factual premises β. However, if we have as factual premise $\beta \wedge \gamma$, then EFD derives from these two obligations $O(\neg\alpha)$. If EFD is accepted then the relation between facts and absolute obligations is identical to the relation between antecedent and consequent of the conditional obligations

The following so-called fence example was introduced in [Prakken and Sergot 1994] It is an extended version of the so-called Forrester Paradox' you should not kill, but if you kill you should do it gently [Forrester, 1984]

Example 3 1 (Forrester 'Paradox) *Consider the premises $O(\neg f \mid \top)$ $O(w \mid f)$ and $O(w \mid c)$ with background knowledge $w \rightarrow f$ where f can be read as the fact that there is a fence around your house w similarly for a white fence and c for a cliff next to your house We assume that the background knowledge is incorporated in the definitions of C_O and C_V in the obvious way Notice that $O(w \mid f)$ is a CTD obligation of $O(\neg f \mid \top)$ and $O(w \mid c)$ is not*

Let $\mathcal{F}$ be the conjunction of all factual premises When there is a fence and a cliff $\mathcal{F} = f \wedge c$ the first premise is intuitively overridden, and therefore it is not violated Hence the obligation $O(\neg f)$ should not *be derivable If there is a fence without a cliff $\mathcal{F} = f$ the first premise is intuitively not overridden and therefore it is violated Hence the obligation $O(\neg f)$ should be derivable*

The obligation $O(\neg f \mid f \wedge c)$ is not derived from $O(\neg f \mid \top)$ by RSA_O, because it is overridden by $O(w \mid c)$ The counterintuitive obligation $O(\neg f)$ can therefore not be derived by RSA_O and EFD from $O(\neg f \mid \top)$ and $\mathcal{F} = f \wedge c$ However the obligation $O(\neg f \mid f)$ is not derived either from $O(\neg f \mid \top)$ by RSA_O because it is overridden by $O(w \mid f)$ according to C_O Because $O(\neg f \mid f)$ is not derivable the intuitive obligation $O(\neg f)$ cannot be derived by RSA_O and EFD from $O(\neg f \mid \top)$ and $\mathcal{F} = f$

The problem in this example is that both $O(w \mid f)$ and $O(w \mid c)$ are treated as more specific obligations that override the obligation $O(\neg f \mid \top)$ i e both are treated as cases of overridden defeasibility However, this is not correct for $O(w \mid f)$ This last obligation should be treated as a CTD obligation, i e as a case of factual defeasibility What is most striking about the Forrester 'Paradox' is the observation that when the premise $O(\neg f \mid \top)$ is violated by f then the obligation $O(\neg f)$ should be derivable but not when $O(\neg f \mid \top)$ is overridden by $f \wedge c$ This means that violation or overriding of $O(\neg f \mid \top)$ 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 as a type of factual defeasibility Hence, also the Forrester Paradox shows that factual and overridden defeasibility lead to different conclusions Moreover it is exactly the difference between cancellation and overshadowing that we discussed in the introduction of this paper Overriding of $O(\neg f \mid \top)$ by $f \wedge c$ means that $O(\neg f)$ is cancelled and has no force anymore Violation of $O(\neg f \mid \top)$ by f means that $O(\neg f)$ has still its force, it is only overshadowed and not cancelled Hence, 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

One obvious way to solve the problem mentioned in Example 3 1 is to say that condition C_O is too strong In [van der Torre 1994] we gave an ad hoc solution of the previous problem by weakening the definition of overridden with an additional condition which represents that a CTD obligation cannot override its primary obligations

> C'_O there is no premise $O(\alpha' \mid \beta')$ such that $\beta \wedge \gamma$ logically implies β', β' logically implies β and not vice versa α and α' are inconsistent and α and β' are consistent [van der Torre 1994]

This definition gives the intuitive conclusions and not the counterintuitive ones with EFD, as the following example shows

Example 3 2 RSA_O *with C'_O derives $O(\neg f \mid f)$ from $O(\neg f \mid \top)$, but it does not derive $O(\neg f \mid f \wedge c)$ The counterintuitive obligation $O(\neg f)$ still cannot be derived by RSA_O and EFD from $O(\neg f \mid \top)$ and $\mathcal{F} = f \wedge c$ The intuitive obligation $O(\neg f)$ can be derived by EFD from $O(\neg f \mid f)$ and $\mathcal{F} = f$ Hence RSA_O with C'_O and EFD derive exactly the intuitive obligations*

Another more sophisticated solution is to change the EFD rule instead of C_O The most important advantage of changing EFD is that this is also a solution when RSA_{OV} is accepted The condition C_V of RSA_{OV} ensures that the consequent and the antecedent of a dyadic obligation are always consistent This consistency is a formalization of the Kantian principle ought implies can , as we discussed in Section 2 3 However with RSA_{OV} and EFD α and $O(\neg\alpha)$ will never be true at the same time so violated obligations are not represented by the absolute obligations Another advantage of changing EFD instead of C_O is that it is less ad hoc In [van der Torre 1994] we had to further adapt the definition of C'_O for another notorious (and highly ambiguous) paradox , the so-called Reykjavic 'Paradox' [Belzer, 1986] A third advantage of changing the EFD rule is that EFD cannot formalize fulfilled obligations satisfactorily A fulfilled obligation is something that ought to be the case and that is also factually the case For example, in the Forrester 'Paradox' the obligation $O(\neg f) \wedge \neg f$ could represent that there is no fence and the obligation that there ought to be no fence is fulfilled However, when the weakening of the consequent rule WC $\frac{O(\alpha|\beta)}{O(\alpha\vee\gamma|\beta)}$ is accepted then $O(f \mid f)$ can be derived from $O(w \mid f)$ From $O(f \mid f)$ and $\mathcal{F} = f$ the absolute obligation $O(f)$ can be derived by EFD, although the fence is certainly not a fulfilled obligation Moreover, if the inference rule F is accepted, then all facts are detached as absolute obligations by EFD We will show that this can be solved by changing the definition of EFD

To formalize a notion of factual detachment that ignores overridden defeasibility in case of violated obligations we introduce the following so-called *retraction test*

(R-test) The test says that if we consider whether α is absolutely obligatory we have to consider possibilities in which α is true and possibilities in which α is false

> R-test α is obligatory ($\mathcal{O}(\alpha)$ is an absolute obligation) iff α ought to be the case *on the assumption that* $\neg\alpha$ *and* α *are not the case* i e on the assumption that α is contingent

The R-test can be considered as a version of the Kantian principle for factual detachment In this interpretation of ought implies can' *ought* refers to the absolute obligations and *can* means that neither $\neg\alpha$ nor α is factually the case The R-test can be formalized as follows, where '$-$' is a retraction operator satisfying the Gardenfors postulates [Gardenfors 1988] For simplicity we write retraction as $\alpha = \beta - \{\gamma_i\}$ where α, β and γ_i are sentences of $\mathcal{L}$ α is the result of the retraction of the γ_i from β and therefore α does not derive any of these γ_i

$$\text{RFD} \quad \frac{\mathcal{O}(\alpha \mid \beta - \{\alpha, \neg\alpha\}),\ \mathcal{A}(\beta)}{\mathcal{O}(\alpha)} \qquad (12)$$

Notice that this formalization inherits problems of retraction i e that it is not unique and computationally complex For this reason we will use a very simple notion of retraction that suffices for our purposes We only consider cases where β is a conjunction of literals (atoms possibly preceded by a negation sign) and α a literal For this simple case the retraction $\beta - \{\alpha, \neg\alpha\}$ is simply the deletion of α and $\neg\alpha$ from β This type of retraction can be illustrated with the following examples If $\alpha = \neg f$ and $\beta = f$ then $\mathcal{O}(\neg f \mid f - \{\neg f, f\}) = \mathcal{O}(\neg f \mid \top)$ If $\alpha = \neg f$ and $\beta = f \wedge c$ then $\mathcal{O}(\neg f \mid f \wedge c - \{\neg f, f\}) = \mathcal{O}(\neg f \mid c)$ So if we want to derive $\mathcal{O}(\neg f)$ with RFD and the premise $\mathcal{A}(f \wedge c)$ then we need $\mathcal{O}(\neg f \mid c)$ as other premise

We reconsider the Forrester Paradox with this simplified notion of retraction and we show that RFD derives exactly the intuitive conclusions

Example 3 3 *Let $\mathcal{F}$ be the conjunction of the factual premises First consider the situation when there is a fence but not a cliff i e $\mathcal{F} = f$ The absolute obligation $\mathcal{O}(\neg f)$ can be derived from $\mathcal{O}(\neg f \mid \top)$ and $\mathcal{A}(f)$ by* RFD

Now consider the situation when there is a fence and a cliff i e $\mathcal{F} = f \wedge c$ The absolute obligation $\mathcal{O}(\neg f)$ cannot be derived by RFD, *because the derivation via $\mathcal{O}(\neg f \mid c)$ from $\mathcal{O}(\neg f \mid \top)$ is blocked by C_O*

In the previous example, RFD and RSA_{OV} yield exactly the same intuitive conclusions as EFD and RSA_O with C_O' The rules RFD and RSA_{OV} are in our opinion more appropriate to model defeasible deontic reasoning than the rules EFD and RSA_O with C_O' because RSA_{OV} is preferred over RSA_O, as we argued in the previous section Another advantage of RFD is that the R-test is very intuitive and not an ad hoc like solution of the problem like the adaptation of C_O Finally, RFD also *formalizes an intuitive notion of fulfilled obligations*, because it deals with fulfilled obligations in exactly the same way as with violated obligations

The relation between EFD and RFD is given by the following lemma

Lemma 1 *Let $\mathcal{F}$ be the conjunction of the factual premises If α and $\neg\alpha$ are not in $Cn[\mathcal{F}]$ where Cn stands for consequence set, then $\mathcal{O}(\alpha)$ is derived by* EFD *iff it is derived by* RFD

Proof *From the Gardenfors postulates follows that $Cn[\mathcal{F} - \{\alpha\}] = Cn[\mathcal{F}]$ when $F \wedge \neg\alpha$ is consistent*

4 Further research

In this piper we only consideied examples in which overridden defeasibility is always of the cancelling t\pe However, Ross gave in [Ross 1930] also examples of so-railed prima facie obligations that can be considered as overridden defeasibility of the overshadowing type we will study how this (an be analyzed in our framework We will also studv the relation between the R-test and the Ramsey test in conditional logic The crucial differeii(e between the R test and the Ramsey test is that in the R test the consequence is taken into account

Acknowledgement

We thank Patrick van der laag for several discussions on the issues raised in this paper and Henery Prakken for his critical comments

References

[Alebourron 191)4] C E Alehourron Philosophical foundations of deontic logic and the logic of defeasible conditionals In *Deontie Loqic in Computer once Normative* System *Specifications* pages 43-84 John Wiley & Sons 1994

[Belzer 1986] M Belzer \ logic of deliberition In Proceedinqs *of tht Fifth National Conference on Artificial Intelligent?* pages 38-43 1986

[Boutiher, 1994a] C Boutiher Conditional logics of normality a modal approach *Artifical Intelligence* G8 87-154 1994

[Boutiher 1994b] C Boutiher Towards a logic for qualitative decision theory In *Proceedings of the Fourth International Conference on Principles of Knowledge Representation and Reasoning (K R 94}* pages 75-36 1994

[Brewka, 1994] G Brevvka Adding specificity and priorities to default logic In *Proceedings of the JELIA 94,* 1994

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

[Chisholm 1963] R M Chisholrn Contrary-to-dutv imperatives and deontic logic *Analysis* 24 33-36, 1963

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

[Gardenfors 1988] P Gardenfors *Knowledge in Flux* MIT Press, Cambridge 1988

[Hansson 1971] B Hansson An analysis of some deontic logics In *Deontic Logic Introductionary and Systematic Readinqs* pages 121-147 D Reidel Publishing Company Dordrecht Holland 1971

[Hansson, 1990] 50 Hansson Preference-based deontic logic (PDL) *Journal of Philosophical Logic* 19 75-93 1991)

[Hort\, 1993] 1 F Horty Deontic logic as founded in nonmonotonic logic *Annals of Mathematics and AT tificial Intelligence* 9 69-91 1993

[Jones and Sergot, 1994] A J I Jones and M Sergot *Proceedings of the Second Workshop on Deontic Logic in Computer Science (Atan 94)* Oslo 1994

[Ronolige 1988] K Konolige On the relation between default and autoepistemir logic *Artificial Intelligence* 35 343 382 1988

[Lewis 1974] D Lewis Semantic analysis for dvadic deonly logic In *Logual Theory and Semantical Analysis* pages 1 14 D Reidel Publishing Companv Dordrecht Holland 1974

[Loewer and Belzer, 1983] B Loewer and M Belzer D\adu deontic detachment *Synthese,* 54 295-318 1983

[Makmson 1993] D Makinson Five fates of minimality *Studia Lagica* 52 339 379, 1993

[McCartv 1992] L T MrCartv Defeasible deontic reasoning In *Fourth International Workshop on Non monotonic Reasoning* Plymouth 1992

[Meyer and Wiennga, 1994] J -J Meyer and R Wiermga *Deontic Logic in Computer Science Normative system Specification Revision of selected papers that were presented at the First Workshop on Deontic Loqic in Computer Science (Aton 91)* John Wiley & Sons 1994

[Mott, 1973] PL Mott On Chisholm s paradox *Journal of Philosophical Logic* 2 1973

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

[Prakken, 1993] H Prakken *Logical Tools for Modelling Legal Argument* PhD thesis, Free University, Amsterdam, 1993

[Reitpr, 1987] R Reiter A theory of diagnosis from first principles *Artificial Intelligence* 32 57-95, 1987

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

[Ryu and Lee 1994] "Y U Rvu and R M Lee Defeasible deontic reasonig A logic programming model In *De onttc Logic in Computer Science Normative System specification* John Wiley & SOFIA, 1994

[Smith 1993] T Smith Violation of norms In *Proceed ings of the Fourth International Conference on AI and Law (ICA1L 93)* pages 60-65 Nev. York 1993 ACM

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

[Tan and van der Torre 1994b) Y -H Tan and L W N van der Torre Multi preference semantic s for a defeasible deontic logic In *Proceedings of the JURIX 94,* Amsterdam 1994

[Tan and van der Torre, 1994c] Y -H Tan and L W N van der Torre Representing deontic reasoning in a diagnostic framework In *Proceedings of tht Work shop on Legal Applications of Logic Programming of the Eleventh International Conference on Logic Pro qramming (ICLP 94),* 1994 An extended version will appear in *Journal of Artifitial Intelligence and Law*

[Tan and van der Torre 1995] Y -H Tan and L W N van der Torre Why defeasible deontic logic needs a multi preference semantics In *Proceedings of the ECSQARU'95,* Fribourg 1995

[Tomberlm 1981] JE Tomberlin Contrary-to-dutyim pcratives and conditional obligation *Nous,* 16 357 375 1981

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

[van Fraasaen, 1973] B C van Fraassen Values and the heart command *Journal of Philosophy,* 70 5-19,1973