

Preprint of a paper appearing in

*Journal of Philosophical Logic*

Volume 23 (1994), pp. 35--65.

# Moral Dilemmas and Nonmonotonic Logic

John F. Horty

Philosophy Department and  
Institute for Advanced Computer Studies

University of Maryland

College Park, MD 20742

(Email: [horty@umiacs.umd.edu](mailto:horty@umiacs.umd.edu))

## Contents

1	Introduction	1
2	Oughts and imperatives	2
3	Default logic	6
4	Imperatives as defaults	10
5	Conditional oughts	17
6	Problems with conditional oughts	21
7	Conclusion	28

# 1 Introduction

The purpose of this paper is to establish some formal connections between deontic and nonmonotonic logics, and to suggest some ways in which the techniques developed in the study of nonmonotonic reasoning and the issues confronted there might help to illuminate deontic ideas. These two subjects have evolved within different disciplines. The field of deontic logic was developed by philosophers and legal theorists as a high level framework for describing valid patterns of normative reasoning. The study of nonmonotonic logic was initiated, much more recently, by researchers in artificial intelligence who felt that ordinary logical techniques could not be applied properly to a number of practical problems arising within that area—most notably, problems involving planning and action, such as the frame problem. Even though the two subjects come from different disciplines, however, it is not really surprising that there should be close connections between them. Both are concerned, very broadly, with formalizing certain aspects of commonsense reasoning. Both recognize that many of the rules governing our commonsense reasoning are *prima facie*, or defeasible. And both must deal, in particular, with clashes among these defeasible rules.

Although I believe that the relations between deontic and nonmonotonic logic may be extensive, I focus here, narrowly, only on two particular theories. The first is the account of obligation sketched by Bas van Fraassen [26], which differs from standard deontic logic in allowing for moral conflicts; the second is Raymond Reiter’s default logic [20], one of the first nonmonotonic formalisms, and one of the most widely applied.

These two theories are reviewed in Sections 2 and 3. In Section 4, I show that van Fraassen’s account of simple (categorical) oughts can be interpreted in a natural way using the notion of default consequence introduced by Reiter. This is the primary positive contribution of the paper, establishing a concrete correspondence between particular theories from the two fields. Section 5 generalizes this correspondence to cover also a treatment of conditional oughts based on that sketched by van Fraassen; but it turns out that this account of conditional oughts is itself problematic. Section 6 describes some of the problems presented by conditional oughts, and shows that there are strong analogies between

these problems and difficulties already studied within the framework of default logic. These analogies suggest that the formal connection established here between nonmonotonic and deontic logics is not just a technical accident, but that it reflects a deeper level of common concern.

## 2 Oughts and imperatives

In standard deontic logic, obligation is interpreted as a kind of necessity, which can be modeled using familiar possible worlds techniques.<sup>1</sup> In the simplest deontic models, each possible world or situation is associated with a single, nonempty set of deontic alternatives—the set of situations in which, relative to the original, things go as they ought to, or in which all oughts in force in the original situation are satisfied. Where  $\bigcirc$  is the connective representing ‘It ought to be that ...’, a statement of the form  $\bigcirc A$  is then supposed to be true at a given situation whenever  $A$  is true at each of its deontic alternatives. The idea behind this valuation rule is that  $\bigcirc A$  should be true whenever  $A$  is a necessary condition for things turning out as they ought.

Let us say that a situation gives rise to a moral dilemma if it presents both of two conflicting propositions as obligatory. We often seem to find ourselves in such situations, and there are a number of very vivid examples in philosophy and literature; but if standard deontic logic is correct, all of this is misleading. Because it assigns to each situation only a single set of deontic alternatives (and because this set must be nonempty), standard deontic logic rules out the possibility of moral dilemmas. No situation can support conflicting oughts;  $\bigcirc A$  and  $\bigcirc \neg A$  are not jointly satisfiable. This conclusion follows at once from the semantics: if  $\bigcirc A$  is true at a situation, so that  $A$  is true at all of its deontic alternatives, then  $\neg A$  cannot be true at any of them. Another way of reaching the same result is to notice that, on the standard semantics,  $\bigcirc A$  and  $\bigcirc B$  imply  $\bigcirc(A \wedge B)$ , but that  $\bigcirc(A \wedge \neg A)$  is unsatisfiable.

There is currently no consensus among moral theorists on the question whether an ideal

---

<sup>1</sup>Standard deontic logic is discussed from a historical perspective in Føllesdal and Hilpinen [10]; it is presented from a more analytic point of view as the system *KD* in Chellas [2].

ethical theory could actually be structured in such a way that moral dilemmas might arise.<sup>2</sup> Still, it can seem like an objectionable feature of standard deontic logic that it rules out this possibility. Because the question is open, and the possibility of moral dilemmas is a matter for substantive ethical discussion, it seems to be inappropriate for a position on this issue to be built into the logic of the subject. And even if it does turn out, ultimately, that research in ethics is able to exclude the possibility of conflicts in a correct moral theory, it may be useful all the same to have a logic that allows for conflicting oughts. For one thing, the task of actually applying a correct moral theory to each of the ethical decisions we face every day would be difficult and time-consuming; and it seems unlikely, for most of us, that such a theory could have any more bearing upon our day to day ethical reasoning than physics has upon our everyday reasoning about objects in the world. Most of our commonsense ethical thinking seems to be guided instead, not by the dictates of moral theory, but by simple rules of thumb—‘Return what you borrow’, ‘Don’t cause harm’—and it is not hard to generate conflicts among these.<sup>3</sup> Moreover, practical reasoning more generally is conditioned by a number of oughts, many of which are founded in a concern with matters other than morality—etiquette, aesthetics, fun—and of course, these lead to other conflicts both among themselves and with the oughts of morality. Even if we do eventually conclude, then, that there can be no clashes among the oughts generated by a correct ethical theory, it still seems necessary to allow for conflicting oughts in any logic that aims to represent either our everyday moral thinking or our normative reasoning more broadly.

The best known proposal for weakening deontic logic to allow conflicts among oughts was set out in van Fraassen [26]. In fact, that paper contains two suggestions, the second a refinement of the first. Both involve a nonstandard mechanism for evaluating ought statements. Rather than assigning a truth value to these statements based on a primitive relation of deontic alternativeness among situations, van Fraassen evaluates them against a set of

---

<sup>2</sup>There is now a large literature on this topic, but a good sample of the conflicting positions can be found in Williams [27] and Donagan [4].

<sup>3</sup>The relation between moral theory and the rules of thumb that guide everyday ethical decisions has recently been discussed by Dennett [3].

background *imperatives*, which are supposed to represent the dictates of various sources of obligation. Of course, since a single agent might recognize conflicting sources of obligation, and since even the same source of obligation can at times issue conflicting commands, this picture must allow for the possibility that an agent might find himself constrained by a set of imperatives that cannot all be fulfilled at once.

In presenting van Fraassen's proposal, we will use an exclamation point as the imperative operator; the sentence  $!(\neg K)$ , for example, might represent the imperative

Thou shalt not kill.

Imperatives can be fulfilled or violated; we will say that an imperative  $!(A)$  is fulfilled in any situation in which  $A$  is true, and violated otherwise. We will use lowercase Greek letters  $(\alpha, \beta, \gamma, \dots)$  to represent the situations, or indices, at which imperatives and other formulas are evaluated. Usually in deontic logic, these situations are identified with possible worlds. However, because we are not concerned with iterated oughts, and in order to avoid extraneous complications when it comes to default logic, the situations will be interpreted here simply as ordinary valuations of the underlying language. Finally, the symbol  $\models$  will stand as usual for the satisfaction relation between situations and formulas; and where  $A$  is a formula and  $\mathcal{S}$  some set of formulas, we will let  $|A| = \{\alpha : \alpha \models A\}$  and  $|\mathcal{S}| = \bigcap\{|A| : A \in \mathcal{S}\}$ .

Now suppose that  $\mathcal{I}$  is the background set of imperatives governing some agent. Van Fraassen's initial suggestion is that a statement  $\bigcirc A$  is true with respect to  $\mathcal{I}$  just in case there is some imperative  $!(B)$  belonging to  $\mathcal{I}$  such that  $|B| \subseteq |A|$ . The idea here is that a proposition is obligatory if it is a necessary condition for fulfilling some imperative. This initial suggestion really contains the heart of the proposal. We can see already how conflicting propositions might both be obligatory, even though their conjunction is not. If the background imperative set is  $\mathcal{I}_1 = \{!(A),!(\neg A)\}$ , for example, then  $\bigcirc A$  and  $\bigcirc \neg A$  will both be true, but  $\bigcirc(A \wedge \neg A)$  will be false. In addition, it is clear that  $\bigcirc B$  will be a consequence of  $\bigcirc A$  whenever  $B$  is a consequence of  $A$ .

The initial suggestion runs into difficulties, however, when it comes to logical interconnections among imperatives. Suppose that an agent is constrained only by two imperatives:

Fight in the army or perform alternative service,  
 Don't fight in the army.

The first of these might issue from some piece of legislation to which the agent is subject; the second from religion or conscience. Let us represent this imperative set as  $\mathcal{I}_2 = \{!(F \vee S), !(\neg F)\}$ . Now it seems intuitively that  $\bigcirc S$  should be true with respect to  $\mathcal{I}_2$ ; given the imperatives governing his action, the agent ought to perform alternative service. Yet the initial suggestion does not yield this result, since there is no *single* imperative  $!(B)$  belonging to  $\mathcal{I}_2$  such that  $|B| \subseteq |S|$ .

To handle this kind of problem, van Fraassen introduces the notion of a situation's *score*, the set of imperatives it fulfills. Formally, where  $\mathcal{I}$  is the background set of imperatives, we let  $score_{\mathcal{I}}(\alpha) = \{!(A) \in \mathcal{I} : \alpha \models A\}$ . Using this new notion, he then refines his original suggestion in a way that leads to the following definition.

**Definition 1** *The formula  $\bigcirc A$  is true with respect to the imperative set  $\mathcal{I}$  just in case there is some  $\alpha \in |A|$  for which there is no  $\beta \in |\neg A|$  such that  $score_{\mathcal{I}}(\alpha) \subseteq score_{\mathcal{I}}(\beta)$ .*

According to this new proposal, a proposition is classified as obligatory if it is a necessary condition for fulfilling, not just a single imperative, but some maximal set of imperatives—a necessary condition for achieving some maximal score. The new proposal preserves the desired features of the earlier version: conflicts among obligations are allowed, without implying an obligation to do the impossible; and any consequence of an obligatory proposition is obligatory. However, the new proposal adds to the simpler version the idea that, whenever it is possible to satisfy more imperatives without violating those already fulfilled, it is best to do so. Because of this, for example, the formula  $\bigcirc S$  turns out to be true with respect to  $\mathcal{I}_2$  above, since any situation fulfilling both  $!(F \vee S)$  and  $!(\neg F)$  must fall within  $|S|$ .<sup>4</sup>

---

<sup>4</sup>The new proposal differs from the preliminary version also in its treatment of some pathological imperative sets. If the background imperative set happens to contain an inconsistent imperative, such as  $!(A \wedge \neg A)$ , then everything is obligatory according to the preliminary version, but on the refined analysis, this imperative has no effect on the agent's obligations. If the background imperative set is empty, then nothing is obligatory according to the preliminary version, but on the refined analysis the logical truths at least are obligatory.

### 3 Default logic

The study of nonmonotonic logics was initiated in the late 1970's, and the field solidified in 1980 with the publication of a special issue of *Artificial Intelligence* [1] devoted to the topic.<sup>5</sup> These logics have found applications in areas as diverse as database theory and automated diagnosis; but an important initial motive in their development was the need felt within artificial intelligence for a formalism more naturally suited than ordinary logic to model the tentative nature of commonsense reasoning. Often, it seems, we want to draw conclusions from a given body of data that we are willing to abandon when the data is supplemented with further information. To take a standard example, if we were told that Tweety is a bird, most of us would conclude that Tweety can fly—since we believe that, as a general rule, birds can fly. However, we would abandon this conclusion, and we would not feel that we had been presented with any kind of inconsistency, if we were then told in addition that Tweety cannot fly.

The particular formalism with which we are concerned here—Reiter's default logic [20]—models this phenomenon by supplementing ordinary logic with new rules of inference, known as *default rules*. In order to characterize the conclusion sets of theories involving these new rules, Reiter then modifies the standard, monotonic notion of logical consequence.

An ordinary rule of inference (with a single premise) can be depicted simply as a premise-conclusion pair, such as  $(A/B)$ . This rule commits a reasoner to  $B$  once  $A$  has been established. By contrast, a default rule is a triple, such as  $(A : C / B)$ . Very roughly, this rule commits the reasoner to  $B$  once  $A$  has been established and, in addition,  $C$  is consistent with the reasoner's conclusion set. The formula  $A$  is referred to as the *prerequisite* of this default rule,  $B$  as its *consequent*, and  $C$  as its *justification*. A *default theory* is a pair  $\Delta = \langle W, \mathcal{D} \rangle$ , in which  $W$  is a set of ordinary formulas and  $\mathcal{D}$  is a set of default rules.

Before going on to set out the new notion of a conclusion set defined by Reiter for default theories, let us see how the information given above about Tweety might be represented

---

<sup>5</sup>Many of the papers from this issue are reprinted in Ginsberg [11]. This collection contains also much of the most important work on nonmonotonic reasoning through 1987 and together with Etherington [8] currently serves as the best introduction to the field.

in default logic. The first case, in which we are told only that Tweety is a bird, can be represented by the default theory  $\Delta_1 = \langle \mathcal{W}_1, \mathcal{D}_1 \rangle$ , where  $\mathcal{W}_1 = \{Bt\}$  and  $\mathcal{D}_1 = \{(Bt : Ft / Ft)\}$ . Here the default rule says that if we know Tweety is a bird, and it is consistent with what we know that Tweety can fly, then we should conclude that Tweety can fly. (The generic statement ‘Birds fly’ can be taken to mean that, once we learn of some object that it is a bird, we should conclude that it flies, unless we happen to know that it does not. The default rule can then be thought of as an instantiation for Tweety of this generic truth.) In this case, because we do know that  $Bt$ , and there is no reason to think that  $Ft$  is inconsistent with what we know, the default rule yields  $Ft$  as a conclusion. Where  $Th$  is a function mapping any set of formulas to its logical closure, then, the appropriate conclusion set based on  $\Delta_1$  seems to be  $Th[\{Bt, Ft\}]$ , the logical closure of what we are told to begin with together with the conclusions of the applicable defaults. In the second case, however, when we are told in addition that Tweety does not fly, we move to the default theory  $\Delta_2 = \langle \mathcal{W}_2, \mathcal{D}_2 \rangle$ , with  $\mathcal{D}_2 = \mathcal{D}_1$  and  $\mathcal{W}_2 = \mathcal{W}_1 \cup \{\neg Ft\}$ . Here the default rule cannot be applied, because its justification is inconsistent with what we know. So the appropriate conclusion set based on  $\Delta_2$  seems to be  $Th[\mathcal{W}_2]$ .

These two examples illustrate, in some simple and natural cases, the kind of conclusion sets desired from given default theories. The task of arriving at a general definition of this notion, however, is not trivial; the trick is to find a way of capturing the intended meaning of the new component—the justification—present in default rules. A default rule is supposed to be applicable only if its justification is consistent with the conclusion set; but what does consistency mean? Consistency is usually defined in terms of logical consequence (a set is consistent if there is no explicit contradiction among its consequences), and so there is a danger of circularity here. In fact, the very application of a default rule might undermine its own justification, or the justification of some other rule that has already been applied. As an example, consider the theory  $\Delta_3 = \langle \mathcal{W}_3, \mathcal{D}_3 \rangle$ , with  $\mathcal{W}_3 = \{A, B \supset \neg C\}$  and  $\mathcal{D}_3 = \{(A : C / B)\}$ . Before any new conclusions are drawn from this information, the rule  $(A : C / B)$  seems to be applicable, since its prerequisite belongs already to the initial data set  $\mathcal{W}_3$ , and its justification is consistent with this set. The effect of applying

this rule, though, is to introduce  $B$  into the conclusion set; just a bit of additional reasoning then shows that the conclusion set must contain  $\neg C$  as well, and so the applicability of the default rule is undermined.

Of course, a chain of reasoning like this showing that some default rule is undermined can be arbitrarily long; and so we cannot really be sure that a default rule is applicable in some context until we have applied it, along with all the other rules that seem applicable, and then surveyed the logical closure of the result. Because of this, the conclusion set associated with a default theory cannot be defined in the usual iterative way, by successively adding to the original data the conclusions of the applicable rules of inference, and then taking the limit of this process.

Instead, Reiter is forced to adopt a fixed point approach in specifying the conclusion sets of default theories. He first defines an operator  $\Gamma$  that uses the information from a particular default theory to map formula sets into formula sets.

**Definition 2** *Where  $\Delta = \langle \mathcal{W}, \mathcal{D} \rangle$  is a default theory and  $\mathcal{S}$  is some set of formulas,  $\Gamma_{\Delta}(\mathcal{S})$  is the minimal set satisfying the following three conditions:*

1.  $\mathcal{W} \subseteq \Gamma_{\Delta}(\mathcal{S})$ ,
2.  $Th[\Gamma_{\Delta}(\mathcal{S})] = \Gamma_{\Delta}(\mathcal{S})$ ,
3. for each  $(A : B / C) \in \mathcal{D}$ , if  $A \in \Gamma_{\Delta}(\mathcal{S})$  and  $\neg B \notin \mathcal{S}$ , then  $C \in \Gamma_{\Delta}(\mathcal{S})$ .

The first two conditions in this definition tell us simply that  $\Gamma_{\Delta}(\mathcal{S})$  contains the information provided by the original theory, and that it is closed under logical consequence; the third condition tells us that it contains the conclusions of the default rules applicable in  $\mathcal{S}$ ; and the minimality constraint prevents unwarranted conclusions from creeping in.

Where  $\Delta = \langle \mathcal{W}, \mathcal{D} \rangle$  is a default theory, the operator  $\Gamma_{\Delta}$  maps any formula set  $\mathcal{S}$  into the minimal superset of  $\mathcal{W}$  that is closed under both ordinary logical consequence and the default rules from  $\mathcal{D}$  that are applicable in  $\mathcal{S}$ . The appropriate conclusion sets of default theories—known as extensions—are then defined as the fixed points of this operator.

**Definition 3** *The set  $\mathcal{E}$  is an extension of the default theory  $\Delta$  if and only if  $\Gamma_{\Delta}(\mathcal{E}) = \mathcal{E}$ .*

As the reader can verify, the default theories  $\Delta_1$  and  $\Delta_2$  above have the advertised conclusion sets as their extensions. In addition, it should be clear that the notion of an extension defined here is a conservative generalization of the corresponding notion of a conclusion set from ordinary logic: the extension of a default theory  $\langle \mathcal{W}, \mathcal{D} \rangle$  in which  $\mathcal{D}$  is empty is simply  $Th[\mathcal{W}]$ .

In contrast to the situation in ordinary logic, however, not every default theory leads to a single set of appropriate conclusions. Some default theories, such as  $\Delta_3$  above, have no extensions; these theories are often viewed as incoherent. More interesting, for our purposes, some lead to multiple extensions. A standard example arises when we try to encode as a default theory the following set of facts:

Nixon is a Quaker,  
 Nixon is a republican,  
 Quakers tend to be pacifists,  
 Republicans tend not to be pacifists.

If we instantiate for Nixon the general statements expressed here about Quakers and republicans, the resulting theory is  $\Delta_4 = \langle \mathcal{W}_4, \mathcal{D}_4 \rangle$ , with  $\mathcal{W}_4 = \{Qn, Rn\}$  and  $\mathcal{D}_4 = \{(Qn : Pn / Pn), (Rn : \neg Pn / \neg Pn)\}$ . This theory allows both  $Th[\mathcal{W}_4 \cup \{Pn\}]$  and  $Th[\mathcal{W}_4 \cup \{\neg Pn\}]$  as extensions. Initially, before we draw any new conclusions, both of the default rules from  $\mathcal{D}_4$  are applicable, but once we adopt the conclusion of either, the applicability of the other is blocked.

In cases like this, when a default theory leads to more than one extension, it is difficult to decide what conclusions a reasoner should actually draw from the information contained in the theory. One option discussed in the literature is to suppose that the reasoner should arbitrarily select one of the theory's several extensions and endorse the conclusions contained in it; another option is to suppose that the reasoner should endorse only those conclusions contained in the intersection of these extensions. For the purpose of modeling commonsense reasoning, the multiple extensions associated with default theories can sometimes seem like an embarrassment: what we really want is a unique conclusion set, and so we are forced

either to select nondeterministically from among these various extensions, or else to combine them somehow into a unique set. As we shall see, however, the multiple extensions provided by default logic are no longer embarrassing when it comes to interpreting deontic ideas; they give us exactly what we need.

## 4 Imperatives as defaults

Often, and in all of our examples so far, default rules seem to represent something like commonsense probabilistic generalizations. The defaults concerning birds or Quakers, for instance, seem to mean simply that a large majority of birds can fly, or that a large majority of Quakers are pacifists. The connection between defaults and generalizations of this kind has suggested to many that default reasoning can best be understood as a kind of qualitative probabilistic reasoning, a view that is most thoroughly developed by Judea Pearl [19].

There are, however, some important examples of default reasoning that do not seem to fit so naturally into the probabilistic framework. In driving along a narrow country road, for instance, it is best, whenever one approaches the crest of a hill, to adopt the default that there will be traffic in the oncoming lane, even if the road is deserted and the actual likelihood of traffic is low. Again, the presumption of innocence in a legal system is a kind of default that overrides probabilistic considerations: even if the most salient reference class to which an individual belongs is one among which the proportion of criminals is very high, we are to presume that he has committed no crime unless there is conclusive evidence to the contrary.<sup>6</sup>

Those who favor a probabilistic understanding of defaults can attempt to account for discrepancies like these between defaults and commonsense generalizations by supposing that default rules might reflect, in addition, information concerning utilities of the outcomes. (For example, it could be argued that the default concerning oncoming traffic is reasonable,

---

<sup>6</sup>The notion of presumption is discussed in detail by Ullman-Margalit [25], who argues that specific presumptions are justified by a mixture of probabilistic and “value-related” considerations, and cites the presumption of innocence as one in whose justification the value-related considerations seem to outweigh those of probability.

even though the likelihood is low, because the cost of a false negative in this case is potentially so high.) But there is also another explanation of the differences here between defaults and commonsense probabilistic generalizations. What these examples suggest is that default rules can be used to represent *norms* quite generally. When the norms involved have a probabilistic basis, it is natural to expect default reasoning to resemble probabilistic reasoning. But default rules can be used also, it seems, to represent other kinds of norms—such as legal or ethical norms—and in that case, any relation with probabilistic reasoning will be more distant.

It is this reading of defaults, as representing norms in general, that motivates the connections developed here between default and deontic logics, and in particular the central observation of this paper: if the norms generated by *imperatives* are represented through default rules, then van Fraassen’s theory of oughts can be interpreted in Reiter’s default logic.

Formally, the interpretation is straightforward: with each imperative set  $\mathcal{I}$  we associate a default theory  $\Delta = \langle \mathcal{W}, \mathcal{D} \rangle$ , where  $\mathcal{W} = \emptyset$  and  $\mathcal{D} = \{(: A / A) : !(A) \in \mathcal{I}\}$ . (A default rule written without a visible prerequisite should be taken to have as its prerequisite the universally true sentence  $\top$ .) It then turns out that the formula  $\bigcirc A$  is true with respect to  $\mathcal{I}$  just in case  $A$  belongs to some extension of  $\Delta$ .

In the course of establishing this result, we will appeal to two background facts about default logic. The first is simply that a default theory  $\langle \mathcal{W}, \mathcal{D} \rangle$  in which  $\mathcal{W}$  is itself consistent has only consistent extensions. This is well known, and was established in [20]. The second fact is more complicated, and we need some notation to state it. Where  $\mathcal{S}$  is some formula set and  $\Delta = \langle \mathcal{W}, \mathcal{D} \rangle$  is a default theory, we can define the *generating defaults* for  $\mathcal{S}$  with respect to  $\Delta$  as

$$GD(\mathcal{S}, \Delta) = \{(A : B / C) \in \mathcal{D} : A \in \mathcal{S} \text{ and } \neg B \notin \mathcal{S}\}.$$

(This definition is due to Reiter, although he restricts it to the case in which  $\mathcal{S}$  is an extension of  $\Delta$ .) Next, where  $\mathcal{D}$  is some set of defaults, we let

$$Con[\mathcal{D}] = \{C : (A : B / C) \in \mathcal{D}\}$$

stand for the consequents  $\mathcal{D}$ .

Now, using these concepts, Reiter shows in [20] that any extension of a default theory can be characterized as the logical closure of the initial information from the theory together with the consequents of the generating defaults for the extension. More formally, what he shows is that whenever  $\mathcal{E}$  is an extension of  $\Delta$ , we then have

$$(*) \quad \mathcal{E} = Th[\mathcal{W} \cup Con[GD(\mathcal{E}, \Delta)]].$$

Unfortunately, the converse of this result does not hold in general: there are cases in which  $\mathcal{E}$  and  $\Delta$  might satisfy  $(*)$  even though  $\mathcal{E}$  is not an extension of  $\Delta$ .<sup>7</sup> Although the converse does not hold in general, however, it can be shown to hold for a special class of default theories—those theories in which the prerequisite of each default rule is entailed by the initial information contained in the theory. This is our second background fact, which we record explicitly as a lemma.

**Lemma 1** *Let  $\Delta = \langle \mathcal{W}, \mathcal{D} \rangle$  be a default theory such that  $A \in Th[\mathcal{W}]$  for each default rule  $(A : B / C) \in \mathcal{D}$ . Then  $\mathcal{E}$  is an extension of  $\Delta$  if and only if  $\mathcal{E} = Th[\mathcal{W} \cup Con[GD(\mathcal{E}, \Delta)]]$ .*

This lemma will not be proved here; it follows almost immediately from the very helpful characterization of extensions found in Marek and Truszczyński [17]. What it tells us is that, whenever  $\Delta$  is a theory belonging to this restricted class, we might as well *define* the extensions of  $\Delta$  as those sets  $\mathcal{E}$  satisfying  $(*)$ . The result is useful for our present purposes, of course, because the default theories associated with imperative sets meet the restrictions of the lemma.

At this point, we can establish the connection between deontic and default logic. We move through three additional lemmas, leading up to the main theorem. Throughout, we assume that  $\mathcal{I}$  is some imperative set, and that  $\Delta$  is the associated default theory. In addition, we

---

<sup>7</sup>As an example, let  $\Delta_5 = \langle \mathcal{W}_5, \mathcal{D}_5 \rangle$ , where  $\mathcal{W}_5 = \emptyset$  and  $\mathcal{D}_5 = \{(A : A / A)\}$ , and let  $\mathcal{E} = Th[\{A\}]$ . Then  $\mathcal{E}$  and  $\Delta_5$  satisfy  $(*)$ , but the only extension of  $\Delta_5$  is  $Th[\emptyset]$ . Interestingly, extensions cannot be characterized even as the *minimal* sets satisfying  $(*)$ . To see this, take  $\Delta_6 = \langle \mathcal{W}_6, \mathcal{D}_6 \rangle$ , where  $\mathcal{W}_6 = \emptyset$  and  $\mathcal{D}_6 = \mathcal{D}_5 \cup \{(\neg A / A)\}$ . Now  $\mathcal{E} = Th[\{A\}]$  is a minimal set satisfying  $(*)$ , but it is not an extension of  $\Delta_6$ ; this theory has *no* extensions.

define  $score_{\Delta}(\alpha) = \{(: A / A) \in \mathcal{D} : \alpha \models A\}$ . Where  $\Delta$  is the default theory associated with  $\mathcal{I}$ , this notion of default-score obviously carries the same information as the notion of imperative-score from Definition 1, and could just as easily have been used there instead.

The first of these additional lemmas shows that whenever a valuation falls within the model class of some extension, the score of that valuation is equivalent to the generating default set of that extension. The second shows that a valuation whose score is never exceeded must fall within the model class of some extension. The third shows that a valuation falling within the model class of some extension cannot be exceeded or matched in score by a valuation falling outside of that model class.

**Lemma 2** *Let  $\alpha \in |\mathcal{E}|$  for some extension  $\mathcal{E}$  of  $\Delta$ . Then  $score_{\Delta}(\alpha) = GD(\mathcal{E}, \Delta)$ .*

**Proof** Suppose  $(: A / A) \in score_{\Delta}(\alpha)$ , so that  $(: A / A) \in \mathcal{D}$  and  $\alpha \models A$ . Since  $\alpha \in |\mathcal{E}|$ , it then follows at once that  $\neg A \notin \mathcal{E}$ ; for otherwise, we would have  $\alpha \models \neg A$ . Hence  $(: A / A) \in GD(\mathcal{E}, \Delta)$ . Next, suppose  $(: A / A) \in GD(\mathcal{E}, \Delta)$ . Since  $\mathcal{E}$  is an extension of  $\Delta$ , we know from Lemma 1 that  $\mathcal{E} = Th[Con[GD(\mathcal{E}, \Delta)]]$ , so that  $A \in \mathcal{E}$ . Therefore  $\alpha \models A$ , and so  $(: A / A) \in score_{\Delta}(\alpha)$ . ■

**Lemma 3** *Let  $\alpha$  be an interpretation such that there is no  $\beta$  for which  $score_{\Delta}(\alpha) \subset score_{\Delta}(\beta)$ . Then  $\alpha \in |\mathcal{E}|$  for some extension  $\mathcal{E}$  of  $\Delta$ .*

**Proof** Where  $\alpha$  is as described, define  $\mathcal{E} = Th[Con[score_{\Delta}(\alpha)]]$ . Of course,  $\alpha \in |\mathcal{E}|$ . We prove that  $\mathcal{E} = Th[Con[GD(\mathcal{E}, \Delta)]]$ , from which it follows by Lemma 1 that  $\mathcal{E}$  is an extension of  $\Delta$ .

To show that  $\mathcal{E} \subseteq Th[Con[GD(\mathcal{E}, \Delta)]]$ , it is enough to show that  $score_{\Delta}(\alpha) \subseteq GD(\mathcal{E}, \Delta)$ . So suppose  $(: A / A) \in score_{\Delta}(\alpha)$ . Then  $A \in Con[score_{\Delta}(\alpha)]$ . Of course  $Con[score_{\Delta}(\alpha)]$  is consistent (since it has  $\alpha$  as a model), and so  $\mathcal{E}$  is consistent. Therefore  $\neg A \notin \mathcal{E}$ , and so  $(: A / A) \in GD(\mathcal{E}, \Delta)$ .

To show that  $Th[Con[GD(\mathcal{E}, \Delta)]] \subseteq \mathcal{E}$ , it is enough to show that  $GD(\mathcal{E}, \Delta) \subseteq score_{\Delta}(\alpha)$ . So suppose  $(: A / A) \in GD(\mathcal{E}, \Delta)$ , but  $(: A / A) \notin score_{\Delta}(\alpha)$ . Since  $(: A / A) \in GD(\mathcal{E}, \Delta)$ , we know that  $\neg A \notin Th[Con[score_{\Delta}(\alpha)]]$ . Therefore,  $Con[score_{\Delta}(\alpha)] \cup \{A\}$  is consistent,

and so there must be some interpretation  $\beta$  such that  $\beta \models \text{Con}[\text{score}_\Delta(\alpha)] \cup \{A\}$ . In that case, however, since  $(: A / A) \notin \text{score}_\Delta(\alpha)$ , we would have  $\text{score}_\Delta(\alpha) \subset \text{score}_\Delta(\beta)$ , contrary to the conditions of the lemma. ■

**Lemma 4** *Let  $\mathcal{E}$  be an extension of  $\Delta$  with  $\alpha \in |\mathcal{E}|$ . Then there is no interpretation  $\beta \notin |\mathcal{E}|$  such that  $\text{score}_\Delta(\alpha) \subseteq \text{score}_\Delta(\beta)$ .*

**Proof** Suppose  $\beta \notin |\mathcal{E}|$  but  $\text{score}_\Delta(\alpha) \subseteq \text{score}_\Delta(\beta)$ . Since  $\beta \notin |\mathcal{E}|$ , there must be some formula  $A \in \mathcal{E}$  such that  $\beta \not\models A$ . Now since  $\mathcal{E}$  is an extension of  $\Delta$ , Lemma 1 tells us that  $\mathcal{E} = \text{Th}[\text{Con}[\text{GD}(\mathcal{E}, \Delta)]]$ , and then since  $\alpha \in |\mathcal{E}|$ , Lemma 2 tells us that  $\mathcal{E} = \text{Th}[\text{Con}[\text{score}_\Delta(\alpha)]]$ . Therefore,  $A \in \text{Th}[\text{Con}[\text{score}_\Delta(\alpha)]]$ . But given the assumption that  $\text{score}_\Delta(\alpha) \subseteq \text{score}_\Delta(\beta)$ , we can conclude also that  $A \in \text{Th}[\text{Con}[\text{score}_\Delta(\beta)]]$ . So  $\beta \models A$ , contrary to hypothesis. ■

With these lemmas in hand, the proof of the main theorem is straightforward.

**Theorem 1** *Where  $\Delta$  is the default theory associated with the imperative set  $\mathcal{I}$ , the formula  $\bigcirc A$  is true with respect to  $\mathcal{I}$  if and only if  $A \in \mathcal{E}$  for some extension  $\mathcal{E}$  of  $\Delta$ .*

**Proof** First, suppose  $A \in \mathcal{E}$ , where  $\mathcal{E}$  is an extension of  $\Delta$ . Then of course  $|\mathcal{E}| \subseteq |A|$ . Now pick any  $\alpha \in |\mathcal{E}|$  (there has to be one since extensions are consistent). Lemma 4 tells us that there is no  $\beta \notin |\mathcal{E}|$  such that  $\text{score}_\Delta(\alpha) \subseteq \text{score}_\Delta(\beta)$ . Since  $|\mathcal{E}| \subseteq |A|$ , however, we can conclude that there is no  $\beta \in |\neg A|$  such that  $\text{score}_\Delta(\alpha) \subseteq \text{score}_\Delta(\beta)$ , and so  $\bigcirc A$  is true with respect to  $\mathcal{I}$ .

Next, suppose  $\bigcirc A$  is true with respect to  $\mathcal{I}$ : there is some  $\alpha \in |A|$  for which there is no  $\beta \in |\neg A|$  such that  $\text{score}_\Delta(\alpha) \subseteq \text{score}_\Delta(\beta)$ . Now pick some interpretation  $\gamma$  whose score contains that of  $\alpha$  and is also maximal—that is, a  $\gamma$  such that  $\text{score}_\Delta(\alpha) \subseteq \text{score}_\Delta(\gamma)$  and for which there is no  $\delta$  such that  $\text{score}_\Delta(\gamma) \subset \text{score}_\Delta(\delta)$ . (It is clear that there must be such an interpretation. Let  $\mathcal{D}$  be the set of defaults from  $\Delta$ . Standard techniques allow us to extend  $\text{Con}[\text{score}_\Delta(\alpha)]$  to a maximal consistent subset of  $\text{Con}[\mathcal{D}]$ , and then any interpretation  $\gamma$  satisfying this extended set will meet the conditions.) By Lemma 3, there is some extension

$\mathcal{E}$  of  $\Delta$  such that  $\gamma \in |\mathcal{E}|$ . It is easy to see that  $|\mathcal{E}| \subseteq |A|$ . For suppose otherwise, that there is some  $\delta \in |\mathcal{E}| - |A|$ . By Lemma 2, we would have  $score_{\Delta}(\gamma) = score_{\Delta}(\delta)$ , but this contradicts the assumption that  $\bigcirc A$  is true with respect to  $\mathcal{I}$ , since we know  $score_{\Delta}(\alpha) \subseteq score_{\Delta}(\gamma)$ , and so we would have  $score_{\Delta}(\alpha) \subseteq score_{\Delta}(\delta)$ . Therefore  $|\mathcal{E}| \subseteq |A|$ , and so  $A \in \mathcal{E}$  since  $\mathcal{E}$  is logically closed. ■

I close this section with two points relating the ideas discussed here to some of the other literature in the area.

The first point concerns nonmonotonic reasoning. Although this paper concentrates on the application of ideas from nonmonotonic reasoning to deontic logic, there is at least some flow in the opposite direction. Among the various interpretations of nonmonotonic reasoning, one of the most intriguing is that suggested by Jon Doyle, who understands the process in terms of an agent's criteria for choosing mental states based on default information. Interpreting mental states as formula sets, Doyle finds a natural way of treating individual defaults as preferences among mental states; and he shows in [5] and [6] that the appropriate mental states, or extensions, can then be characterized as those satisfying a requirement of Pareto optimality based on these preferences. This interpretation is conceptually very similar to van Fraassen's treatment of obligation; and Theorem 1 shows that, at least for a simple class of defaults, the match is exact. In more recent work, Doyle and Michael Wellman [7] have developed this interpretation by applying results from group decision theory to the study of nonmonotonic reasoning. The connection suggested here between Doyle's understanding of default reasoning and a previously existing deontic logic seems to buttress the interpretation, and to indicate another direction in which it might be developed.

The second point concerns deontic logic. Most attempts by deontic logicians to accommodate the phenomenon of moral conflict remain within the traditional modal framework, but simply rely on weaker, non-normal modal logics. For example, Brian Chellas [2, Sections 6.5, 10.2] recommends the logic  $D$ , which results from supplementing ordinary classical logic

with the rule schema

$$\frac{A \supset B}{\bigcirc A \supset \bigcirc B}$$

and the axiom

$$\neg \bigcirc \perp$$

(where  $\perp$  stands for the universally false proposition). In fact, as van Fraassen points out, the system  $D$  provides a sound and complete axiomatization for the *first* of his two proposals concerning the evaluation of ought statements, described here in Section 2 as his initial suggestion. The relationship can be stated precisely as follows: where  $\mathcal{I}$  is a set of imperatives each of which is itself consistent, and  $B$  is an  $\bigcirc$ -free formula, then  $\bigcirc B$  is true with respect to  $\mathcal{I}$  according to the initial suggestion if and only if  $\bigcirc B$  can be derived in the logic  $D$  from the set  $\{\bigcirc A : !(A) \in \mathcal{I}\}$ .<sup>8</sup>

What this result shows is that, in a sense, the initial suggestion does not really move beyond what can be accomplished using familiar modal techniques: the background imperatives could just as easily be coded as ought statements, and then the other oughts generated by these imperatives derived in a well-defined modal logic. It is now necessary to ask whether van Fraassen's refined proposal, presented here in Definition 1, can likewise be subsumed using ordinary modal techniques. Is there some axiomatizable modal logic in which, when a set of background imperatives is coded into ought statements, exactly the oughts generated by the refined proposal can then be derived?

The answer to this question appears to be No—for according to the refined proposal, it is possible for a recursive set of imperatives to generate a set of ought statements that is not even recursively enumerable. Perhaps the simplest example of this results if we imagine a neurotic agent who feels that *everything* is imperative; his imperative set is thus  $\mathcal{I} = \{!(A) : A \in \mathcal{L}\}$ , where  $\mathcal{L}$  is his background language, which we can take as recursive.

---

<sup>8</sup>The restriction of  $B$  to an  $\bigcirc$ -free formula is necessary because truth with respect to imperatives does not allow nested oughts. The restriction that the individual imperatives of  $\mathcal{I}$  must themselves be satisfiable can be dropped if we relate the initial proposal not to  $D$  but instead to the system  $EM$ , which results when the axiom  $\neg \bigcirc \perp$  is dropped from  $D$ .

This imperative set maps into the default theory  $\Delta = \langle \emptyset, \{(: A / A) : A \in \mathcal{L}\} \rangle$ , which has as its extensions exactly the maximal consistent sets of  $\mathcal{L}$ -sentences. A formula  $B$  belongs to some extension of  $\Delta$ , therefore, just in case  $B$  is a consistent  $\mathcal{L}$ -sentence; and so it follows from Theorem 1 that a statement  $\bigcirc B$  will be true with respect to  $\mathcal{I}$  just in case  $B$  is consistent. But of course, if  $\mathcal{L}$  is a sufficiently rich language, such as that of first-order logic, the set of its consistent sentences will not be recursively enumerable.<sup>9</sup>

As we recall, van Fraassen’s initial suggestion for accommodating moral dilemmas, which can be cast naturally as a modal logic, yields unintuitive results in certain cases. The refined proposal, which appears to be correct, can be interpreted in a natural way within default logic—a particular nonmonotonic formalism—but not, it seems, within modal logic. What this suggests, most generally, is that certain techniques developed within the field of nonmonotonic reasoning can be used to provide a theoretical framework superior in some ways to the usual modal framework for studying the logic of conflicting obligations.

## 5 Conditional oughts

So far we have been concerned only with simple (categorical) oughts. According to van Fraassen’s theory, these are engendered by simple imperatives, interpreted here as prerequisite-free default rules. In this section, we turn our attention to conditional oughts.

Unfortunately, van Fraassen does not actually present in [26] a finished account of condi-

---

<sup>9</sup>This argument shows that the relation set out in Definition 1 between imperatives and the ought statements they support cannot correspond to the consequence relation of an axiomatizable modal logic; but it is still conceivable that this relation between imperatives and their supported oughts might correspond to some modal consequence relation that can be defined only using semantic techniques. In fact, given the variety of semantically definable modal consequence relations, and the unlimited syntactic possibilities for coding imperatives using modal operators, I can think of no general reason why this should not be possible. However, the simple idea of coding each imperative of the form  $!(A)$  into a ought statement of the form  $\bigcirc A$  will not work, at least in a logic whose consequence relation extends the classical consequence relation. In any such logic, the formula  $\bigcirc(A \wedge \neg A)$  would have to be a semantic consequence of  $\bigcirc(A \wedge \neg A)$ , of course; but according to van Fraassen’s refined proposal, the ought statement  $\bigcirc(A \wedge \neg A)$  is not supported by any imperative set, even if it contains  $!(A \wedge \neg A)$ .

tional oughts. As we have seen, he sets out in this paper two accounts of simple oughts—a preliminary version that grounds these oughts in imperatives, and then a refined version incorporating the idea that it is best to satisfy as many imperatives as possible. Only the preliminary version is actually generalized to the more complicated topic of conditional oughts; these are now supposed to be founded on conditional (or hypothetical) imperatives. Nevertheless, by analogy with the treatment of simple oughts, we can see how this preliminary account of conditional oughts can be refined to include the idea of satisfying maximal sets of conditional imperatives; and it turns out that the refined theory is again interpretable within default logic, with conditional imperatives taken as default rules containing non-trivial prerequisites.

Conditional oughts will be represented here in the standard way; a statement of the form ‘It ought to be that  $A$ , given  $B$ ’ is symbolized  $\bigcirc(A/B)$ . We use analogous notation to represent conditional imperatives; for example, the imperative

If you go to the Everglades, watch out for alligators

might be represented as  $!(W/E)$ . A conditional imperative of this kind can be fulfilled or violated only in those situations in which its antecedent is satisfied; if its antecedent is satisfied, the imperative is said to be fulfilled if its consequent is also satisfied, and to be violated otherwise.

Now according to the preliminary treatment of simple oughts, as we recall, a proposition is obligatory if it is a necessary condition for satisfying some single imperative: where  $\mathcal{I}$  is the background imperative set,  $\bigcirc A$  is supposed to be true with respect to  $\mathcal{I}$  if there is some imperative  $!(B)$  in  $\mathcal{I}$  for which  $|B| \subseteq |A|$ . In generalizing this treatment to the conditional case, van Fraassen allows the imperative set  $\mathcal{I}$  to contain conditional as well as simple imperatives, and he defines a formula  $\bigcirc(A/C)$  to be true with respect to  $\mathcal{I}$  just in case there is some imperative  $!(B/C)$  belonging to  $\mathcal{I}$  such that  $|C| \cap |B| \subseteq |A|$ . The idea behind these modifications is this: in evaluating a conditional ought statement, we restrict our attention to those situations satisfying its antecedent; the statement is then true whenever, within this restricted range of situations, satisfying the consequent of the

ought is a necessary condition for fulfilling some imperative whose antecedent matches that of the ought. Notice that, if we interpret both simple oughts and simple imperatives as themselves conditional upon the universally true  $\top$ , this preliminary account of conditional oughts absorbs the preliminary account of simple oughts as a special case.

It is important to emphasize that according to this treatment, when we evaluate conditional oughts, we are supposed to consider *only* those imperatives governed by identical conditions; in evaluating an ought of the form  $\bigcirc(A/C)$  we are supposed to consider only those imperatives of the form  $!(B/C)$ . The aim of this restriction, evidently, is to bring into play exactly the right set of background imperatives, allowing us to avoid false conflicts. For example, suppose that an agent is subject to the imperatives

Don't eat with your fingers,

If you are served asparagus, eat it with your fingers.<sup>10</sup>

We represent these here through the imperative set  $\mathcal{I}_3 = \{!(\neg F),!(F/A)\}$ . According to the theory as it stands, only the second of these imperatives is considered in evaluating an ought conditional on the assumption that the agent is served asparagus; the first is ignored. Because of this,  $\bigcirc(F/A)$  is true with respect to  $\mathcal{I}_3$ , but  $\bigcirc(\neg F/A)$  is false. If we were to relax this restriction and consider the entire imperative set, both of these formulas would be true; these imperatives would then lead to a conflict under the assumption that the agent is served asparagus.

Let us see how to arrive at a refined version of the preliminary treatment presented so far of conditional oughts. In the case of simple oughts, van Fraassen refines his preliminary account by introducing the notion of score—the set of imperatives fulfilled in some situation—and then classifying a proposition as obligatory if it is a necessary condition for achieving some maximal score. Since in evaluating a conditional ought we are supposed to be concerned only with those imperatives governed by the same condition, this concept of score must be modified. The new notion is

$$score_{\mathcal{I},B}(\alpha) = \{!(A/C) \in \mathcal{I} : |C| = |B| \text{ and } \alpha \models A\},$$

---

<sup>10</sup>See Martin [18, p. 143].

which gives the score of a situation  $\alpha$  relative to (the proposition expressed by) a condition  $B$ . Using this new notion of score, the valuation rule for conditional oughts can be presented as follows.

**Definition 4** *The formula  $\bigcirc(A/B)$  is true with respect to the imperative set  $\mathcal{I}$  if and only if either  $|B| = \emptyset$  or there is some  $\alpha \in |B| \cap |A|$  for which there is no  $\beta \in |B| \cap |\neg A|$  such that  $\text{score}_{\mathcal{I},B}(\alpha) \subseteq \text{score}_{\mathcal{I},B}(\beta)$ .*

The idea here is that a conditional ought is true just in case either the antecedent is impossible, or assuming the antecedent to be true, the consequent is a necessary condition for achieving an antecedent-relative maximal score. Again, if we imagine that simple oughts and imperatives are themselves conditional on  $\top$ , this definition yields the earlier Definition 1 as a special case.

As with simple oughts, the interpretation of this refined theory into default logic is straightforward. We first associate with each imperative set  $\mathcal{I}$  a default theory  $\Delta = \langle \mathcal{W}, \mathcal{D} \rangle$ , where  $\mathcal{W} = \emptyset$  and  $\mathcal{D} = \{(B : A / A) : !(A/B) \in \mathcal{I}\}$ . So far this is only a slight generalization of the earlier interpretation of simple imperatives, treating the antecedents of conditional imperatives as prerequisites of the associated default rules. It is also necessary, however, to accommodate the fact that, in evaluating a conditional ought of the form  $\bigcirc(A/B)$ , we are supposed to look only at those situations in which  $B$  is true, and we are supposed to consider only those imperatives themselves conditional upon  $B$ . What this suggests is that we should relate such a conditional ought, not directly to the extensions of  $\Delta$ , but instead to extensions of the theory  $\Delta[B] = \langle \mathcal{W}[B], \mathcal{D}[B] \rangle$ , where  $\mathcal{W}[B] = \mathcal{W} \cup \{B\}$  and  $\mathcal{D}[B] = \{(C : A / A) \in \mathcal{D} : |C| = |B|\}$ . The formula  $B$  is true in each extension of  $\Delta[B]$ , and these extensions are sensitive only to those default rules whose prerequisites are equivalent to  $B$ .

The relation between default logic and the account in Definition 4 of conditional oughts can now be stated.

**Theorem 2** *Where  $\Delta$  is the default theory associated with the imperative set  $\mathcal{I}$ , the formula  $\bigcirc(A/B)$  is true with respect to  $\mathcal{I}$  if and only if  $A \in \mathcal{E}$  for some extension  $\mathcal{E}$  of  $\Delta[B]$ .*

The proof of this theorem is just a slightly more complicated version of the argument behind Theorem 1, and it will not be given here. The only important fact to note in developing the argument is that, like the prerequisite-free default theories considered earlier, theories of the form  $\Delta[B]$  also satisfy the restrictions of Lemma 1, and so their extensions can be characterized as explained there.

## 6 Problems with conditional oughts

Although Theorem 2 establishes the relevant correspondence between default logic and the refined theory of conditional oughts suggested by van Fraassen’s paper, this result has, I feel, less interest than the earlier Theorem 1. The reason for this is that, while the analysis of simple oughts underlying the earlier theorem appears to be sound, its generalization to conditional oughts is problematic.

It is easy to see the problem. As we recall, in evaluating a conditional ought of the form  $\bigcirc(A/B)$ , we are supposed to consider only those imperatives themselves conditional upon  $B$ . The point of this restriction is to avoid false conflicts: in the asparagus example,  $\mathcal{I}_3 = \{!(\neg F),!(F/A)\}$  above, the restriction allowed us to avoid concluding  $\bigcirc(\neg F/A)$  along with the desired  $\bigcirc(F/A)$ . It turns out, however, that the restriction is too severe, as we can see by supplementing our simple etiquette with one additional imperative:

Put your napkin on your lap.

Let us take  $\mathcal{I}_4 = \mathcal{I}_3 \cup \{!(N)\}$  as the new imperative set. Intuitively, we would want an agent to conclude from these imperatives that he should put his napkin on his lap even when he is eating asparagus. But the theory does not give us this result;  $\bigcirc(N/A)$  is false with respect to  $\mathcal{I}_4$ . Because the new imperative  $!(N)$  is not itself explicitly conditional upon  $A$ , it disappears from consideration as soon as we assume that the agent is served asparagus.

This problem can be seen as an instance of a general dilemma involved in evaluating a conditional ought on the basis of imperatives. If we consider only those imperatives explicitly triggered by the special conditions of the ought, then we lose track of the more general imperatives that should apply even under these special conditions. On the other hand,

we cannot evaluate a conditional ought against the entire set of background imperatives considered uniformly, because we want those imperatives explicitly triggered by its special conditions to override more general imperatives in case of conflict.

It is worth pointing out that this problem appears to be very general, and not dependent upon the particular mechanisms used here for relating imperatives to conditional oughts. In fact, the problem does not even depend upon the idea that oughts are grounded in imperatives, but can be stated entirely in the language of a dyadic deontic logic.

To see this, we can code the imperatives from  $\mathcal{I}_4$  into statements of conditional obligation; representing unconditional imperatives as oughts explicitly conditional upon  $\top$ , this gives us as premises the three formulas

$$\bigcirc(\neg F/\top),$$

$$\bigcirc(F/A),$$

$$\bigcirc(N/\top).$$

From these premises, we wish to derive the statement  $\bigcirc(N/A)$ , but not the statement  $\bigcirc(\neg F/A)$ . It seems that the only way to derive  $\bigcirc(N/A)$  is by strengthening the antecedent of the third premise; any theory that simply rules out this kind of strengthening—such as the kind of logics surveyed in Lewis [16], for example—will not allow us to derive this conclusion. On the other hand, a system that admits a rule of the form

$$\frac{\bigcirc(P/Q)}{\bigcirc(P/Q \wedge R)},$$

allowing unrestricted strengthening in the antecedent, will incorrectly yield  $\bigcirc(\neg F/A)$  from the first premise. What is needed, apparently, is a certain amount of strengthening, but not too much: we want to allow oughts formulated explicitly only for general circumstances to apply also by default in more specific situations, unless they are overridden in those situations. As far as I know, there is no standard philosophical logic that exhibits this behavior.<sup>11</sup>

---

<sup>11</sup>It would seem to follow, for example, that the consequence relation associated with any such theory

Let us turn now to look at the dilemma concerning conditional oughts that we have been considering from the standpoint of default logic; as we will see, problems of this kind have already been confronted within that framework.

We begin by formulating the default theory associated with the imperative set  $\mathcal{I}_4$ ; this theory is  $\Delta_7 = \langle \mathcal{W}_7, \mathcal{D}_7 \rangle$ , where  $\mathcal{W}_7 = \emptyset$  and  $\mathcal{D}_7 = \{(: \neg F / \neg F), (A : F / F), (: N / N)\}$ . Now according to the analysis underlying Theorem 2, an ought statement conditional upon  $A$  is supposed to be true against the background of this theory if the consequent of that statement belongs to some extension of  $\Delta_7[A] = \langle \{A\}, \{(A : F / F)\}$ —the default theory arrived at from  $\Delta_7$  by assuming  $A$  true, and then attending only to those default rules that are themselves conditional upon  $A$ . This conditioned theory  $\Delta_7[A]$  has  $Th[\{A, F\}]$  as its only extension. So if we adopt this approach, there is no conflict between the statements  $\bigcirc(F/A)$  and  $\bigcirc(\neg F/A)$ ; only the first is true, since we ignore the default rule that would have given rise to the second. However, we also ignore the default rule that would have supported  $\bigcirc(N/A)$ , and the statement is false. This choice corresponds to the first branch of our dilemma.

To see the second branch, suppose that we decide to modify  $\Delta_7$  somewhat differently in evaluating statements conditional upon  $A$ . Instead of assuming  $A$  true and then trimming off the default rules without  $A$  as prerequisite, we assume  $A$  true and leave the set of default rules unchanged, considering them all uniformly. This leads to the theory  $\langle \{A\}, \mathcal{D}_7 \rangle$ , which has two extensions:  $Th[\{A, F, N\}]$  and  $Th[\{A, \neg F, N\}]$ . Since  $N$  belongs to some extension of this theory (both of them, in fact), we are now able to conclude that  $\bigcirc(N/A)$  is true. However, we are also forced to conclude that both  $\bigcirc(F/A)$  and  $\bigcirc(\neg F/A)$  hold, since  $F$  and  $\neg F$  are each true in some extension; the false conflict is reintroduced.

Apparently, what is needed in order to resolve this dilemma is a way of representing and attending to all the background imperatives at once, while still allowing some to override would have to be nonmonotonic. Suppose the formula  $\bigcirc(F/A)$  were deleted from our premise set above. In that case, since the general injunction against eating with one's fingers is not explicitly overridden in the particular situation in which asparagus is served, it should apply here by default also; and so we *would* want to derive  $\bigcirc(\neg F/A)$ . But with  $\bigcirc(F/A)$  present as a premise, the general injunction is overridden; and so  $\bigcirc(\neg F/A)$  is no longer acceptable as a conclusion.

others. As it turns out, a similar problem arose early on in the use of default logic for representing ordinary defeasible information. The problem, which was first noticed by Reiter and Criscuolo [21], grew out of the attempt to provide a reasonable formalization within default logic of knowledge bases such as the following:

Tom is a bat,  
 All bats are mammals,  
 Bats usually can fly,  
 Mammals usually can't fly.

The most straightforward representation of this information gives us  $\Delta_8 = \langle \mathcal{W}_8, \mathcal{D}_8 \rangle$ , with  $\mathcal{W}_8 = \{Bt, \forall x(Bx \supset Mx)\}$  and  $\mathcal{D}_8 = \{(Bt : Ft / Ft), (Mt : \neg Ft / \neg Ft)\}$ . This theory uses *normal* default rules—rules whose justifications and consequents match—to represent the instantiations for Tom of the generic truths that bats can fly and that mammals cannot. Since the normal representation places the two conflicting defaults on a par, the theory leads to two extensions:  $Th[\mathcal{W}_8 \cup \{Ft\}]$  and  $Th[\mathcal{W}_8 \cup \{\neg Ft\}]$ . But this result is intuitively undesirable; only the second of these two extensions is legitimate. In contrast to the earlier Nixon example, there seems to be a reason here for preferring one of these two defaults over the other. Since bats are a particular kind of mammal, it seems that default information about bats should override conflicting default information about mammals in general.

In order to avoid unwanted extensions like these, Reiter and Criscuolo began to study some more general techniques for representing generic statements within default logic, without restricting themselves only to normal default rules. One of their most interesting proposals, applied to the example at hand, would involve replacing the rule  $(Mt : \neg Ft / \neg Ft)$  in the theory  $\Delta_8$  with the new default  $(Mt : [\neg Ft \wedge \neg Bt] / \neg Ft)$ . This new rule belongs to the class of *semi-normal* defaults—those whose justifications entail their consequents. The replacement does have the effect of eliminating the unwanted extension; the new default is no longer applicable, since its justification is inconsistent with any set including  $\mathcal{W}_8$ . However, the new, semi-normal default no longer seems like an instantiation for Tom of the generic statement that mammals usually cannot fly, but instead, of a statement like

Mammals usually can't fly, unless they're bats,

which explicitly mentions a class of exceptions.

This idea of avoiding unwanted extensions by coding exceptions explicitly into semi-normal default rules can be applied also in the present domain, where the default rules are supposed to represent imperatives. Returning to our example, let us take  $\mathcal{D}'_7$  as the set of defaults resulting from  $\mathcal{D}_7$  when the normal rule  $(: \neg F / \neg F)$  is replaced with the semi-normal rule  $(: [\neg F \wedge \neg A] / \neg F)$ , which represents an imperative like

Don't eat with your fingers, unless you're served asparagus;

and let us take  $\Delta'_7 = \langle \mathcal{W}_7, \mathcal{D}'_7 \rangle$  as the default theory associated with the imperative set  $\mathcal{I}_4$ . We now get the right conditional oughts from this default theory by following the *second* branch of the dilemma set out above. To see if an ought statement conditional upon  $A$  is true with respect to  $\Delta'_7$ , we see if its consequent belongs to any extension of  $\langle \{A\}, \mathcal{D}'_7 \rangle$ —the theory that results from  $\Delta'_7$  by assuming  $A$  true but attending to all of the defaults in  $\mathcal{D}'_7$ . This theory has  $Th[\{A, F, N\}]$  as its only extension; and so  $\bigcirc(F/A)$  and  $\bigcirc(N/A)$  are both true as desired, but  $\bigcirc(\neg F/A)$  is false.

When default rules are taken to represent imperatives, in fact, the strategy of incorporating exceptions explicitly into these rules is reminiscent of some ideas found in the early work of R. M. Hare. In Section 4.3 of [13], for example, Hare presents a picture according to which we are supposed to be guided in much of our action, moral and otherwise, by principles analogous to the imperatives discussed here; but he imagines that these principles would be heavily laden with qualifications. Much of the process of learning to act properly is supposed to involve getting the qualifications right. Initially, we learn very general principles, such as

Signal before you stop or turn the car.

But as our range of activity becomes more varied and sophisticated, these initial principles are supposed to be modified in a way that yields more complex precepts like

Signal before you stop or turn the car, except in an emergency.

On Hare's view, we become competent actors in a given domain (becoming a good driver is the example he uses) once we have picked up the appropriately modified set of principles and practice them habitually.

It seems, then, that the problem of deriving the right conditional oughts from a set of background imperatives might be solvable if we are able to require that the imperatives should be properly qualified to account for exceptional circumstances; and this strategy for achieving a solution has historical precedents elsewhere in moral theory. Is it the right strategy?

We can shed some light on this question by focusing again on the problem of exceptions to defeasible generalizations as it arises in knowledge representation. Here, the strategy of encoding these exceptions explicitly into semi-normal defaults has been subject to serious criticism—most notably by David Touretzky [22]. He objects to the idea for two reasons. First, any working knowledge representation system must have the ability to accommodate updates in some simple way. However, if exceptions to generic statements were to be listed explicitly in defaults, then as new information is added to a knowledge base, the default rules themselves would have to be continually modified in order to reflect the new exceptions introduced; and this would make the update operation far too difficult. Second, the number of exceptions to most generic truths is substantial, and so the resulting defaults would be unwieldy.

These objections of Touretzky's are powerful, and for the most part they have been accepted by researchers in the field in knowledge representation and nonmonotonic reasoning. However, the objections do rest on a framework of critical assumptions appropriate for work in artificial intelligence: they are grounded in the idea that a logical formalization is to be evaluated, ultimately, by its prospects for helping us to construct, or at least understand, a workable implementation. Since it is not obvious that work in deontic logic should be judged by these same standards, it is hard to see exactly how Touretzky's objections to the explicit exceptions approach in knowledge representation might bear upon the analogous proposal in normative theory. Perhaps we should agree with the early Hare that the imperatives guiding our action in any realistic range of situations must be extensively qualified. And it

is perhaps too much to expect that the process of incorporating a new moral principle into our background imperative set should be as simple as updating a knowledge base.

Nevertheless, it does not seem that these objections of Touretzky's can simply be dismissed in the case of imperatives. As for the first criticism concerning updates, this leads to other problems even when implementational issues are put aside. The idea that the formalization of some particular imperative might have to be modified as new imperatives are introduced, so that its proper rendering would vary depending on the imperative set in which it is embedded, suggests a holistic representational strategy that at least some people find objectionable.<sup>12</sup> And as for the second objection, concerning the complexity of the default rules resulting from the explicit exceptions approach, there are, in fact, good reasons for thinking that the moral principles guiding our everyday actions *should* be simple. Some of these reasons are brought out, surprisingly, in Hare's own later work, such as Chapter 2 of [14]. There, a distinction is drawn between the intuitive or everyday level of moral thinking and a more critical or reflective level. Hare allows that moral principles operating at the critical level can be of unlimited complexity, but he argues that the intuitive principles governing our everyday moral life must be relatively straightforward and free of qualification. The most persuasive of his arguments are based on psychological considerations about the limitations on our ability to learn extremely complicated moral principles, or to apply them effectively in the kind of situations that call for everyday moral decisions.

Although the issues involved are complicated, I feel that Touretzky's objections to the idea of encoding exceptions explicitly into defaults are sufficiently compelling—even when those defaults are supposed to represent imperatives—that we should seek another strategy for resolving the dilemma described here concerning conditional oughts. Fortunately, a good deal of research within the field of nonmonotonic reasoning has recently been devoted to the project of developing alternatives to the idea of encoding exceptions explicitly within

---

<sup>12</sup>The strategy is holistic because individual default rules could not then be said to represent individual imperatives; at best, these default rules could be said to represent the meaning of an imperative only within a particular imperative set. The relation between the strategy of encoding exceptions explicitly into semi-normal defaults and holism about meaning is discussed more thoroughly in Horty [15].

defaults.<sup>13</sup> The work is diverse, and it is dangerous to generalize; but very roughly, the aim of this research is to articulate broadly applicable principles on the basis of which one default should override another. It is hoped that these principles might somehow be incorporated into the logic, allowing individual defaults to be represented more simply.

Because of the positive correspondence established in this paper between deontic and nonmonotonic logics, and because of the similar problems faced by each in the treatment of exceptional information, it seems reasonable to expect that some of the general techniques for handling exceptions currently being explored within nonmonotonic reasoning might apply also to the analogous problems in deontic logic—allowing us, for example, to derive the right conditional oughts from a set of background imperatives, while keeping the representation of imperatives simple. In this way, nonmonotonic logics may provide a new set of tools for understanding the logical aspects of *prima facie* obligation.

## 7 Conclusion

From a philosophical standpoint, the work presented here is based on van Fraassen [26]. The bulk of that paper is organized around a series of arguments against the assumption, built into standard deontic logic, that moral dilemmas are impossible; and van Fraassen only briefly sketches his alternative approach. His paper ends with the conclusion that “the problem of possibly irresolvable moral conflict reveals serious flaws in the philosophical and semantic foundations of ‘orthodox’ deontic logic, but also suggests a rich set of new problems and methods for such logic.” My goal has been to suggest that some of these methods might be found in current research on nonmonotonic reasoning, and that some of the problems may have been confronted there as well.

---

<sup>13</sup>Much of the research on this broad topic has been focused on two particular problems. The first is the task of reasoning with the kind of interacting generic statements found in defeasible inheritance hierarchies. This problem was first studied systematically by Etherington and Reiter [9] and by Touretzky [23]; recent surveys can be found in Touretzky et al. [24] and Horty [15]. The second problem, of more long-run importance, is the task of resolving the kinds of difficulties first pointed out by Hanks and McDermott [12] about temporal projection and the frame problem. Several papers on this topic can be found in Section 5.3 of [11].

I have shown that nonmonotonic logics provide a natural framework for reasoning about moral dilemmas, perhaps even more useful than the ordinary modal framework, and that the issues surrounding the treatment of exceptional information within these logics run parallel to some of the problems posed by conditional oughts. However, there is also another way in which deontic logic might benefit from a connection to nonmonotonic reasoning. A familiar criticism among ethicists of work in deontic logic is that it is too abstract, and too far removed from the kind of problems confronted by real agents in moral deliberation. It must be said that similar criticisms of abstraction and irrelevance are often lodged against work in nonmonotonic reasoning by more practically minded researchers in artificial intelligence; but here, at least, the criticisms are taken seriously. Nonmonotonic logic aims at a qualitative account of commonsense reasoning, which can be used to relate planning and action to defeasible goals and beliefs; and at least some of the theories developed in this area have been tested in realistic situations. By linking the subject of deontic logic to this research, it may be possible also to relate the idealized study of moral reasoning typical of the field to a more robust treatment of practical deliberation.

## Acknowledgments

I owe a special debt to Richmond Thomason for extensive comments on an earlier version of this paper that substantially improved the final result. I have benefited also from discussions with or comments from Nuel Belnap, Craig Boutilier, Jon Doyle, David Etherington, David Makinson, Judea Pearl, Michael Slote, V.S. Subrahmanian, and David Touretzky.

This work has been supported by the National Science Foundation under Grants No. IRI-9003165 and IRI-8907122, and by the Army Research Office under Grant No. DAAL-03-88-K0087.

## References

- [1] D. Bobrow (ed.). Special issue on nonmonotonic logic. *Artificial Intelligence*, vol. 13 (1980), 174 pp.

- [2] B. Chellas. *Modal Logic: An Introduction*. Cambridge University Press (1980), xii+295 pp.
- [3] D. Dennett. The moral first aid manual. In *The Tanner Lectures on Human Values, vol. VIII*. Cambridge University Press (1988), pp. 119–147.
- [4] A. Donagan. Consistency in rationalist moral systems. *The Journal of Philosophy*, vol. 81 (1984), pp. 291–309.
- [5] J. Doyle. Reasoned assumptions and Pareto optimality. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence (IJCAI-85)*, Morgan Kaufmann Publishers (1985), pp. 87–90.
- [6] J. Doyle. On universal theories of defaults. Technical Report CMU-CS-88-111, Computer Science Department, Carnegie Mellon University (1988), 21 pp.
- [7] J. Doyle and M. Wellman. Impediments to universal preference-based default theories. *Artificial Intelligence*, vol. 49 (1991), pp. 97–128.
- [8] D. Etherington. *Reasoning with Incomplete Information*. Morgan Kaufmann Publishers (1988), viii+240 pp.
- [9] D. Etherington and R. Reiter. On inheritance hierarchies with exceptions. In *Proceedings of the National Conference on Artificial Intelligence (AAAI-83)*, Morgan Kaufmann Publishers (1983), pp. 104–108.
- [10] D. Føllesdal and R. Hilpinen. Deontic logic: an introduction. In *Deontic Logic: Introductory and Systematic Readings*, R. Hilpinen (ed.), D. Reidel Publishing Company (1971), pp. 1–35.
- [11] M. Ginsberg (ed.). *Readings in Nonmonotonic Reasoning*. Morgan Kaufmann Publishers (1987), viii+481 pp.

- [12] S. Hanks and D. McDermott. Default reasoning, nonmonotonic logics, and the frame problem. *Proceedings of the Fifth National Conference on Artificial Intelligence (AAAI-86)*, Morgan Kaufmann Publishers (1986), pp. 328–333.
- [13] R. Hare. *The Language of Morals*. Oxford University Press (1952), vii+202 pp.
- [14] R. Hare. *Moral Thinking: Its Levels, Method, and Point*. Oxford University Press (1980), ix+242 pp.
- [15] J. Horty. Some direct theories of nonmonotonic inheritance. In *Handbook of Logic in Artificial Intelligence and Logic Programming, vol. 2: Nonmonotonic Reasoning*, D. Gabbay and C. Hogger (eds.), Oxford University Press (forthcoming).
- [16] D. Lewis. Semantic analyses for dyadic deontic logic. In *Logical Theory and Semantic Analysis*, S. Stenlund (ed.), D. Reidel Publishing Company (1974), pp. 1–14.
- [17] W. Marek and M. Truszczyński. Relating autoepistemic and default logics. In *Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning (KR-89)*, Morgan Kaufmann Publishers (1989).
- [18] J. Martin. *Miss Manners' Guide to Excruciatingly Correct Behavior*. Atheneum Publishers (1892), xvii+745 pp.
- [19] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann Publishers (1988).
- [20] R. Reiter. A logic for default reasoning. *Artificial Intelligence*, vol. 13 (1980), pp. 81–132.
- [21] R. Reiter and G. Criscuolo. On interacting defaults. In *Proceedings of the Seventh International Joint Conference on Artificial Intelligence (IJCAI-81)*, Published by the American Association for Artificial Intelligence (1981), pp. 270–276.
- [22] D. Touretzky. Implicit ordering of defaults in inheritance systems. In *Proceedings of the Fourth National Conference on Artificial Intelligence (AAAI-84)*, Morgan Kaufmann Publishers (1984), pp. 322–325.

- [23] D. Touretzky. *The Mathematics of Inheritance Systems*. Morgan Kaufmann Publishers (1986), 220 pp.
- [24] D. Touretzky, J. Horty, and R. Thomason. A clash of intuitions: the current state of nonmonotonic multiple inheritance systems. In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence (IJCAI-87)*, Morgan Kaufmann Publishers (1987), pp. 476–482.
- [25] E. Ullman-Margalit. On presumption. *The Journal of Philosophy*, vol. 80 (1983), pp. 143–163.
- [26] B. van Fraassen. Values and the heart’s command. *The Journal of Philosophy*, vol. 70 (1973), pp. 5–19.
- [27] B. Williams. Ethical consistency. In *Proceedings of the Aristotelian Society*, supp. vol. 39 (1965), pp. 103–124. A revised version appears in B. Williams, *Problems of the Self: Philosophical Papers 1956–1972*, Cambridge University Press (1973), pp. 166–186.