Prepint of a paper appearing in

Annals of Mathematics and Artificial Intelligence

Volume 9 (1993), pp. 69--91.

Deontic Logic as Founded on 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)

Contents

1 Introduction

Deontic logic was originally developed as a tool for formalizing normative reasoning in ethical and legal contexts, and has since been explored primarily by philosophical logicians and a few legal theorists. As it turns out, however, the subject also has a bearing on artificial intelligence, for at least three reasons. First, deontic logic is often employed as a formalism for knowledge representation in the legal domain; references to some of these applications, as well as others in computer science, can be found in Wieringa and Meyer [39]. Second, as Davis [7, p. 448] points out, we would eventually want to encode at least a rudimentary system of norms in any intelligent system, particularly an autonomous system; and the language of deontic logic seems a promising candidate for providing a declarative statement of the appropriate prohibitions, permissions, and obligations. And third, the formalism of deontic logic can be applied not only to legal and ethical reasoning, but also to the kind of normative reasoning about goals studied, for example, in Wellman and Doyle [38], and may thus have a role in the theory of planning as well.

Ever since its inception in the work of von Wright [37], deontic logic has been developed primarily as a species of modal logic. I argue in this paper, however, that the techniques of nonmonotonic logic may provide a better theoretical framework—at least for the formalization of commonsense normative reasoning—than the usual modal treatment. After reviewing some standard approaches to deontic logic, I focus on two areas in which nonmonotonic techniques promise improved understanding: reasoning in the presence of conflicting obligations, and reasoning with conditional obligations.

2 Modal techniques in deontic logic

2.1 Standard deontic logic

On the usual approach to deontic logic, obligation is interpreted as a kind of necessity, which can be modeled using possible worlds techniques. The most familiar theory of this kind, known as *standard deontic logic*, is based on models of the form $\mathcal{M} = \langle W, f, v \rangle$, with W a set of possible worlds, v a valuation mapping sentence letters into sets of worlds at which they are true, and f a function mapping each world into a nonempty set of worlds. Where α is an individual world, $f(\alpha)$ can be thought of as the set of worlds ideal from the standpoint of α, those in which all the oughts in force at α are satisfied; or if we follow the common practice of identifying propositions with sets of worlds, $f(\alpha)$ can then be viewed as a proposition expressing the standard of obligation at work in α.

Against the background of these standard deontic models, the valuation rule for the connective $\bigcirc$, representing 'It ought to be the case that ...', is given as

$$\mathcal{M}, \alpha \models \bigcirc A \text{ if and only if } f(\alpha) \subseteq |A|,$$

with $|A|$ representing the set of worlds in which A is true. The idea is that $\bigcirc A$ should hold just in case A is a necessary condition for things turning out as they should—just in case A is entailed by the relevant standard of obligation.

Let us say that a situation gives rise to a *normative conflict* if it presents both of two conflicting propositions as obligatory—if, for example, it supports the truth of both $\bigcirc A$ and $\bigcirc \neg A$. We often seem to face conflicts like this in everyday life, and there are a number of vivid examples in philosophy and literature. Perhaps the best known of these is Sartre's description in [29] of a student during the Second World War who felt for reasons of patriotism and vengeance (his brother had been killed by the Germans) that he ought to leave home in order to join the Free French, but who felt also, for reasons of sympathy and personal devotion, that he ought to stay at home in order to care for his mother.

Sartre presents this student's situation in a compelling way that really does make it seem as if he had been confronted with conflicting, and perhaps irreconcilable, moral standards. However, if standard deontic logic is correct, Sartre is mistaken: the student did not face a normative conflict—no one ever does, because according to standard deontic logic, such a conflict is impossible. This is easy to see: in order for $\bigcirc A$ and $\bigcirc \neg A$ to hold jointly at a world α, we would need both $f(\alpha) \subseteq |A|$ and $f(\alpha) \subseteq |\neg A|$, from which it follows that $f(\alpha)$ would have to be empty; but in these standard models, the only requirement on f is that it should map each world into a nonempty set. Apart from what is presupposed by the background framework of normal modal logic, then, the entire content of standard deontic logic seems to be simply that there are no normative conflicts; and in fact, validity in these standard models can be axiomatized by supplementing the basic modal logic K with

$$\neg(\bigcirc A \wedge \bigcirc \neg A)$$

as an additional axiom schema. The resulting system is known as KD.

Now this feature of standard deontic logic—that it rules out normative conflicts—has received extensive discussion in the philosophical literature, by writers such as Donagan [10], Foot [11], Lemmon [20], Marcus [23], and Williams [40]. The bulk of this literature is concerned with the special case in which the norms of interest are the oughts generated by an ideal ethical theory. It is often argued, as by Donagan, for example, that an ideal moral theory could not be structured so as to generating conflicting oughts; and it is sometimes concluded from this that it is a desirable feature of standard deontic logic that it rules out the possibility of normative conflict. However, such a conclusion is surely unjustified if we think of the oughts represented in a deontic logic as including, not simply the norms generated by an ideal ethical theory, but also those involved in our everyday, commonsense normative reasoning.

For one thing, the task of actually applying an ideal 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. Much 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.[1] Moreover, our normative reasoning more generally is concerned, not only with ethical matters, but also with the dictates of "Small Moralls" (etiquette, aesthetics); and of course, these lead to other conflicts both among themselves and with the oughts of morality. Therefore, even if it does turns out

[1]The relation between moral theory and the rules of thumb that guide everyday ethical decisions has recently been discussed by Dennett [9].

that there can be no clashes among the oughts generated by an ideal ethical theory, it still seems necessary to allow for conflicting oughts in any logic that aims to represent either our commonsense ethical thinking or our normative reasoning more broadly.

The need for a deontic logic that tolerates conflicting norms is even clearer if we imagine an intelligent system that is designed to reason about and achieve certain goals supplied to it by its users, and that represents those goals declaratively as ought statements in a deontic logic. It is always possible for different users (or even for the same user) to supply the system with conflicting goals; and in such a case, we would not want the mechanisms for reasoning about goals to break down entirely, as it would if it were guided by standard deontic logic.

This kind of situation is analogous to that envisioned by Belnap [2, 3] as a way of motivating the applicability of contradiction tolerating logics (in particular, a relevance logic) in the area of automated reasoning. Belnap imagines a computer designed to reason from data supplied by its users; and he argues that there are situations in which, even if the users inadvertently supply the machine with inconsistent information—say, A and $\neg A$—we would not want it to conclude that everything is true. In the same way, we can easily imagine a situation in which, even if a machine happens to be supplied by its users with inconsistent goals—say, $\bigcirc A$ and $\bigcirc\neg A$—we would not want it to conclude, as in standard deontic logic, that it should regard every proposition as a goal. Indeed, if the module for reasoning about goals in such a machine is integrated with the module for reasoning about facts, then the present situation is actually a special case of Belnap's: since $\bigcirc A$ and $\bigcirc\neg A$ are logically inconsistent in the standard deontic logic, a reasoner guided by this system would have to conclude from this information, not only that everything is obligatory, but also that everything is true.

2.2 A weak modal logic

One strategy for adapting deontic logic to reason sensibly in the face of conflicting norms is to continue the attempt to develop the subject within a modal framework, but simply to move to a weaker, non-normal modal logic. The clearest example of this is Chellas's suggestion, in [5] and [6, Sections 6.5 and 10.2], that we adopt as our deontic models certain minimal models for modal logics, in which the accessibility relation maps individual worlds, not into sets of worlds, but into sets of propositions—sets of sets of worlds. More exactly, Chellas recommends a deontic logic based on models of the form $\mathcal{M} = \langle W, N, v \rangle$, with W and v as before, but with N a function from W into $\mathcal{P}(\mathcal{P}(W))$, subject to the condition that, for each of the propositions X and Y in $\mathcal{P}(W)$, if $X \in N(\alpha)$ and $X \subseteq Y$, then $Y \in N(\alpha)$.[2] Intuitively, the various propositions belonging to $N(\alpha)$ can be thought of as expressing the variety of different ways in which things ought to turn out at α, the variety of different standards of obligation at work in α.

In these models, the truth conditions for ought statements can be presented through the rule

$$\mathcal{M}, \alpha \models \bigcirc A \text{ if and only if there is an } X \in N(\alpha) \text{ such that } X \subseteq |A|;$$

the idea is that $\bigcirc A$ should hold just in case A is a necessary condition for satisfying some standard of obligation in force at α. And validity is axiomatized by the system *EM*, which results from supplementing ordinary propositional logic with the rule schema

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

In fact, this logic is weak enough to tolerate normative conflicts: the statements $\bigcirc A$ and $\bigcirc \neg A$ are jointly satisfiable, without entailing $\bigcirc B$. However, in weakening standard deontic logic to allow conflicts, it seems that we have now arrived at a system that is too weak: it fails to validate intuitively desirable inferences. Suppose, for example, that an agent is subject to the following two norms, the first issuing perhaps from some legal authority, the second from religion or conscience:

[2] Chellas recommends also the further condition that $\emptyset \notin N(\alpha)$. We ignore this condition because it seems like an overly strong constraint for many application areas, particularly the case in which the oughts of a deontic logic represent goals supplied to an intelligent system by its users. We would not want to rule out the possibility that a fallible user might present an intelligent system with an impossible goal ("Find a rational root for this equation"), or to abandon sensible reasoning in such a case.

You ought either to fight in the army or perform alternative service,

You ought not to fight in the army,

We can represent these norms through the formulas $\bigcirc(F \vee S)$ and $\bigcirc\neg F$. Now it seems intuitively that the agent should conclude from these premises that he ought to perform alternative service. However, the inference from $\bigcirc(F \vee S)$ and $\bigcirc\neg F$ to $\bigcirc S$ is not valid in the logic *EM*.

Let us look at this problem a bit more closely. Any logical consequence of an ought derivable in *EM* is itself derivable as an ought in this system; and of course, S is a logical consequence of $(F \vee S) \wedge \neg F$. Therefore, we would be able to derive $\bigcirc S$ from our premise set if we could somehow merge the individual oughts $\bigcirc(F \vee S)$ and $\bigcirc\neg F$ together into a joint ought of the form

$$\bigcirc((F \vee S) \wedge \neg F).$$

But how could we get this latter statement? It seems possible to derive it from our premises only through a rule of the form

$$\frac{\bigcirc A \qquad \bigcirc B}{\bigcirc(A \wedge B),}$$

dubbed by Williams [40] as the rule of *agglomeration*. However, such a rule is not admissible in *EM*, and in fact, it is exactly the kind of thing that this logic is designed to avoid: from $\bigcirc A$ and $\bigcirc\neg A$, agglomeration would allow us to conclude $\bigcirc(A \wedge \neg A)$, and so $\bigcirc B$ for arbitrary B, due to closure of ought under logical consequence.

Evidently, the issue of agglomeration is crucial for a proper logical understanding of normative conflicts. We do not want to allow unrestricted agglomeration, as in the standard deontic logic *KD*; this would force us to treat conflicting oughts as incoherent. On the other hand, we do not want to block agglomeration entirely, as in the weak deontic logic *EM*; we would then miss certain desirable consequences in cases in which conflict is not a problem.

3 An approach based on nonmonotonic logic

As far as I know, the only intuitively adequate account of reasoning in the presence of normative conflicts set out in the literature so far occurs in van Fraassen's [36], a paper that is largely devoted to more broadly philosophical issues. Suppose that Γ is a set of oughts, possibly conflicting. The basic idea behind van Fraassen's suggestion is that $\bigcirc A$ should follow from Γ just in case satisfying A is a necessary condition for fulfilling, not just a single ought from Γ, but some maximal set of these.

Formally, the account relies on a notion of score. Where $\mathcal{M}$ is an (ordinary, classical) model of the underlying, ought-free language, the score of $\mathcal{M}$, relative to Γ, is defined as the set of oughts from Γ that it fulfills: $score_{\Gamma}(\mathcal{M}) = \{\bigcirc B \in \Gamma : \mathcal{M} \models B\}$. In this non-modal framework, we now let $|A|$ represent the ordinary model class of A: $|A| = \{\mathcal{M} : \mathcal{M} \models A\}$. Van Fraassen's the notion of deontic consequence, which we represent as the relation $\vdash_F$, is then defined as follows:

Definition 1 $\Gamma \vdash_F \bigcirc A$ *if and only if there is a model* $\mathcal{M}_1 \in |A|$ *for which there is no model* $\mathcal{M}_2 \in |\neg A|$ *such that* $score_{\Gamma}(\mathcal{M}_1) \subseteq score_{\Gamma}(\mathcal{M}_2)$.

As in the logic *EM*, this notion of consequence is weak enough that conflicting oughts do not imply arbitrary oughts: we cannot derive $\bigcirc B$ from $\bigcirc A$ and $\bigcirc \neg A$. However, unlike *EM*, this way of characterizing deontic consequence does allow what seems to be the right degree of agglomeration: we can agglomerate individual oughts as long as this does not lead to the introduction of an inconsistent formula within the scope of an ought. For example, although we do not get

$$\bigcirc A, \bigcirc \neg A \vdash_F \bigcirc (A \wedge \neg A),$$

we do have

$$\bigcirc (F \vee S), \bigcirc \neg F \vdash_F \bigcirc ((F \vee S) \wedge \neg F);$$

and then, since any logical consequence of an ought is itself an ought, this tells us that

$$\bigcirc (F \vee S), \bigcirc \neg F \vdash_F \bigcirc S.$$

Although this suggestion of van Fraassen's does appear to capture an intuitively attractive and stable account of reasoning in the presence of conflicting norms, and although the general topic of normative conflict has been an issue of intense concern in philosophy for over a decade, it is hard to find any discussion of this proposal in either the philosophical or the logical literature on the topic. I feel that part of the reason for this is that both philosophers and logicians are accustomed to approaching deontic logic from the theoretical perspective of modal logic; and as we will see, van Fraassen's proposal does not fit naturally within this framework. It turns out, however, that the proposal can be accommodated within the framework of nonmonotonic logic. In fact, it can be interpreted in a natural way within theories exemplifying two of the most popular techniques developed for the study of nonmonotonic reasoning—the fixed point and model preference techniques.

3.1 Default logic

The best known of the fixed point approaches to nonmonotonic reasoning is Reiter's default logic [28], a theory that supplements ordinary logic with new rules if inference, known as *default rules*, and then modifies the ordinary notion of logical consequence in order to accommodate these new rules.

An ordinary rule of inference (with a single premise) can be depicted simply as a premise-conclusion pair, such as (A/B); such a 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 if A has been established and, in addition, C is consistent with the reasoner's conclusion set. A *default theory* is a pair $\Delta = \langle \mathcal{W}, \mathcal{D} \rangle$, in which is $\mathcal{W}$ is a set of ordinary formulas and $\mathcal{D}$ is a set of default rules.

In specifying the conclusions derivable from a default theory, Reiter first defines an operator , that uses the information from a particular default theory to map formula sets into formula sets. Where $\Delta = \langle \mathcal{W}, \mathcal{D} \rangle$ is a default theory and $\mathcal{S}$ is some set of formulas, $,_{\Delta}(\mathcal{S})$ is the minimal set satisfying the following three conditions:

1. $\mathcal{W} \subseteq \, ,_{\Delta}(\mathcal{S})$;

2. $Cn[\Gamma_{\Delta}(\mathcal{S})] = \Gamma_{\Delta}(\mathcal{S})$;

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 operator Γ_{Δ} maps any formula set $\mathcal{S}$ into the minimal superset of $\mathcal{W}$ that is closed under both ordinary 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: the set $\mathcal{E}$ is an *extension* of the default theory Δ if and only if $\Gamma_{\Delta}(\mathcal{E}) = \mathcal{E}$.

Default logic is a conservative extension of ordinary classical logic, in the sense that the extension of a default theory $\langle \mathcal{W}, \mathcal{D} \rangle$ in which $\mathcal{D}$ is empty is simply $Cn[\mathcal{W}]$, the ordinary consequence set of $\mathcal{W}$. In contrast to the situation in ordinary logic, however, not every default theory leads to a single set of appropriate conclusions. Some 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 $\Delta = \langle \mathcal{W}, \mathcal{D} \rangle$, with $\mathcal{W} = \{Q(n), R(n)\}$ and $\mathcal{D} = \{(Q(n) : P(n) \,/\, P(n)), (R(n) : \neg P(n) \,/\, \neg P(n))\}$. This theory allows as extensions both $Cn[\mathcal{W} \cup \{P(n)\}]$ and $Cn[\mathcal{W} \cup \{\neg P(n)\}]$.

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. For this reason, the multiple extensions associated with default theories often seem like an embarrassment: most of the time, 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 (usually by taking their intersection) into a unique set. When it comes to interpreting deontic ideas, however, these multiple extensions are no longer embarrassing: they give us exactly what we need for understanding the logic of normative conflict.

Formally, the interpretation of van Fraassen's theory within default logic is straightforward. Where , is some set of ought statements, we can define the default theory $\Delta_\Gamma = \langle \mathcal{W}, \mathcal{D} \rangle$, where $\mathcal{W} = \emptyset$ and $\mathcal{D} = \{(\top : B \;/\; B) : \bigcirc B \in \, , \}$ (with $\top$ the universal truth). It can then be shown that

Theorem 1 *,* $\vdash_F \bigcirc A$ *if and only if* $A \in \mathcal{E}$ *for some extension* $\mathcal{E}$ *of* Δ_Γ*.*

A proof can be found in [16].

3.2 A model preference logic

A different approach to nonmonotonic reasoning is taken by theories falling within the model preference framework. Here, the idea is that the standard notion of logical consequence, according to which a formula is a consequence of some premise set just in case it is true in all models of that premise set, is too severe as a representation of commonsense consequence, because it forces us to consider, in addition to the most plausible models of those premises, also others that are intuitively more bizarre. Model preference logics proceed by specifying a preference ordering on the background models, and then defining a new notion of consequence according to which a formula follows from a premise set just in case it holds in all the most preferred models of those premises.

The original and best known approach falling within this framework is the semantic theory underlying McCarthy's circumscription [25], in which models are ordered according to the extension assigned to a particular predicate, usually a predicate representing instances of some abnormality. On this approach, the information

> Tweety is a bird,
>
> Birds tend to fly,

for example, could be represented through the formulas

$B(t)$,

$\forall x[B(x) \wedge \neg Ab(x) \supset F(x)]$,

where the second of these premises states, more exactly, that birds fly unless they are abnormal. Models (sharing the same domain) might then be ordered according to the extension assigned to the abnormality predicate Ab, with a model $\mathcal{M}_1$ preferred to a model $\mathcal{M}_2$ whenever the extension assigned by $\mathcal{M}_1$ to Ab is a subset of that assigned by $\mathcal{M}_2$. In the most preferred models, of course, the extension of Ab will be empty. Therefore, even though $F(t)$—representing the conclusion that Tweety flies—does not hold in all models of the above premises, and so is not logically entailed, it does hold in all of the most preferred models of these premises, and can be shown to follow from the premises through a simple application of McCarthy's circumscription schema.

There have been a number of variations of McCarthy's original notion of circumscription, involving increasingly sophisticated preference relations among models; and the idea of model preference has been explored in other contexts as well. Another early example is Minker's [24] generalization of the closed world assumption to the context of disjunctive databases, which takes the most preferred models of a database to be those in which the fewest atomic sentences are true. Preference orderings involving temporal considerations have been considered by researchers such as Kautz [18] and Shoham [30]. And the theory of model preference logics has been studied from a very general point of view in a series of papers by Lehmann and his colleagues, beginning with [19].

It was pointed out by van Benthem [34] that van Fraassen's theory, as he originally states it, can also be seen as a certain kind of model preference logic, in which the preference order over models is not absolute, as in the theories previously mentioned, but dependent on the background premise set. Where , is a set of ought statements, van Benthem suggests an ordering $\leq_\Gamma$ on models, defined so that $\mathcal{M}_1 \leq_\Gamma \mathcal{M}_2$ just in case $\mathcal{M}_2 \models B$ implies $\mathcal{M}_1 \models B$ for each $\bigcirc B \in$, ; the idea is that the more preferred models relative to , are those that satisfy more of the oughts from , . It can then be seen that:

Theorem 2 , $\vdash_F \bigcirc A$ *if and only if there is a model* $\mathcal{M}$ *such that* $\mathcal{M} \models A$ *and* $\mathcal{M}' \models A$ *for all models* $\mathcal{M}' \leq_\Gamma \mathcal{M}$.

The proof is trivial, of course, since the present characterization of deontic consequence is simply a reformulation of van Fraassen's original definition that makes the idea of model preference more explicit.

4 Exploring the theory

4.1 The consequence relation

Although, as we have seen, van Fraassen's notion of deontic consequence fits naturally within the framework of nonmonotonic logic, the consequence relation $\vdash_F$ is itself monotonic: $, \vdash_F \bigcirc A$ implies $, ,, ' \vdash_F \bigcirc A$. This result follows at once from our Theorem 1 together with Theorem 3.2 of Reiter [28], and also, more directly, from Theorem 3 below; what it suggests is that, in relating van Fraassen's account of oughts to nonmonotonic logics, we are not actually relying on the nonmonotonicity of these theories, but only on their ability to yield multiple, mutually inconsistent sets of sentences as consequence sets for a given set of premises. This will change in Section 6, where we do appeal to nonmonotonicity in our treatment of conditional oughts.

It is easy to see also both that the logical truths follow as oughts from any premise set, and, as mentioned, that any logical consequence of a generated ought is itself generated as an ought: $\vdash A$ implies $, \vdash_F \bigcirc A$; and $, \vdash_F \bigcirc A$ and $A \vdash B$ together imply $, \vdash_F \bigcirc B$. Moreover, van Fraassen's consequence relation allows us to derive only consistent formulas as oughts (a form of ought implies can), no matter what ought statements it is supplied with as premises: if $, \vdash_F \bigcirc A$, then A is consistent.

Because only consistent formulas are derivable as oughts, we can see at once that the consequence relation $\vdash_F$ is not reflexive. Although an inconsistent ought might appear among some set of premises, it cannot appear as a conclusion of those premises; and so we do not have

$$\bigcirc(A \wedge \neg A) \vdash_F \bigcirc(A \wedge \neg A),$$

for example. From this, it follows that van Fraassen's theory cannot be captured as a conventional modal logic, since any such logic carries a reflexive consequence relation.

In addition, the $\vdash_F$ relation fails to satisfy the cut rule; for example, although we have

$$\bigcirc(A \wedge B) \vdash_F \bigcirc A$$

and

$$\bigcirc A, \bigcirc \neg B \vdash_F \bigcirc(A \wedge \neg B),$$

we do not have

$$\bigcirc(A \wedge B), \bigcirc \neg B \vdash_F \bigcirc(A \wedge \neg B).$$

This observation, along with the example, is again due to van Benthem [34].

Finally, it is perhaps obvious, but just worth pointing out, that the set of oughts supported by a particular premise set is sensitive, not only to the total model theoretic content of the various ought statements in the premises, but to the presentation of this content, the way it is divided up among the various oughts. If Γ is a background set of ought statements, let $\overline{\Gamma} = \{B : \bigcirc B \in \Gamma\}$ represent the *model theoretic content* of Γ. Then it is possible to have two sets of ought statements Γ_1 and Γ_2 equivalent in content in the sense that $|\overline{\Gamma_1}| = |\overline{\Gamma_2}|$, and also to have $\Gamma_1 \vdash_F \bigcirc A$, without having $\Gamma_2 \vdash_F \bigcirc A$. For example, let $\Gamma_1 = \{\bigcirc A, \bigcirc \neg A\}$ and $\Gamma_2 = \{\bigcirc(A \wedge \neg A)\}$. Then although $|\overline{\Gamma_1}| = |\overline{\Gamma_2}|$, we have $\Gamma_1 \vdash_F \bigcirc A$, but not $\Gamma_2 \vdash_F \bigcirc A$.

4.2 Relation to familiar deontic logics

As might be expected, van Fraassen's consequence relation $\vdash_F$ generally lies between $\vdash_{EM}$ and $\vdash_{KD}$, the consequence relations associated with *EM* and *KD*; it generally allows us to derive more oughts from a given set of premises than *EM* and fewer than *KD*. But there are exceptions to this general rule, and we need to introduce some technical vocabulary in order to state the matter exactly.

First, let us officially characterize an *ought statement* as a statement of the form $\bigcirc A$ in which A is $\bigcirc$-free. Since the standard modal theories allow for iterated deontic operators and van Fraassen's theory does not, we must restrict ourselves in comparisons to the shared sub-language of ought statements. Next, let us define a set of ought statements Γ as *unit consistent* if each individual ought belonging to the set is itself satisfiable—that is, if B is consistent for each $B \in \overline{\Gamma}$. And let us say that Γ is not just unit consistent but *consistent* if the oughts belonging to Γ are jointly satisfiable—that is, if $\overline{\Gamma}$ itself is consistent.

Before working out the exact relations among these different deontic logics, we first offer yet another characterization of the consequence relation $\vdash_F$, which is perhaps the most straightforward.

Theorem 3 *Let Γ be a set of ought statements. Then $\Gamma \vdash_F \bigcirc A$ if and only if there a consistent subset $\mathcal{G}$ of $\overline{\Gamma}$ such that $\mathcal{G} \vdash A$.*

Proof First, suppose $\Gamma \vdash_F \bigcirc A$. Let $\mathcal{M}_1$ be as in Definition 1, and let $\mathcal{G} = Th(\mathcal{M}_1) \cap \overline{\Gamma}$. Clearly, $\mathcal{G}$ is consistent and a subset of $\overline{\Gamma}$; and it is clear also that $score_\Gamma(\mathcal{M}) = score_\Gamma(\mathcal{M}')$ for any $\mathcal{M}, \mathcal{M}' \in |\mathcal{G}|$. To see that $\mathcal{G} \vdash A$, suppose otherwise: then there exists a model $\mathcal{M}_2 \in |\mathcal{G}| \cap |\neg A|$; but in that case we have $score_\Gamma(\mathcal{M}_2) = score_\Gamma(\mathcal{M}_1)$, contrary to the definition of $\vdash_F$. Next, suppose $\mathcal{G} \vdash A$ for some consistent subset $\mathcal{G}$ of $\overline{\Gamma}$. Standard techniques allow us to define a maximal consistent subset $\mathcal{G}^*$ of $\overline{\Gamma}$ containing $\mathcal{G}$. Since $\mathcal{G}^*$ is consistent, and since it must also entail A, we have some model $\mathcal{M}_1 \in |\mathcal{G}^*| \subseteq |A|$; and then since $\mathcal{G}^*$ is maximal, it is easy to see that there can be no $\mathcal{M}_2 \in |\neg A|$ such that $score_\Gamma(\mathcal{M}_1) \subseteq score_\Gamma(\mathcal{M}_2)$. So $\Gamma \vdash_F \bigcirc A$. ∎

We consider first the relations between van Fraassen's theory and *EM*. If a set of ought statements Γ is not even unit consistent, we must have $\Gamma \vdash_{EM} \bigcirc A$ for every A; and so *EM* is stronger than van Fraassen's theory, since this theory allows us to derive only consistent oughts. As we have seen from the army example discussed in Sections 2 and 3, however, van Fraassen's theory does allow us to draw conclusions from certain unit consistent sets that cannot be derived in *EM*; and together with the following theorem, this shows that the theory is properly stronger than *EM* for unit consistent sets of oughts.

Theorem 4 *Let Γ be a unit consistent set of ought statements. Then if $\Gamma \vdash_{EM} \bigcirc A$, it follows that $\Gamma \vdash_F \bigcirc A$.*

Proof We begin by constructing a model for the modal language in which the possible worlds are ordinary models of the underlying classical language. Let $\mathcal{M} = \langle W, N, v \rangle$, where

W is the set of models of the underlying classical language, and in which $N(\alpha) = \{X : |B| \subseteq X$ and $\bigcirc B \in \,, \}$ for each $\alpha \in W$, and $v(p) = |p|$ for each proposition letter p. It is clear that $\mathcal{M}$ is a minimal model satisfying the condition that, if $X \in N(\alpha)$ and $X \subseteq Y$, then $Y \in N(\alpha)$; and clear also that $\mathcal{M}, \alpha \models \,,$ for each $\alpha \in W$. Therefore, since $, \vdash_{EM} \bigcirc A$, we know that $\mathcal{M} \models \bigcirc A$; that is, for each $\alpha \in W$, there is an $X \in N(\alpha)$ such that $X \subseteq |A|$. From this and the definition of N, we can conclude that $|B| \subseteq |A|$ for some $\bigcirc B \in \,,$. However, since , is unit consistent, $\{B\}$ is then a consistent subset of $\overline{,}$ that entails A; and so we can conclude that $, \vdash_F \bigcirc A$ from Theorem 3. ∎

We turn now to *KD*. Of course, anything can be derived in *KD* from an inconsistent set of oughts; and so, together with the following theorem, this shows that, as expected, *KD* is properly stronger than van Fraassen's theory.

Theorem 5 *Let , be a set of ought statements. Then if* $, \vdash_F \bigcirc A$*, it follows that* $, \vdash_{KD} \bigcirc A$.

Proof Suppose $, \vdash_F \bigcirc A$. By Theorem 3, it follows that $\mathcal{G} \vdash A$ for some subset $\mathcal{G}$ of $\overline{,}$; and so $\vdash (B_1 \wedge \ldots \wedge B_n) \supset A$, for some $B_1, \ldots, B_n \in \overline{,}$. Since *KD* is a normal modal logic, we can conclude from this that $\vdash_{KD} (\bigcirc B_1 \wedge \ldots \wedge \bigcirc B_n) \supset \bigcirc A$; and so $, \vdash_{KD} \bigcirc A$, since $\bigcirc B_1, \ldots, \bigcirc B_n \in \,,$. ∎

It is reassuring to see, however, that, unlike *EM*, van Fraassen's theory differs from *KD* only when applied to an inconsistent set of ought statements; otherwise, the two theories yield exactly the same results.

Theorem 6 *Let , be a consistent set of ought statements. Then if* $, \vdash_{KD} \bigcirc A$*, it follows that* $, \vdash_F \bigcirc A$.

Proof As in the proof of Theorem 4, we construct a model for the modal language with the ordinary models of the underlying classical language as its possible worlds. Let $\mathcal{M} = \langle W, f, v \rangle$, with W and v as before, but in which $f(\alpha) = |\overline{,}|$ for each $\alpha \in W$. Since , is consistent, $f(\alpha)$ is always a nonempty set; and so $\mathcal{M}$ is a standard deontic model. Moreover, $\mathcal{M} \models \,,$, and so since $, \vdash_{KD} \bigcirc A$, we have $\mathcal{M} \models \bigcirc A$; that is, $f(\alpha) \subseteq |A|$ for each $\alpha \in W$. From this and the definition of f, we can conclude that $|\overline{\Gamma}| \subseteq |A|$; and since $\overline{\Gamma}$ is itself consistent, Theorem 3 allows us to conclude that $\Gamma \vdash_F \bigcirc A$. ∎

4.3 Some variations

Although van Fraassen's account embodies an intuitively coherent and stable approach to reasoning in the presence of normative conflicts, it is not the only such approach. In this section, I simply mention a couple of variations on van Fraassen's original account—one that generates fewer oughts from a given premise set, and one that generates more.

First, suppose an agent is given $\bigcirc A$ and $\bigcirc\neg A$ as premises. We have assumed so far that the agent should draw from this information both the conclusions $\bigcirc A$ and $\bigcirc\neg A$, though not the agglomerate $\bigcirc(A \wedge \neg A)$. But there is another option. It is possible to imagine in this case that the agent might want to resist the conclusion that $\bigcirc A$ precisely because he has reason to believe that $\bigcirc\neg A$, and that he might likewise want to resist the conclusion that $\bigcirc\neg A$ because he has reason to believe that $\bigcirc A$. And in general, it is possible to imagine that the agent might want to conclude that a proposition ought to hold just in case he has reason for thinking that it ought to hold, and no reason for thinking that it ought not to.[3] In the present environment, we can capture this approach to deontic reasoning quite simply, by modifying the idea underlying Theorem 1 to reflect a "skeptical" treatment of extensions. We can suppose that $\bigcirc A$ follows from a set Γ of ought statements just in case A belongs, not simply to some extension of Δ_Γ, but to each such extension.

Second, we have noted that van Fraassen's treatment ignores inconsistent oughts that

[3]It may be a view along these lines that lies behind the following passage by Foot:

> What we must ask ... is whether in cases of irresolvable moral conflict we have to back both the judgment in favor of doing a and the judgment in favor of b, although doing b involves not doing a. Is it not possible that we should rather declare that the two are incommensurable, so that we have nothing to say about the overall merits of a and b, whether because there is nothing that *we* can say or because there is no truth of the matter and therefore nothing to *be* said [11, p. 395–396].

occur among the premises: from $\bigcirc((A \wedge \neg A) \wedge B)$, for example, no oughts can be concluded except the logical truths. There is another option, however. We can imagine that an agent provided with an inconsistent ought, which cannot be satisfied entirely, might still wish to satisfy "as much" of this formula as possible. This approach can be captured by adapting an idea set out in another context in Anderson et al. [1, Section 82.4]: we first articulate the premise set of ought statements into a larger set representing its intended meaning more explicitly, and then apply van Fraassen's approach to this articulated set of premises in order to draw the appropriate conclusions.

The procedure suggested by Anderson et al. for articulating the premise set, and defended in detail there, is as follows (we confine ourselves to the propositional case). Implication is first eliminated from ought statements, so that the resulting formulas are written in $\wedge$, $\vee$, and $\neg$; and an occurrence of a subformula in an ought statement is defined as *positive* or *negative* depending on whether it lies within the scope of an even or odd number of negations. Given a premise set of ought statements , , the articulated set , * is then defined as the smallest superset of , that contains both $\bigcirc(\ldots B \ldots)$ and $\bigcirc(\ldots C \ldots)$ whenever it contains either $\bigcirc(\ldots (B \wedge C) \ldots)$ with the occurrence of the conjunction positive, or $\bigcirc(\ldots (B \vee C) \ldots)$ with the occurrence of the disjunction negative. As an example, where , $= \{\bigcirc((A \wedge \neg A) \wedge B)\}$, then , $^* =$, $\cup \{\bigcirc(A \wedge B), \bigcirc(\neg A \wedge B), \bigcirc A, \bigcirc \neg A, \bigcirc B\}$.

Given this idea of articulation, the new notion of deontic consequence can be defined in the obvious way: $\bigcirc A$ can be said to follow from , in the present sense just in case it follows from , * according to van Fraassen's original definition—that is, just in case , $^* \vdash_F \bigcirc A$. With , as above, for example, we can reach the following conclusions in the present sense, none of which follows according to van Fraassen's original definition: $\bigcirc(A \wedge B)$, $\bigcirc(\neg A \wedge B)$, $\bigcirc A$, $\bigcirc \neg A$, and $\bigcirc B$.

Of course, the two variations suggested here on van Fraassen's original account run in orthogonal directions, and they can be combined without mutual interference: in reasoning from a premise set , , an agent might first extend this to the articulated set , *, and then draw only those conclusions contained in each extension of the associated default theory Δ_{Γ^*}. Where again , $= \{\bigcirc((A \wedge \neg A) \wedge B)\}$, an agent reasoning in this way would have to abandon all the conclusions listed above except $\bigcirc B$.

5 Conditional oughts

Much of our normative reasoning involves ought statements that are not absolute but conditional, as in 'Given A, it ought to be that B', which we represent through the standard notation $\bigcirc(B/A)$.

In the literature, two general styles of analysis have been proposed for this kind of conditional ought. First, some writers have proposed an analysis involving some combination of an ordinary ought and an ordinary material conditional. Von Wright [37] originally suggested, for example, that the conditional ought should be analyzed through a formula of the form $\bigcirc(A \supset B)$, and Prior [27] suggested $A \supset \bigcirc B$ (these two suggestions are compared in Hintikka [15]). Others—such as Hansson [14], Lewis [21, Section 5.1], and van Fraassen [35]—have suggested that $\bigcirc(B/A)$ should be analyzed instead as a primitive dyadic modal construction within the general framework of conditional logic (these various suggestions are compared in Lewis [22]). As usual in conditional logic, this kind of analysis relies on a background ordering of the possible worlds, intended to represent a relation of similarity; the basic idea is then that $\bigcirc(B/A)$ should be true at a world if B is obligatory at the nearest or most similar worlds in which A is true.

As it turns out, there are problems with each of these two general lines of approach, concerning the degree of strengthening, or monotonicity, to be allowed in the antecedent of a conditional ought. If $\bigcirc(B/A)$ is analyzed either as $A \supset \bigcirc B$ or as $\bigcirc(A \supset B)$, then the conditional ought allows unrestricted strengthening in the antecedent: $\bigcirc(B/A \wedge C)$ then follows from $\bigcirc(B/A)$ for any statement C. This is easy to see. First, $(A \wedge C) \supset \bigcirc B$ is a consequence of $A \supset \bigcirc B$; and second, since $(A \wedge C) \supset B$ is a consequence of $A \supset B$ and oughts are closed under consequence, $\bigcirc((A \wedge C) \supset B)$ follows from $\bigcirc(A \supset B)$. On the other hand, the analysis of conditional oughts based on the semantic framework of conditional logic blocks strengthening in the antecedent completely. There is no way to derive $\bigcirc(B/A \wedge C)$ from $\bigcirc(B/A)$; there is no reason to think, just because B is obligatory in the nearest worlds in which A holds, that it should be obligatory also in the nearest worlds in which $A \wedge C$ holds.

It seems, however, that neither of these extreme approaches to strengthening in the antecedent of a conditional ought is correct, as we can see through an example. Suppose that an agent, hoping to abide by the proprieties, decides that his behavior should be governed by the following three oughts:

> You ought not to eat with your fingers,
>
> You ought to put your napkin on your lap,
>
> If you are served asparagus, you ought to eat it with your fingers.

Taking an unconditional ought, in the usual way, as an ought conditional on the universal truth $\top$, we can represent these three statements as $\bigcirc(\neg F/\top)$, $\bigcirc(N/\top)$, and $\bigcirc(F/A)$. Now it seems, intuitively, that the third of these oughts should override the first in case asparagus is served, so that in that case, the agent should not conclude that he ought not to eat with his fingers; but even if asparagus is served, nothing interferes with the second of these oughts, and so the agent should still conclude that he ought to put his napkin on his lap. That is: from the given premises, we want to derive $\bigcirc(N/A)$, but not $\bigcirc(\neg F/A)$.

The only way to derive $\bigcirc(N/A)$ in this situation, it seems, is by strengthening the antecedent of the second premise; a treatment of conditional oughts that simply rules out this kind of strengthening, such as those based on conditional logic, will not allow us to derive this conclusion. On the other hand, a treatment that allows unrestricted strengthening, such as those suggested by von Wright and Prior, 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 very general circumstances to apply also by default in more specific situations, unless they are overridden in those situations.

As far as I know, no treatment of conditional oughts based on any of the standard philosophical logics is able to model this kind of reasoning. It seems, for example, that the consequence relation associated with any appropriate theory 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. Adding a premise leads us to withdraw a conclusion.

The idea of analyzing conditional oughts within the general semantic framework of conditional logic led to certain departures from the earlier treatment that involved mixing ordinary oughts with material conditionals; but in retrospect, it seems that these departures may have been both too radical and too conservative. The way in which the departures seem too radical is by forcing us entirely to abandon strengthening, or monotonicity, in the antecedent of a conditional ought; for it appears that we may want to admit a certain amount of antecedent monotonicity. But the departures also seem too conservative because, although they do abandon antecedent monotonicity, they nevertheless treat conditional oughts within an ordinary logical framework, with a monotonic consequence relation; and it appears that the consequence relation that governs our reasoning about conditional oughts is itself nonmonotonic.

6 A nonmonotonic approach to conditional oughts

Because it seems to demand a nonmonotonic consequence relation, it is natural to hope that a useful theory of conditional oughts might be developed within the framework of nonmonotonic logic. This section first describes a preliminary attempt at developing such a theory, which generalizes the theory of simple oughts set out earlier, and then explores some problems with the preliminary proposal.

6.1 Conditioned extensions

We focus first on *ought contexts*: structures of the form $\langle W, , \rangle$, like default theories, except that the set of defaults is replaced by a set , of conditional ought statements, and the set of ordinary formulas is replaced by a single formula W. The two components of an ought context are supposed to represent both the background set of conditional oughts and the particular facts relevant to an agent's normative reasoning in that context.

Let us say that a conditional ought $\bigcirc(B/A)$ is *overridden* in the context $\langle W, , \rangle$ just in case there is a statement $\bigcirc(D/C) \in$, such that (i) $|W| \subseteq |C| \subset |A|$, (ii) $W \cup \{D, B\}$ is inconsistent, and (iii) $W \cup \{D\}$ is consistent. The idea here is that a conditional ought should be overridden in some context whenever another ought is applicable, more specific than the original, and inconsistent with the original. In the definition, clause (i) tells us that $\bigcirc(D/C)$ is both applicable in the context and more specific than $\bigcirc(B/A)$, while clause (ii) tells us that that the two oughts are inconsistent in the context. The point of clause (iii) is to prevent a conditional ought from being overridden by others that are themselves inconsistent in a particular context; for example, it prevents any conditional ought from being overridden in any context by a statement of the form $\bigcirc(\perp/C)$, with $\perp$ universally false.

Using this characterization of the circumstances under which conditional oughts are overridden, let us now define a set of sentences $\mathcal{E}$ as a *conditioned extension* of the context $\langle W, , \rangle$ just in case there is another set $\mathcal{F}$ such that

$$
\begin{aligned}
\mathcal{F} = \{B :\ & \bigcirc(B/A) \in , , \\
& |W| \subseteq |A|, \\
& \bigcirc(B/A) \text{ is not overridden in } \langle W, , \rangle, \\
& \neg B \notin \mathcal{E}\},
\end{aligned}
$$

and $\mathcal{E} = Cn[\{W\} \cup \mathcal{F}]$. This is, of course, a fixed point definition; and so there is reason to suspect, just as certain default theory lack conventional extensions, that certain ought contexts might lack conditioned extensions. Fortunately, the suspicion turns out to be unfounded.

Theorem 7 *Every ought context $\langle W, , \rangle$ has a conditioned extension $\mathcal{E}$.*

Proof Given $\langle W, , \rangle$, first define

$$\begin{aligned}\mathcal{F}_1 = \{B : \ & \bigcirc(B/A) \in , , \\ & |W| \subseteq |A|, \\ & \bigcirc(B/A) \text{ is not overridden in } \langle W, , \rangle\},\end{aligned}$$

and then let $\mathcal{F}_2$ be some maximal subset of $\mathcal{F}_1$ that is consistent with W; these are guaranteed to exist. Let $\mathcal{E} = Cn[\{W\} \cup \mathcal{F}_2]$. Evidently, $\mathcal{E}$ is a conditioned extension of $\langle W, , \rangle$ if and only if $\mathcal{F}_2 = \mathcal{F}$ (where $\mathcal{F}$ is as defined in the text); and it is clear from the definition of $\mathcal{F}_2$ that $\mathcal{F}_2 = \mathcal{F}$ just in case $\mathcal{F}_2 = \mathcal{F}_1 \cap \{B : \neg B \notin \mathcal{E}\}$. So suppose first that $B \in \mathcal{F}_1$ and $\neg B \notin \mathcal{E}$. Then B is consistent with $\{W\} \cup \mathcal{F}_2$, and so $B \in \mathcal{F}_2$, since $\mathcal{F}_2$ is maximal. Next, suppose $B \in \mathcal{F}_2$. Of course, $B \in \mathcal{F}_1$; and we must have $\neg B \notin \mathcal{E}$ as well, for otherwise we would have both B and $\neg B$ in $Cn[\{W\} \cup \mathcal{F}_2]$, and so $\mathcal{F}_2$ would not be consistent with W.
∎

Because conditioned extensions exist for every ought context, we can define a relation $\vdash_{CF}$ of conditional deontic consequence in the following way: where , is a set of conditional oughts, we let

Definition 2 , $\vdash_{CF} \bigcirc(B/A)$ *if and only if* $B \in \mathcal{E}$ *for some conditioned extension* $\mathcal{E}$ *of* $\langle A, , \rangle$.

This notion of conditional deontic consequence yields the correct results in the asparagus case: where

$$, = \{\bigcirc(\neg F/\top), \bigcirc(N/\top), \bigcirc(F/A)\},$$

the unique conditioned extension of $\langle A, , \rangle$ is $Cn[\{A, F, N\}]$; and so we have , $\vdash_{CF} \bigcirc(N/A)$, as desired, but we do not have , $\vdash_{CF} \bigcirc(\neg F/A)$. Just as in those theories based on the semantic framework of conditional modal logics, the present account of conditional oughts is nonmonotonic in the antecedent of the conditional. The unique conditioned extension of $\langle \top, , \rangle$, for example, is $Cn[\{\neg F, N\}]$, and so we have , $\vdash_{CF} \bigcirc(\neg F/\top)$; but, as mentioned, we do not have , $\vdash_{CF} \bigcirc(\neg F/A)$. In addition, however, the consequence relation $\vdash_{CF}$— unlike the consequence relation associated with conditional modal logics, and also unlike the earlier $\vdash_F$—is itself nonmonotonic. For example, let $\Gamma' = \Gamma - \{\bigcirc(F/A)\}$. Then the unique extension of $\langle A, \Gamma' \rangle$ is $Cn[\{A, \neg F, N\}]$, and so we have $\Gamma' \vdash_{CF} \bigcirc(\neg F/A)$; but again, although $\Gamma' \subseteq \Gamma$, we do not have $\Gamma \vdash_{CF} \bigcirc(\neg F/A)$.

The present account exhibits, also, several properties desirable in a conditional deontic logic. Since conditioned extensions are closed under logical consequence, the consequents of supported ought statements are closed under consequence as well: if $\Gamma \vdash_{CF} \bigcirc(B/A)$ and $B \vdash C$, then $\Gamma \vdash_{CF} \bigcirc(C/A)$. Again, by examining the definition of conditioned extensions we can see that conditional oughts are sensitive only to the propositions expressed by their antecedents, not to the particular sentences expressing those propositions: if $|A| = |B|$, then $\Gamma \vdash_{CF} \bigcirc(C/A)$ just in case $\Gamma \vdash_{CF} \bigcirc(C/B)$. And finally, an ought context $\langle W, \Gamma \rangle$ will have an inconsistent extension if and only if the formula W is itself inconsistent; and from this we conclude that $\Gamma \vdash_{CF} \bigcirc(\bot/A)$ if and only if $|A| = |\bot|$.[4]

It turns out, moreover, that the consequence relation $\vdash_{CF}$ is a conservative extension of the relation $\vdash_F$ described earlier, in the following sense:

Theorem 8 *Where Γ is a set of conditional oughts, let $\Gamma' = \{\bigcirc B : \bigcirc(B/A) \in \Gamma$ and $|A| = |\top|\}$. Then $\Gamma' \vdash_F \bigcirc C$ if and only if $\Gamma \vdash_{CF} \bigcirc(C/\top)$.*

Proof (sketch) We know by Theorem 1 that $\Gamma' \vdash_F \bigcirc C$ if and only if $C \in \mathcal{E}$ for some extension $\mathcal{E}$ of the default theory $\Delta_{\Gamma'}$. Reflection on the construction underlying Theorem 2.1 of Reiter [28] shows that $\mathcal{E}$ is an extension of $\Delta_{\Gamma'}$ just in case there is a set $\mathcal{F}$ such that

$$\begin{aligned}\mathcal{F} = \{B :\ & \bigcirc B \in \Gamma', \\ & \neg B \notin \mathcal{E}\},\end{aligned}$$

and $\mathcal{E} = Cn[\mathcal{F}]$. It is easy to see that no conditional ought can be overridden in any context of the form $\langle \top, \Gamma \rangle$; and of course $|\top| \subseteq |A|$ if and only if $|A| = |\top|$. Therefore, we can

[4]The three properties described in this paragraph can be compared to the rules RCOEA, RCOM, and COD from Chellas[6, Section 10.2].

conclude that $\mathcal{E}$ is an extension of $\Delta_{\Gamma'}$ just in case there is a set $\mathcal{F}$ such that

$$\begin{aligned}\mathcal{F} = \{B : \quad & \bigcirc(B/A) \in , , \\ & |\top| \subseteq |A|, \\ & \bigcirc(B/A) \text{ is not overridden in } \langle \top, , \rangle, \\ & \neg B \notin \mathcal{E}\},\end{aligned}$$

and $\mathcal{E} = Cn[\{\top\} \cup \mathcal{F}]$; that is, just in case $\mathcal{E}$ is a conditioned extension of $\langle \top, , \rangle$. The theorem then follows at once from the definition of the relation $\vdash_{CF}$. ∎

From this result and the discussion in Section 4, we can conclude that the consequence relation $\vdash_{CF}$, like $\vdash_F$, satisfies neither reflexivity nor cut.

6.2 Problems with the theory

This account of conditional deontic consequence exhibits a number of advantages not found in the usual modal approaches. The consequence relation $\vdash_{CF}$ is itself nonmonotonic, as is the antecedent place in derived conditional oughts; but unlike those accounts based on the the semantic framework of conditional logic, the current account does allow for a certain amount of strengthening, or monotonicity, in the antecedent of these derived oughts. And the theory generalizes the earlier treatment of reasoning in the presence of normative conflict, which already lies beyond the scope of modal approaches to deontic logic.

However, the present account of conditional deontic consequence is beset by several problems, and so can be taken, at best, only as a preliminary. We close simply by listing four of these problems.

First and most important, the account does not allow any kind of transitivity, or chaining, across conditional oughts. We cannot derive $\bigcirc(C/A)$ from a premise set consisting of $\bigcirc(C/B)$ and $\bigcirc(B/A)$; and in particular, taking simple oughts as oughts conditional upon $\top$, we cannot derive $\bigcirc B$ from $\bigcirc(B/A)$ and $\bigcirc A$. Of course, this situation is no worse than the situation in those accounts based on conditional logics, which also forbid transitivity of the conditional, and of the deontic conditional. However, the nonmonotonic framework allows for a new possibility that is not present in these standard logics—the possibility that transitivity should hold as a defeasible rule, subject to override. This is, in fact, exactly how transitivity is supposed to work in a number of application areas of nonmonotonic logics, such as the kind of reasoning supported by nonmonotonic inheritance hierarchies. Here, we would want to conclude, for example, that Elton is underpaid given only the premises that Elton is a musician and that musicians tend to be underpaid; but we would allow this conclusion to be overridden by the additional information that Elton is a rock star, where rock stars are a particular class of musicians that tend not to be underpaid.

I think that it would be natural to incorporate this kind of defeasible transitivity also into an account of conditional deontic reasoning; but I have not attempted to do so here because the task of combining defeasible transitivity with a proper treatment of overriding (known in the inheritance literature as "preemption") presents significant technical and conceptual problems. In spite of the efforts of a number of researchers—including Boutilier [4], Delgrande [8], Geffner [12], and Pollock [26]—I know of no solution to these problems for a language as expressive as propositional calculus that is generally accepted; and the matter is not settled even for the very simple language of inheritance hierarchies, as can be seen from Horty [17] and Touretzky et al. [32].

The second problem faced by the present account of conditional deontic consequence concerns the matter of reasoning with disjunctive antecedents. If , $= \{\bigcirc(C/A), \bigcirc(C/B)\}$, for example, it seems that we should be able to conclude from , that $\bigcirc(C/A \vee B)$; but in fact, the only conditioned extension of $\langle A \vee B, , \rangle$ is $Cn[\{A \vee B\}]$, and so we do not have get this result. This kind of problem is, of course, well known in the context of default logic; and several proposals, such as that of Gelfond et al. [13], have been put forth for modifying standard default logic so that it yields the desirable conclusions in the presence of disjunctive information. Given the similarity between conditional extensions of ought contexts and ordinary extensions of default theories, it should not be too difficult to adapt these proposals to the present case; but it is not simply an exercise, since the adaptation would have to involve extending the notion of overriding to apply properly to disjunctive antecedents.

The third problem with the present theory concerns a detail in the treatment of overridden oughts. According to the present theory, an conditional ought can be overridden only by a single opposing statement, which is both applicable in the context and more specific. However, there appear to be cases in which it is natural to suppose that an ought, although not overridden by a single opposing rule, might be overridden by a set of opposing rules. Suppose, for example, that , $= \{\bigcirc(Q/\top), \bigcirc(\neg(P \wedge Q)/A), \bigcirc(P/A)\}$. Here, it seems that in the context $\langle A, , \rangle$, the first rule should be overridden by the second two taken together, although it is not overridden by either individually.

The final problem concerns yet another detail in the present treatment of overriding. Suppose an ought statement is overridden by another which is itself overridden. What is the status of the original? According to the present treatment, it remains out of play; but it is also possible to imagine that the original rule should then be reinstated. As an example, let , $= \{\bigcirc(Q/\top), \bigcirc(P/A), \bigcirc(\neg P/A \wedge B)\}$, and consider the context $\langle W, , \rangle$, where W is the formula $(A \wedge B) \wedge \neg(P \wedge Q)$. Of course, the first rule in , is overridden in this context by the second, but the second is likewise overridden by the third. Since overridden rules remain out of play, according to the the present treatment, this context has $Cn[\{W, \neg P\}]$ as its only conditioned extension; and so we do not have $\bigcirc(Q/W)$. But it does not seem unreasonable to modify the present treatment so that the statement $\bigcirc(Q/\top)$ is reinstated in this context, since the rule that overrides it is itself overridden. In that case, we would have $Cn[\{W, \neg P, Q\}]$ as a conditioned extension; and so we would be able to derive $\bigcirc(Q/W)$ from , . The issue of reinstatement in inheritance hierarchies is explored in detail in Horty [17] and in Touretzky et al. [33].

The problems pointed out here with the present account of conditional deontic consequence are serious, but I do not feel that they should lead us to abandon the project of designing a conditional deontic logic using the techniques of nonmonotonic logic. In fact, none of these problems is unique to the deontic interpretation of the background nonmonotonic theory; instead, they reflect more general difficulties in nonmonotonic reasoning, which surface here just as they surface elsewhere. Of course, it is impossible to offer a final evaluation of the nonmonotonic approach to conditional deontic reasoning until these issues with the underlying logical framework are resolved. But the approach does seem to be promising;

and it may be that, in bringing the techniques of nonmonotonic logic into contact with the new data provided by normative reasoning, we will not only discover new possibilities for the construction of deontic logics, but gain a deeper understanding of the underlying nonmonotonic logics as well.

Acknowledgments

I am grateful to Johan van Benthem for showing me in [34] that the connections between deontic and nonmonotonic logic are more extensive than I had imagined; and to Nuel Belnap, Jon Doyle, Matt Ginsberg, Michael Slote, and Rich Thomason for a variety of helpful discussions. Anyone familiar with Thomason's [31] will notice that I have plagiarized the title of his paper. This work has been supported by the National Science Foundation under Grant No. IRI-9003165.

References

[1] A. Anderson, N. Belnap, and J.M. Dunn. *Entailment: The Logic of Relevance and Necessity, volume 2*, Princeton University Press (1992).

[2] N. Belnap. How a computer should think. In *Contemporary Aspects of Philosophy*, G. Ryle (ed.), Oriel Press (1977), pp. 30–56.

[3] N. Belnap. A useful four-valued logic. In *Modern Uses of Multiple-valued Logic*, J.M. Dunn and G. Epstein (eds.), D. Reidel Publishing Company (1977), pp. 8–37.

[4] C. Boutilier. *Conditional Logics for Default Reasoning and Belief Revision.* PhD Dissertation, Computer Science Department,University of Toronto (1992). Available as Technical Report KRR-TR-92-1.

[5] B. Chellas. Conditional obligation. In *Logical Theory and Semantic Analysis*, S. Stenlund (ed.), D. Reidel Publishing Company (1974), pp. 23–33.

[6] B. Chellas. *Modal Logic: An Introduction.* Cambridge University Press (1980), xii+295 pp.

[7] E. Davis. *Representations of Commonsense Knowledge.* Morgan Kaufmann Publishers (1990).

[8] J. Delgrande. An approach to default reasoning based on a first-order conditional logic: revised report. *Artificial Intelligence*, vol. 36 (1988), pp. 63–90.

[9] D. Dennett. The moral first aid manual. In *The Tanner Lectures on Human Values, vol. VIII.* Cambridge University Press (1988), pp. 119–147.

[10] A. Donagan. Consistency in rationalist moral systems. *The Journal of Philosophy*, vol. 81 (1984), pp. 291–309.

[11] P. Foot. Moral realism and moral dilemma. *Journal of Philosophy*, vol. 80 (1983), pp. 379–398.

[12] H. Geffner. *Default Reasoning: Causal and Conditional Theories.* PhD Dissertation, Computer Science Department, UCLA (1989). Available as Technical Report 137.

[13] M. Gelfond, V. Lifschitz, H. Przymusinska, and M. Truszczynski. Disjunctive defaults. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference* (KR-91), Morgan Kaufmann (1991).

[14] B. Hansson. An analysis of some deontic logics. In *Deontic Logic: Introductory and Systematic Readings*, R. Hilpinen (ed.), D. Reidel Publishing Company (1971), pp. 121–147.

[15] J. Hintikka. Deontic logic and its philosophical morals. In J. Hintikka, *Models for Modalities: Selected Essays*, D. Reidel Publishing company (1969), pp. 184–214.

[16] J. Horty. Moral dilemmas and nonmonotonic logic. Forthcoming in *Journal of Philosophical Logic.* A preliminary version appears in *Proceedings of the First International Workshop on Deontic Logic in Computer Science*, J.-J. Ch. Meyer and R. Wieringa

(eds.), Technical Report, Computer Science Department, Free University, Amsterdam, The Netherlands (1991).

[17] J. Horty. Some direct theories of nonmonotonic inheritance. Forthcoming in *Handbook of Logic in Artificial Intelligence and Logic Programming, vol. 2: Nonmonotonic Reasoning and Uncertain Reasoning*, D. Gabbay and C. Hogger (eds.), Oxford University Press.

[18] H. Kautz. The logic of persistence. In *Proceedings of the Fifth National Conference on Artificial Intelligence* (AAAI-86), Morgan Kaufmann (1986), pp. 401–405.

[19] S. Kraus, D. Lehmann, and M. Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, vol. 44 (1990), pp. 167–207.

[20] E. J. Lemmon. Moral dilemmas. *Philosophical Review*, vol. 70 (1962), pp. 139–158.

[21] D. Lewis. *Counterfactuals*. Harvard University Press (1973).

[22] 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.

[23] R. B. Marcus. Moral dilemmas and consistency. *Journal of Philosophy*, vol. 77 (1980), pp. 121–136.

[24] J. Minker. On indefinite databases and the closed world assumption. In *Proceedings of the Sixth Conference on Automated Deduction*, Lecture Notes in Computer Science, Springer Verlag (1982), pp. 292–308.

[25] J. McCarthy. Circumscription—a form of nonmonotonic reasoning. *Artificial Intelligence*, vol. 13 (1980). pp. 27–39.

[26] J. Pollock. How to reason defeasibly. Manuscript, Philosophy Department, University of Arizona (1991).

[27] A. N. Prior. *Formal Logic*. Oxford University Press (1962).

[28] R. Reiter. A logic for default reasoning. *Artificial Intelligence*, vol. 13 (1980), pp. 81–132.

[29] J. P. Sartre. *L'Existentialisme est un Humanisme*. Nagel (1946). Translated as "Existentialism is a Humanism" in *Existentialism from Dostoevsky to Sartre*, W. Kaufmann (ed.), Meridian Press (1975).

[30] Y. Shoham. *Reasoning about Change*. MIT Press (1988).

[31] R. Thomason. Deontic logic as founded on tense logic. In *New Studies in Deontic Logic*, R. Hilpinin and D. Follesdal (eds.), D. Reidel Publishing Company (1981), pp. 141–152.

[32] 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.

[33] D. Touretzky, R. Thomason, and J. Horty. A skeptic's menagerie: conflictors, preemptors, reinstaters, and zombies in nonmonotonic inheritance. In *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence* (IJCAI-91), Morgan Kaufmann Publishers (1991), pp. 478–483.

[34] J. van Benthem. Personal correspondence (March 1992).

[35] B. van Fraassen. The logic of conditional obligation. *Journal of Philosophical Logic*, vol. 1 (1972), pp. 417–438.

[36] B. van Fraassen. Values and the heart's command. *The Journal of Philosophy*, vol. 70 (1973), pp. 5–19.

[37] G. H. von Wright. Deontic logic. *Mind*, vol. 60 (1951), pp. 1–15.

[38] M. Wellman and J. Doyle. Preferential semantics for goals. In *Proceedings of AAAI-91*, MIT Press (1991), pp. 698–703.

[39] R. J. Wieringa and J.-J. Ch. Meyer. Applications of deontic logic in computer science: a concise overview. In *Proceedings of the First International Workshop on Deontic Logic in Computer Science*, J.-J. Ch. Meyer and R. Wieringa (eds.), Technical Report, Computer Science Department, Free University, Amsterdam, The Netherlands (1991).

[40] 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.