An Update Semantics for Prima Facie Obligations

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

Abstract

The deontic logic DUS is a Deontic Update Semantics for prescriptive obligations based on the update semantics of Veltman. In DUS the definition of logical validity of obligations is not based on static truth values but on dynamic action transitions. In this paper prescriptive prima facie obligations are formalized in update semantics. The logic formalizes the specificity principle, has reinstatement and does not have an irrelevance problem. Moreover, it handles the diagnostic problem by distinguishing between overridden, conflict and factual defeasibility.

Contents

1 The logic of norms

Deontic logic is a modal logic in which Op is read as 'p ought to be (done).' Deontic logic has traditionally been used by philosophers to analyze the structure of the normative use of language. In the eighties deontic logic had a revival, when it was discovered by computer scientists that this logic can be used for the formal specification and validation of a wide variety of topics in computer science (for an overview and further references see [WM93]). For example, deontic logic can be used to formally specify soft constraints in planning and scheduling problems as norms. The advantage is that norm violations do not create an inconsistency in the formal specification, in contrast to violations of hard constraints. Recently new interest in deontic logic has arisen, because it could be fruitful for the analysis and specification of security issues about electronic networks such as authorization, access regulation, and privacy maintenance [CF97], in particular for electronic commerce [FvdT98], and the relation between deontic logic and logics of desires (and goals) suggests that (extensions of) deontic logic can be used in qualitative decision theory [Pea93, Bou94, Lan96, vdT98].

In [vdTT98b] the deontic logic DUS is introduced, a deontic update semantics for prescriptive obligations, based on the update semantics of Veltman [Vel96]. In DUS meaning becomes a dynamic notion: you know the meaning of a normative sentence, if you know the change it brings about in the deontic state of anyone who is subjected to the news conveyed by it. In this paper we introduce a deontic update semantics for *prima facie* obligations [Ros30] and we show that the dynamic approach not only gives a better analysis of the traditional deontic problems, but it also gives a better analysis of the problems discussed in defeasible deontic logic. An example of reasoning with prima facie obligations is that you have a prima facie obligation to go to a birthday party if you promised to go, but this prima facie obligation does not turn into a proper obligation when you have to save a child from drowning. Prima facie obligations can be overridden or cancelled by other, stronger obligations and are thus defeasible. However, not all defeasible obligations are prima facie obligations! For example, 'prima facie α should be (done)' is different from the defeasible obligation 'normally α should be (done)' (based on respectively weak and strong overridden defeasibility [vdTT97]). The distinctive property is that the obligation to go to the party is not completely cancelled if you have to save a child from drowning, but it is still in force as a prima facie obligation. Consequently, saving the child from drowning is not an ideal situation, because the prima facie obligation to go to the party is violated, whereas defeasible obligations can be completely cancelled by exceptional circumstances.

It has been argued that more specific defeasible obligations are stronger than more general defeasible obligations, and therefore override them in case of conflict [Hor93, vdT94, AB96, Mor96]. Unfortunately, the analysis of the specificity principle in logics of defeasible reasoning does not apply to defeasible deontic logic, because it may interfere with the violability of norms [vdT94, vdTT95, vdTT97, vdT97]. This interference is illustrated by the following diagnostic problem.

1. You should not speed.
2. To prevent a possible disaster, you should speed.
3. If you are speeding, then you should speed safely.
4. You are speeding safely.

Is the fact 'you are speeding safely' a violation or an exception? Obviously, this is a crucial question for legal knowledge-based systems. In case of a possible disaster, you should speed according to the second line and the first obligation is cancelled. Moreover, if you are speeding, then you should speed safely according to the third line and the first obligation is overshadowed. Only in absence of a possible disaster you have to pay a penalty for speeding, because only in that case the first obligation is a violated proper obligation. Note that the second obligation has to be stronger than the first obligation to cancel it, but the third obligation does not have to be stronger than the first one to overshadow it.

The layout of this paper is as follows. First, we introduce prescriptive prima facie obligations in update semantics and we show that the logic formalizes the specificity principle without introducing an irrelevance problem. Second, we show how different types of defeasibility are distinguished to handle the diagnostic problem.[1]

2 Prima facie obligations in DUS

Ross [Ros30] introduced the notion of so-called prima facie obligations. In his own words: "I suggest '*prima facie* duty' or 'conditional duty' as a brief way of referring to the characteristic (quite distinct from that of being a duty proper) which an act has, in virtue of being of a certain kind (e.g. the keeping of a promise), of being an act which would be a duty proper if it were not at the same time of another kind which is morally significant" [Ros30, p.19]. A prima facie duty is a duty proper when it is not overridden by another prima facie duty. When a prima facie duty is overridden, it is not a duty proper, but it is still in force: "When we think ourselves justified in breaking, and indeed morally obliged to break, a promise [...] we do not for the moment cease to recognize a prima facie duty to keep our promise" [Ros30, p.28]. Consequently, a prima facie duty is again a duty proper when its overriding duties are violated or themselves overridden [vdTT97]. In such a case we say that the obligation is reinstated.[2]

[1] Makinson observed in 1993 [Mak93] that 'at the present state of play, it would not seem advisable to try to cover all complicating factors [of deontic logic] at once, but rather to get an initial appreciation of them few at a time, only subsequently putting them together and investigating their interactions.' In [Mak98, vdTT98b] only non-defeasible obligations in a propositional setting have been studied. The language of DUS now can be extended with permissions (following the suggestions in [vdT97]) and prohibitions, a first-order base language, nested conditionals (following [Wey92]), background knowledge (following [vdT94]), authorities, agents, actions, time and exceptions. In this paper we only consider the latter extension.

[2] Compare Prakken and Sergot's cottage housing regulations [PS96, PS97]: there should not be a fence, but if there is a dog, then there should be a white fence. In contrast to prima facie obligations, these defeasible obligations do not have reinstatement [vdTT97]. If there is

In this section we define prescriptive prima facie obligations in update semantics. The logic handles conflicts of hierarchic obligations, which prima facie exist, but might be dynamically re-evaluated. Two characteristic properties of the logic are that obligations are overridden by more specific and conflicting obligations, and that unresolvable strong conflicts like 'p ought to be (done) and $\neg p$ ought to be (done)' are 'inconsistent,' in the sense that they derive all sentences of the deontic language.

We start with the basic definitions of Veltman's update semantics [Vel96]. To define a deontic update semantics for a deontic language L, one has to specify a set Σ of relevant deontic states (called information states in [Vel96]), and a function $[\,]$ that assigns to each sentence ϕ an operation $[\phi]$ on Σ. If σ is a state and ϕ a sentence, then we write '$\sigma[\phi]$' to denote the result of updating σ with ϕ. We can write '$\sigma[\psi_1]\ldots[\psi_n]$' for the result of updating σ with the sequence of sentences $\psi_1, \ldots, \psi_n$. Moreover, one of the deontic states has to be labeled as the minimal deontic state, written as $\mathbf{0}$, and another one as the absurd state, written as $\mathbf{1}$.

Definition 1 (Deontic update system) *A deontic update system is a triple $\langle L, \Sigma, [\,]\rangle$ consisting of a logical language L, a set of relevant deontic states Σ and a function $[\,]$ that assigns to each sentence ϕ of L an operation. Σ contains the elements $\mathbf{0}$ and $\mathbf{1}$.*

Veltman explains what kind of semantic phenomena may successfully be analyzed in update semantics and he gives a detailed analysis of one such phenomenon: default reasoning. To define obligations in update semantics we have to define the deontic language, the deontic states and the deontic updates. The deontic language is a propositional language with the dyadic operator $\mathsf{oblige}(\alpha|\beta)$, read as '$\alpha$ ought to be (done), if β is (done).'

Definition 2 (Deontic language) *Let A be a set of atoms and L_0^A a propositional language with A as its non-logical symbols. A string of symbols ϕ is a sentence of L_1^A if and only if either ϕ is a sentence of L_0^A or there are two sentences ψ_1 and ψ_2 of L_0^A such that $\phi = \mathsf{oblige}(\psi_1|\psi_2)$. We write $\mathsf{oblige}\ \psi$ for $\mathsf{oblige}(\psi|\top)$, where $\top$ stands for any tautology.*

A deontic state is a possible worlds model $\langle W, R, V\rangle$, where W is a set of worlds, R is a ranking function on ordered pairs of worlds of W (see below), and V a valuation function for propositions at the worlds. The rank of a pair of worlds (w_1, w_2) represents the strength of the prima facie obligation that prefers world w_2 to w_1. If there is no such obligation then its rank is 0, and if there are several of such obligations, then its rank is the strength of the strongest of the obligations. Note that the numbers are only used to codify a qualitative ordering, like numbers on a temperature scale, because we do not calculate with the numbers (i.e. the additive property of numbers is not exploited). We call an ordered pair of worlds a link. In particular, we call an ordered pair (w_1, w_2)

a dog and there cannot be a white fence, then it does not follow that there should not be a fence.

an (α_1, α_2) link if $w_1 \models \alpha_1$ and $w_2 \models \alpha_2$. We add the following features to these deontic states.

Explicit sub-state. We extend the possible worlds model with a second deontic state, which is a sub-state of the first one. The complete state is used for the context of justification and the sub-state is used for the context of deliberation, see [vdTT98b]. Whereas in Kripke semantics a unique world is singled out, called the actual world, we single out a set of worlds, called the context of deliberation.

Full models. We define an update system for a specific A, W and V. In this paper, we assume that the deontic state contains a world for each interpretation of L_0^A. If we want to represent background knowledge, then this assumption has to be dropped [vdT94, Lan96].

Infinity ∞. The ranking function can also have the value ∞, which is larger than all integers and which has the special property that the addition of any integer number to it results again in ∞. Once a link is ranked ∞, it cannot be updated to another value. In the absurd state, all links are ranked ∞. Formally, the ranking R is a mapping of $W \times W$ to the set of positive integers $I\!N$ plus infinity, $I\!N \cup \{\infty\}$, with infinity larger than any element of $I\!N$, i.e. $\forall x \in I\!N(x \neq \infty \rightarrow x < \infty)$.

Definition 3 (Deontic state) *Let L_1^A be a deontic language. Assume a set of worlds W and a valuation function V for L_0^A such that for every interpretation of L_0^A there is at least one corresponding $w \in W$. A deontic state is a tuple $\Sigma = \langle W, W^*, R, V\rangle$ consisting of the set of worlds W, a possibly empty subset $W^* \subseteq W$, an integer (or ∞) valued ranking function R on $W \times W$ and the valuation function V.*

0, *the* minimal state, *is given by* $\langle W, W, W \times W \rightarrow 0, V\rangle$, *and*

1, *the* absurd state, *is given by* $\langle W, \emptyset, W \times W \rightarrow \infty, V\rangle$.

The deontic updates are operations on the deontic states that either zoom in on the deontic state (for facts), or increase the ranks of links (for obligations). The prescriptive obligations have the dynamic component of creating a new deontic state. The general principle is that in case of conflict later obligations are stronger than earlier ones. For the update with obligation $\mathsf{oblige}(\alpha|\beta)$ there is a conflict if all the reverse links (i.e. $(\alpha \wedge \beta, \neg\alpha \wedge \beta)$ links) are non-zero. If there is no conflict then the rank of the $(\neg\alpha \wedge \beta, \alpha \wedge \beta)$ links is 1. Otherwise, their rank is higher than the minimum of the reverse links. Finally, von Wright's contingency principle, i.e. the obligation 'α ought to be (done) if β is (done)' implies the consistency of $\alpha \wedge \beta$ and $\neg\alpha \wedge \beta$, is formalized by a test on the existence of $\alpha \wedge \beta$ and $\neg\alpha \wedge \beta$ worlds.[3]

Definition 4 (Deontic updates) *Let $\sigma = \langle W, W^*, R, V\rangle$ be a deontic state, and let $\min(\alpha|\beta)$ be the minimum of $\{R(w_1, w_2) \mid \sigma, w_1 \models \alpha \wedge \beta$ and $\sigma, w_2 \models$*

[3] We can also define obligations $\mathsf{oblige}^*(\alpha|\beta)$ that refer to W^* instead of W, see [vdTT98b]. These obligations are called factually detached. Moreover, we can define non-defeasible obligations by giving the relevant links the value ∞.

$\neg\alpha \wedge \beta\}$ *if this set is non-empty, undefined otherwise. The update function* $\sigma[\phi]$ *is defined as follows.*

- *if* ϕ *is a factual sentence of* L_0^A*, then*
  - *if* $W' = \{w \in W^* \mid \sigma, w \models \phi\} \neq \emptyset$*, then* $\sigma[\phi] = \langle W, W', R, V\rangle$*;*
  - *otherwise,* $\sigma[\phi] = \mathbf{1}$*.*
- *if* $\phi = \mathsf{oblige}(\alpha|\beta)$*, then*
  - *if there are* $w_1, w_2 \in W$ *such that* $\sigma, w_1 \models \neg\alpha \wedge \beta$ *and* $\sigma, w_2 \models \alpha \wedge \beta$*, then* $\sigma[\phi] = \langle W, W^*, R', V\rangle$ *with for all* $w_1, w_2 \in W$
    - *if* $\sigma, w_1 \models \neg\alpha \wedge \beta$ *and* $\sigma, w_2 \models \alpha \wedge \beta$ *then* $R'(w_1, w_2) = \max(R(w_1, w_2), \min(\alpha|\beta) + 1)$*;*
    - *otherwise* $R'(w_1, w_2) = R(w_1, w_2)$*;*
  - *otherwise,* $\sigma[\phi] = \mathbf{1}$*.*

A crucial notion of update systems is acceptance. The formula ϕ is accepted in a deontic state σ, written as $\sigma \Vdash \phi$, if the update by ϕ results in the same state. In that case, the information conveyed by ϕ is already subsumed by σ. Acceptance is the counterpart of satisfaction in standard semantics.

Definition 5 (Acceptance) *Let* σ *be an deontic state and* ϕ *a formula of the logical language* L*.* $\sigma \Vdash \phi$ *if and only if* $\sigma[\phi] = \sigma$*.*

If an update is accepted, then the deontic state usually has a specific content. For example, it is easily checked that a fact α is accepted if all the worlds of $W^* \neq \emptyset$ satisfy α, or $\sigma = \mathbf{1}$. Moreover, an obligation $\mathsf{oblige}(\alpha|\beta)$ is accepted if the ranking of all $(\neg\alpha \wedge \beta, \alpha \wedge \beta)$ links is higher than the smallest rank of the reversed links. Different notions of validity can be based on the notion of acceptance (see Veltman's paper [Vel96] for an overview). In Definition 7 we will use the following one. An argument is valid if updating the minimal state $\mathbf{0}$ with the premises $\psi_1, \ldots, \psi_n$, in that order, yields a deontic state in which the conclusion is accepted.

Definition 6 (Validity) *Let* $\psi_1, \ldots, \psi_n$ *and* ϕ *be formulas of the logical language* L*. The argument of* ϕ *from the premises* $\psi_1, \ldots, \psi_n$ *is valid, written as* $\psi_1, \ldots, \psi_n \Vdash_1 \phi$*, if and only if* $\mathbf{0}[\psi_1] \ldots [\psi_n] \Vdash \phi$*.*

Finally, we give the non-monotonic validity relation $|\!\!\sim$ where the premises are a set (not given in [Vel96]). An argument is valid if all deontic states constructed by updating the minimal state $\mathbf{0}$ with the premises $\psi_1, \ldots, \psi_n$ *in some order such that the premises are accepted*, also accept the conclusion. We show that one of its features is that more specific and conflicting obligations are only accepted if they are later than more general ones. Hence, more specific and conflicting obligations are stronger than more general ones and override them.

Definition 7 (Validity, continued) *Let* $\psi_1, \ldots, \psi_n$ *and* ϕ *be sentences of the deontic language* L_1^A*.* $\psi_1, \ldots, \psi_n \,|\!\!\sim \phi$ *iff for all permutations* π *of* $1 \ldots n$ *such that* $\psi_{\pi(1)}, \ldots, \psi_{\pi(n)} \Vdash_1 \psi_i$ *for* $1 \leq i \leq n$ *we have* $\psi_{\pi(1)}, \ldots, \psi_{\pi(n)} \Vdash_1 \phi$*.*

The following example illustrates how the logic formalizes the specificity principle, without creating an irrelevance problem.

Example 1 (Speeding) *Consider the two obligations* $\mathsf{oblige}(\neg s|\top)$ *and* $\mathsf{oblige}(s|p)$*, where* s *stands for speeding and* p *for a potential disaster if you are not speeding. If the more specific obligation comes first then all* $(s, \neg s)$ *and* $(p \wedge \neg s, p \wedge s)$ *links have rank 1 and both obligations are equally strong. Otherwise, the* $(s, \neg s)$ *links have rank 1 and the* $(p \wedge \neg s, p \wedge s)$ *links have rank 2, i.e. the more specific obligation is stronger and overrides the more general one. The latter case is illustrated in Figure 1 below. For readability only positive atoms and non-zero links are represented.*

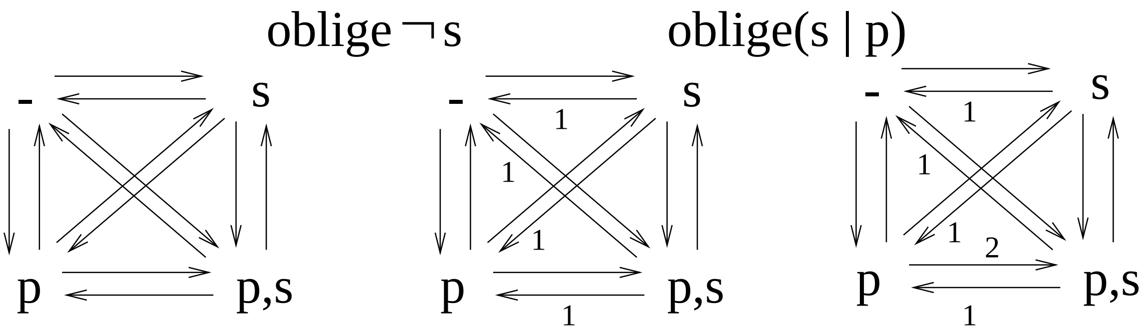

Figure 1: Speeding to prevent a disaster

The only accepted order of the premises in Definition 7 is that the more general obligation comes before the more specific one, because only in that case the premises are accepted.

$$\mathsf{oblige}(\neg s|\top), \mathsf{oblige}(s|p) \Vdash_1 \mathsf{oblige}(\neg s|\top)$$
$$\mathsf{oblige}(\neg s|\top), \mathsf{oblige}(s|p) \Vdash_1 \mathsf{oblige}(s|p)$$
$$\mathsf{oblige}(s|p), \mathsf{oblige}(\neg s|\top) \nVdash_1 \mathsf{oblige}(s|p)$$

The second deontic state in Figure 1 accepts $\mathsf{oblige}(\neg s \,|\, p)$*, but the third does not. Consequently we have the following.*

$$\mathsf{oblige}(\neg s|\top) \mid\!\mid\!\!\sim \mathsf{oblige}(\neg s|p)$$
$$\mathsf{oblige}(\neg s|\top), \mathsf{oblige}(s|p) \mid\!\mid\!\!\not\sim \mathsf{oblige}(\neg s|p)$$

Hence, the logic formalizes the specificity principle.[4] *Moreover, the logic does not have an irrelevance problem, because we have for example that you should not speed in the weekend (*w*).*

$$\mathsf{oblige}(\neg s|\top), \mathsf{oblige}(s|p) \mid\!\mid\!\!\sim \mathsf{oblige}(\neg s|w)$$

Our second example illustrates more properties of the logic.

[4] If we also want that $O(s \wedge a|p)$ overrides $O(\neg s|\top)$, then we have to add an additional test on 'preferred' worlds, see [vdTT98b].

Example 2 *Let A be the set of propositional atoms $\{p, q, r, s\}$. We have the following.*

SA	$\mathsf{oblige}\ p \mid\!\sim \mathsf{oblige}(p\|q)$
RI	$\mathsf{oblige}\ p, \mathsf{oblige}(\neg p \wedge \neg q\|r) \mid\!\sim \mathsf{oblige}(p\|r \wedge q)$
RIO	$\mathsf{oblige}\ p, \mathsf{oblige}(\neg p \wedge \neg q\|r), \mathsf{oblige}(q\|r \wedge s) \mid\!\sim \mathsf{oblige}(p\|r \wedge s)$
DD	$\mathsf{oblige}(p\|q \wedge r) \wedge \mathsf{oblige}(q\|r) \mid\!\sim \mathsf{oblige}(p \wedge q\|r)$
AND	$\mathsf{oblige}\ p, \mathsf{oblige}\ q \mid\!\sim \mathsf{oblige}(p \wedge q)$
OR	$\mathsf{oblige}\ p, \mathsf{oblige}\ q \mid\!\sim \mathsf{oblige}(p \vee q)$
D	$\mathsf{oblige}\ p, \mathsf{oblige}\ \neg p \mid\!\sim \bot$
DS	$\mathsf{oblige}\ p, \mathsf{oblige}(\neg p\|q) \mid\!\not\sim \bot$
WC	$\mathsf{oblige}\ p \mid\!\not\sim \mathsf{oblige}(p \vee q)$
FC	$\mathsf{oblige}\ p, \mathsf{oblige}(\neg p\|q) \mid\!\not\sim \mathsf{oblige}\ \neg q$

*The logic has strengthening of the antecedent (***SA***), reinstatement (***RI*** and* **RIO***), a kind of deontic detachment (***DD***) and a conjunction and disjunction rule (***AND*** and* **OR***). Obviously, all these derivations are only valid non-monotonically. In particular, Example 1 illustrates that strengthening of the antecedent is blocked by a more specific and conflicting prima facie obligation. Moreover, a strongly conflicting set derives a contradiction (***D***), but a specificity set does not (***DS***). Finally, the logic does not derive weaker obligations* **WC** *or the obligation to evade conflicts* **FC***.*

3 The diagnostic problem

To handle the diagnostic problem different types of defeasibility can be distinguished in the logic of prescriptive prima facie obligations. If there are conflicting links, then we call it overridden defeasibility (ranks of conflicting links are unequal) or conflict defeasibility (equally ranked conflicting links). Otherwise we call it factual defeasibility.[5] All three lead to a restriction on strengthening of the antecedent, the characteristic property of defeasible conditionals [Alc93], but in this paper only the first two lead to non-monotonicity.

The diagnostic problem is the most interesting for ought-to-do obligations, which in contrast to ought-to-be obligations refer to actions. Since we are not interested in the problems of formalizing actions, we make two simplifying assumptions. First, we only consider deterministic actions (see e.g. [vdTT98a]

[5]Several cases of apparent dilemmas of prima facie obligations can be solved by analyzing them as different types of defeasibility. Examples of conflict defeasibility are 'be polite' and 'be honest' ($\{\mathsf{oblige}\ p, \mathsf{oblige}\ h\}$), and 'the window ought to be closed if it rains' and 'it ought to be open if the sun shines' ($\{\mathsf{oblige}(c|r), \mathsf{oblige}(\neg c|s)\}$). Examples of factual defeasibility are the contrary-to-duty paradoxes, e.g. 'Smith should not kill Jones,' 'if Smith kills Jones, then he should do it gently' and 'Smith kills Jones' ($\{\mathsf{oblige}\ \neg k, \mathsf{oblige}(g|k), k\}$), given that gentle killing implies killing ($\vdash g \to k$), and 'a certain man should go to the assistance of his neighbors,' 'if the man goes to their assistance, then he should tell them that he will come,' 'if the man does not go to the assistance, then he should not tell them he will come' and 'the man does not go' ($\{\mathsf{oblige}\ a, \mathsf{oblige}(t|a), \mathsf{oblige}(\neg t|\neg a), \neg a\}$).

for the indeterministic case). Consequently, we only have to extend our deontic language with the distinction, well-known from decision theory, between decision variables and parameters or events [Lan96] (called controllable and uncontrollable propositions in [Bou94]). Second, we assume complete knowledge of the parameters (see [Bou94, Lan96] for the general case). Given these two assumptions, we can restrict ourselves to the distinction between worlds which still can be realized and worlds which can no longer be realized due to some action of the agent. They are respectively the worlds in which the parameters and the facts are both true W_{p+f}, and the worlds in which the parameters are true but some fact is false W_{p-f}. There is a violation of a duty proper if and only if for every world that still can be realized there is a better world which can no longer be realized, i.e. for every world $w \in W_{p+f}$ there is a world $w' \in W_{p-f}$ such that the rank of the (w, w') link is higher than the rank of the (w', w) link. The following example illustrates how this distinction handles the diagnostic problem discussed in the introduction.

Example 3 (Speeding, continued) *Consider the obligations*

$$S = \{\mathsf{oblige}(\neg s|\top), \mathsf{oblige}(s|p), \mathsf{oblige}(s \wedge a|s)\}$$

, where $s \wedge a$ stands for speeding safely. First we distinguish the different types of defeasibility. It is easily shown that the only constraint on the ordering of the premises is that the second obligation comes later than the first one, and that the links of the third obligation do not conflict with the links of the first two obligations. The $(s, \neg s)$ links have rank 1, the $(p \wedge \neg s, p \wedge s)$ links have rank 2, and the $(s \wedge \neg a, s \wedge a)$ links have rank 1. Hence, the second obligation gives rise to overridden defeasibility and the third obligation gives rise to factual defeasibility. This is exactly the desired behavior, because this third obligation is never cancelling the first one, it can only overshadow it.

Now consider the fact 'you are speeding safely' $(s \wedge a)$. There is a violation of a duty proper, because for every $s \wedge a$ world w there is a $\neg(s \wedge a)$ world w' such that the rank of the (w, w') link is higher than the rank of the (w', w) link. Moreover, consider the fact 'you are speeding safely to prevent a disaster' $(s \wedge a \wedge p)$ where p is a parameter. There is no violation of a duty proper, only an exception, because it is not the case that for every $s \wedge a \wedge p$ world w there is a $\neg(s \wedge a) \wedge p$ world w' such that the rank of the (w, w') link is higher than the rank of the (w', w) link. On the contrary, $\neg(s \wedge a) \wedge p$ is a violation.

Finally, consider the fact 'you are speeding to prevent a disaster' $(s \wedge p)$ where p is a decision variable. Again there is a violation of a duty proper, because for every $s \wedge p$ world w there is a $\neg(s \wedge p)$ world w' such that the rank of the (w, w') link is higher than the rank of the (w', w) world. It is argued in [vdTT97] that this is exactly the desired result, because if you can evade a conflict then you should (it is even argued that the theorem **FC** *should be valid). The underlying idea is that agents cannot escape their responsibilities by creating exceptional circumstances, because exceptional circumstances are sub-ideal. Although there is no violation of a duty proper, there is still a violation of a prima facie duty.*

Summarizing, deontic states are complex entities that encode the distinction between the different types of defeasibility and thus can be used to handle the diagnostic problem. This is in contrast to simple preference orderings, but in line with multi-preference semantics [TvdT95, vdTT95, vdTT97] and so-called frames [Vel96]. However, the logics based on the latter two types of states have strong overridden defeasibility (no reinstatement, establishing a conflict is not a violation) and they therefore cannot be used for prima facie obligations.

4 Conclusions

In this paper we extended our dynamic approach to formalizing norms [vdTT98b]. Moreover, in [Mak98] an iterative approach is proposed. Besides taking into account the philosophical issue that norms do not have a truth value,[6] the dynamic and iterative approaches also give better analyses of the benchmark examples of deontic logic, in particular the deontic paradoxes.[7] In this paper we introduced a logic of prescriptive prima facie obligations, in which more specific obligations override more general ones. The specificity problem is solved by giving more specific and conflicting obligations a higher strength. An analogous mechanism is used in Geffner and Pearl's conditional entailment [GP92]. Moreover, the diagnostic problem is handled by distinguishing between overridden, conflict and factual defeasibility.

We now investigate utilitarian semantics with additive utilities [Lan96], which gives several new possibilities for the update function. For example, we can add the same number to all relevant links. Another extension is the formalization of test operators in the deontic update semantics, e.g. $\mathsf{ideal}(\alpha|\beta)$ for the test 'ideally, α is (done), if β is (done).' (It is shown in [TvdT96] that we have to introduce a separate operator for weakening of the consequent to combine it with strengthening of the antecedent.) The interaction between oblige and ideal is analogous to the interaction between normally and presumably operators in Veltman's update semantics [Vel96]. A final extension with descriptive obligations $O\alpha$ can be used to discuss the relation between prescriptive and descriptive obligations. We can write formulas like $[\mathsf{oblige}\ \alpha]O\alpha$, as in dynamic

[6]One of the first topics discussed in the development of deontic logic was the question whether norms have truth values, see [Mak98, vdTT98b]. For example, von Wright was hesitant to call deontic formulas 'logical truths,' because "it seems to be a matter of extra-logical decision when we shall say that 'there are' or 'are not' such and such norms'." Alchourrón and Bulygin distinguish between statements that describe a normative system, and statements that prescribe a certain behavior or state of affairs. In the first sense, the sentence 'it is obligatory to keep right on the streets' is a description of the fact that a certain normative system contains an obligation to keep right on the streets. In the second sense this statement is the obligation of traffic law itself.

[7]Recently it was discovered that approaches based on temporal and action concepts are not able to formalize the benchmark examples satisfactorily [PS96, PS97, vdTT98a]. Several approaches based on complex inductive definitions were proposed, usually applying techniques developed in non-monotonic reasoning. The dynamic and iterative approaches are natural successors and generalizations of Horty's non-monotonic approach [Hor93, Hor94] based on Reiter's default logic (Reiter's default rules do not have a truth value either and they iteratively construct extensions), Prakken and Sergot's contextual reasoning [PS96, PS97], labeled deontic logic [vdTT95, vdTT97, vdT98] and phased deontic logic [TvdT96].

logic, representing that after α has been promulgated, the obligation that 'α should be (done)' is true.

Acknowledgments. We thank Henry Prakken and Emil Weydert for discussing the issues raised in this paper, and an anonymous referee of the Fourth International Workshop of Deontic Logic in Computer Science (ΔEON'98) for several helpful comments.

References

[AB96] N. Asher and D. Bonevac. Prima facie obligation. *Studia Logica*, 57:19–45, 1996.

[Alc93] C.E. Alchourrón. Philosophical foundations of deontic logic and the logic of defeasible conditionals. In J.-J. Meyer and R. Wieringa, editors, *Deontic Logic in Computer Science: Normative System Specification*, pages 43–84. John Wiley & Sons, 1993.

[Bou94] C. Boutilier. Toward a logic for qualitative decision theory. In *Proceedings of the KR'94*, pages 75–86, 1994.

[CF97] R. Conte and R. Falcone. Icmas'96: Norms, obligations, and conventions. *AI Magazine*, 18,4:145–147, 1997.

[FvdT98] B.S. Firozabadi and L.W.N. van der Torre. Towards a formal analysis of control systems. In *Proceedings of the ECAI'98*, pages 317–318, 1998.

[GP92] H. Geffner and J. Pearl. Conditional entailment: bridging two approaches to default reasoning. *Artificial Intelligence*, 53:209–244, 1992.

[Hor93] J.F. Horty. Deontic logic as founded in nonmonotonic logic. *Annals of Mathematics and Artificial Intelligence*, 9:69–91, 1993.

[Hor94] J.F. Horty. Moral dilemmas and nonmonotonic logic. *Journal of Philosophical Logic*, 23:35–65, 1994.

[Lan96] J. Lang. Conditional desires and utilities - an alternative approach to qualitative decision theory. In *Proceedings of the ECAI'96*, pages 318–322, 1996.

[Mak93] D. Makinson. Five faces of minimality. *Studia Logica*, 52:339–379, 1993.

[Mak98] D. Makinson. On a fundamental problem of deontic logic. In *Proceedings of the ΔEON'98*, pages 3–42, 1998.

[Mor96] M. Morreau. *Prima Facie* and seeming duties. *Studia Logica*, 57:47–71, 1996.

[Pea93] J. Pearl. From conditional oughts to qualitative decision theory. In *Proceedings of the UAI'93*, pages 12–20, 1993.

[PS96] H. Prakken and M.J. Sergot. Contrary-to-duty obligations. *Studia Logica*, 57:91–115, 1996.

[PS97] H. Prakken and M.J. Sergot. Dyadic deontic logic and contrary-to-duty obligations. In D. Nute, editor, *Defeasible Deontic Logic*, pages 223–262. Kluwer, 1997.

[Ros30] D. Ross. *The Right and the Good*. Oxford University Press, 1930.

[TvdT95] Y.-H. Tan and L.W.N. van der Torre. Why defeasible deontic logic needs a multi preference semantics. In *Symbolic and Quantitative Approaches to Reasoning and Uncertainty (ECSQARU'95)*, LNAI 946, pages 412–419. Springer, 1995.

[TvdT96] Y.-H. Tan and L.W.N. van der Torre. How to combine ordering and minimizing in a deontic logic based on preferences. In *Deontic Logic, Agency and Normative Systems (ΔEON'96)*, Workshops in Computing, pages 216–232. Springer, 1996.

[vdT94] L.W.N. van der Torre. Violated obligations in a defeasible deontic logic. In *Proceedings of the ECAI'94*, pages 371–375, 1994.

[vdT97] L.W.N. van der Torre. *Reasoning about Obligations: Defeasibility in Preference-based Deontic Logic*. PhD thesis, Erasmus University Rotterdam, 1997.

[vdT98] L.W.N. van der Torre. Labeled logics of conditional goals. In *Proceedings of the ECAI'98*, pages 368–369, 1998.

[vdTT95] L.W.N. van der Torre and Y.H. Tan. Cancelling and overshadowing: two types of defeasibility in defeasible deontic logic. In *Proceedings of the IJCAI'95*, pages 1525–1532, 1995.

[vdTT97] L.W.N. van der Torre and Y.H. Tan. The many faces of defeasibility in defeasible deontic logic. In D. Nute, editor, *Defeasible Deontic Logic*, pages 79–121. Kluwer, 1997.

[vdTT98a] L.W.N. van der Torre and Y.-H. Tan. The temporal analysis of the Chisholm paradox. In *Proceedings of the AAAI'98*, 1998.

[vdTT98b] L.W.N. van der Torre and Y.-H. Tan. An update semantics for deontic reasoning. In *Proceedings of the ΔEON'98*, pages 409–426, 1998.

[Vel96] F. Veltman. Defaults in update semantics. *Journal of Philosophical Logic*, 25:221–261, 1996.

[Wey92] E. Weydert. Hyperrational conditional logic. In *Proceedings of the ECAI'92 Workshop on Theoretical Foundations of Knowledge Representation and Reasoning*. Springer, 1992.

[WM93] R.J. Wieringa and J.-J.Ch. Meyer. Applications of deontic logic in computer science: A concise overview. In J.-J. Meyer and R. Wieringa, editors, *Deontic Logic in Computer Science*, pages 17–40. John Wiley & Sons, Chichester, 1993.