DefDiode: a Diagnostic Model for Defeasible Deontic Logic*†

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

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 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 (see [11] and [8]). Deontic logic is a modal logic in which the modal operator O is used to express that something is obliged, see [3]. 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 [3]. It satisfies the axioms K: $O(p) \wedge O(p \rightarrow q) \rightarrow O(q)$ and D: $\neg(O(p) \wedge O(\neg p))$.

Three decades ago, Chisholm described in [4] 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 [3]). Two decades later, Forrester described in [5] his version of the 'paradox', the so-called Forrester 'Paradox', which could not be solved by any of these new deontic logics. In [17] and [16] we introduced Diode; a DIagnostic framework for DEontic reasoning. In these papers we showed how one can solve the Chisholm and Forrester 'Paradoxes' in Diode. Diode is based on the observation that there is a striking resemblance between deontic reasoning and diagnostic reasoning from first principles as it was introduced by Reiter in [14]. 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. Details of this semantics will be explained later. Recently, Boutilier [2] also proposed a preferential semantics for deontic logic, based on an observed resemblance between deontic logic and qualitative decision theory.

*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: European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty (Ecsqaru'95), 1995.

‡Erasmus University Research Institute for Decision and Information Systems (Euridis). Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands. E-Mail: {ytan,ltorre}@euridis.fbk.eur.nl.

In this paper we address the question how we can extend DIODE to DEFDIODE such that it can model *defeasible* deontic logic. It is well-known that defeasible reasoning is a very important aspect of legal reasoning (see [7] and [12]). We will argue that defeasible deontic logic can only be modeled in DIODE if we introduce two independent preference orderings in the underlying semantics of DEFDIODE; one to model the deontic aspects and another to model the normality aspects that are used to model defeasible reasoning. This multi preference semantics generalizes the preferential semantics for non-monotonic logics initiated by Shoham [15] and Kraus, Lehmann and Magidor [9] in the same way that multi-modal logics generalizes classical modal logic. Interestingly, it appears that these preference orderings interfere in a complicated way, thus generating new and interesting questions about preferential semantics. This supports Makinson's conjecture that defeasible deontic logic is the most complicated of the five faces of minimality [10].[1]

2 DIODE

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.

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 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 deontic rule has a distinct violation constant V_i. For a detailed description of the syntax and proof theory of DIODE, see [17] and [16].

DIODE contains a preferential semantics that defines a preference ordering on models 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 [13] 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.

[1] Boutilier [2] 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.

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.*

The preferred violation sets correspond to the diagnosis sets of faulty components in Reiter's sense. The set of background knowledge sentences and deontic rules of T corresponds to the set of system description rules (SD) and the set of facts of T corresponds to the set of observations (OBS). If some of the obligations are violated by some of the facts, then the preferred violation set is the minimal set of violations that restores the conflict between obligations and facts. 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 (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' [5]: you should not kill, but if you kill you should do it gently.

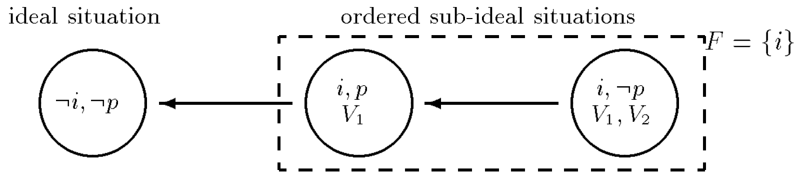

Figure 1: Preference relation of the Forrester 'Paradox'

Example 1 (Forrester 'Paradox') *Consider the following sentences of a theory T:*

1. $\neg V_1 \to \neg i$: *You should not insult someone;*

2. $i \wedge \neg V_2 \to p$: *If you insult someone you should do it in private;*

3. $p \to i$: *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 $p \to i$) of the Forrester 'Paradox' is given in Figure 1. The circles denote equivalence classes of the preference ordering defined in Definition 1 of models that satisfy the deontic rules and background knowledge sentences, in this case the first three sentences above. 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 shown in the picture. Hence, in an equivalence class, the propositional constants are shown which are true (or false) in all relevant *models of the class.*

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 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. In the original representation in modal deontic logics, there was an obligation for $\neg i$ which *represented the violation and caused the inconsistency. In* DIODE, *there is no contextual obligation for* $\neg i$ *since* $\neg i$ *refers to the ideal state and is not considered by the process of zooming in; the violation is now represented by the violation constant* V_1. *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.*

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. 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 [1] 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* 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 defeated by another conditional obligation. Various authors, e.g. [6, 10], have investigated the formalization of *defeasible* conditional obligations (traditionally called *prima facie* obligations), deontic rules which are subject to exceptions. Horty [6] 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.

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 [17] 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 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, $h \rightarrow Ex_1$. When exceptions are explicit, in case of a conflict between the two independent minimization processes (for the defeasible conditional obligation $\alpha \wedge \neg V_1 \wedge \neg Ex_1 \rightarrow \beta$ this is the case 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 preference orderings, and it means that defeasible conditional obligations are only defeated when there is an explicit defeater rule that states this. Semantically, the normality ordering is a subset ordering on exception constants Ex_i just like the ideality ordering on violation constants V_i.

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 [6].[2]

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 $\alpha_2 \rightarrow Ex_1$ should be added. 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 [6].

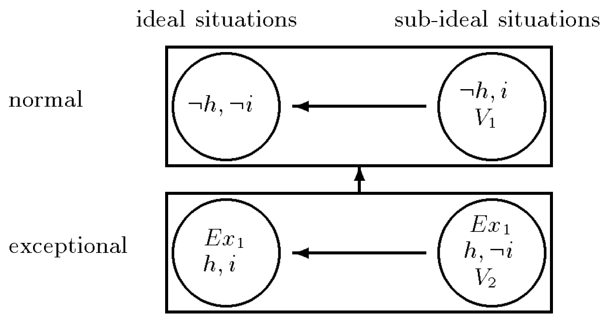

Figure 2: Preference relation of Public Interest

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 circles denote again equivalence classes in the deontic ordering and the 'horizontal' arrows the deontic preference ordering.[3] *The boxes denote equivalence classes in the normality ordering and the 'vertical' arrow the normality 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 $\neg h$-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.*

[2] 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. This could be solved, for instance, by representing deontic rules as Reiter default rules (like in Horty's logic [6]).

[3] Actually, the two leftmost circles belong to the same ideality equivalence class. For representational convenience we have split them.

In Example 4 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 [18], a strong definition of overridden like Horty's definition [6] 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. 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 [18] 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 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 since the last condition is not satisfied.

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

Example 4 (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.*

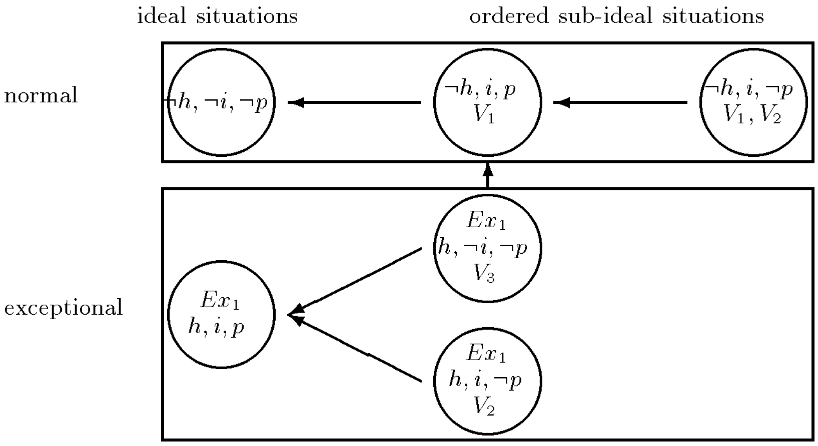

Figure 3: Preference relation of the Forrester 'Paradox' with Public Interest

According to Definition 6 we have that rule (1.) is only overridden by rule (4.) for h, but not overridden by rule (2.). Hence we should only add (5.) $h \rightarrow Ex_1$. The crucial observation is that although both (2.) and (4.) conflict with (1.) and are more specific than (1.), it is only (4.) that overrides (1.). Rule (2.), which is a CTD of (1.), has a different effect, and is treated differently from (4.).

The preference ordering of the deontic rules (together with the rule $p \rightarrow i$) is given in Figure 3. 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.

4 Further Research

In non-monotonic logics based on conditional logic, exceptions are determined implicitly by a kind of 'built in' non-monotonicity of conditional logic [1]. 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?* This is subject of our current research. The interesting thing about it is that we cannot simply use definitions or algorithms from existing non-monotonic logics since the two orderings can interfere, as we already saw in the simple example of the Forrester 'Paradox'.

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

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

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

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

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

[6] 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.

[7] 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.

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

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

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

[11] J.-J. Meyer and R. Wieringa. *Deontic Logic in Computer Science: Proceedings of the First Workshop on Deontic Logic in Computer Science (Δeon'91).* John Wiley & Sons, 1993.

[12] H. Prakken. *Logical Tools for Modelling Legal Argument, Ph-D thesis.* Amsterdam, 1993.

[13] 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.

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

[15] Y. Shoham. *Reasoning About Change.* MIT Press, 1988.

[16] 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.

[17] 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)*, Italy, 1994. (An extended version of this article will appear in the *Journal of Artificial Intelligence and Law.*)

[18] 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.