Legal knowledge based systems
JURIX 94
The Relation with Legal Theory

The Foundation for Legal Knowledge Systems
Editors:
H. Prakken
A.J. Muntjewerff
A. Soeteman

Y-H Tan and L.W.N. Van der Torre, *Multi Preference Semantics for a Defeasible Deontic Logic*, in: A. Soeteman (eds.), Legal knowledge based systems JURIX 94: The Foundation for Legal Knowledge Systems, Lelystad: Koninklijke Vermande, 1994, pp. 115-126, ISBN 90 5458 190 5.

More information about the JURIX foundation and its activities can be obtained by contacting the JURIX secretariat:

Mr. C.N.J. de Vey Mestdagh
University of Groningen, Faculty of Law
Oude Kijk in 't Jatstraat 26
P.O. Box 716
9700 AS Groningen
Tel: +31 50 3635790/5433
Fax: +31 50 3635603
Email: sesam@rechten.rug.nl

 http://jurix.bsk.utwente.nl/

Multi Preference Semantics for a Defeasible Deontic Logic

Yao-Hua Tan & Leendert W.N. van der Torre
Erasmus University Research Institute for Decision and Information Systems (EURIDIS)
Erasmus University, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands
E-Mail: {tan,torre}@cs.few.eur.nl

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 propose a multi preference semantics for a defeasible deontic logic that is based on this fundamental difference. The semantics contains one preferential relation for ideality, which can be used to formalize deontic 'paradoxes' like the Chisholm and Forrester 'Paradoxes', and another preferential relation for normality, which can be used to formalize exceptions.
Keywords: deontic logic, non-monotonic logic, knowledge representation

1 Introduction

In the field of AI-and-law it is widely acknowledged that formal representations of legal knowledge are useful. For example, such formal representations may reveal ambiguities in the legal knowledge or even plain contradictions. Deontic logic is the logic about obligations and permissions. Hence, it might be expected that deontic logic is particularly appropriate for representing legal knowledge. In recent years many researchers have pointed out how a wide variety of legal arguments can be modeled in deontic logic, see for example [Jones & Sergot, 1992]. However, it was also discovered that deontic logic itself suffers from certain 'paradoxes' of which the so-called Chisholm 'Paradox' and Forrester 'Paradox' are the most notorious ones. The problem with these 'paradoxes' is that they cannot be represented in deontic logic in such a way that the logic yields the intuitive conclusions. The following example, taken from Dutch criminal law, shows that, for example, Forrester type of paradoxes can also occur in legal arguments.

1. Libel: He, who insults someone, will be punished with a jail sentence of at most six months.

2. Libel (writing): If this occurs by the means of a piece of writing shown in public, the offender will be punished with a jail sentence of at most one year.

3. Exception: Libel does not exist when the offender acted out of general interest and could reasonably assume that the insult was true.

The first line implies that a person should not insult someone, which follows from the general guideline that a person should not do something he could be punished for. The problem we discuss in this paper is what obligations are implied by the second line. It seems that the second line implies the obligation not to insult someone in public. This obligation, however, is just a special instance of the general obligation not to insult someone, and this obligation is therefore already implied by the first line. We argue that this second line refers to the sub-ideal situation where a person does insult someone; in that case, insulting him in private is preferred over insulting him in public. Given this reading, the two lines together describe three different states, representing different degrees of the offence of libel. In the ideal state there is no libel, in the sub-ideal state there is libel in private and in the worst state there is libel in public. These sub-ideal states are characteristic for so-called Contrary-To-Duty (CTD) obligations. An example of a CTD obligation is given by the second line in the following formal representation of the legal knowledge implicit in the libel article:

1. $O(\neg i)$: You should not insult someone.
2. $i \rightarrow O(p)$: If you insult someone, you should do it in private.
3. $p \rightarrow i$: Insulting someone in private logically implies that you insult him.

These three logical sentences are an instance of the so-called Forrester 'Paradox' of deontic logic, as will be explained later. An example of the Chisholm 'Paradox' in legal reasoning has been discovered by Jones and Sergot [Jones & Sergot, 1992]. We think that the kind of reasoning about sub-ideal states is typical for legal reasoning and that therefore the Forrester and Chisholm 'Paradoxes' are typical problems of representing legal knowledge. Since it is well-known that these 'paradoxes' cannot be represented satisfactorily in standard deontic logics, we introduce in this paper a new defeasible deontic logic in which these 'paradoxes' can be represented in a much better way.

In our formalization of the example above, the law is seen as an instrument to control behavior. A rational person is supposed to minimize the total sum of his penalties and therefore (ideally) not to commit an offence; but if he commits an offence he should do so in a way that is not too bad. For example, a thief should not use violence, and a kidnapper should not kill his victim. Bench-Capon also argues that in many areas of the law, regulations serve as what he calls 'a tariff that enables rational behavior' [Bench-Capon, 1994]. He considers deontic logic as 'essentially concerned with the representation and analysis of reasoning about a fundamental distinction that arises naturally in law: the distinction between what *ideally* is the case on the one hand, and what *actually* is the case on the other' [Jones & Sergot, 1992].[1] Such a deontic logic, Bench-Capon argues, can only be used for norms prohibiting an act that is in itself sub-ideal (like the obligation not to kill), but not for 'norms which are rather means to encourage behavior which has a tendency towards realising the ideal'. We fully agree with this criticism and we will argue that the defeasible deontic logic presented in this

[1] This distinction reflects that facts about what actually is the case neither entail nor are entailed by facts about what ought to be the case. In the syntax of deontic logics this distinction is usually represented by a modal operator. In the associated (Kripke) semantics a distinction is made between actual and ideal worlds.

paper – characterized by the distinction between ideal and sub-ideal states – is more adequate to represent such a 'tendency towards realising the ideal' than the standard deontic logics. The seriousness of the penalties for offences reflects something of an ordering on deontic states; the worse the offence, the higher the penalty. For example, in the libel example it was the case that libel in public is worse than libel in private, hence for the first offence the maximum penalty is twelve months. Our defeasible deontic logic formalizes these deontic ordering aspects of legal knowledge, while still retaining the basic intuitions of standard deontic logics.

The idea of making a distinction between ideal and sub-ideal states to represent CTD obligations has already been proposed in the literature, for example [Jones & Porn, 1985; Prakken & Sergot, 1994]. Some of the proposed logics also have a preferential ordering on ideal and sub-ideal states. The main problem considered in this article, however, is what happens with the deontic preferential ordering in a *defeasible* deontic logic. We argue in this paper that deontic logics that can represent defeasible reasoning structures need two distinct orderings in the semantics: one for ideality and one for normality. The ordering of normality can be used to formalize exceptions, for example given by the specificity principle (lex specialis). We show that the two preferential orderings may interfere, which leads to new, yet unsolved problems.

2 DIODE

In [Tan & van der Torre, 1994b] and [Tan & van der Torre, 1994a] DIODE – a diagnostic framework for deontic reasoning – is introduced. The logic is motivated by an observed resemblance between deontic reasoning and the formal analysis of diagnostic reasoning introduced in [Reiter, 1987]. We think that this motivation is preferable over a more general resemblance between deontic and non-monotonic reasoning [Horty, 1993], since our approach motivates which non-monotonic techniques should be used for the formalization of deontic reasoning. This avoids the confusion between CTD structures and kinds of defeasible reasoning structures discussed in [Prakken & Sergot, 1994].

Standard deontic logic makes a binary distinction between ideal and nonideal states. We propose a more refined notion of ideality, discriminating between ideal and varying sub-ideal states. From a technical point of view such a more refined relation of ideality on worlds has a striking resemblance with the graded 'normality relations' of several logics for non-monotonic reasoning [Shoham, 1988; Kraus *at al*, 1990; Boutilier, 1994]: in these logics the possible worlds are marked as more or less normal, and in deriving conclusions it is assumed that the actual world is as normal as possible, given the facts that are known. In this section we adapt existing techniques for defeasible reasoning in order to model reasoning with CTD obligations. In addition to the modeling of reasoning with CTD obligations, we will study in the next section what happens when the two things are put together: i.e. we study the combination of CTD reasoning and reasoning with exceptions.

The basic idea of DIODE is to formalize 'if α is the case then it ought to be that β is the case' by $\alpha \wedge \neg V_i \rightarrow \beta$. 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.

Definition 1 *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 typically consists of a set of factual sentences of L (denoted by the set F in Fig. 1 and 2), 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 deontic rule has a distinct violation constant V_i.

DIODE contains a preferential semantics that defines a preferential ordering on models using the V_i constants. This preferential 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 the Contrary-To-Duty (CTD) obligations which were already mentioned in the Introduction. They are well-known from the notorious Chisholm and Forrester 'Paradoxes'. In [Prakken & Sergot, 1994] several other examples of sub-ideal states and CTD obligations are given.

Definition 2 *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) for $M_1 \sqsubseteq M_2$ and not $M_2 \sqsubseteq M_1$.*

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

Definition 3 *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 4 *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.*

In the deontic context given by a DIODE theory T, the sentences of L which are true in the preferred models are contextually obliged.

Definition 5 *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 define a preferential ordering on the models which orders all ideal and sub-ideal states. The facts (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 some examples from 'standard' deontic logic (SDL, a normal modal system of type KD according to the Chellas classification [Chellas, 1980]) which are translated to DIODE.

Example 1 (Forrester 'Paradox') *Consider the following sentences of a* SDL *theory T, which have a similar logical structure as the Forrester 'Paradox' [Forrester, 1984]:*

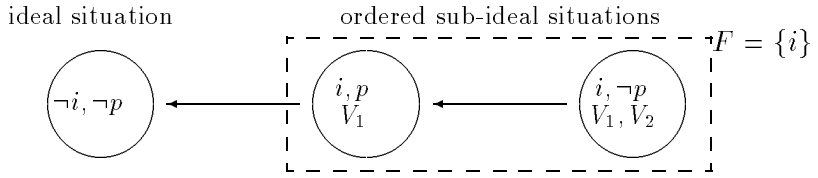

Figure 1: Preference relation of the Forrester 'Paradox'

1. $O(\neg i)$: *You should not insult someone;*
2. $i \rightarrow O(p)$: *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.*

From this theory follows $T \models_{SDL} O(\neg i)$ *and* $T \models_{SDL} O(i)$. *The last obligation is derived with the derived inference rule* $\frac{\vdash p \rightarrow q}{\vdash O(p) \rightarrow O(q)}$ *(which derives* $O(p) \rightarrow O(i)$ *from* $p \rightarrow i$*) from* $O(p)$. *This derivation is possible since* $p \rightarrow i$ *has the status of a theorem; for the details, see Forrester's paper. The main problem of this 'paradox' is that* $O(\neg i)$ *and* $O(i)$ *are inconsistent in* SDL.

In DIODE, *the sentences are translated to (see [Tan & van der Torre, 1994b]) the following sentences of a* DIODE *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 preferential ordering of the deontic rules (together with the background rule $p \rightarrow i$*) of the Forrester 'Paradox' is given in Fig. 1. The circles denote equivalence classes of models; only models which are preferred for some factual situation are given. For example, models satisfying* $\neg i$ *and* V_1 *are never preferred and are therefore not depicted.*

In the ideal situation, given by the left circle, you do not insult someone. If you insult someone, i.e. for $F = \{i\}$, *the relevant models are restricted to the sub-ideal models containing* V_1 *and the sub-ideal models containing* V_1 *and* V_2. *In Fig. 1 this zooming in on the ordering is depicted by a dashed box. The optimal sub-ideal state, represented by the left most 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. *There is no contextual obligation for* $\neg i$ *(which caused the inconsistency in the* SDL *representation) since* $\neg i$ *refers to the ideal state and is not considered by the process of zooming in. The worst state reflects, in a sense, two violations: the first one is the offence of libel and the second one is doing it in public.*

A more complicated 'paradox' was given by Chisholm [Chisholm, 1963]. An example of the 'paradox' in The United Nations Convention on Contracts for the International Sale of Goods has been given by Jones and Sergot [Jones & Sergot, 1992], involving

permissions. Here we give an example adapted from [Jones, 1989], that represents the deontic reasoning of a judge who considers legal rules as obligations of his behavior to sentence offenders.

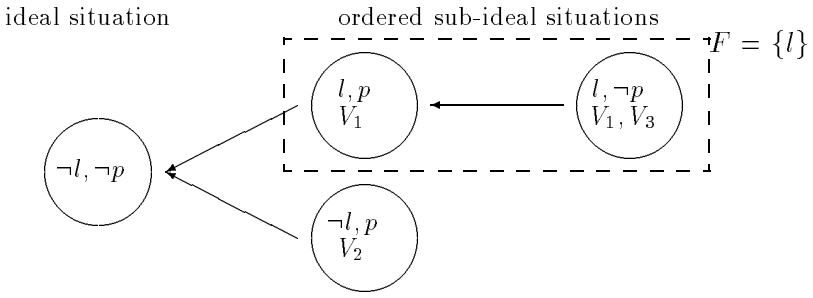

Figure 2: Preference relation of the Chisholm 'Paradox'

Example 2 (Chisholm 'Paradox') *In* SDL*, the Chisholm 'Paradox' is given by the following sentences of a* SDL *theory* T*:*

1. $O(\neg l)$*: A person should not commit the offence of libel.*
2. $O(\neg l \rightarrow \neg p)$*: It should be that if a person does not commit the offence of libel, he is not punished for libel.*
3. $l \rightarrow O(p)$*: If a person does commit the offence, he should be punished for it.*
4. l*: A suspect commits the offence of libel.*

Since SDL *allows a kind of so-called deontic detachment as a result of the* K*-axiom, i.e.* $(O(\alpha) \wedge O(\alpha \rightarrow \beta)) \rightarrow O(\beta)$*, we have* $T \models_{SDL} O(\neg p)$ *from the first two sentences.* SDL *also allows factual detachment, i.e.* $(\alpha \wedge (\alpha \rightarrow O(\beta))) \rightarrow O(\beta)$*, and therefore we have* $T \models_{SDL} O(p)$ *from the last two sentences. And again these two derived obligations are inconsistent.*

In DIODE*, the sentences are translated to (see [Tan & van der Torre, 1994b]) the following sentences of a* DIODE *theory* T*:*

1. $\neg V_1 \rightarrow \neg l$*: A person should not commit the offence of libel.*
2. $\neg V_2 \rightarrow (\neg l \rightarrow \neg p)$*: It should be that if a person does not commit the offence of libel, he is not punished for libel.*
3. $l \wedge \neg V_3 \rightarrow p$*: If a person does commit the offence, he should be punished for it.*
4. l*: A suspect commits the offence of libel.*

The preferential ordering of the deontic rules of the Chisholm 'Paradox' is given in Fig. 2. In the ideal situation, there is no offence and no punishment. For $F = \{l\}$*, the models are restricted to the models containing* V_1 *and the models containing* V_1 *and* V_3*, which is again depicted by a dashed box. Therefore* $\{V_1\}$ *is the only preferred violation set and* T *provides a contextual obligation for* p*.*

The previous two examples showed the two-phase mechanism of DIODE. The first phase consists of building a preferential 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. A drawback of this approach is that there has to be a distinction between facts and background knowledge, a well-known phenomena within non-monotonic logics; see [Boutilier, 1994] for a discussion of this distinction.

Two similar phases exist in the defeasible variant of DIODE which will be developed in the next section. However, in the first phase *two* preferential orderings will be constructed; not only one for ideality but also one for normality.

2.1 Discussion

The logic developed thus far has two serious defects which require further research. The major defect of DIODE is that all facts are contextually obliged. Contextual reasoning can be considered as reasoning 'within ideal worlds' of SDL semantics. In SDL semantics, an obligation $O(\alpha)$ is true in a world w in the model iff α is true in all deontically ideal worlds accessible from w. A preferred violation set characterizes, so to say, a set of deontically sub-ideal worlds, and within this context all formulas are obliged. The set of all contextual obligations also contains facts from the theory T, i.e. facts are also contextually obliged. This is less remarkable than one might think *prima facie*, once one realizes that the preferred violation sets indicate which (new) obligations hold given certain facts. In this context such facts acquire the status of a kind of factual laws. One of the solutions to the problem that facts are contextually obliged is to restrict the contextual obligations to formulas which cannot be derived from the facts F. In that case, contextual obligations can be derived from obligations or from obligations and facts, but not from facts alone.

A more serious defect of the DIODE semantics is that permissions cannot be represented. In SDL, permissions can easily be defined in terms of obligations: $P(\alpha) = \neg O(\neg\alpha)$. In [Prakken, 1994] a distinction between weak and strong permissions is made. A weak permission is the absence of an obligation and can also be formalized easily in DIODE in terms of obligations. Strong permissions however, capturing the notion of an explicit permission, cannot be defined in terms of obligations. To represent strong permissions, the logic and its semantics have to be extended which is subject of future research.

Because of space limitations we refer to our paper [Tan & van der Torre, 1994a] for a comparison with alternative logics representing sub-ideal states.

3 DIODE and exceptions

There is a fundamental difference between a conditional obligation being violated by a fact, and a conditional obligation being defeated by another conditional obligation. Various authors [Prakken, 1994; Horty, 1993; Makinson, 1993] have investigated the formalization of *defeasible* conditional obligations (traditionally called *prima facie* obligations), deontic rules which are subject to exceptions. Horty [Horty, 1993] gives his well-known example of being served asparagus; in that specific case, you should eat with your fingers

and the obligation not to eat with your fingers is defeated. 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.

Given a set of defeasible conditional obligations in DIODE, the question remains how to determine when there are exceptional circumstances, i.e. when an exception constant is true. In [Tan & van der Torre, 1994b] it is suggested that all exceptions should be 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 the this obligation is that every journalist should expose Nixon in the Watergate affair. In that case, a defeater rule must be added that states that a situation of public interest is an exception to the rule not to insult someone, $h \rightarrow Ex_1$. When exceptions are explicit, in case of a conflict between the two independent minimization processes (for the defeasible conditional obligation above when α is true, β is false and Ex_1 and V_1 can be either true or false), violations should be preferred over exceptions; i.e. Ex_1 should be false and V_1 should be true. This is the simplest way to deal with conflicts between the preferential orderings, and it means that defeasible conditional obligations are only defeated when there is an explicit defeater rule that states this.

The following definition of overridden is a formalization of the notion of specificity (lex specialis). 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 [Horty, 1993] definition of overridden.[2]

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 $\alpha_2 \wedge \neg V_2 \wedge \rightarrow \beta_2 \in T$) 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 $\alpha_2 \rightarrow Ex_1$ should be added. The next example is an instance of Horty's asparagus example:

[2] Notice that Definition 6 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 , which is the classical material implication. This could be solved, for instance, by representing deontic rules as Reiter default rules (like in Horty's logic).

Example 3 (Public Interest) *Consider the following sentences:*

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 preferential ordering on normality is that the models with exceptional circumstances (public interest) are semantically seperated from the normal situation. The intended preferential semantics are given in Fig. 3. The circles denote again equivalence classes in the deontic ordering and the 'horizontal' arrows the deontic preferential ordering. The boxes denote equivalence classes in the normality ordering and the 'vertical' arrow the normality preferential 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 $\neg h$-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.[3] *Because of the exceptional circumstances, the models are not subject to the obligation that usually, you should not insult someone.*

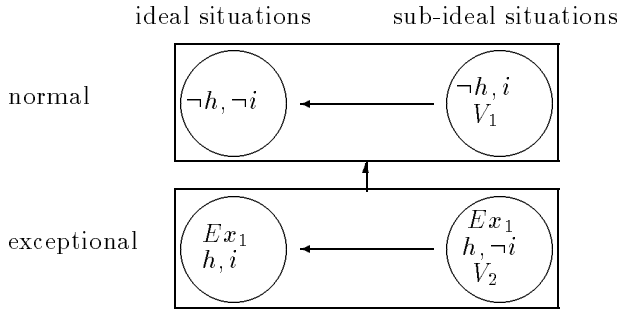

Figure 3: Preference relation of Public Interest

In Example 5 we will merge the Forrester 'Paradox' with the previous example. Before we can do this, however, we have to reconsider the Forrester 'Paradox' in a defeasible deontic setting. As we showed in [van der Torre, 1994], a strong definition of overridden like Horty's definition [Horty, 1993] or Def. 6 will give unintuitive results.

Example 4 *Reconsider the sentences of the Forrester 'Paradox' 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.*

[3] Notice that this obligation is not implied by the article of criminal law given in the Introduction. There only the *permission* to insult someone is implied by the exceptional circumstances. Such a permission could be represented in the semantics by the fact that all h models are equivalent in the deontic ordering. But, since permissions have not yet been formalized in , we use this example.

Given Def. 6 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. 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 [van der Torre, 1994] 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.

Definition 7 *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 $\alpha_2 \wedge \neg V_2 \wedge \rightarrow \beta_2 \in T$) 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 Def. 7 since the last condition is not satisfied.

Now we can combine the two examples. The Forrester 'Paradox' was a typical example of CTD reasoning and the Public Interest example was typical for a defeasible rule being overridden.

Example 5 (Forrester 'Paradox' and Public Interest) *Consider the following sentences:*

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;*
4. *$h \wedge \neg V_3 \rightarrow i$: If someone does harm the public interest, then you should insult him.*

The preferential ordering of the deontic rules (together with the rule $p \rightarrow i$) is given in Fig. 4. Just like in the Public Interest example, there is a distinction between normal circumstances (when there is not a situation of public interest) and exceptional circumstances (when there is such a situation). In the normal circumstances, the ordering is identical to the ordering given by the Forrester 'Paradox'. In the exceptional circumstances however, the preferred situation is that you insult him. Since this is exceptional there is no violation of the first deontic rule V_1.

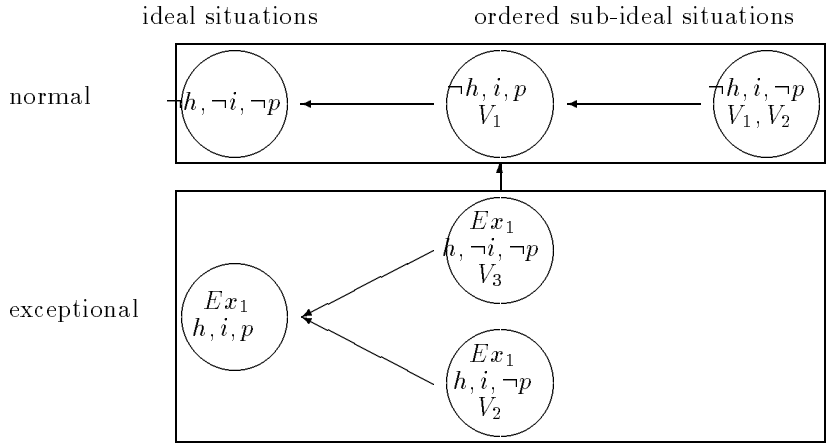

Figure 4: Preference relation of the Forrester 'Paradox' with Public Interest

3.1 Discussion

In non-monotonic logics based on conditional logic, exceptions are determined implicitly by a kind of 'built in' non-monotonicity of conditional logic [Boutilier, 1994]. In this article we have given the intended semantics of several examples, but the interesting new problem that has emerged is *how can exceptions be determined implicitly in a defeasible deontic logic?* The interesting thing about it is that we cannot simply use definitions or algorithms from existing non-monotonic logics since the two orderings may (and will) interfere, as we already saw in the simple example of the Forrester 'Paradox'.

Another interesting aspect of this multiple preference semantics is that it generalizes the preferential semantics for non-monotonic logics initiated by Shoham [Shoham, 1988] and Kraus, Lehmann and Magidor [Kraus *at al*, 1990] in the same way that multi-modal logics generalizes classical modal logic. This seems to follow Makinson's suggestion that defeasible deontic logic is the most complicated of the five faces of minimality [Makinson, 1993].

References

T. Bench-Capon. Deontic logic: Who needs it? In *Proceedings of workshop 'Artificial normative reasoning' of the Eleventh European Conference on Artificial Intelligence (ECAI'94)*, Amsterdam, 1994.

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

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

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

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

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.

A.J.I. Jones. Deontic logic and legal knowledge representation. In *Expert Systems in Law*, Bologna, 1989.

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

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.

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

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

H. Prakken. Two approaches to defeasible deontic reasoning. In *Proceedings of the Second Workshop on Deontic Logic in Computer Science (Δeon'94)*, Oslo, 1994.

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 (Δeon'94)*, Oslo, 1994.

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

Y. Shoham. *Reasoning About Change*. MIT Press, 1988.

Y.-H. Tan and L.W.N. van der Torre. DIODE: deontic logic based on diagnosis from first principles. In *Proceedings of workshop 'Artificial normative reasoning' of the Eleventh European Conference on Artificial Intelligence (ECAI'94)*, Amsterdam, 1994.

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.

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.

Acknowledgements

Thanks to Patrick van der Laag and Henry Prakken for several discussions on the issues raised in this paper and to Mark-Jan van der Torre for a discussion on the legal examples. 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.