Cancelling and Overshadowing: two types of defeasibility in defeasible deontic logic*†

Leendert W.N. van der Torre‡ Yao-Hua Tan‡

March 1, 1997

Abstract

In this paper we give a general analysis of dyadic deontic logics that were introduced in the early seventies to formalize deontic reasoning about subideal behavior. Recently it was observed that they are closely related to non-monotonic logics, theories of diagnosis and decision theories. In particular, we argue that two types of defeasibility must be distinguished in a defeasible deontic logic: overridden defeasibility that formalizes cancelling of an obligation by other conditional obligations and factual defeasibility that formalizes overshadowing of an obligation by a violation. 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'.

Keywords: knowledge representation, defeasible deontic logic, non-monotonic logic, legal reasoning

1 Introduction

In recent years defeasible deontic logic has become increasingly popular as a tool to model legal reasoning in expert systems [13, 18, 19], because defeasible reasoning is an important aspect of legal reasoning [12, 21]. Deontic logic is a modal logic in which the modal operator $\mathcal{O}$ is used to express that something is obligatory.[1] For example, if the proposition r stands for the fact that you are robbed, then $\mathcal{O}(\neg r)$ 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

*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.

†Submitted to IJCAI'95.

‡Erasmus University Research Institute for Decision and Information Systems (EURIDIS). Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands. E-Mail: {ltorre,ytan}@euridis.fbk.eur.nl. Tel: (+31)10-4082601. Fax: (+31)10-4526134. Http://www.euridis.fbk.eur.nl/Euridis/welcome.html.

[1]The best known deontic logic is so-called 'standard' deontic logic (SDL), a normal modal system of type KD according to the Chellas classification [5]. It satisfies, besides the inference rules modes ponens $\frac{p, p\rightarrow q}{q}$ and necessitation $\frac{\vdash p}{\vdash \mathcal{O}(p)}$, the propositional tautologies and the axioms K: $(\mathcal{O}(p \rightarrow q) \wedge \mathcal{O}(p)) \rightarrow \mathcal{O}(q)$ and D: $\neg(\mathcal{O}(p) \wedge \mathcal{O}(\neg p))$.

'Paradox' that we will discuss later. An example of a conditional obligation in a dyadic modal logic is $\mathcal{O}(h \mid r)$, which expresses that you ought to be helped (h) when you are robbed (r). If both $\mathcal{O}(\neg r \mid \top)$ 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 [11, 18, 25].

In defeasible reasoning one can distinguish two types of defeasibility. To illustrate the difference between the two we consider the default rule $\frac{:p}{p}$. This default can be defeated by the fact $\neg p$, or it can be overridden by another more specific default $\frac{q:\neg p}{\neg p}$, for example in Brewka's prioritized default logic [4]. We call the first case *factual defeasibility* and the last one *overridden defeasibility*. In both these cases the default $\frac{:p}{p}$ is cancelled either by the fact $\neg p$, or by the default rule $\frac{q:\neg p}{\neg p}$ respectively. By *cancellation* we mean, for example, that if $\neg p$ is true, then the default assumption that p is true is null and void. The truth of $\neg p$ implies that the default assumption about p is completely falsified. To say that a fact is inconsistent with a default rule makes no sense literally, because a default rule has no truth value. However, if we consider the autoepistemic translation [14] $\neg L \neg p \rightarrow p$ of the default $\frac{:p}{p}$ (with the autoepistemic belief operator L), then $\neg L\neg p \wedge \neg p$ is inconsistent. In other words, the default assumption $\neg L \neg p$ is not consistent with the fact $\neg p$. The fundamental difference between deontic logic and logics for defeasible reasoning is that $\mathcal{O}(p \mid \top) \wedge \neg p$ is *not* inconsistent. That is the reason why the deontic operator $\mathcal{O}$ 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 at the same time. Although the obligation $\mathcal{O}(p \mid \top)$ is violated by the fact $\neg p$, the obligation still has its force. 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 *overshadowing* to distinguish it from cancellation in the case of defeasible logics.

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 already been defended by other authors before us. Secondly, and this is a new claim, we argue that an analogous distinction between factual defeasibility and overridden defeasibility also holds for *defeasible* obligations. A defeasible obligation $\mathcal{O}(p \mid \top)$ can be violated by a fact $\neg p$, or overridden by 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 only overriding leads to cancellation, while violation leads to overshadowing. Because of this difference between cancellation in the first case and overshadowing in the second case, it becomes essential not to confuse the two types of defeasibility in analyzing the 'paradoxes'. We show that if they are confused, counterintuitive conclusions follow for the Chisholm and Forrester 'Paradoxes'. The distinction between these two kinds of defeasibility is in our opinion the real paradox of defeasible deontic logic, because they 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 defeasibility in deontic reasoning, and we show that the

Chisholm 'Paradox' can better be analyzed as a case of factual defeasibility rather than overridden defeasibility, as is usually done. In Section 3 we focus on the overshadowing aspect of factual defeasibility and the cancellation aspect of overridden defeasibility, and we show that in an adequate analysis of the Forrester 'Paradox' both these aspects have to be combined. In Section 4 we discuss further research.

2 Overridden versus Factual Defeasibility

In this section we give a general analysis of defeasible deontic logic by analyzing some intuitive inference patterns.[2] We show the fundamental difference between factual and overridden defeasibility.

2.1 Contrary-To-Duty obligations

Deontic logic is plagued by many 'paradoxes', intuitively consistent sentences which are formally inconsistent, or derive counterintuitive sentences. The most notorious 'paradoxes' are caused by so-called Contrary-To-Duty (CTD) obligations, obligations that refer to subideal situations. For example, Lewis [15] describes the CTD obligation that you ought to 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.'* [15]

In the early seventies, several dyadic modal systems were introduced to formalize CTD obligations, see [15] for an overview. Unfortunately, several technical problems related to CTD reasoning persisted in the dyadic logics [29]. A dyadic obligation $\mathcal{O}(\alpha \mid \beta)$ can be read as 'if β (the antecedent) is the case then α (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 we 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 by $\beta \wedge \neg\alpha$, the best thing to do is γ. Recently it was observed that the violation can be formalized in non-monotonic logics [11], theories of diagnosis [26, 28] and decision theories [3]. We say that dyadic obligations satisfy the Kantian principle 'ought implies can' when ought refers to 'the best of those that remain'. This will be explained in more detail in Section 2.3.

[2] See [27] for a semantic analysis of the two types. In the multi preference semantics, there are two distinct preference 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 overridden.

Since the late seventies, several temporal deontic logics and deontic action logics were introduced [1, 16, 17], which formalize satisfactorily 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' [16]. 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 are 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 can be overridden by other obligations. 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 you should eat with your fingers [11]. We say that an obligation is cancelled by exceptional circumstances when it is overridden. For example, the obligation not to eat with your fingers is cancelled by the exceptional circumstances that you 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' by analyzing the problematic CTD obligation that occurs in it as a type of overridden defeasibility (see e.g. [18, 25]). 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 argue in this paper that this 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'. First we present the 'paradox' in a normal dyadic deontic logic, to show its paradoxical character. Subsequently, we analyze the 'paradox' in a defeasible deontic logic in which there is only overridden defeasibility, and discuss the shortcomings of this approach. Finally, we give an analysis of the Chisholm 'Paradox' in terms of factual defeasibility. 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 paper in terms of inference patterns are, in principle, applicable to any defeasible deontic logic.

Assume a deontic logic with a finite propositional base logic $\mathcal{L}$ and dyadic modal obligations $\mathcal{O}(\alpha \mid \beta)$, where β (the antecedent) and α (the consequent) are sentences of $\mathcal{L}$. Assume further the unrestricted strengthening of the antecedent rule SA.

$$\mathrm{SA} : \frac{\mathcal{O}(\alpha \mid \beta)}{\mathcal{O}(\alpha \mid \beta \wedge \gamma)} \tag{1}$$

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

$$\mathrm{DD} : \frac{\mathcal{O}(\alpha \mid \beta), \mathcal{O}(\beta \mid \gamma)}{\mathcal{O}(\alpha \mid \gamma)} \tag{2}$$

The notorious Chisholm 'Paradox' (alias the CTD 'paradox', alias the 'paradox' of deontic detachment) is as follows:[3]

Example 2.1 (Chisholm 'Paradox') [6] *Consider the premises $\mathcal{O}(a \mid \top)$, $\mathcal{O}(t \mid a)$ and $\mathcal{O}(\neg t \mid \neg a)$, 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 $\mathcal{O}(\neg t \mid \neg a)$ is a CTD obligation of the (primary) obligation $\mathcal{O}(a \mid \top)$, because its antecedent is inconsistent with the consequent of the latter.*

The intuitive obligation $\mathcal{O}(t \mid \top)$ 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. *Hence, if he does not tell them, the ideal situation is no longer reachable. However, from $\mathcal{O}(t \mid \top)$ the counterintuitive $\mathcal{O}(t \mid \neg a)$ can be derived by* SA. *This is counterintuitive, because there is no reason to tell them he is coming when the man does not go. Moreover, in many deontic logics $\mathcal{O}(\neg t \mid \neg p)$ and $\mathcal{O}(t \mid \neg p)$ are inconsistent.*

This counterintuitive obligation cannot 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 by the following restricted strengthening of the antecedent rule RSA_O. RSA_O contains the so-called non-overridden condition C_O which represents that $\mathcal{O}(\alpha \mid \beta)$ is not overridden by some $\mathcal{O}(\alpha' \mid \beta')$ for $\beta \wedge \gamma$. It is based on a simplified notion of specificity, because background knowledge is not taken into account and an obligation cannot be overridden by more than one obligation.

$$\mathrm{RSA}_O : \frac{\mathcal{O}(\alpha \mid \beta), C_O}{\mathcal{O}(\alpha \mid \beta \wedge \gamma)} \tag{3}$$

where condition C_O is defined as follows:

> C_O: there is no premise $\mathcal{O}(\alpha' \mid \beta')$ 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':

[3] The original 'paradox' was given in a monadic modal logic, here we give the obvious formalization in a dyadic logic. See [29] for a discussion of the Chisholm 'Paradox' in several conditional deontic logics.

Example 2.2 *The intuitive obligation $\mathcal{O}(t \mid \top)$ can still be derived by* DD *from the first two obligations. From $\mathcal{O}(t \mid \top)$ the counterintuitive $\mathcal{O}(t \mid \neg a)$ cannot be derived by* RSA_O*, because $\mathcal{O}(t \mid \top)$ is overridden for $\neg a$ by the CTD obligation $\mathcal{O}(\neg t \mid \neg a)$, 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 anymore when the first premise is $\mathcal{O}(a \mid i)$, where i can be read as the fact that the man is personally invited for the assistance. 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 $\mathcal{O}(a \mid \top)$ and $\mathcal{O}(t \mid a)$. Again the intuitive obligation $\mathcal{O}(t \mid \top)$ can be derived by* DD*. From this derived obligation the counterintuitive $\mathcal{O}(t \mid \neg a)$ can again be derived by* RSA_O*, because there is no CTD obligation which cancels the counterintuitive obligation.*

In [28] we dubbed the intuition that the obligation $\mathcal{O}(t \mid \top)$ is intuitive and the obligation $\mathcal{O}(t \mid \neg a)$ is counterintuitive as 'deontic detachment as a defeasible rule'. The obligation $\mathcal{O}(t \mid \top)$, 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.[4]

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

This is shown by the following variant of the example, where the conditional obligation is represented as an absolute obligation:[5]

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

[4] For examples of deontic logics *not* satisfying the CC rule, see Chellas' CKD [5] (a nonnormal modal deontic logic) and Hansson's Preference Deontic Logic (PDL) [10].

[5] It may be argued that the premise $\mathcal{O}(a \to t \mid \top)$ does not represent the conditional obligation correctly. However, Horty [11] showed (by adapting an example of Van Fraassen [31]) that exactly the same inferences are made from the premises $\mathcal{O}(a \vee s \mid \top)$ and $\mathcal{O}(\neg a \mid \top)$, where a can be read as the fact that you are in the army and s as the fact that you perform alternative service. When DD is not accepted and CC is, then this leads to semantic problems as Tomberlin [29] showed for Mott's solution of the Chisholm 'Paradox' [20].

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 [26, 28]. DIODE is a diagnostic model for deontic reasoning based on Reiter's theory of diagnosis [23]. 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 $\mathcal{O}(\alpha \mid \beta)_L$ 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 $\mathcal{O}(\alpha \mid \beta)$. 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.

$$\mathrm{RSA}_V : \frac{\mathcal{O}(\alpha \mid \beta)_L, C_V}{\mathcal{O}(\alpha \mid \beta \wedge \gamma)_L}, C_V : L \wedge \beta \wedge \gamma \text{ is consistent} \tag{5}$$

$$\mathrm{RSA}_{OV} : \frac{\mathcal{O}(\alpha \mid \beta)_L, C_O, C_V}{\mathcal{O}(\alpha \mid \beta \wedge \gamma)_L}, C_V : L \wedge \beta \wedge \gamma \text{ is consistent} \tag{6}$$

$$\mathrm{DD}_V : \frac{\mathcal{O}(\alpha \mid \beta)_{L_1}, \mathcal{O}(\beta \mid \gamma)_{L_2}, C_V}{\mathcal{O}(\alpha \mid \gamma)_{L_1 \wedge L_2}}, C_V : L_1 \wedge L_2 \wedge \gamma \text{ is consistent} \tag{7}$$

$$CC_V : \frac{\mathcal{O}(\alpha_1 \mid \beta)_{L_1}, \mathcal{O}(\alpha_1 \rightarrow \alpha_2 \mid \beta)_{L_2}, C_V}{\mathcal{O}(\alpha_2 \mid \beta)_{L_1 \wedge L_2}}, C_V : L_1 \wedge L_2 \wedge \beta \text{ is consistent} \tag{8}$$

It can easily be checked that for example RSA_V is better than RSA_O, 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.

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 Kantian principle 'ought implies can' states essentially that a derived obligation cannot be contradicted by the facts. This is the meaning of 'can' here. An analogy that illustrates that the Kantian Principle induces factual defeasibility is to compare the 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 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).

Example 2 shows 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 [31], which was formalized by Horty [11] in Reiter's default logic. They introduced it to represent moral dilemmas, deontic inconsistencies like $\mathcal{O}(p \mid \top)$ and $\mathcal{O}(\neg p \mid \top)$. See [11] for the motivation and [26] for the comparison with DIODE. For other proposals of factual defeasibility, see Hansson' dyadic deontic logic [9] (no strengthening of the antecedent), Boutilier's extension of Hansson's logic [3] and Prakken and Sergot's so-called Down axiom [22].

2.4 Facticity

As discussed in Section 2.1, in dyadic deontic logic ought refers to 'the best of those that remain'. Dyadic obligations $\mathcal{O}(\alpha \mid \beta)$ 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. [1], 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'.

$$\mathrm{F} : \frac{}{\mathcal{O}(\alpha \mid \alpha)} \tag{9}$$

The next example shows a possible use of F.

Example 2.5 *Consider the single premise $\mathcal{O}(a \to t \mid \top)$. The obligation $\mathcal{O}(a \to t \mid a)$ is derived from the premise by RSA_V and the obligation $\mathcal{O}(a \mid a)$ is derived by F. From these two obligations the intuitive obligation $\mathcal{O}(t \mid a)$ 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 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 $\mathcal{O}(\alpha)$. No further properties of the monadic operator are assumed.

The simplest definition of factual detachment is the following rule FD, alias deontic modes ponens.

$$\text{FD}: \frac{\mathcal{O}(\alpha \mid \beta), \beta}{\mathcal{O}(\alpha)} \tag{10}$$

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 [2]). $\mathcal{A}(\alpha)$ is true when α is logically equivalent with the conjunction of all factual premises that are given.

$$\text{EFD}: \frac{\mathcal{O}(\alpha \mid \beta), \mathcal{A}(\beta)}{\mathcal{O}(\alpha)} \tag{11}$$

Note that EFD yields a kind of overridden defeasibility with respect to factual detachment. If we have, for example, as premises $\mathcal{O}(\alpha \mid \beta)$ and $\mathcal{O}(\neg\alpha \mid \beta \wedge \gamma)$, then EFD derives the conclusion $\mathcal{O}(\alpha)$ 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 $\mathcal{O}(\neg\alpha)$. 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.

An interesting example of defeasible deontic logic, as discussed in [30] in Horty's logic, is the following extended version of the so-called Forrester 'Paradox'. In its original formulation: you should not kill, but if you kill you should do it gently [7].

Example 3.1 (Forrester 'Paradox) *Consider the premises $\mathcal{O}(\neg f \mid \top)$, $\mathcal{O}(w \mid f)$ and $\mathcal{O}(w \mid c)$ 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 $\mathcal{O}(w \mid f)$ is a CTD obligation of $\mathcal{O}(\neg f \mid \top)$ and $\mathcal{O}(w \mid c)$ 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 $\mathcal{O}(\neg f)$ 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 $\mathcal{O}(\neg f)$ should be derivable.*

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

The problem in this example is that both $\mathcal{O}(w \mid f)$ and $\mathcal{O}(w \mid c)$ are treated as more specific obligations that override the obligation $\mathcal{O}(\neg f \mid \top)$, i.e. both are treated as cases of overridden defeasibility. However, this is not correct for $\mathcal{O}(w \mid f)$. 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 $\mathcal{O}(\neg f \mid \top)$ is violated by f, then the obligation $\mathcal{O}(\neg f)$ should be derivable, but not when $\mathcal{O}(\neg f \mid \top)$ is overridden by $f \wedge c$. This means that violation or overriding of $\mathcal{O}(\neg f \mid \top)$ 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 $\mathcal{O}(\neg f \mid \top)$ by $f \wedge c$ means that $\mathcal{O}(\neg f)$ is cancelled and has no force anymore. Violation of $\mathcal{O}(\neg f \mid \top)$ by f means that $\mathcal{O}(\neg f)$ 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 [30] 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 $\mathcal{O}(\alpha' \mid \beta')$ such that $\beta \wedge \gamma$ logically implies β', β' logically implies β and not vice versa, α and α' are inconsistent and β and α' are consistent. [30]

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* $\mathcal{O}(\neg f \mid f)$ *from* $\mathcal{O}(\neg f \mid \top)$, *but it does not derive* $\mathcal{O}(\neg f \mid f \wedge c)$. *The counterintuitive obligation* $\mathcal{O}(\neg f)$ *still cannot be derived by* RSA$_O$ *and* EFD *from* $\mathcal{O}(\neg f \mid \top)$ *and* $\mathcal{F} = f \wedge c$. *The intuitive obligation* $\mathcal{O}(\neg f)$ *can be derived by* EFD *from* $\mathcal{O}(\neg f \mid f)$ *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 $\mathcal{O}(\neg\alpha)$ 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 [30], we had to further adapt the definition of C'_O for another notorious (and highly ambiguous) 'paradox', the so-called Reykjavic 'Paradox' [16]. 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 $\mathcal{O}(\neg f) \wedge \neg f$ 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: $\frac{\mathcal{O}(\alpha|\beta)}{\mathcal{O}(\alpha\vee\gamma|\beta)}$ is accepted, then $\mathcal{O}(f \mid f)$ can be derived from $\mathcal{O}(w \mid f)$. From $\mathcal{O}(f \mid f)$ and $\mathcal{F} = f$ the absolute obligation $\mathcal{O}(f)$ can be derived by EFD, although the fence is certainly not a fulfilled obligation. Moreover, if 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 Gärdenfors postulates [8]. 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.

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

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 adaption 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 Gärdenfors postulates follows that $Cn[\mathcal{F} - \{\alpha\}] = Cn[\mathcal{F}]$ when $F \wedge \neg\alpha$ is consistent.*

4 Further research

In this paper we only considered examples in which overridden defeasibility is always of the cancelling type. However, Ross gave in [24] also examples of so-called prima facie obligations that can be considered as overridden defeasibility of the overshadowing type. We will study how this can be analyzed in our framework. We will also study the relation between the R-test and the Ramsey test in conditional logic. The crucial difference between the R-test and the Ramsey test is that in the R-test the consequence is taken into account.

Acknowledgement Thanks to Patrick van der Laag for several discussions on the issues raised in this paper.

References

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

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

[3] C. Boutilier. Toward a logic for qualitative decision theory. In *Proceedings of the KR'94*, 1994.

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

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

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

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

[8] P. Gärdenfors. *Knowledge in Flux.* MIT Press, Cambridge, 1988.

[9] B. Hansson. An analysis of some deontic logics. In *Deontic Logic: Introductionary and Systematic Readings*, pages 121–147. D. Reidel Publishing Company, Dordrecht, Holland, 1971.

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

[11] J.F. Horty. Nonmonotonic techniques in the formalization of commonsense normative reasoning. In *Proceedings of the Workshop on Nonmonotonic Reasoning*, pages 74–84, Austin, Texas, 1993.

[12] A.J.I. Jones and M. Sergot. Deontic logic in the representation of law: Towards a methodology. *Artificial Intelligence and Law*, 1:45–64, 1992.

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

[14] K. Konolige. On the relation between default and autoepistemic logic. *Artificial Intelligence*, 35:343–382, 1988.

[15] D. Lewis. Semantic analysis for dyadic deontic logic. In *Logical Theory and Semantical Analysis*, pages 1–14. D. Reidel Publishing Company, Dordrecht, Holland, 1974.

[16] B. Loewer and M. Belzer. Dyadic deontic detachment. *Synthese*, 54:295–318, 1983.

[17] D. Makinson. Five faces of minimality. *Studia Logica*, 52:339–379, 1993.

[18] L.T. McCarty. Defeasible deontic reasoning. In *Fourth International Workshop on Nonmonotonic Reasoning*, Plymouth, 1992.

[19] J.-J. Meyer and R. Wieringa. *Deontic Logic in Computer Science: Normative system Specification. Revision of selected papers that were presented at the First Workshop on Deontic Logic in Computer Science (*Δ*eon'91)*. John Wiley & Sons, 1994.

[20] P.L. Mott. On Chisholm's paradox. *Journal of Philosophical Logic*, 2, 1973.

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

[22] H. Prakken and M.J. Sergot. Contrary-to-duty imperatives, defeasibility and violability. In *Proceedings of the Second Workshop on Deontic Logic in Computer Science (*Δ*eon'94)*, Oslo, 1994.

[23] R. Reiter. A theory of diagnosis from first principles. *Artificial Intelligence*, 32:57–95, 1987.

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

[25] Y.U. Ryu and R.M. Lee. Defeasible deontic reasonig: A logic programming model. In *Deontic Logic in Computer Science: Normative System Specification*. John Wiley & Sons, 1994.

[26] Y.-H. Tan and L.W.N. van der Torre. 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 (ECAI'94)*, Amsterdam, 1994.

[27] Y.-H. Tan and L.W.N. van der Torre. Multi preference semantics for a defeasible deontic logic. In *Proceedings of the JURIX'94*, Amsterdam, 1994.

[28] Y.-H. Tan and L.W.N. van der Torre. 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 (ICLP'94)*, 1994. An extended version of this paper will appear in the Journal of Artificial Intelligence and Law.

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

[30] 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.

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