Why Defeasible Deontic Logic needs a Multi Preference Semantics*

Yao-Hua Tan[1] and Leendert W.N. van der Torre[1,2]

[1] EURIDIS
Erasmus University Rotterdam
P.O. Box 1738, 3000 DR Rotterdam, The Netherlands
{ytan,ltorre}@euridis.fbk.eur.nl
[2] Tinbergen Institute and Department of Computer Science
Erasmus University Rotterdam
P.O. Box 1738, 3000 DR Rotterdam, The Netherlands

Abstract. There is a fundamental difference between a conditional obligation being violated by a fact, and a conditional obligation being overridden by another conditional obligation. In this paper we analyze this difference in the multi preference semantics of our defeasible deontic logic DEFDIODE. The semantics contains one preference relation for ideality, which can be used to formalize deontic paradoxes like the Chisholm and Forrester paradoxes, and another preference relation for normality, which can be used to formalize exceptions. The interference of the two preference orderings generates new questions about preferential semantics.

1 Introduction

In recent years deontic logics has become increasingly popular as a tool to model legal reasoning in expert systems [7, 10]. Deontic logic is a modal logic in which the modal operator O is used to express that something is obliged, see [2]. For example, if the proposition i stands for the fact that you insult someone, then $O(\neg i)$ means that you should not insult someone. The sentence $O(\neg i) \wedge i$ is consistent and expresses that the obligation not to insult someone is violated by the fact i that you insult someone. The most well-known deontic logic is so-called 'standard' deontic logic (SDL), a normal modal system of type KD according to the Chellas classification [2]. It satisfies, besides the propositional tautologies and the inference rules modus ponens $\frac{p, p \rightarrow q}{q}$ and necessitation $\frac{\vdash p}{\vdash O(p)}$, the axioms K: $O(p) \wedge O(p \rightarrow q) \rightarrow O(q)$ and D: $\neg(O(p) \wedge O(\neg p))$.

It is well-known that defeasible reasoning is a very important aspect of legal reasoning (see [6, 11]). In this paper we argue that in case of defeasible deontic logic, one needs two preference orderings in the semantics of such a logic. In a defeasible deontic logic, two kinds of defeasibility can be distinguished, so-called *factual defeasibility* and *overridden defeasibility*, see [16] for an analysis in terms of inference patterns. Factual defeasibility can be used to represent that an obligation is *overshadowed by a violating fact* and overridden defeasibility can be used to represent that an obligation is *cancelled*

* 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.

by another obligation. The semantics contains a preferential ordering to model the deontic aspects and another ordering to model the normality aspects, which are used to model exceptional circumstances. Interestingly, it appears that these preference orderings interfere in a complicated way, thus generating new and interesting questions about preferential semantics.[3]

In this paper we use DEFDIODE to analyze the interference between the two preferential orderings. In DEFDIODE, the preferential orderings are very simple (they are subset orderings, defined on abnormality predicates). These orderings do not model all the subtleties of the individual orderings (see [8] for a detailed description of the subtle distinctions), but they are sufficient to analyze the interference problems. We will illustrate the interference of the two orderings by a simple example. In this example there is a situation which can be considered as a kind of overshadowing and as a kind of cancelling. The semantics clearly show that, in this example, overshadowing is preferred over cancelling.

2 DIODE

Three decades ago, Chisholm described in [3] a notorious paradox of deontic logic, the so-called Chisholm Paradox, which has led to the development of new deontic logics that were meant to solve the Chisholm paradox (see [2]). Two decades later, Forrester described in [4] his version of the paradox, the so-called Forrester paradox, which could not be solved by any of these new deontic logics. A set of sentences is called a paradox of a deontic logic when the (most obvious) formalization in the deontic logic is inconsistent. In [14] we introduced DIODE; a DIagnostic framework for DEontic reasoning. In these papers we showed how one can solve certain aspects of the Chisholm and Forrester paradoxes in DIODE. From a semantic point of view one could say that in DIODE the deontic modal operator is replaced by a preferential semantics as this was initially developed for conditional and non-monotonic logics. In this section, the details of this semantics will be explained.

The basic idea of DIODE is to translate a conditional obligation 'if α is the case, then it ought to be that β is the case' into the propositional formula $\alpha \wedge \neg V_i \rightarrow \beta$.[4] V_i is a propositional constant denoting whether the obligation is violated; the conditional obligation can be read as 'if α is the case and the obligation is not violated then β is the case'. For example, the obligation not to insult someone is formalized in DIODE by $\neg V_1 \rightarrow \neg i$ where i stands for insulting someone.

Let L be a propositional logic. L_V is L extended with (a finite number of) violation constants V_i. We write $\models$ for entailment in L_V. A deontic theory T of L_V consists of a set of factual sentences of L (denoted by the set F in Figure 1), a set of background knowledge sentences of L and a set of absolute and conditional obligations (deontic rules) of L_V, typically given by $\neg V_i \rightarrow \beta$ or $\alpha \wedge \neg V_i \rightarrow \beta$ with $\alpha, \beta \in L$. Every distinct

[3] Boutilier [1] also argues for a second normality preference ordering. This ordering is used in his logic to model factual defaults, not the defeasibility of conditional obligations. Therefore, his two orderings give rise to completely different problems than our two orderings.

[4] Usually such a conditional obligation is translated into either $\alpha \rightarrow O(\beta)$ or $O(\beta \mid \alpha)$, where O is a monadic or dyadic modal operator (see [2]).

deontic rule has a distinct violation constant V_i. For a detailed description of the syntax and proof theory of DIODE and related work, see [14].

DIODE contains a preferential semantics that defines a preference ordering on models (see e.g. [13]) using the V_i constants. This preference ordering orders all ideal and sub-ideal states. The motivation of the distinction between ideal and sub-ideal states is that not all obligations refer to an ideal situation, but also often to sub-ideal situations. These obligations are so-called *Contrary-To-Duty* (CTD) obligations. For example, if you are obliged not to insult someone $O(\neg i)$, then the conditional obligation that if you insult someone, you should apologize $i \rightarrow O(a)$ is a CTD obligation. A CTD obligation describes the *optimal* subideal state. They are well-known from the notorious Chisholm and Forrester paradoxes. In [12] several other examples of sub-ideal states and CTD obligations are given.

Definition 1. Let T be a theory of L_V and M_1 and M_2 two models of T. M_1 is preferred over M_2, written $M_1 \sqsubseteq M_2$, iff $M_1 \models V_i$ then $M_2 \models V_i$ for all i. We write $M_1 \sqsubset M_2$ (M_1 is strictly preferred over M_2) iff $M_1 \sqsubseteq M_2$ and not $M_2 \sqsubseteq M_1$.

Given this partial pre-ordering, we use the following basic definitions:

Definition 2. An interpretation *M preferentially satisfies A* (written $M \models_{\sqsubset} A$) iff $M \models A$ and there is no other interpretation M' such that $M' \sqsubset M$ and $M' \models A$. In this case we say that M is a *preferred model* of A. A preferentially entails B (written $A \models_{\sqsubset} B$) iff for any M, if $M \models_{\sqsubset} A$ then $M \models B$.

The notion of preferential entailment can be used to identify minimal (with respect to set inclusion) violation sets.

Definition 3. Let T be a theory of L_V and M a preferred model of T, i.e. $M \models_{\sqsubset} T$. The set $\{V_i \mid M \models V_i\}$ is a *preferred violation set* of T.

A deontic theory can have more than one preferred violation set. In the deontic context given by a DIODE theory T, the sentences of L which are true in all preferred models are called contextually obliged.

Definition 4. Let T be a theory of L_V. T provides a contextual obligation for α iff $T \models_{\sqsubset} \alpha$ and $\alpha \in L$.

Semantically, the deontic rules (together with the background knowledge sentences) define a preference ordering on the models which orders all ideal and sub-ideal states. The facts (a subset of T, represented by F) zoom in on this partial ordering by selecting the (sub)ideal states where the facts are true. This zooming in will be demonstrated by an instance of the Forrester paradox [4]: you should not kill, but if you kill you should do it gently.

Example 1. **(Forrester paradox)** Consider the following sentences of a theory T:

1. $\neg V_1 \rightarrow \neg i$: You should not insult someone;
2. $i \wedge \neg V_2 \rightarrow p$: If you insult someone you should do it in private;
3. $p \rightarrow i$: Insulting someone in private logically implies that you insult him;

4. i: You insult someone.

The preference ordering of the deontic rules (together with the background rule $p \rightarrow i$) of the Forrester 'Paradox' is given in Figure 1. This figure must be read as follows. The models are ordered by the subset relation on the violation constants V_i. The circles denote equivalence classes of this ordering (all models in a circle satisfy the same violation constants) and the arrows indicate which models are strictly preferred. The set of obligations which are violated in this equivalence class are written in the circle. Moreover, the circles also contain a set of propositions. These propositions are true in all the models of the equivalence class which are preferred for some set of factual sentences. Hence, only models which are relevant, i.e. which are minimal for the set of formulas of L they make true, are shown in the figure. For example, the models that satisfy $\neg i$ and V_1 are never preferred and are therefore not represented; all *relevant* models that satisfy V_1, also satisfy i. Equivalence classes without such relevant models, e.g. the equivalence class of V_2, are not shown. When the facts contain all the propositions that are written in some circle, then the preferred violation set is the set of obligations in this circle. Hence, the circles contain the minimal set of violated obligations that are consistent with the propositions in the circle. For example, in the i, p-circle, V_1 has to be true due to the obligation $\neg V_1 \rightarrow \neg i$.

In the ideal situation, given by the left circle, you do not insult someone. If you insult someone, i.e. for $F = \{i\}$, you only consider equivalence classes that contain i. Hence, the relevant models are restricted to the sub-ideal models containing V_1 and the sub-sub-ideal models containing V_1 and V_2. In Figure 1 this zooming in on the ordering is depicted by a dashed box. The optimal sub-ideal state, represented by the leftmost circle within the dashed box, represents the fact that you insult him in private. This means that $\{V_1\}$ is the only preferred violation set and T provides a contextual obligation for p. The worst state reflects, in a sense, two violations: the first one is the offense of insulting someone and the second one is doing it in public.

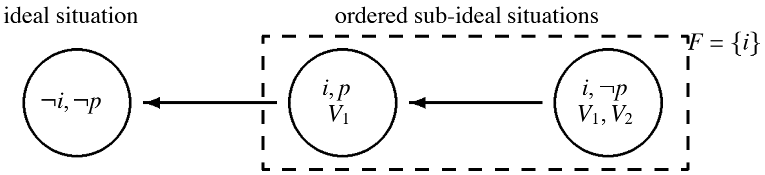

Fig. 1. Preference relation of the Forrester paradox

The previous example showed the two-phase mechanism of DIODE. The first phase consists of building a preference ordering on all models, given by the deontic rules and background knowledge (like $p \rightarrow i$ in the Forrester paradox). The second phase zooms in on this ordering by selecting the models where the facts are true. Two similar phases exist in the defeasible variant of DIODE which will be developed in the next section. However, in the first phase *two* preference orderings will be constructed; not only one for ideality but also one for normality.

3 DefDiode: Diode with exceptions

There is a fundamental difference between a conditional obligation being violated by a fact, and a conditional obligation being overridden by another conditional obligation. See [16] for a discussion of this difference in terms of inference patterns. For example, in a legal setting, when an obligation is violated you have to pay a fine for it, but when it is cancelled by another obligation, you cannot be fined for it. Horty [5] gives his well-known example of being served asparagus. You should not eat with your fingers. But if you are served asparagus, then you should eat with your fingers. In the special case where you are served asparagus, the first obligation is less specific and hence cancelled by the second one. Various authors, e.g. [5, 8, 9], have investigated the formalization of *defeasible* conditional obligations (traditionally called *prima facie* obligations), deontic rules which are subject to exceptions. Explicit exceptions can be introduced in Diode by formalizing a defeasible conditional obligation 'if α is the case then *usually* it ought to be that β is the case' by $\alpha \wedge \neg V_i \wedge \neg Ex_i \rightarrow \beta$, where Ex_i is a propositional constant denoting whether the defeasible conditional obligation is defeated (by some exceptional circumstances). For example, a defeasible conditional obligation that usually you should not insult someone can be formalized by $\neg V_1 \wedge \neg Ex_1 \rightarrow \neg i$. The Ex_i abnormalities are used to control the preferences between two conflicting defeasible conditional obligations. Hence, the rules that determine when an abnormality Ex_i holds are quite different from the rules that determine when a violation V_i holds. From a semantic point of view there are two independent preference relations on the models; one for minimizing the V_i constants and one for minimizing the Ex_i constants.

DefDiode is an extension of Diode in the sense that in DefDiode obligations might contain an exception constant Ex_i. Given a set of defeasible conditional obligations in DefDiode, the question remains how to determine when there are exceptional circumstances, i.e. when an exception constant is true. In this paper, we make the assumption that all exceptions are given explicitly and that in case of a conflict, violations are preferred over exceptions. Obviously, there is no a priori reason to prefer violations over exceptions; it follows from the assumption that *all* exceptions are given explicitly. For example, assume there is a second deontic rule that states that you should insult someone when he does harm the public interest, formalized by $h \wedge \neg V_2 \rightarrow i$ where h stands for someone harming public interest. An example of this obligation is that every journalist should expose Nixon in the Watergate affair. In that case, a so-called defeater rule must be added that states that a situation of public interest is an exception to the rule not to insult someone, $h \rightarrow Ex_1$. Semantically, the normality ordering is a subset ordering on exception constants Ex_i just like the ideality ordering on violation constants V_i (though the rules which determine when an abnormality Ex_i holds are quite different from the rules that determine when a violation V_i holds!).

The following definition of overridden is a formalization of the notion of specificity. This definition can be used in our framework to identify exceptional circumstances. The definition is borrowed from non-monotonic logics. However, as we will see later, this definition has to be adapted for defeasible *deontic* logic since it is too strong. In spirit it is similar to Horty's definition of overridden [5].

Definition 5. Let $F_b \subseteq T$ be the set of background knowledge sentences of T. A defeasible conditional obligation $\alpha_1 \wedge \neg V_1 \wedge \neg Ex_1 \rightarrow \beta_1 \in T$ is *overridden for* α_2 by $\alpha_2 \wedge \neg V_2 \wedge \neg Ex_2 \rightarrow \beta_2 \in T$ (or $\alpha_2 \wedge \neg V_2 \rightarrow \beta_2 \in T$) iff:

1. $F_b \wedge \beta_1 \wedge \beta_2$ is inconsistent, and
2. $F_b \wedge \alpha_2 \models \alpha_1$ and $F_b \wedge \alpha_1 \not\models \alpha_2$.

In all cases where a defeasible conditional obligation $\alpha_1 \wedge \neg V_1 \wedge \neg Ex_1 \rightarrow \beta_1$ is *overridden for* α_2 by $\alpha_2 \wedge \neg V_2 \wedge \neg Ex_2 \rightarrow \beta_2$ (or $\alpha_2 \wedge \neg V_2 \rightarrow \beta_2$), the explicit defeater rule $\alpha_2 \rightarrow Ex_1$ is added to T.[5] The next example is an instance of Horty's asparagus example: you should not eat with your fingers, but if you eat asparagus you should eat with your fingers [5].

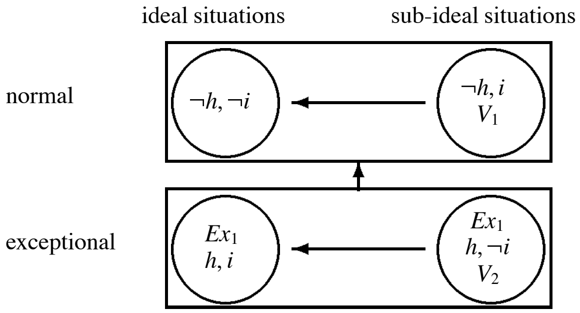

Fig. 2. Preference relation of Public Interest

Example 2. **(Public Interest)** Consider the following sentences of a theory T:

1. $\neg V_1 \wedge \neg Ex_1 \rightarrow \neg i$: Usually, you should not insult someone.
2. $h \wedge \neg V_2 \rightarrow i$: When someone does harm the public interest, you should insult him.

The second obligation overrides the first one for h so the clause $h \rightarrow Ex_1$ should be added. The idea of the preference ordering on normality is that the models with exceptional circumstances (public interest) are semantically separated from the normal situation. The intended preferential semantics are given in Figure 2. The boxes denote equivalence classes in the normality ordering and the 'vertical' arrow the normality preference ordering. The circles denote equivalence classes in the deontic ordering *within an equivalence class of the normality ordering*, and the 'horizontal' arrows the deontic preference ordering. The upper box represents the 'normal' models, which is determined by the fact that h is false, i.e. there is no situation of public interest. Deontically, the

[5] Notice that Definition 5 is syntax-dependent, since the logically equivalent $\alpha \wedge \neg V_i \rightarrow \beta$ and $\neg V_i \rightarrow (\alpha \rightarrow \beta)$ are treated differently. This is the consequence of the strong notion of implication used in DEFDIODE, which is the classical material implication. Notice that we can still use classical models, because the explicit defeater rules are added before the preferential orderings are built. There are several ways to solve this syntax dependence. Horty [5] solves this, for instance, by representing deontic rules as Reiter default rules.

$\neg h$-models are ordered according to the obligation that usually, you should not insult someone. The lower box contains the models where h is true and which are therefore exceptional, which is also denoted by Ex_1. These models are deontically ordered by the obligation that in this situation, you should insult him. Because of the exceptional circumstances, the models are not subject to the obligation that usually, you should not insult someone.

Without the explicit defeater $h \rightarrow Ex_1$, there is a conflict when h is true, because the first obligation implies that you should not insult, and the second obligation implies that you should. In the semantics, this conflict would be represented by the fact that h, i models are incomparable with $h, \neg i$ models. The introduction of the explicit defeater, and hence the exceptionality level in the multi preference semantics, results in two normality classes. Within these classes, all models are comparable. Hence, there is no conflict anymore: the explicit defeater has resolved the conflict.

Now we reconsider the Forrester paradox in a defeasible deontic setting. As we showed in [15], a strong definition of overridden like Horty's definition [5] or our Definition 5 above will give unintuitive results.

Example 3. **(Forrester paradox)** Reconsider the sentences of Example 1 in a defeasible setting:

1. $\neg V_1 \wedge \neg Ex_1 \rightarrow \neg i$: Usually, you should not insult someone;
2. $i \wedge \neg V_2 \rightarrow p$: If you insult someone, you should do it in private;
3. $p \rightarrow i$: Insulting someone in private implies that you insult him.

Given Definition 5 of overridden, the first sentence is overridden by the second one for i; i.e. we should add the formula $i \rightarrow Ex_1$. However, the addition of the formula is highly counterintuitive since it implies that the first obligation can never be violated. In our semantic analysis, we can see that the introduction of an explicit defeater is counterintuitive, because there is no conflict to be resolved. In the picture without explicit defeaters, i.e. Figure 1, there are no incomparable models! The intuitive reading of the example is that the second obligation is a CTD obligation of the first one, and hence the first and more general obligation should hold and not be overridden.

The problem here is that the CTD obligation is considered as an exception because the conclusions of the deontic rules are inconsistent and the condition of the second rule is more specific. For a defeasible deontic logic, this condition is too strong.

The previous example showed the interesting situation where a definition borrowed from non-monotonic logic is too strong for a defeasible deontic logic. In [15] we introduced therefore the following weaker notion of overridden which excludes this possibility by introducing a test (the third condition) whether the second sentence is a CTD obligation of the first sentence. The additional condition for the definition is very natural. A CTD obligation is an obligation where the reference situation is in contradiction with duty, namely $\alpha_1 \wedge \beta_1$. The condition is that $F_b \not\models \alpha_2 \rightarrow \alpha_1 \wedge \beta_1$. It can easily be seen that this reduces to condition 3.

Definition 6. Let F_b be the set of background knowledge sentences of T. A defeasible conditional obligation $\alpha_1 \wedge \neg V_1 \wedge \neg Ex_1 \rightarrow \beta_1 \in T$ is *overridden for* α_2 by $\alpha_2 \wedge \neg V_2 \wedge \neg Ex_2 \rightarrow \beta_2 \in T$ (or $\alpha_2 \wedge \neg V_2 \rightarrow \beta_2 \in T$) iff:

1. $F_b \wedge \beta_1 \wedge \beta_2$ is inconsistent, and
2. $F_b \wedge \alpha_2 \models \alpha_1$ and $F_b \wedge \alpha_1 \not\models \alpha_2$, and
3. $F_b \wedge \beta_1 \wedge \alpha_2$ is consistent.

For the Public Interest example the conditions are still satisfied. In the Forrester paradox, the defeasible deontic rule not to insult someone is no longer overridden for *i* according to Definition 6, because the last condition is not satisfied.

Acknowledgements. Thanks to Patrick van der Laag, Henry Prakken, Marek Sergot and Tina Smith for several discussions on the issues raised in this paper.

References

1. C. Boutilier. Toward a logic for qualitative decision theory. In *Proceedings of KR'94*, 1994.
2. B.F. Chellas. *Modal Logic: An Introduction*. Cambridge University Press, 1980.
3. R.M. Chisholm. Contrary-to-duty imperatives and deontic logic. *Analysis*, 24:33–36, 1963.
4. J.W. Forrester. Gentle murder, or the adverbial Samaritan. *Journal of Philosophy*, 81:193–197, 1984.
5. J.F. Horty. Deontic logic founded in nonmonotonic logic. *Annals of Mathematics and Artificial Intelligence*, 9:69–91, 1993.
6. 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.
7. A.J.I. Jones and M. Sergot. *Proceedings of the Second Workshop on Deontic Logic in Computer Science (Deon'94)*. Oslo, 1994.
8. D. Makinson. Five faces of minimality. *Studia Logica*, 52:339–379, 1993.
9. L.T. McCarty. Defeasible deontic reasoning. In *Fourth International Workshop on Nonmonotonic Reasoning*, Plymouth, 1992.
10. J.-J. Meyer and R. Wieringa. *Deontic Logic in Computer Science: Proceedings of the First Workshop on Deontic Logic in Computer Science (Deon'91)*. John Wiley & Sons, 1993.
11. H. Prakken. *Logical Tools for Modelling Legal Argument, Ph-D thesis*. Amsterdam, 1993.
12. 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 (Deon'94)*, Oslo, 1994.
13. Y. Shoham. *Reasoning About Change*. MIT Press, 1988.
14. 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.
15. 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.
16. L.W.N. van der Torre and Y.-H. Tan. Cancelling and overshadowing: two types of defeasibility in defeasible deontic logic. Technical Report WP 95.02.01, EURIDIS, 1995. To appear in: *Proceedings of the IJCAI-95*.

This article was processed using the LaTeX macro package with LLNCS style