Two Approaches to the Formalisation of Defeasible Deontic Reasoning

Henry Prakken*
Computer/Law Institute
Faculty of Law, Free University
De Boelelaan 1105, 1081 HV Amsterdam
The Netherlands
email: henry@rechten.vu.nl

April 26, 1996

Abstract

This paper compares two ways of formalising defeasible deontic reasoning, both based on the view that the issues of conflicting obligations and moral dilemmas should be dealt with from the perspective of nonmonotonic reasoning. The first way is developing a special nonmonotonic logic for deontic statements. This method turns out to have some limitations, for which reason another approach is recommended, viz. combining an already existing nonmonotonic logic with a deontic logic. As an example of this method the language of Reiter's default logic is extended to include modal expressions, after which the argumentation framework in default logic of [20, 22] is used to give a

*An earlier version of this article was written while the author was working at the Department of Computing, Imperial College London, supported by ESRC/MRC/SERC Joint Council Initiative Project G9212036. Work on the present version was supported by a research fellowship of the Royal Netherlands Academy of Arts and Sciences, and by Esprit WG 8319 'Modelage'. I thank one of the referees for his interesting comments. Also, many thanks are due to Marek Sergot for valuable discussions on the topic of this paper.

plausible logical analysis of moral dilemmas and prima facie obligations.

1 Introduction

In recent years the study of deontic logic has received new impulses by the use of techniques of nonmonotonic logics. The focus has in particular been on the problem of moral dilemmas, or conflicting obligations, arising from what W.D. Ross [25] has called 'prima facie' principles. A collection of papers on this issue is [6].

What exactly is the problem? In standard deontic logic (SDL), a normal modal system of type KD in the classification of Chellas [2] the validity of the scheme D^*: [1]

$$\neg(\mathrm{O}A \wedge \mathrm{O}\neg A)$$

is the only way of obtaining the desirable validity of the scheme D:

$$\neg\mathrm{O}\bot$$

However, many regard the validity of D^* as unacceptable, on the ground that in life's daily circumstances people are often faced with what seem to be genuine conflicts of obligations, arising from general 'prima facie' principles, whereas D^* is taken to mean that such genuine conflicts do not exist.

Various modal deontic logics have been proposed in which D^* is invalid but D valid ([2, 26, 16]) but, as argued by Horty in [8, 9, 10], these logics seem to be too weak. Consider, for example, 'You should either serve in the army or do alternative service' and 'You ought not to serve in the army': intuitively, this seems to imply 'You should perform alternative service', but in all proposed logics without D^* also the scheme

$$(\mathrm{O}(A \vee B) \wedge \mathrm{O}\neg A) \rightarrow \mathrm{O}B$$

is invalid. Horty observes that if O validates the rule RK of consequential closure:

$$\frac{A_1 \wedge \cdots \wedge A_n \rightarrow B}{\mathrm{O}A_1 \wedge \cdots \wedge \mathrm{O}A_n \rightarrow \mathrm{O}B} \qquad (n \geq 0)$$

[1] The names of the schemes in this article are based on [2].

then this inference would be valid if the following principle, called C, were valid as well:

$$(\mathrm{O}A \wedge \mathrm{O}B) \rightarrow \mathrm{O}(A \wedge B)$$

However, it is easy to see that if D and RK are valid, the invalidity of D^* invalidates C as well. Horty concludes - and I agree - that what we would want to have is a restricted version of C, valid only if it does not violate D.

In order to obtain this, Horty shifts the perspective: instead of designing a modal logic weaker than SDL, he regards deontic rules as *defeasible*, i.e. as subject to unforeseen exceptions and to running into conflicts with other rules. This point of view makes it possible to apply results from artificial-intelligence research on common-sense reasoning, in the form of so-called nonmonotonic logics. A basic feature of these logics is that they allow for 'jumping to conclusions' on the basis of general, defeasible rules if no conflicting information is available; the other side of this is that nonmonotonic inferences may have to be withdrawn if we come to know more.

The nonmonotonic perspective is particularly realistic for rules meant for everyday life. In formulating or discovering the rules on which to act in their daily circumstances, people have to cope with their limited abilities and resources: for humans it is impossible to foresee the entire future, in particular to anticipate every possible exception to or every possible collison between the rules. Therefore people often have to jump to their conclusions on the basis of general, defeasible rules, subject to possible conflicts and exceptions.

Before the rise of artificial–intelligence research, these observations had already been made by practical philosophers (e.g. [25]) and legal theorists (e.g. [7]). In the field of artificial intelligence and law several researchers have advocated the use of nonmonotonic logics (see e.g. [4, 5] and, for an overview, [20]). In [22, pp. 331–2] I have, moreover, suggested to apply this view to the issue of moral dilemmas and prima facie obligations. Horty's proposal is included in the same ΔEON–91 proceedings ([8]); there he also presents a first formalisation involving deontic operators.

Although thus the starting points of my suggestion and Horty's analysis are the same and although, moreover, both are based on Reiter's [24] default logic, there is an important difference. While the logic developed by Horty is in fact (like e.g. [12]) a special defeasible logic for deontic reasoning, I suggest (like e.g. [15] and [17]) to combine a deontic logic with an already

existing general nonmonotonic formalism. The aim of this paper is to assess the merits of these two strategies of formalising defeasible deontic reasoning. I will compare the approaches in two stages. First I will carry out a case study, by developing my suggestion of [22] in more detail and comparing the result to Horty's logic. After that I will evaluate the two approaches in more general terms.

2 Horty's nonmonotonic deontic logic

2.1 The logic

One of the best known formalisations of nonmonotonic reasoning is Reiter's [24] default logic; this is the system used by Horty in developing his nonmonotonic deontic logic, which in turn is inspired by his logical reconstruction in [8, 10] of a proposal of van Fraassen [3]. First I give a very brief outline of default logic (using the '(F, Δ)' notation of [22]). It is based on a set F of first-order formulas and a set Δ of *defaults*, which are inference rules of the form $A : B/C$, in which A is the *prerequisite*, B the *justification*, and C the *consequent*. Informally, this reads as 'If A holds and B may be consistently assumed, C may be inferred'. New beliefs can be derived by using ground instances of any default of Δ, as long as consistency is preserved. If as many defaults as possible are thus used, i.e. if applying any new default would cause an inconsistency, sets result which are called *extensions* of (F, Δ). Since defaults can conflict, a default theory may have several, mutually inconsistent, extensions.

I will discuss Horty's views as they are presented in their most advanced form in [9]. In his system Horty assumes as given an 'ought context' $\langle F, \Delta_\Gamma \rangle$: F is a single first-order formula, representing the factual information, and Γ is a set of conditional ought sentences of the form $\mathrm{O}(A/B)$, standing for 'A ought to be in case of B'. These definitions are a further development of his analysis of unconditional oughts $\mathrm{O}A$ as being Reiter-defaults $\top : A/A$. Therefore $\mathrm{O}(A/B)$ can be read as a Reiter-default $B : A/A$.

In order to let more specific conditionals override more general ones, Horty defines a conditional ought $\mathrm{O}(A/B)$ to be *overridden* in the context $\langle F, \Delta_\Gamma \rangle$ just in case there is an ought $\mathrm{O}(A'/B')$ such that

1. $F \vdash B', B' \vdash B$ and $B \nvdash B'$; and

2. $\{F, A', A\}$ is inconsistent; and

3. $\{F, A'\}$ is consistent.

Then a *conditioned extension* of $\langle F, \Delta_\Gamma \rangle$ is defined as a set E such that there is another set S such that

$$S = \{A \mid \begin{array}{l} \mathrm{O}(A/B) \in \Gamma, \\ F \vdash B, \\ \mathrm{O}(A/B) \text{ is not overridden in } \langle F, \Delta_\Gamma \rangle, \\ \neg A \notin E\}, \end{array}$$

and $E = Th(S \cup \{F\})$.

Then, after showing that every ought context has a conditioned extension, Horty defines

$\mathrm{O}(A/B)$ *is true with respect to* Γ *iff* $A \in E$ *for some conditioned extension* E *of* $\langle B, \Delta_\Gamma \rangle$.

Finally, Horty defines $\mathrm{O}A$ as $\mathrm{O}(A/\top)$.

It is easy to see that this account of obligation satisfies Horty's requirements (including their conditional counterparts). That D is valid is immediate from the last condition on S, and the validity of the restricted form of C follows from the fact that if two formulas are in the same conditioned extension, also their conjunction is in that extension. Finally, the invalidity of D^* and of the unrestricted version of C follows from the possibility, just as in default logic, of multiple, mutually inconsistent extensions. For instance, the ought context $\Gamma = \langle B \wedge C, \{\mathrm{O}(A/B), \mathrm{O}(\neg A/C)\} \rangle$ has two conditioned extensions: $Th\{A, B, C\}$ and $Th\{\neg A, B, C\}$. Then by definition both $\mathrm{O}(A/B \wedge C)$ and $\mathrm{O}(\neg A/B \wedge C)$ are true with respect to Γ, whereas $\mathrm{O}(A \wedge \neg A/B \wedge C)$ is false with respect to Γ.

2.2 Criticism

Granted that a nonmonotonic perspective on moral dilemmas is fruitful, is Horty's system a promising approach to the formalisation of defeasible deontic reasoning? Horty, although listing a number of open problems, claims it

is. In my opinion, however, there are serious reasons for taking another approach, since the above-sketched logic (let us call it H) seems to be inherently unable to deal with some very common forms of deontic reasoning. A first problem is that in H no satisfactory analysis of explicit permissions seems to be possible. In standard deontic logic PA is defined as $\neg \mathrm{O} \neg A$, which seems reasonable. However, if in H we allow explicit permissions of this kind in the set Γ of background oughts, there is a problem, since Horty's treatment of oughts makes them collapse into non-deontic Reiter- defaults, for which reason negated background oughts have no reasonable Reiter-default counterpart. Therefore, in H permissions cannot be expressed as premises. There is a way to express a kind of permission as a conclusion, viz. as $\neg \mathrm{O}(\neg A/B)$; however, this is what is often called *weak permission*, the absence of an obligation to A, and that is not the same as SDL's permission, which captures the notion of an *explicit* permission of A. An important difference between these two kinds of permissions is that if A is weakly permitted, then a subsequent prohibition of A just regulates something that was not yet regulated, while if A is explicitely permitted, a prohibition introduces a normative conflict.

Another problem is that in its present form H can only be used for defining relations between ought statements, since it does not distinguish between what is and what should be the case. One thing that cannot be expressed is 'factual detachment', i.e. the derivation of an unconditional obligation OA from a conditional obligation O(A/B) and B. Moreover, it is also impossible to express that an obligation has been violated (in fact, H even validates O(A/A)); among other things this means that H cannot express contrary-to-duty imperatives, another well–studied topic in deontic logic.[2]

A third problem has to do with combining factual and deontic defaults. In both legal and moral reasoning deontic conditionals are not the only defeasible conditionals; very often their antecedent is itself derived by another rule, often called a 'classification rule' or an 'interpretation rule'. Particularly in the legal field it is widely accepted that also these rules are defeasible (see e.g. [7, 4]). To extend Hart's standard example on a park regulation forbidding vehicles to enter the park: not only this rule itself may turn out to be defeasible, for example, if the vehicle is an ambulance, but also rules

[2]In [28] H is extended to cope with this problem but in a way which is not quite satisfactory: firstly, violations cannot be expressed in the logical object language; and second, the resulting system still suffers from the other problems discussed in this section.

on when something counts as a vehicle may be defeasible: imagine that a court says that objects on wheels that are meant for normal transport are vehicles: then roller skates used by people on their way to the office might be recognised as an exception. Now since, as just described, H makes deontic defeasible conditionals collapse into factual Reiter-defaults, there seems to be no simple way to capture combined reasoning with classification rules and deontic rules, which is clearly a severe restriction of the logic when it has to be applied to realistic examples.

The heart of the problems seems to be that in the present form of H, default logic can only be used for the formalisation of defeasible conditionals which are deontic. Although one might, of course, try to extend H, also an alternative and perhaps easier route suggests itself. Why not use default logic as it is, with the only change that the language on which it is based, first-order predicate logic, is extended with modal deontic operators? Then deontic conditionals do not collapse into factual conditionals. An additional advantage is that this solution does not depend on the existence of a special deontic kind of defeasibility, for which as yet there is no convincing evidence. In fact, this is the solution I shall explore in the rest of this paper. Before that, however, one possible objection should be discussed, viz. that when we combine default logic with standard deontic logic, D^* is valid again or in other words, we have no satisfactory way of maintaining that genuine conflicts of duties are possible.

3 Moral dilemmas and consistency

Obviously, combining default logic with SDL makes no sense if SDL unjustly validates D^*. However, in my view it is not at all obvious that D^* should be given up; in this section I will suggest some reasons why, at least for certain kinds of ought statements, also the validity of this principle can be reasonably defended. Let us look more closely at the most common objections to the validity of D^*.

Of course, it cannot be denied that in the practice of everyday life people's actions are governed by many legal or quasilegal regulations, moral codes and so on. However, as argued by Alchourrón in [1], one should not confuse the consistency of a *description* of such a state of affairs with the consistency of the described norms themselves: it is perfectly possible to consistently

express the fact that contradicting rules apply to a certain situation, in the same way as it is possible to consistently say that people have contradicting factual beliefs. I have the impression that sometimes when people refer to the existence of moral dilemmas, they do not clearly distinguish these two situations.

However, the factual existence of conflicting obligations is not the only argument given against D^*; sometimes it is argued that when people find themselves in a moral dilemma they really feel bound by both of the conflicting obligations and this feeling is not accounted for by a theory which regards at most one of the obligations as holding and which thus, so the argument goes, dismisses such feelings as unmotivated or even irrational.

I doubt whether an explanation of such feelings really requires the invalidity of D^*; in my view also an alternative view is possible, employing a more pragmatic view on the effect of contradictions. That in one particular situation a rule is dropped to maintain consistency does not mean that it has no binding force at all, since in other, unproblematic situations it can still be applied. I see no compelling reasons why the binding force of a deontic rule should be equated with its application to every single occasion. Here we can benefit from artificial-intelligence research on nonmonotonic reasoning. The general result of this work has been a more flexible view on the role and effect of contradictions: many logics have been developed in which a contradiction does not make a body of information completely useless, and in which it is possible to reason with preferences on how to resolve the contradiction. If we use one of these logics (as I will do in Section 4), we can also explain the binding force of a deontic rule involved in a conflict as: 'if there had been no conflict, we would have accepted the rule's consequent without discussion, and so we will in future situations without conflict'.

It might be argued that it is nevertheless desirable to have the syntactic means to express the binding force of conflicting obligations in the language itself. This is indeed an interesting point but if the O-operator has to fulfil this purpose, then in my view it captures a weak notion of 'ought', which does not cover all uses of this term, certainly not the one of legal reasoning: OA then means something like 'there is a reason to do A, even if we might be obliged to do $\neg A$'. I agree that for this notion of 'ought' D^* should indeed be invalid; however, it is very important to realise that this notion, rather than being the usual notion of 'ought', which should determine 'the' logic of obligation, concerns only certain types of ought–statements; other types

may very well validate D^*.

To summarise this section, I have tried to give some reasons why accepting D^* is at least defensible: firstly, one should be careful in distinguishing between *describing* and *expressing* norms; secondly, research in nonmonotonic logic allows for a more flexible view on the effects of inconsistent information; and finally, there seem to be different senses of 'ought', some of which validate and some of which invalidate D^*.

4 An argumentation framework in deontic default logic

I will now discuss the second strategy of formalising defeasible deontic reasoning: combining an existing deontic logic with an existing nonmonotonic logic. As the deontic logic I will use SDL; this choice is for convenience only: the story is the same for other deontic logics validating C and D^*. As the nonmonotonic formalism I shall use my formal argumentation framework based on default logic ([20, 22]). Essentially, this framework adds two things to default logic: it uses the notion of an argument and it provides a way for expressing *preferences* between conflicting arguments. Although for making the general points also other formalisms may be suitable, the use of the framework is attractive for at least two reasons: the link with default logic facilitates a transparent comparison with Horty's logic, while the notion of an argument seems to fit nicely with the informal notion of a moral dilemma.

In this section I will first consider the extension of default logic to modal logics, then I sketch my framework, and finally I will apply it to the issues in deontic reasoning that are discussed by Horty.

4.1 Modal default logic

Although Reiter [24, pp. 93–4] explicitly wants to stay within the setting of first-order predicate logic, nothing in his system prevents the use of other underlying logics. I now consider an extension to modal predicate logics at least as strong as K. It is easy to verify that this extension is straightforward, apart from one technical problem. In Reiter's treatment of so-called open defaults, i.e. defaults containing free variables, an essential element is the skolemisation of first-order formulas (see [24, pp. 115–18]), and skolemisation

is problematic for modal predicate logics invalidating the so-called Barcan formula

$$\Diamond\exists x Px \rightarrow \exists x \Diamond Px \ .$$

The reason is that if we skolemise $\Diamond\exists x Px$ as $\Diamond Pa$ where a is a Skolem constant, we can subsequently derive $\exists x \Diamond Px$. Therefore the extension of default logic to modal logics only works for modal logics validating the Barcan formula.

4.2 The framework

I now give an outline of the theory developed in [20, 22]. This system is one of several formal argumentation systems that have been developed in the past few years (for an overview and comparison see [20, Ch. 9] or [29, Ch.9]). Unlike most other systems, my framework was designed with applications to normative, in particular legal argumentation, in mind.

The framework is based on the normal part of default logic, i.e. the justification and the consequent of a default are assumed to be identical. Below normal defaults $A : B/B$ will be written as $A \Rightarrow B$; formulas of the form $\Rightarrow A$, which is shorthand for $\top \Rightarrow A$, are used for representing unconditional defeasible rules. The framework consists of four parts: the first concerns the notion of an *argument*, the second says when arguments are in *conflict*, the third is about ways of *comparing* arguments and the final part defines what it means that an argument is *justified*. The framework will be sketched against the background of a fixed default theory (F, Δ). The interesting case is, of course, when this default theory has multiple extensions.

To start with *arguments*, they are essentially the same as Reiter's [24] default proofs.[3] Let for any set D of defaults $PRE(D)$ and $CONS(D)$ respectively be the sets of all prerequisites and of all consequents of D. Then an argument is defined as a finite sequence $D_0, \ldots, D_n$ of sets of ground instances of D, such that

1. For $1 \leq i \leq n, F \cup CONS(D_i) \vdash \varphi$ for all $\varphi \in PRE(D_{i-1})$;

[3] Alternative definitions are possible, but for the present purposes this is not essential; my present concern is to have a formalism based on the general approach that can be compared to H in its application to defeasible deontic reasoning.

2. $D_n = \emptyset$;

3. $\cup_{i=0}^{n} CONS(D_i) \cup F$ is consistent.

Informally, D_1 collects all defaults that are 'directly' used to derive the conclusion, i.e. of which the joint consequents (together with the facts) deductively imply the conclusion; D_2 then collects those defaults that are in the same way used to 'directly' derive the prerequisites of the defaults in D_1, and so on, until we can 'ground' our argument in the facts.

For any argument $A = D_0, \ldots, D_n$ a *subargument* of A is an argument $A' = D'_0, \ldots, D'_n$ such that for each $i (0 \leq i \leq n), D'_i \subseteq D_i$. Furthermore, A' is a *proper* subargument of A iff $A' \neq A$. Every formula implied by $\cup_{i=0}^{n} CONS(D_i) \cup F$ is a *conclusion* of A, while it is also a *final* conclusion of A iff it is not a conclusion of a proper subargument of A.

The second main notion of the framework is that of a conflict between arguments. This is defined in terms of the *final conclusions* of an argument. An argument A_1 *attacks* an argument A_2 iff A_1 and A_2 have contradictory final conclusions and A_1 (or A_2) does not have any conclusion φ such that $\neg\varphi$ is a conclusion of a subargument of A_2 (or A_1)[4]. Note that this implies that A_1 attacks A_2 iff A_2 attacks A_1.

The third building block is a way of comparing pairs of arguments. What is very important is that this is assumed to be done according some *unspecified* standard of defeat, provided by the user of the framework, and only having some minimal formal properties, for instance that it is asymmetric and noncircular. In particular, the defeat relation is not required to express a notion of specificity (although it can be used for that purpose; cf. [18, 20]). The reason for this is that in reality arguments are compared on many different grounds: for example, in law arguments are also, and even with higher priority, compared with respect to the hierarchical status of the rules involved and with respect to the time of their enactment. This is not typical for legal reasoning, nor even for deontic reasoning in general; for instance, in London's underground stations hand-written factual information on movable whiteboards is intended to override information printed on fixed signs, regardless of whether the hand-written information is more specific or not. In conclusion, although testing for specificity may be a logical matter, deciding

[4]The second condition of this definition prevents the possibility of saving a defeated argument by extending it in some suitable way.

to *prefer* the most specific argument is a matter of content. Accordingly, the framework regards the criteria for comparing arguments, in addition to the facts and defaults, as a third category of 'input' provided by the user.

The final main element of the framework is the definition of a justified argument, i.e. of an argument with which a dispute can be won. In order to reflect the step-by-step nature of argumentation this notion is defined inductively: the idea is that in each inductive step arguments attacking each other are only compared with respect to their final conclusions, which idea is captured by clause **??** of the definition; intermediate conclusions should already have been justified at earlier steps in the induction, which is expressed by clause **??**. A further idea captured by clause **??** is that an argument which is not itself better than a counterargument can still be saved, or reinstated by another argument which *is* better than this counterargument. To avoid the definition being circular, it is expressed as stating conditions on *sets* of arguments rather than on individual arguments.

The set of *justified arguments* is the smallest set JA of arguments such that $A \in JA$ iff

1. All proper subarguments of A are in JA; and

2. A defeats all arguments A' that attack A and that are such that neither A' nor a subargument of A' are defeated by another argument in JA.

A very important aspect of this definition is that it divides arguments into three classes. The first class is, of course, that of *justified* arguments. Furthermore, if there are arguments which defeat other arguments, there are, of course, also arguments which are *overruled*; formally, they are defined as the arguments which are attacked by a justified argument. Finally, the definition leaves room for a nonempty class of arguments which are neither justified, nor overruled, but merely *defensible*. Technically, the significance of the notion of defensible arguments is that an argument needs not itself be justified in order to prevent a counterargument from being justified; it needs merely be defensible. Philosophically, the notion will be important in the analysis of moral dilemmas.

The following example illustrates the definitions. Consider a bureaucratic institution of which one official says that requests of customers need not be answered, while a higher official says that written requests must be answered.

Moreover, there are two conflicting precedents on what counts as a written request. This is formalised as follows, where $F = \{f\}$ and $\Delta = \{d_1 - d_4\}$.

d_1: $By\_fax \Rightarrow Written$
d_2: $\neg By\_mail \Rightarrow \neg Written$
d_3: $Request \Rightarrow \neg \mathrm{O} Answer$
d_4: $Written \Rightarrow \mathrm{O} Answer$
f: $Request \wedge By\_Fax \wedge (By\_fax \rightarrow \neg By\_mail)$

Assume that the defeat relations between the arguments are as follows. No defeat relation holds between $A_1 = \{d_1\}, \emptyset$ and $A_2 = \{d_2\}, \emptyset$, while, given the ranking of the officials, the argument $A_4 = \{d_4\}, \{d_1\}, \emptyset$ defeats its counterargument $A_3 = \{d_3\}, \emptyset$. Note that of A_4 the proper subarguments are A_1 and $\emptyset$, while of A_1, A_2 and A_3 the only proper subargument is $\emptyset$. It is easy to see that $\emptyset$ is trivially justified. However, since A_1 and A_2 do not defeat each other and are not reinstated by other arguments, they are both not justified. Then by clause 1 also A_4 is not justified. One the other hand, since none of the arguments are attacked by a justified argument, they are all defensible.

Assume now that some authority decides that A_1 defeats A_2: then A_1 is justified, which in turn makes A_4 justified, since this argument defeats A_3 while now also all its subarguments are justified. Then A_2 and A_3 are overruled.

4.3 Comparison with Horty's logic

Let us now reconsider the problems discussed in Sections 1 and 2. First the scheme D^*: even if it holds for the deontic logic used in the framework, it does not hold for the notion of a defensible argument (although it does hold for the notion of a justified argument): if $F = \emptyset$ and $\Delta = \{\Rightarrow \mathrm{O}A, \Rightarrow \mathrm{O}\neg A\}$ then, even if the arguments $\{\Rightarrow \mathrm{O}A\}$ for $\mathrm{O}A$ and $\{\Rightarrow \mathrm{O}\neg A\}$ for $\mathrm{O}\neg A$ are defensible[5], there is no defensible argument for $\mathrm{O}(A \wedge \neg A)$. On the other hand, if A is replaced by a formula B such that $\{A, B\} \cup F$ is consistent, then there is a defensible (and in this case also justified) argument for $\mathrm{O}(A \wedge B)$. Hence, for defensible arguments the framework validates the restricted version of C desired by Horty.

[5]Here and below I leave $\emptyset$ implicit.

Next, the framework has no problems with expressing (defeasible) factual detachment: if there is an argument $D_1, \ldots, D_n$ for A and D contains $A \Rightarrow B$, then $\{A \Rightarrow B\}, D_1, \ldots D_n$ is an argument for B. Furthermore, in the present analysis deontic defeasible conditionals do not collapse into factual ones, which prevents the other two problems of H. Firstly, since defaults of the form $A \Rightarrow \mathrm{P}B$ can be expressed, there is no problem at all in expressing explicit permission. This makes it possible to express conflicts between obligations and permissions, such as 'You ought not to kill', versus 'You may kill in self-defence'. Secondly, there are also no problems with chaining factual and deontic defaults, as needs hardly be explained: from $F = \{A\}, \Delta = \{A \Rightarrow B, B \Rightarrow \mathrm{O}C\}$ an argument for $\mathrm{O}C$ can be constructed.

Let us now consider specificity. As explained above, there are good reasons for regarding it as only one of the many possible criteria which might (but need not!) be applied. My framework reflects this view, but in H specificity is the only criterion determining whether an obligation is overridden by another one. It must be said, however, that H's definition of 'overridden' can easily be adapted to capture other criteria.

Next the issue of transitivity, or chaining, of defeasible deontic conditionals should be discussed. Horty ([9, p. 82]) remarks that from two defeasible conditionals 'if A then ought B' and 'if B then ought C' a new conditional 'if A then ought C' should be nonmonotonically derivable and he regrets the invalidity of this derivation in his logic. Now note first that, although in default logic and therefore also in my framework it is impossible to derive from two defaults $A \Rightarrow B$ and $B \Rightarrow C$ a new default $A \Rightarrow C$, what can be done, at least if F contains A, is creating an argument for C with the first two defaults. Note also, however, that it is impossible to construct an argument for $\mathrm{O}C$ from the default theory $F = \{A\}, \Delta = \{A \Rightarrow \mathrm{O}B, B \Rightarrow \mathrm{O}C\}$. Now, the latter is in fact what Horty wants to be possible, but clearly this form of chaining is not the same as what is usually regarded as chaining, in which the consequent of the first and the antecedent of the second default are the same. In fact, the kind of chaining proposed by Horty requires the validity of 'deontic detachment', which is the derivation of 'it ought to be that B' from 'It ought to be that A' and 'given A it ought to be that B'. Whatever one's views are on this inference rule (see section 5 for a discussion), it seems reasonable to demand that a deontic logic can at least formally distinguish deontic from factual detachment. Because of the collapse of deontic into factual rules, H also fails in this respect.

Finally, let us return to the question how to account for the feeling of being bound by both horns of a moral dilemma. Consider

d_1: $Ax \Rightarrow \mathrm{O}Bx$
d_2: $Cx \Rightarrow \mathrm{O}\neg Bx$

Recall that open defaults serve as schemes for all their ground instances. Now if $F = \{Aa, Ca\}$ then we have an argument for $\mathrm{O}Ba$ and one for $\mathrm{O}\neg Ba$. Assume first that both arguments are defensible: then the binding force of both obligations can be accounted for by saying so. Assume now that the argument for $\mathrm{O}\neg Ba$ defeats the argument for $\mathrm{O}Ba$; then the binding force can be explained in two ways. One thing we can say is that if we had only known Aa, we would have had a justified argument for $\mathrm{O}Ba$. Moreover, we can say that since d_2 is still in Δ, all its ground instances not involved in a conflict are still applicable. For example, if we add the formula Ab to F, then we still have a justified argument for Bb (and even one for $\neg Ba \wedge Bb$).

In conclusion, we can say that the framework of this section is more succesful in meeting Horty's demands than Horty's own logic. In particular, my framework has succeeded in preserving results of traditional deontic logic, rather than starting all over again from scratch. The only demand of Horty that has not been met is the invalidity of D^* but some reasons have been offered why this is not neccessarily a drawback.

Of course, my formalism is not the final answer to all the problems; for example, it inherits all the well-known problems of default logic and also as an argumentation framework it still needs further development (cf. [20, p. 207]). However, the system was not presented for its own sake; my aim has been to show by way of a concrete example how easy it is to meet Horty's demands if the 'general' approach is chosen. I now turn to a more general evaluation of the two approaches.

5 General evaluation

The aim of this paper is more general than comparing two particular logical systems: its purpose is to contrast two strategies of formalising defeasible deontic reasoning: designing a special nonmonotonic logic for deontic reasoning, and combining an existing nonmonotonic logic with an existing deontic logic. So far we have only criticised Horty's particular way of formalising the

'special' approach. However, the comparison also gives rise to more general observations. In particular, from the criticism of the collapse in H of deontic into factual defaults we can learn that any theory of defeasible deontic reasoning will have to combine an account of defeasible conditionals with an analysis of deontic operators. It is this issue that Horty's logic fails to address.

What can we say about such an analysis? To start with, in section 3 I have already argued that if nonmonotonic methods are used for the conditional part, in the deontic-operator part more of the traditional deontic logics can be retained than Horty and some others claim. Next the analysis has to face the question whether defeasible conditionals that are used in deontic contexts have different properties than conditionals that are used in other contexts of practical reasoning. If we cannot find such differences, it seems more attractive to follow the 'general' approach and to profit from results obtained in other fields.

There is one respect in which it has been claimed that deontic defeasible conditionals are special, and that is the issue of deontic detachment. Some, while acknowledging that the world is rarely perfect, argue that this principle should be understood as a defeasible inference rule, employed on the assumption that people at least *tend* to fulfil their obligations. In cases in which evidence to the contrary emerges, this assumption has to be retracted and this, then, is a form of nonmonotonic reasoning. This analysis has been defended by [13] and [28] (although in [13] still within a monotonic framework).

What are the merits of this way of looking at deontic detachment? Firstly, it should be remarked that it is debatable whether this principle is really a property of deontic *conditionals*: it seems that its underlying intuition can best be formalised by accepting a defeasible inference rule

$$\mathrm{O}A \sim\!\!> A$$

(where $\sim\!\!>$ denotes a defeasible inference relation). However, also apart from the question how deontic detachment should be formalised, in my view the defeasible interpretation of this principle is philosophically flawed. Let us look at the assumption underlying deontic detachment as thus conceived, viz. that the world is as *ideal* as possible given what we know. This assumption is radically different from the one underlying the usual forms of defeasible

reasoning, which is that the world is as *normal* as possible given what we know. The latter has a rational justification: assuming that things are as normal as possible reduces the likelihood of error. However, it is highly doubtful whether the same justification can be given for the assumption that the world is as ideal as possible.

In my view a better account of deontic detachment is as giving rise to obligations in a different sense, i.e. as reflecting that if something is the case, something else has gone wrong. To use the well-known Chisholm paradox: if you ought to go to help your neigbours, and if you ought to tell them that you are coming if you are coming, then if you don't tell them that you are coming, it can be inferred that you have not fulfilled all your obligations. This seems a fundamentally weaker notion than the usual sense of ought. For a formal analysis in this spirit see [11].

6 Conclusion

Summarising our comparison of the two approaches to formalising defeasible deontics, we have found no convincing philosophical evidence that there is a special deontic kind of defeasibility. Firstly, we have just seen that the interpretation of deontic detachment as a defeasible inference rule is philosophically flawed. Let us next go back to Horty's claim that a nonmonotonic perspective leads to a different deontic logic: what I have aimed to show is that this perspective, on the contrary, tells us that the traditional account of the deontic operators is in itself tenable, if only it is combined with a nonmonotonic logic and a related, more pragmatic attitude towards contradictions. I have shown that thus Horty's requirements for a logical analysis of moral dilemmas can very well be satisfied without the need to give up all the results of traditional deontic logic.

Yet historically the 'special' approach is understandable: within deontic logic moral dilemmas and prima facie obligations have long been the subject of debate and then the impression can easily arise that these are genuine issues of deontic reasoning. However, with hindsight we can say that these debates were forerunners of the later discussions within artificial intelligence on the logical nature of common-sense reasoning. Now if these issues within moral philosophy and legal theory coincide with more general issues within artificial intelligence, then it seems natural that also the more general *results*

of artificial intelligence are used.

I do not claim that the general strategy is completely free from technical problems. Above we have already seen that the combination of default logic and standard deontic logic works well only if the deontic logic validates the Barcan formula. Also other problems can arise, for example if both the defeasible and the deontic logic use preference relations on worlds. For instance, both [23] and [27] have suggested a way of analysing contary-to-duty imperatives with a graded preference ordering on worlds, as to how well they satisfy the standards of ideality. Now, as has been pointed out by [27], if such a deontic logic is combined with a nonmonotonic logic that uses normality relations on worlds (e.g. with [17]) then the issue of the interaction between the two relations should be addressed.

What I do claim, however, is that the 'general' approach to formalising defeasible deontics is not only philosophically but also methodologically more adequate: problems belonging to only one of the fields can be studied in isolation, and solutions to these problems are immediately available in the combination. In conclusion, it seems that in the absence of evidence for a special deontic kind of defeasibility the 'general' approach is preferable for the simple reason that it makes life easier.

References

[1] C.E. Alchourrón, Logic of norms and logic of normative propositions, *Logique ét Analyse* 12, 1969, 242-268.

[2] B. Chellas, *Modal logic: an introduction.* Cambridge University Press, 1980.

[3] B.C. van Fraassen, The logic of conditional obligation. *The Journal of Philosophical logic*, 1972, 417-438.

[4] A. von der L. Gardner, *An Artificial Intelligence approach to legal reasoning.* MIT press, 1987.

[5] T.F. Gordon, The importance of nonmonotonicity for legal reasoning. In H. Fiedler, F. Haft, R. Traunmüller (eds.), *Expert systems in law.* Tübingen, 1988, 110- 126.

[6] C.W. Gowans (ed.), *Moral dilemmas*. Oxford 1987.

[7] H.L.A. Hart, *The concept of law*. Clarendon Press, Oxford, 1961.

[8] J.F. Horty, Moral dilemmas and nonmonotonic logic (preliminary report). *Proceedings of the First International Workshop on Deontic Logic and Computer Science*, Amsterdam 1991, 212-231.

[9] J.F. Horty, Nonmonotonic techniques in the formalisation of commonsense normative reasoning. *Proceedings Workshop on Nonmonotonic Reasoning*, Austin, TX 1993, 74-84.

[10] J.J. Horty, Moral dilemmas and nonmonotonic logic. *Journal of Philosophical Logic* 23 (1994), 35-65.

[11] A.J.I. Jones and I. Pörn, Ideality, sub- ideality and deontic logic. *Synthese* 65 (1985), 275-290.

[12] A.J.I. Jones, Towards a formal theory of defeasible deontic conditionals. *Annals of Mathematics and Artificial Intelligence* 9 (1993), 151-166.

[13] B. Loewer and M. Belzer, Dyadic deontic detachment. *Synthese* 54 (1983), 295-318.

[14] D. Makinson, Five faces of minimality. *Studia Logica* Vol. 52, no. 3, 1993, 339-379.

[15] L.T. McCarty, Defeasible deontic reasoning. *Fundamenta Informaticae* 21 (1994), 125-148.

[16] J.-J.Ch. Meyer, A different approach to deontic logic: deontic logic viewed as a variant of dynamic logic. *Notre Dame Journal of Formal Logic* 29 (1), 1988, 109-136.

[17] M. Morreau, Prima facie and seeming duties. *Proceedings of the second International Workshop on Deontic Logic and Computer Science*, Tano, Oslo, 1994, 221-251.

[18] H. Prakken, A tool in modelling disagreement in law: preferring the most specific argument. *Proceedings of the Third International Conference on Artificial Intelligence and Law*, Oxford 1991. ACM Press 1991, 165-174.

[19] H. Prakken, Reasoning with normative hierarchies. *Proceedings of the First International Workshop on Deontic Logic and Computer Science*, Amsterdam 1991, 315-334.

[20] H. Prakken, *Logical tools for modelling legal argument.* Doctoral dissertation Free University Amsterdam, 1993.

[21] H. Prakken, A logical framework for modelling legal argument. *Proceedings of the fourth International Conference on Artificial Intelligence and Law*, Amsterdam 1993, ACM Press, 1993, 1-10.

[22] H. Prakken, An argumentation framework in default logic. *Annals of Mathematics and Artificial Intelligence* 9 (1993), 93-132.

[23] H. Prakken and M.J. Sergot, Contrary-to-duty obligations. This volume.

[24] R. Reiter, A logic for default reasoning. *Artificial Intelligence* 13 (1980), 81-132.

[25] W.D. Ross, *The right and the good.* Oxford University Press, Oxford, 1930.

[26] P.K. Scotch and R.E. Jennings, Non-kripkean deontic logic. In R. Hilpinen (ed.): *New studies in deontic logic.* Reidel, Dordrecht, 1982, 149-162.

[27] Y.-H. Tan and L.W.N. van der Torre, Multi preference semantics for a defeasible deontic logic. In H. Prakken, A. Muntjewerff, A. Soeteman, R.F. Winkels (eds.): *Legal knowledge based systems. The relation with legal theory.*. Koninklijke Vermande, Lelystad, 1994, 115-126.

[28] L.W.N. van der Torre, Violated obligations in a defeasible deontic logic. *Proceedings of the 11th European Conference on Artificial Intelligence (ECAI-94)*, 371-375.

[29] G. Vreeswijk, *Studies in defeasible argumentation.* Doctoral dissertation Free University Amsterdam, 1993.