# A Preference-Based Approach to Defeasible Deontic Inference

**James Delgrande**
Simon Fraser University, Burnaby, B.C., Canada V5A 1S6
jim@cs.sfu.ca

**Abstract**

In this paper we present an approach to defeasible deontic inference. Given a set of rules **R** expressing conditional obligations and a formula $\gamma$ giving contingent information, the goal is to determine the most desirable outcome with respect to this information. Semantically, the rules **R** induce a partial preorder on the set of models, giving the relative desirability of each model. Then the set of minimal $\gamma$ models characterises the best that can be attained given that $\gamma$ holds. A syntactic approach is also given, in terms of maximal subsets of material counterparts of rules in **R**, and that yields a formula that expresses the best outcome possible given that $\gamma$ holds. These approaches are shown to coincide, providing an analogue to a soundness and completeness result. Complexity is not unreasonable, being at the second level of the polynomial hierarchy when the underlying logic is propositional logic. The approach yields desirable and intuitive results, including for the various "paradoxes" of deontic reasoning. The approach also highlights an interesting difference in how specificity is dealt with in nonmonotonic and deontic reasoning.

## 1 Introduction

With the advent and expected proliferation of artificial agents, it is crucial that these agents operate not just to accomplish their goals and aims, but that they do so in conformity with commonsense guidelines and norms, and with respect to various moral and ethical criteria. Deontic logic is the area of logic that addresses notions such as these, dealing with concepts such as obligation, permission, prohibition, and the like. It can be thought of as stipulating what an agent should or should not do, or what it may do or ought to do. Thus, an assertion such as *one shouldn't eat with the hands* states what is preferable for an agent to do, even though the agent may be perfectly capable of eating with its hands. In the general case, a deontic assertion may be overruled by a more specific one, such as when eating pizza one should eat with the hands. While the focus here is on the behaviour of agents, the area has also seen significant activity in a variety of other fields, including linguistics, philosophy, and law.

Approaches to deontic reasoning have often been expressed as a modal logic, with an operator $O(\phi)$, read as "$\phi$ ought to be done (or ought to be the case)," or a binary modal operator $O(\phi|\psi)$, read as "if $\psi$ is the case (or is done) then $\phi$ ought to be the case (done)". Other formalisms have also been employed, including approaches to nonmonotonic reasoning. However, it has proven to be very difficult to provide an appropriate system for reasoning about obligation.

In this paper the goal is not to provide a logic of obligation per se. Instead, we are interested in using the information in a set of deontic assertions to determine what may ideally be attained. The question we are concerned with is, given a set of deontic assertions (perhaps obtained in part via reasoning in some deontic logic) and contingent information about a domain, what is the best that can be expected or attained? For example, given that *one should not eat with the hands* but *if eating pizza, one should eat with the hands*, and that one is eating out on a Thursday, the best outcome is to not eat with the hands; given the additional information that a veggie pizza is served, the best option is that one eats with the hands. The overall goal is to provide a justifiable, computational account for dealing with a set of such assertions.

We begin with an underlying logic that contains classical propositional logic. A set of rules **R** provides assertions of conditional obligation, where the antecedent and consequent of these rules are formulas in the underlying logic. Two means of reasoning with the set of rules are developed, corresponding to semantic and proof-theoretic aspects of the approach. First, the rules induce a partial preorder on the set of models of the language, where the ordering gives the relative "goodness" of a model. Then, given that $\gamma$ is the case, the minimal $\gamma$ models in the preorder determine what ought to be the case or, equivalently, the best possible overall outcome. Second, a syntactic approach is given that yields a formula that expresses the best outcome possible, given that $\gamma$ is the case. This is carried out in terms of maximal subsets of material counterparts of rules in **R** consistent with $\gamma$. These approaches are shown to coincide, in that for any formula $\gamma$, the set of least $\gamma$ models in the ordering exactly characterises the "expansion" of $\gamma$ according to **R**.

The resulting approach is simple, but arguably works well: a wide variety of examples are handled appropriately, including the standard "paradoxes" of deontic reasoning. The specificity of a rule's antecedent is taken into account, so that more specific rules override less specific rules. As well, reasoning in the case of violated obligations is appropriately handled. In each part of the approach, a single outcome is obtained: semantically a single preorder over models is induced while syntactically one obtains a single formula. Complexity is manageable, with the main decision

problems at the second level of the polynomial hierarchy for propositional logic.

The next section goes over background material. Section 3 describes the approach informally and Section 4 gives the formal details. The following section briefly considers extensions of the approach. Section 6 discusses how the approach differs from conditional accounts of nonmonotonicity, after which we briefly conclude.

# 2 Background

## 2.1 Notation

We will generally work with an arbitrary logic that contains classical propositional logic, expressed in a language $\mathcal{L}$ that contains the propositional connectives $\neg$, $\wedge$, $\vee$, $\supset$ and $\equiv$. (The assumption of propositional logic is not a requirement, but it makes the presentation easier.) Examples will be given using classical propositional logic. Lower case English letters will denote atoms, and a propositional model may be given as a juxtaposition of literals. Formulas will be denoted by lower case Greek letters $\phi$, $\psi$, ..., possibly primed or subscripted. $\mathcal{M}$ is the set of models, or possible worlds. Individual possible worlds are denoted by $w$, possibly primed or subscripted. The fact that formula $\phi$ is true at possible world $w$ is denoted $w \models \phi$. For a formula (set of formulas) $\phi$, $[\phi]$ is the set of models of $\phi$, i.e. $[\phi] = \{w \in \mathcal{M} \mid w \models \phi\}$. All other notions, such as logical entailment, validity, etc., are standard. $\top$ is taken to be some tautology, while $\bot$ is defined to be $\neg\top$.

We will deal with a finite set of rules $\mathbf{R}$ where $r \in \mathbf{R}$ is of the form $\phi \rightarrow \psi$ for $\phi, \psi \in \mathcal{L}$. For a rule $r : \phi \rightarrow \psi$ define the *body* $b(r) = \phi$ and *head* $h(r) = \psi$. These rules express conditional obligations, and so $\phi \rightarrow \psi$ expresses that, all other things being equal, if $\phi$ then it is better (obligatory, etc.) that $\psi$. We will later see how to express conditional permission in the framework.

The relation $\preceq$ will be a partial preorder (i.e. a reflexive and transitive binary relation) on possible worlds, with $\prec$ as its strict part. Generally $\preceq$ will be induced from the set $\mathbf{R}$, and denoted $\preceq_{\mathbf{R}}$. Define:

$$\min(\phi, \preceq) \doteq \{w \mid w \models \phi \text{ and } \nexists w' \text{ s.t. } w' \prec w, w' \models \phi\}$$

Last, in examples the following notation will be convenient: For worlds $w_1, \ldots, w_m$,

$$\langle w_1, \ldots, w_n \rangle \prec \langle w_{n+1}, \ldots, w_m \rangle$$

will indicate that $w_i \prec w_j$ for $1 \leq i \leq n < j \leq m$.

## 2.2 Related Work

Deontic logic has generally been investigated as a type of modal logic, beginning with (von Wright 1951). The most familiar of these logics is called *Standard Deontic Logic* (SDL); see (Hilpinen and McNamara 2013) for a survey. SDL uses a unary modal operator $O\phi$, read as "it is obligatory that $\phi$" or "it ought to the the case that $\phi$". Other operators, such as for permission or prohibition, are expressed in terms of $O$. SDL coincides with the modal logic KD (Chellas 1980), which is characterised by serial Kripke structures. In proof theoretic terms, the logic extends propositional logic by the axioms and rule of inference:

**K:** $O(\phi \supset \psi) \supset (O\phi \supset O\psi)$

**D:** $O\phi \supset \neg O\neg\phi$

**N:** From $\vdash \phi$ infer $\vdash O\phi$

While the logic is simple and rather weak,[1] it also allows what are argued to be unintuitive results. For example, conflicting obligations of the form $O\phi \wedge O\neg\phi$ are inconsistent. As well, one obtains what has been called Ross's paradox:

**Example 2.1** *1. You should mail the letter.*

*2. You should mail the letter or burn it.*

The second assertion is a logical consequence of the first. However, this seems counterintuitive; for example, if I'm unable to mail the letter, then I can at least satisfy the second, derived, obligation by burning the letter. As a consequence, yet weaker logics have been proposed. Unsurprisingly such logics have been criticised as being overly weak.

Another issue deals with what should be done in the case of a violated (or *contrary-to-duty*) obligation. Consider the so-called *Chisholm paradox* (Chisholm 1963):

**Example 2.2** *1. You ought to help your neighbour.*

*2. If you help your neighbour you should tell them.*

*3. If you don't help your neighbour, you shouldn't tell them.*

*4. You don't help your neighbour.*

While each assertion appears independent of the other three, in an encoding with a unary operator the fourth is derivable from the first three, see for example (McNamara 2019). McNamara (2019) goes on to argue that this example cannot be adequately represented by any combination of a unary deontic operator and a material conditional.

Another, related difficulty is known as Forrester's paradox (Forrester 1984). Consider the assertions[2]

**Example 2.3** *1. Smith should not kill Jones.*

*2. If Smith kills Jones, he should do so gently.*

*3. Smith kills Jones.*

In the standard account of deontic logic these sentences (together with the implied assertion that if someone is killed gently then they are killed) are jointly inconsistent.

As a consequence, there has been substantial work regarding *conditional obligation*. A conditional obligation "if $\phi$ then it should be that $\psi$" is usually written $O(\psi/\phi)$. Semantically this can be interpreted as saying that in the "best" $\phi$ worlds, $\psi$ is also true. A model then is most often composed of an ordering $\preceq$ on possible worlds, where for worlds $w_1$, $w_2$ if $w_1 \preceq w_2$ then $w_1$ is "better" than $w_2$. The original work here is (Hansson 1971). Lewis (1973), in an exploration of counterfactuals, suggests such a preference-based semantics for deontic logic. Hansson (1990; 2004) also addresses plausibility and preference in unconditional deontic logic. Goble (2003) similarly presents results for monadic

[1] KD is also proposed as a logic of probabilistic certainty, according to Gärdenfors (1975), which is to say KD (and so SDL) does not distinguish obligation from probabilistic certainty.

[2] A less dramatic version of this problem reflecting the kind of application we have in mind is: *The door should not be closed*; *if the door is closed it should be done gently*; *the door is to be closed.*

and conditional deontic logics based on a preference ranking given by a total preorder on worlds.

Binary modal operators for conditional obligation allow a representation of not just contrary-to-duty obligations, but also of *prima facie* obligations or defeasible obligations. Consider the following example:[3]

**Example 2.4** *1. You should not eat with your hands.*

*2. If eating pizza, you should eat with your hands.*

Thus if one is eating pizza, then one ought to eat with the hands; while the antecedent of the first obligation is also satisfied, it is "overridden" by the second. That is, as with normality defaults, one wants to apply the more specific rule. Contrast this with Example 2.3, where the second obligation states what should be the case when the first is violated.

Horty (1993; 2014) has suggested that deontic reasoning be situated in a nonmonotonic framework, not just for addressing conflicting norms and handling specificity, but also for deriving unconditional obligations, or so-called *dyadic detachment*. (Thus, in this last case, from $O\phi$ and $O(\psi/\phi)$ obtaining $O\psi$ "when justified".) In his account, a conditional obligation $O(\psi/\phi)$ is treated as an inference rule, much as in default logic (Reiter 1980). One obligation $O(\psi_1/\phi_1)$ *overrides* another $O(\psi_2/\phi_2)$ just if $\phi_1 \supset \phi_2$ and $\psi_1 \wedge \psi_2$ is inconsistent with contingent domain knowledge. A fixed-point definition is given that specifies a set of *extensions*, or maximal sets of formulas describing what one ought to do. Specificity is taken into account, and so for Example 2.4, if one is eating pizza, only the second obligation is considered. Related work includes (McCarty 1994). van der Torre (1994) analyzes violated obligations and defeasibility in this framework, and extends the approach to cover violated obligations; this is extended to a general conditional logic framework in (van der Torre and Tan 1995).

Factual detachment is also considered in (Straßer 2011), which proposes an approach with focus on specificity and contrary-to-duty instances. Normative reasoning has also been addressed in abstract argumentation systems; for example, (Straßer and Arieli 2015) model deontic conflict, contrary-to-duty and specificity via argumentative attacks. Ryu (1995) augments a conditional deontic logic (due to van Fraassen) with defeasible reasoning, also taking specificity into account. Other work, including (Ryu and Lee 1995; Antoniou, Dimaresis, and Governatori 2009), bases defeasible deontic reasoning on *defeasible logic* (Nute 1994).

Bartha (1999) presents an approach for defeasible detachment in the branching time formalism of (Horty and Belnap 1995). He adopts the same definition of one conditional obligation overriding another, but uses it to induce a preference order over histories. Minimal elements in the ordering determine the definite obligations of an agent. Kowalski and Satoh (2018) addresses defeasible obligation using abductive logic programs, in which obligations are treated as a means of goal satisfaction.

In a different direction, Makinson and van der Torre (2000) introduce *input-output logic*. A *conditional norm* is represented as a pair $(a, x)$ of formulas in propositional logic, where $a$ is the *input* and represents a condition or situation, while $x$ is the *output* and expresses what is desirable, or what should hold. Various systems are specified by adding rules governing such norms. The basic system contains strengthening of antecedents, and so is too strong to handle defeasible obligations of the pizza-eating variety. This is addressed by employing constraints, using notions from belief change and nonmonotonic reasoning.

[3]This is a syntactic variant of an example due to John Horty, in which one eats asparagus instead of pizza.

## 3 The Approach: Intuitions

It has proven to be very difficult to provide a logic of obligation, that is, a formal account of what obligations follow logically from others. Here we deal with an orthogonal problem: Given a set of deontic rules (perhaps obtained in part via reasoning in some understood deontic logic) along with domain-specific information, we wish to determine the best overall outcome that can be attained. This distinction between the goals of a deontic logic and the present approach is well illustrated by Forrester's paradox. As a problem of *logic*, it has been called "the deepest problem of all" (Goble 1991). However, for commonsense reasoning it is (at least seemingly) straightforward: Given the assertions in Example 2.3, it is clearly not acceptable that Smith kill Jones. If Smith nonetheless does kill Jones, then the best outcome is that he does it gently. No problem arises since in the latter case it is a given that a person is going to be killed and the issue is to determine the best thing that can be done in these circumstances.

The general problem we consider is the following: We are given an agent in some domain; we have a set of general rules to guide the agent's behaviour; and we want to determine what the best is that can be attained in any set of circumstances. An important point to note is that, given a set of conditional obligations $\mathbf{R}$ and contingent information $\gamma$, the goal is not to determine what is obligatory with respect to $\gamma$. Rather, the goal is to determine what should be the case, or what would be best, given $\mathbf{R}$ and $\gamma$. To illustrate, consider the rule "if the sun is shining, you should put on sunglasses".[4] Then, given that one is not wearing sunglasses, it is (deontically) preferable that the sun is not shining, since otherwise the obligation would be violated. In contrast, it doesn't make sense to say that if one is not wearing sunglasses then it is obligatory that the sun is not shining.

To this end, a logic with language $\mathcal{L}$ containing classical propositional logic is given for describing a domain. In addition, we have a set of rules $\mathbf{R}$ where $r \in \mathbf{R}$ is of the form $\phi \rightarrow \psi$ for $\phi, \psi \in \mathcal{L}$. A rule $\phi \rightarrow \psi$ has the reading "if $\phi$ is true then $\psi$ is obligatory (or should be the case or should be done)."[5] We also allow rules of the form $\phi \rightarrow_p \psi$ with the reading "if $\phi$ is true then $\psi$ is permissible"; however, as will be shown in the next section, rules of permission can be reduced to obligations. A key stance of the approach is that

[4]I thank a referee for suggesting this example.

[5]There is an issue as to whether an obligation applies to an action, outcome, or both. We assume this is dealt with in the language $\mathcal{L}$, which could refer to time points, actions, etc.

the rules in **R** *constrain* an agent's behaviour;[6] this stance can be expressed informally as: *Anything is permissible unless ruled out by an obligation in* ***R***. Thus, if $\mathbf{R} = \emptyset$ an agent is constrained by no obligations. This also has the effect that rules expressing permissions serve only to cancel obligations; again this is covered in the next section.

Two (equivalent) approaches are given: In the first, the rules **R** induce a partial preorder over worlds, and the minimal $\gamma$ worlds characterise what the best is that may be attained. In the second, the rules are mapped to material conditionals, and a procedure is given for determining maximal consistent (with respect to $\gamma$) sets of rules, again expressing the best that may be attained. For the first approach, a key intuition is that one wants to avoid (or not-prefer) possible worlds with violated obligations. Thus, for $\phi \rightarrow \psi$, if $w_1 \models \phi \supset \psi$ and $w_2 \models \phi \wedge \neg\psi$ then, all other things being equal, $w_1$ is preferred to $w_2$, that is, $w_1 \prec w_2$. To see why this is desirable, consider the assertion *if it's raining one should take an umbrella*, say $r \rightarrow u$. Clearly a $r \wedge u$ world is preferable to a $r \wedge \neg u$ world. However a $\neg r$ world is also preferable to a $r \wedge \neg u$ world: the former violates no norm and is thus acceptable while the latter is undesirable. This leads to the key intuition (or perhaps stance) in specifying a preference ordering, and that is that one wants to avoid worlds with violated obligations; worlds with verified obligations (e.g. with $r \wedge u$) and inapplicable obligations (viz. $\neg r$) are deontically "the same".

This differs from approaches involving normality conditionals (e.g. (Boutilier 1994; Geffner and Pearl 1992)) as well as most approaches to deontic conditionals (e.g. (Horty 1993; Bartha 1999; van der Torre and Tan 1999; Goble 2003)). In these approaches for $\phi \rightarrow \psi$, if $w_1 \models \phi \wedge \psi$ and $w_2 \models \phi \wedge \neg\psi$ then, all else being equal, $w_1 \prec w_2$; if $w_3 \models \neg\phi$, then there is no preference relation between this world and $w_1$ or $w_2$. We return to this point in Section 6.

A further consideration is that for conflicting rules, a stronger, or more specific, rule takes precedence over a weaker or less specific rule. For example, consider a first-aid agent with rules that if someone is overheated ($h$), they should be given water ($d$, for "drink"), unless they have a decreased level of consciousness ($\neg c$) (since they might aspirate the water):

$$r_1 : h \rightarrow d, \quad r_2 : h \wedge \neg c \rightarrow \neg d. \tag{1}$$

It is better to give a hot, conscious subject water than not to do so; i.e. we would have the preference between worlds: $hcd \prec hc\neg d$. However consider the two worlds $h\neg c\neg d$ and $h\neg cd$. In both worlds a person is hot and not conscious, and is given water in one world but not in the other. Each world violates a rule; however the more specific rule $r_2$ should "override" the less specific rule $r_1$ and we would have the preference: $h\neg c\neg d \prec h\neg cd$.

There is a second means by which one rule may take precedence over another, in the case of violated obligations. Consider the Forrester paradox, which can be expressed by the rules $\top \rightarrow \neg m$ and $m \rightarrow g$. All other things being equal, a world satisfying $\neg m$ will be most preferred (since no obligation is violated); a world with $mg$ would be less preferred (since the first obligation is violated) and a world with $m\neg g$ would be least preferred since both obligations are violated.

Given these considerations, a preference order on possible worlds can be defined by, for $w_1, w_2 \in \mathcal{M}$, $w_1$ is *at least as preferred as* $w_2$, $w_1 \preceq w_2$, just if the set of violated "applicable" rules at $w_1$ is a subset of those at $w_2$. This is a rather simple notion, since all rules are treated as being equally important, but it provides a basic approach from which further elaborations can be addressed (see Section 5). This yields a preference order over worlds. Then, given contingent information $\gamma$, we have that $\psi$ is a preferred outcome if it is true at all minimal $\gamma$ worlds.

This provides a semantic characterization and justification of the best that can be attained in different circumstances. However it is not particularly suited for inference. Consequently, an equivalent computational account is also provided, which allows a direct implementation of the approach. Informally, rules are "compiled" into material conditionals taking specificity into account and, given a formula $\gamma$ representing what is contingently true, maximal consistent subsets of the rules with respect to $\gamma$ are defined. The disjunction of these sets is then shown to exactly characterise the first approach. The next section gives the formal details.

## 4 The Approach: Formal Details

As described, we begin with a base logic with language $\mathcal{L}$ containing classical propositional logic for describing a domain. A set of rules **R** expresses conditional obligations where $r \in \mathbf{R}$ is of the form $\phi \rightarrow \psi$ for $\phi, \psi \in \mathcal{L}$. No restriction is placed on the form of the rules in **R**; for example, $a \rightarrow \neg a$ is fine, as is the pair $a \rightarrow b$, $a \rightarrow \neg b$.

For dealing with specificity, we identify those pairs of rules whose heads are jointly inconsistent, and where the body of one implies the body of the other. (Recall that Horty (1993) uses the same definition.)

**Definition 4.1**

$r_2 \trianglelefteq r_1$ *if* $\vdash b(r_1) \supset b(r_2)$ *and* $\vdash \neg(h(r_1) \wedge h(r_2))$

$r_2 \triangleleft r_1$ *if* $r_2 \trianglelefteq r_1$ *and not* $r_1 \trianglelefteq r_2$

$r_2 \triangleleft r_1$ is read as $r_1$ *(strictly) dominates* $r_2$. Intuitively, if the bodies of both rules are satisfied, the more specific rule, $r_1$, takes precedence over $r_2$. Clearly $\triangleleft$ is irreflexive and is not transitive.

Possible worlds with violated obligations are less desirable. Consider a rule $r \in \mathbf{R}$ and world $w \in \mathcal{M}$ where $w \models b(r) \wedge \neg h(r)$. If there is no rule $r'$ that dominates $r$ (i.e. no $r'$ is such that $r \triangleleft r'$) then this counts against the overall preferability of $w$. If there is a rule $r' \in \mathbf{R}$ such that $r \triangleleft r'$ and $w \models b(r') \wedge h(r')$, then this "overrides" the falsified obligation $r$. This isn't the case if $w \models \neg b(r')$. Consequently we want to consider those falsified rules like $r$ such that for all $r \triangleleft r'$ we have $w \not\models b(r')$.

Consider our earlier example (1), in which if a person is overheated then you should given them water, $h \rightarrow d$, and let $w$ be a world where $w \models h \wedge \neg d$. Considering just the rule

[6] This is not the only role or view regarding norms; for example, they have been employed in multi-agent systems, and provide mechanisms that facilitate collaboration and cooperation.

$h \rightarrow d$, this would count against the overall preferability of $w$. However, we also have the obligation that if someone is overheated and has a decreased level of consciousness then one should not give them water: $h \wedge \neg c \rightarrow \neg d$. There are now two possibilities:

- $w \models c$. Then at $w$ a subject is overheated, conscious, and not given water. This is undesirable: the first rule is falsified, and the stronger rule is inapplicable.
- $w \models \neg c$. At $w$ a subject is overheated, not fully conscious, and not given water. In this case, while the first rule is falsified, it is overridden by the more specific rule.

In general, a world is less desirable depending on its set of falsified obligations, where each falsified obligation is not overridden by a more specific rule. This leads to the following definition.

**Definition 4.2** *For* $w \in \mathcal{M}$, $F_m(w, \boldsymbol{R}) = \{r \in \boldsymbol{R} \mid w \models b(r) \wedge \neg h(r)$ *and* $\forall r'$ *such that* $r \lhd r', w \models \neg b(r')\}$

Then a preference order on possible worlds based on "relevant" falsified obligations can be defined:

**Definition 4.3** *For set of rules* $\boldsymbol{R}$*, and* $w_1, w_2 \in \mathcal{M}$*,*

$$w_1 \preceq_{\boldsymbol{R}} w_2 \text{ iff } F_m(w_1, \boldsymbol{R}) \subseteq F_m(w_2, \boldsymbol{R})$$

Definition 4.3 provides a basic and intuitive notion of preference between worlds. It is based on the assumption that all rules in $\mathbf{R}$ have equal weight. While this is often an oversimplification, it provides an appropriate point for examining properties of the approach and considering various examples. We later consider in Section 5 how the approach can be augmented, so that the rules in $\mathbf{R}$ come with a ranking, reflecting a rule's relative (deontic) importance.

Given the above definitions, we can now describe how a conditional permission $\phi \rightarrow_p \psi$ is handled within the present framework. We define:

$$\phi \rightarrow_p \psi \doteq \phi \wedge \psi \rightarrow \psi.$$

Consider the rule $\phi \wedge \psi \rightarrow \psi$. On the one hand, it appears to be vacuous, in that it cannot be falsified, and so is never a member of any $F_m(w, \mathbf{R})$ in Definition 4.2. On the other hand, it can override (or *defeat*) another obligation, as given in Definition 4.1, and so can "rule out" members of $F_m(w, \mathbf{R})$ in Definition 4.2. Consider a variant of the pizza-eating example, in which one should not eat with the hands, but in eating pizza it is *permissible* to use the hands: $r_1 : \top \rightarrow \neg h$, $r_2 : p \rightarrow_p h$ (i.e. $r_2 : p \wedge h \rightarrow h$). Clearly a $\neg p \neg h$ world is minimal according to Definition 4.3. However, it can also be verified that a $p \neg h$ world is also minimal; if rule $r_2$ were $p \rightarrow h$ this would not be the case.

Given a preference ordering on worlds from Definition 4.3, one can determine what is (deontically) the best that can be attained, or is most preferable, in the case that $\gamma$ is true, by examining the minimum $\gamma$-worlds in the ordering.

**Definition 4.4** *Given a set of rules* $\boldsymbol{R}$*,*
$\psi$ *is* (deontically) preferred *given* $\gamma$*,* $\gamma \mathrel{|\!\sim} \psi$*, iff*

$$\min(\preceq_{\boldsymbol{R}}, [\gamma]) \subseteq [\psi]$$

The following are essentially observations:

**Proposition 1**

1. $\mathrel{|\!\sim}$ *is a* preference relation *(Kraus, Lehmann, and Magidor 1990).*
2. *If* $\not\vdash \neg\gamma$ *then* $\gamma \mathrel{|\!\sim} \psi$ *implies* $\gamma \not\mathrel{|\!\sim} \neg\psi$.

While Item 1 states that $\mathrel{|\!\sim}$ is a preference relation, there is a key difference: the rules $\mathbf{R}$ induce a *single* partial preorder on worlds whereas in (Kraus, Lehmann, and Magidor 1990), a consequence relation has a corresponding *set* of preorders on worlds, comprising the models in their approach. This means that the present approach accommodates *irrelevant* properties. For example, given a vocabulary $\{a, b, c\}$, from $\mathbf{R} = \{a \rightarrow b\}$ we obtain $a \wedge c \mathrel{|\!\sim} b$.

This brings up an important distinction between the rules in $\mathbf{R}$ and the obtained relation given by $\mathrel{|\!\sim}$. We have, for instance, that $\phi \mathrel{|\!\sim} \psi \wedge \chi$ implies $\phi \mathrel{|\!\sim} \psi$ and $\phi \mathrel{|\!\sim} \chi$; this follows immediately from the definition of $\mathrel{|\!\sim}$, which in turn was defined from $\preceq_{\mathbf{R}}$. However, the two sets of rules:

$$\mathbf{R}_1 = \{a \rightarrow b, a \rightarrow c\} \quad \text{and} \quad \mathbf{R}_2 = \{a \rightarrow (b \wedge c)\}$$

are in no sense equivalent, since the resulting preference orderings, $\preceq_{\mathbf{R}_1}$ and $\preceq_{\mathbf{R}_2}$, are different (see the first example in Section 4.2).

Similarly, from $\mathbf{R}_3 = \{a \rightarrow b, b \rightarrow c\}$ we obtain $a \mathrel{|\!\sim} c$. This is a defeasible notion; if we added the rule $a \rightarrow \neg c$ then $a \mathrel{|\!\sim} c$ would no longer hold. For rules $\mathbf{R}_4 = \{a \rightarrow c, b \rightarrow c\}$, we get that $c$ holds in the minimal $a \vee b$ worlds, so $a \vee b \mathrel{|\!\sim} c$. For $\mathbf{R}_5 = \{a \rightarrow b\}$ we get $\neg b \mathrel{|\!\sim} \neg a$. (For example, if pizza should be eaten with the hands and one is not eating with the hands, then in the deontically best worlds, one is not eating pizza.) Further examples are discussed in Section 4.2.

## 4.1 The Computational Approach

Definition 4.3 provides a semantic account for a set of rules $\mathbf{R}$, from which inferences regarding obligation can be obtained (Definition 4.4). While in principle this is all one needs, in practice it is infeasible to work with a preorder over all possible worlds. In this section we develop an equivalent formulation based on maximal sets of formulas.

We start with a formula $\gamma$ that represents what is contingently known about a domain; the rules in $\mathbf{R}$ are then used to augment this information to express what should be the case, given that $\gamma$ is true. To this end, the rules in $\mathbf{R}$ are transformed into material conditionals in which the information in more specific, conflicting, rules is taken into account. An expansion of $\gamma$ by a maximal consistent set of such rules gives a maximal set of beliefs that could be held by the agent; the disjunction of these expansions then specifies what is deontically best in the case that $\gamma$ holds.

To this end the set of rules that dominate $r$ is defined by:

$$O(r) = \{r' \in \mathbf{R} \mid r \lhd r'\}$$

In the next definition, for a rule $r$, $r^{\lhd}$ is $r$ but with the additional assertion in the body that no dominating rule is applicable. Then $\mathbf{R}^{\lhd}$ is the resulting set of rules in which this specificity information is incorporated.

**Definition 4.5** *For a set of rules* $\boldsymbol{R}$ *and* $r \in \boldsymbol{R}$*, define:*

$$r^{\lhd} = (b(r) \wedge \wedge_{r' \in O(r)} \neg b(r')) \supset h(r)$$
$$\boldsymbol{R}^{\lhd} = \{r^{\lhd} \mid r \in \boldsymbol{R}\}$$

The next definition specifies for rules $\mathbf{R}$ and formula $\gamma$, the sets of rules that are $maxcon$ with respect to $\mathbf{R}$ and $\gamma$; this is the subsets of $\mathbf{R}^{\lhd}$ that are maximal consistent with $\gamma$.

**Definition 4.6**

$$MC(\gamma, \boldsymbol{R}) = \{\boldsymbol{R'} \subseteq \boldsymbol{R}^{\lhd} \mid \boldsymbol{R'} \textit{ is maxcon wrt } \gamma\}$$

*where $\boldsymbol{R'}$ is maxcon with respect to $\gamma$ iff*

1. $\boldsymbol{R'} \cup \{\gamma\} \not\vdash \bot$ *and*
2. $\forall \boldsymbol{R''}$ *where* $\boldsymbol{R'} \subset \boldsymbol{R''} \subseteq \boldsymbol{R}^{\lhd}$, *we have* $\boldsymbol{R''} \cup \{\gamma\} \vdash \bot$.

The set of contingent deontic outcomes that may justifiably be held is given by $\gamma$ along with the set of maximal sets of applicable rules:

**Definition 4.7** $E(\gamma, \boldsymbol{R}) = \gamma \wedge \bigvee_{\boldsymbol{R'} \in MC(\gamma, \boldsymbol{R})} (\bigwedge \boldsymbol{R'})$

Given this, it can be shown that the set of models of $E(\gamma, \mathbf{R})$ is exactly the minimal models of $\gamma$ in $\preceq_{\mathbf{R}}$. To this end, Definition 4.2 can be rewritten as:

$$F_m(w, \mathbf{R}) = \{r \in \mathbf{R} \mid w \models b(r) \wedge \neg h(r) \wedge \wedge_{r' \in O(r)} \neg b(r')\}$$

The following dual notion will be used:

$$\text{For } w \in \mathcal{M}, \quad S_m(w, \mathbf{R}) = \mathbf{R} \setminus F_m(w, \mathbf{R}).$$

**Lemma 4.1** $S_m(w, \boldsymbol{R}) = \{r \in \boldsymbol{R} \mid w \models r^{\lhd}\}$

**Proof of Lemma:**

$$\begin{aligned}
S_m(w, \mathbf{R}) &= \mathbf{R} \setminus F_m(w, \mathbf{R}) \\
&= \mathbf{R} \setminus \{r \in \mathbf{R} \mid w \models b(r) \wedge \neg h(r) \wedge (\wedge_{r' \in O(r)} \neg b(r'))\} \\
&= \{r \in \mathbf{R} \mid w \not\models b(r) \wedge \neg h(r) \wedge (\wedge_{r' \in O(r)} \neg b(r'))\} \\
&= \{r \in \mathbf{R} \mid w \models \neg(b(r) \wedge \neg h(r) \wedge (\wedge_{r' \in O(r)} \neg b(r')))\} \\
&= \{r \in \mathbf{R} \mid w \models \neg(b(r) \wedge (\wedge_{r' \in O(r)} \neg b(r')) \wedge \neg h(r))\} \\
&= \{r \in \mathbf{R} \mid w \models \neg(b(r) \wedge (\wedge_{r' \in O(r)} \neg b(r')) \vee h(r)\} \\
&= \{r \in \mathbf{R} \mid w \models (b(r) \wedge (\wedge_{r' \in O(r)} \neg b(r'))) \supset h(r)\} \\
&= \{r \in \mathbf{R} \mid w \models r^{\lhd}\} \qquad \square
\end{aligned}$$

We obtain:

**Theorem 4.1** $[E(\gamma, \boldsymbol{R})] = \min(\gamma, \preceq_{\boldsymbol{R}})$

**Proof:**

($\Rightarrow$)

Let $w \in [E(\gamma, \mathbf{R})]$ and so $w \models \gamma \wedge \bigvee_{\mathbf{R'} \in MC(\gamma, \mathbf{R})} (\bigwedge \mathbf{R'})$. Thus $w \models \gamma$. Also for some $\mathbf{R'} \in MC(\gamma, \mathbf{R})$ we have $w \models \mathbf{R'}$.

We are to show that $w \in \min(\gamma, \preceq_{\mathbf{R}})$, i.e. that there is no $w'$ such that $w' \models \gamma$ and $w' \prec w$. Toward a contradiction, assume otherwise, and let $w'$ be a world such that $w' \models \gamma$ and $w' \prec w$.

Since $w' \prec w$, we have by definition that $F_m(w', \mathbf{R}) \subset F_m(w, \mathbf{R})$, and hence $\mathbf{R} \setminus F_m(w, \mathbf{R}) \subset \mathbf{R} \setminus F_m(w', \mathbf{R})$ and thus from the definition of $S_m$ that $S_m(w, \mathbf{R}) \subset S_m(w', \mathbf{R})$.

From Lemma 4.1 we have that $S_m(w, \mathbf{R})$ is the maximum subset of $\mathbf{R}^{\lhd}$ (Definition 4.5) satisfied at $w$. But this is just $\mathbf{R'}$ above, where $\mathbf{R'} \in MC(\gamma, \mathbf{R})$.

But $\mathbf{R'} = S_m(w, \mathbf{R}) \subset S_m(w', \mathbf{R})$ along with $w \models \gamma$ and $w' \models \gamma$, implies that $\mathbf{R'}$ is not maximal, contradicting $\mathbf{R'} \in MC(\gamma, \mathbf{R})$. Hence our assumption that $w \notin \min(\gamma, \preceq_{\mathbf{R}})$ is incorrect, and so $w \in \min(\gamma, \preceq_{\mathbf{R}})$.

($\Leftarrow$)

Assume that $w \notin [E(\gamma, \mathbf{R})]$. We are to show that $w \notin \min(\gamma, \preceq_{\mathbf{R}})$.

Trivially if $w \not\models \gamma$ then $w \notin \min(\gamma, \preceq_{\mathbf{R}})$, so we can assume that $w \models \gamma$.

Since we have $w \models \gamma$ and $w \notin [E(\gamma, \mathbf{R})]$ by Definition 4.7 we get $w \notin [\bigvee_{\mathbf{R'} \in MC(\gamma, \mathbf{R})} (\bigwedge \mathbf{R'})]$, or that $w \not\models \mathbf{R'}$ for every $\mathbf{R'} \in MC(\gamma, \mathbf{R})$.

Consider

$$\begin{aligned}
\mathbf{R}^o &= \{r \in \mathbf{R} \mid w \models (b(r) \wedge \bigwedge_{r' \in O(r)} \neg b(r')) \supset h(r)\} \\
&= \{r \in \mathbf{R} \mid w \models r^{\lhd}\}.
\end{aligned}$$

From Lemma 4.1 we have $\mathbf{R}^o = S_m(w, \mathbf{R})$.

Since $w \models \gamma \wedge \mathbf{R}^o$ it follows that $\gamma \wedge \mathbf{R}^o$ is consistent. Thus according to Definition 4.6 there is a set $\mathbf{R''}$ where $\mathbf{R}^o \subseteq \mathbf{R''}$ and $\mathbf{R''} \in MC(\gamma, \mathbf{R})$. Since we have that $w \not\models \mathbf{R'}$ for every $\mathbf{R'} \in MC(\gamma, \mathbf{R})$ this means that in fact $\mathbf{R}^o \subset \mathbf{R''}$.

Let $w' \in [\mathbf{R''}]$; so $\mathbf{R''} = S_m(w', \mathbf{R})$ by Lemma 4.1.

Then, since $\mathbf{R}^o \subset \mathbf{R''}$, $\mathbf{R}^o = S_m(w, \mathbf{R})$, and $\mathbf{R''} = S_m(w', \mathbf{R})$, we have that $S_m(w, \mathbf{R}) \subset S_m(w', \mathbf{R})$. This in turn means that $\mathbf{R} \setminus F_m(w, \mathbf{R}) \subset \mathbf{R} \setminus F_m(w', \mathbf{R})$ or that $F_m(w', \mathbf{R}) \subset F_m(w, \mathbf{R})$.

Consequently, $w' \prec_{\mathbf{R}} w$ and since $w' \models \gamma$, $w \models \gamma$, we have $w \notin \min(\gamma, \preceq_R)$, as desired. $\square$

While there are computational challenges, these are no worse than those of similar approaches in nonmonotonic and preferential reasoning.

**Theorem 4.2** *Let the underlying logic be classical propositional logic. Deciding $E(\gamma, \boldsymbol{R}) \vdash \psi$ is $\Pi_2^p$-complete.*

**Proof Outline:** Deciding for $r, r' \in \mathbf{R}$ whether $r' \lhd r$ holds is in $\Delta_2^p$ and consequently so is deciding if $r \in \mathbf{R}^{\lhd}$.

Then the complementary problem can be proved, that $E(\gamma, \mathbf{R}) \not\vdash \psi$ is $\Sigma_2^p$-complete, analogous to (Nebel 1998)[Theorem 5.2].

We conclude this part with a result concerning background knowledge. Often an agent's knowledge can be divided into two parts: background information holding across all instances of the domain (like domain constraints), and instance-specific information, such as one block happens to be on top of another. These two parts are combined in the formula $\gamma$ in Definitions 4.4 and 4.7. An alternative is to define $\preceq_{\mathbf{R}}$ with respect to just those worlds that are possible according to the background knowledge, and not to all logically-possible worlds. Then $\gamma$ in Definition 4.4 just needs to refer to instance-specific information. The next result shows that these two approaches coincide. For the next result only we introduce the following notation. For $\mathcal{W} \subseteq \mathcal{M}$:

$$[\gamma]_{\mathcal{W}} \doteq [\gamma] \cap \mathcal{W} \text{ and: } \gamma \mid\!\sim_{\mathcal{W}} \psi \text{ iff } \min(\preceq_{\mathbf{R}}, [\gamma]_{\mathcal{W}}) \subseteq [\psi]_{\mathcal{W}}.$$

Below, $\Delta$ will represent general background knowledge.

**Theorem 4.3** *Let $\Delta \subseteq \mathcal{L}$. Then for a set of rules $\boldsymbol{R}$, we have*

$$\gamma \wedge \Delta \mid\!\sim \psi \textit{ iff } \gamma \mid\!\sim_{[\Delta]} \psi.$$

**Proof:**

Let $\mathcal{D} = [\Delta]$.

($\Rightarrow$) Assume that $\gamma \wedge \Delta \mathrel{|\!\sim} \psi$, which is to say that $\min(\preceq_{\mathbf{R}}, [\gamma \wedge \Delta]) \subseteq [\psi]$. $\min(\preceq_{\mathbf{R}}, [\gamma \wedge \Delta])$ is $\{w \in \mathcal{M} \mid w \models \gamma \wedge \Delta$ and $\forall w'$ s.t. $w' \models \gamma \wedge \Delta, w' \preceq_{\mathbf{R}} w \Rightarrow w \preceq_{\mathbf{R}} w'\}$. This is readily seen to be the same as $\{w \in \mathcal{D} \mid w \models \gamma$ and $\forall w'$ s.t. $w' \models \gamma, w' \preceq_{\mathbf{R}} w \Rightarrow w \preceq_{\mathbf{R}} w'\}$ which is just $\min(\preceq_{\mathbf{R}}, [\gamma]_{\mathcal{D}})$. Consequently, we have $\min(\preceq_{\mathbf{R}}, [\gamma]_{\mathcal{D}}) \subseteq [\psi]_{\mathcal{M}}$. Now, since $\min(\preceq_{\mathbf{R}}, [\gamma]_{\mathcal{D}}) \subseteq \mathcal{D}$, we obtain that $\min(\preceq_{\mathbf{R}}, [\gamma]_{\mathcal{D}}) \subseteq [\psi]_{\mathcal{M}} \cap \mathcal{D}$. Since $[\psi]_{\mathcal{M}} \cap \mathcal{D} = [\psi]_{\mathcal{D}}$ we obtain $\min(\preceq_{\mathbf{R}}, [\gamma]_{\mathcal{D}}) \subseteq [\psi]_{\mathcal{D}}$, and so $\gamma \mathrel{|\!\sim}_{\mathcal{D}} \psi$.

($\Leftarrow$) Assume that $\gamma \mathrel{|\!\sim}_{\mathcal{D}} \psi$ or $\min(\preceq_{\mathbf{R}}, [\gamma]_{\mathcal{D}}) \subseteq [\psi]_{\mathcal{D}}$. $\min(\preceq_{\mathbf{R}}, [\gamma]_{\mathcal{D}})$ is just $\{w \in \mathcal{D} \mid w \models \gamma$ and $\forall w'$ s.t. $w' \models \gamma, w' \preceq_{\mathbf{R}} w \Rightarrow w \preceq_{\mathbf{R}} w'\}$ which can be rewritten as $\{w \in \mathcal{M} \mid w \models \gamma \wedge \Delta$ and $\forall w'$ s.t. $w' \models \gamma \wedge \Delta, w' \preceq_{\mathbf{R}} w \Rightarrow w \preceq_{\mathbf{R}} w'\}$ which is $\min(\preceq_{\mathbf{R}}, [\gamma \wedge \Delta]_{\mathcal{M}})$.

Consequently we have that $\min(\preceq_{\mathbf{R}}, [\gamma]_{\mathcal{D}}) = \min(\preceq_{\mathbf{R}}, [\gamma \wedge \Delta]_{\mathcal{M}})$ and so we get that $\min(\preceq_{\mathbf{R}}, [\gamma \wedge \Delta]_{\mathcal{M}}) \subseteq [\psi]_{\mathcal{D}}$.

Since $[\psi]_{\mathcal{D}} \subseteq [\psi]$, we obtain that $\min(\preceq_{\mathbf{R}}, [\gamma \wedge \Delta]_{\mathcal{M}}) \subseteq [\psi]$ or that $\gamma \wedge \Delta \mathrel{|\!\sim} \psi$. $\square$

## 4.2 Examples

Several examples were given following Definition 4.4. Here we consider in detail further illustrative examples, including those in Section 2.2.

Consider first the simple set of rules $\mathbf{R}_1 = \{r_1 : \top \rightarrow a,\ r_2 : \top \rightarrow b\}$. We obtain the ordering:[7]

$$ab \ \prec_{\mathbf{R}_1} \ \langle a\neg b, \neg ab\rangle \ \prec_{\mathbf{R}_1} \ \neg a\neg b$$

So, it is best if both rules are satisfied, and worst if neither are satisfied. We obtain $\top \mathrel{|\!\sim} a$. Also $\neg b \mathrel{|\!\sim} a$, so even when $\top \rightarrow b$ is falsified, it is better that $a$ be the case than $\neg a$.

In contrast, consider $\mathbf{R}_2 = \{\top \rightarrow a \wedge b\}$. We obtain:

$$ab \ \prec_{\mathbf{R}_2} \ \langle a\neg b, \neg ab, \neg a\neg b\rangle.$$

In $\mathbf{R}_1$ we have two separate obligations whereas in $\mathbf{R}_2$ we have a single composite obligation. In both cases we obtain that $\top \mathrel{|\!\sim} a \wedge b$.

Consider next the pizza example, first that you should not eat with your hands, but if you eat pizza you should eat with your hands:

$$r_1 : \top \rightarrow \neg h \text{ and } r_2 : p \rightarrow h$$

1. For $\mathbf{R} = \{r_2\}$ we obtain $\langle ph, \neg ph, \neg p\neg h\rangle \ \prec_{\mathbf{R}} \ p\neg h$
   Thus if one is not eating pizza, it doesn't matter what one is doing with their hands; if eating pizza, eating with the hands is better than not.
2. For $\mathbf{R} = \{r_1, r_2\}$ we get $\langle ph, \neg p\neg h\rangle \ \prec_{\mathbf{R}} \ \langle p\neg h, \neg ph\rangle$
   Thus it is best to eat with the hands iff one is eating pizza. (See Section 6 for a discussion.)
3. Consider next the addition of a rule that if you eat with your hands you should wash them afterwards: $r_3 : h \rightarrow w$
   For $\mathbf{R} = \{r_1, r_2, r_3\}$ we obtain:
   $$\langle phw, \ \neg p\neg hw, \ \neg p\neg h\neg w\rangle \ \prec_{\mathbf{R}} \ \langle ph\neg w, \ p\neg hw, \ p\neg h\neg w, \ \neg phw\rangle$$
   $$\langle ph\neg w, \ \neg phw\rangle \ \prec_{\mathbf{R}} \ \neg ph\neg w$$
   Inferences can be "read off" the ordering, but among others, if you eat pizza it's best to use your hands and wash them afterwards, and if you don't eat pizza but eat with your hands again it's best to wash them afterwards.
4. Last, consider where one should not eat with the hands, but if eating pizza it is *permissible* to eat with the hands. We have the rules:
   $$r_1 : \top \rightarrow \neg h \text{ and } r_2 : p \rightarrow_p h$$
   but where in fact $r_2$ is $p \wedge h \rightarrow h$. For $\mathbf{R} = \{r_1, r_2\}$ we obtain
   $$\langle ph, \neg p\neg h, p\neg h\rangle \ \prec_{\mathbf{R}} \ \neg ph.$$
   Thus, if eating pizza, it is acceptable to either eat with the hands or not. The world in which pizza isn't eaten but one eats with the hands is worse than other worlds.

The various "paradoxes" described in Section 2.2 are handled appropriately.[8] We discuss each in turn.

**Ross's Paradox (Ex. 2.1):** Let the set of propositional atoms be $\mathcal{P} = \{m, b\}$ with $m$ for "mail the letter" and $b$ for "burn the letter". Consider the sets of rules:

$$\mathbf{R}_1 = \{\top \rightarrow m\} \quad \text{and} \quad \mathbf{R}_2 = \{\top \rightarrow m, \top \rightarrow m \vee b\}$$

For $\mathbf{R}_1$ we obtain the ordering

$$\langle mb, m\neg b\rangle \ \prec_{\mathbf{R}_1} \ \langle \neg mb, \neg m\neg b\rangle$$

Thus, in the best states of affairs one mails the letter; there is no result concerning burning, beyond the fact that burning and not burning are both permissible. The same result holds if the letter is not mailed. For $\mathbf{R}_2$ we obtain

$$\langle mb, m\neg b\rangle \ \prec_{\mathbf{R}_2} \ \neg mb \ \prec_{\mathbf{R}_2} \ \neg m\neg b$$

Here, quite reasonably, if the letter is not mailed, then the second obligation can be satisfied by burning it.

**The Chisholm Paradox (Ex. 2.2):** We have the encoding:

$$\mathbf{R} = \{\top \rightarrow h, \ h \rightarrow t, \ \neg h \rightarrow \neg t\}$$

along with contingent information $\gamma = \neg h$. We obtain:

$$ht \ \prec_{\mathbf{R}} \ h\neg t \quad \text{and} \quad ht \ \prec_{\mathbf{R}} \ \neg h\neg t \ \prec_{\mathbf{R}} \ \neg ht$$

The best thing to do is to help your neighbour and to tell them. In the instance at hand, where you don't help your neighbour, the best thing to do is to not tell them that you will. This is given in the fact that the minimum $\neg h$ model is $\neg h\neg t$, so we obtain $\neg h \mathrel{|\!\sim} \neg t$.

[7] Recall we use the notation $\langle w_1, \ldots, w_n\rangle \prec \langle w_{n+1}, \ldots, w_m\rangle$ to indicate that $w_i \prec w_j$ for $1 \leq i \leq n < j \leq m$.

[8] By "handled" I don't want to suggest that any sort of overarching solution is provided for these problems. Rather, the claim is that appropriate nonmonotonic inferences are obtained for these examples which have proven problematic for logics of obligation.

**The Forrester Paradox (Ex. 2.3):** The encoding is:

$$\mathbf{R} = \{\top \rightarrow \neg m,\ m \rightarrow g\}$$

We might interpret the rule $m \rightarrow g$ as "if you murder someone, you should be gentle". Since $m$ and $g$ are independent here, we obtain the ordering on possible worlds:

$$\langle \neg m g, \neg m \neg g \rangle \ \prec\ m g \ \prec\ m \neg g$$

In the most preferred worlds one does not murder, since at these worlds $\neg m$ is true. If you do murder someone then in the best such worlds one is gentle (i.e. in the minimum $m$ world $g$ is true.)

However, this is not the standard interpretation, which instead is "if you murder someone ($m$) they should be gently murdered ($g$)". That is, there is the implicit constraint that $g \supset m$. This can be handled in two ways. First, we can consider what holds in the minimum $g \supset m$ worlds in the above ordering; we obtain $\neg m \neg g$. Second, we can consider $g \supset m$ as a domain constraint, and so disregard $g \neg m$ worlds. The leadup to Theorem 4.3 shows how this can be expressed, and the theorem itself shows that we obtain the same results as in the first alternative.

**Potentially Troublesome Examples:** We next consider some examples that might appear to be problematic, but that are arguably handled appropriately.

- Mutually defeating obligations: $\mathbf{R} = \{a \rightarrow \neg a, \neg a \rightarrow a\}$. There is no preference between $a$ and $\neg a$ worlds.
- $\mathbf{R} = \{a \rightarrow \neg a\}$. In the preference ordering, every $\neg a$ world is preferred to every $a$ world.
- $\mathbf{R} = \{a \rightarrow b, b \rightarrow \neg a\}$. In the resulting preference ordering, every $\neg a$ world is preferred to every $a$ world.
- $\mathbf{R} = \{a \rightarrow b, \neg a \rightarrow b\}$. Every $b$ world is preferred to every $\neg b$ world.
- Conflicting obligations: $\mathbf{R} = \{a \rightarrow b, a \rightarrow \neg b\}$. In the deontically best worlds $\neg a$ holds (i.e. we have $\top \mid\!\sim \neg a$) since no obligation is violated when $a$ is false. Otherwise, $ab$ and $a\neg b$ worlds are equally ranked.
- $\mathbf{R} = \{\top \rightarrow a, \top \rightarrow \neg b, a \rightarrow b\}$. (Horty 2007) Given the specificity relation between the last two rules, we obtain the ordering: $ab \prec \langle a\neg b, \neg a \neg b \rangle \prec \neg ab$.

**Preferences on Obligations:** We have seen how a more specific rule may take precedence over a less specific rule. It seems that in a given context we can also use the form of a rule to encode preferences among obligations. Consider the set of rules:

$$\{a \rightarrow b, a \rightarrow b \vee c\}$$

These rules induce the following ordering on worlds:

$$\langle abc, ab\neg c \rangle \ \prec\ a\neg bc \ \prec\ a\neg b\neg c$$

In the best $a$ worlds we have that $b$ is true, but in the best $a \wedge \neg b$ worlds we have that $c$ is true. Hence these rules can be interpreted as expressing the preference:

"If $a$ then it should be that $b$, but if not $b$ then $c$."

Next, consider the rules:

$$\{a \rightarrow b, a \rightarrow b \wedge c\}$$

These rules induce the following ordering on worlds:

$$abc \ \prec\ ab\neg c \ \prec\ \langle a\neg bc, a\neg b\neg c \rangle$$

In the best $a$ worlds we have that $b \wedge c$ is true, but in the best $a \wedge \neg c$ worlds we have that $b$ is true. Hence these rules can be interpreted as expressing the preference:

"If $a$ then it should be that $b \wedge c$, but if $b$, $c$ can't both hold, it should be that $b$.

This suggests (in the realm of future work) that it may be feasible to specify a higher-level language for obligation and permission-style constraints that can be translated (or "compiled") into the present approach.

## 5 Adding Weights to Rules

Clearly, not all obligations are of equal importance. Here we extend the approach so that a rule may have an associated weight specifying that rule's importance or perhaps degree of strength. Two possibilities are considered, *quantitative* and *qualitative* weights. In a quantitative approach, the idea is that enough violated lower-ranked obligations can eventually outweigh a higher-ranked obligation. For example, while it is (presumably) a major gaffe to disparage a host's choice of wine, it would be better to do this than it would be to violate a number of lesser obligations, such as eating pizza with knife and fork, not putting a napkin on your lap, and so on.

On the other hand, there may be preferences that no number of lower-ranked obligations will ever jointly outweigh. For example, one should never kill another person, and no number of violated polite-society obligations would ever outweigh such an obligation.

**Quantitative Weights** Here, positive integers are attached to rules, indicating a rule's importance. Given a function from rules to positive integers, $W : \mathbf{R} \mapsto \mathbb{N}^+$, we define

$$w_1 \preceq_{\mathbf{R}} w_2 \quad \text{iff} \quad \sum_{r \in F_m(w_1, \mathbf{R})} W(r) \leq \sum_{r \in F_m(w_2, \mathbf{R})} W(r)$$

This variant is a *modification* of the original approach. If the weights are uniformly 1, then the above equation reduces to:

$$w_1 \preceq_{\mathbf{R}} w_2 \ \text{ iff } \ |F_m(w_1, \mathbf{R})| \leq |F_m(w_2, \mathbf{R})|.$$

Compare this with Definition 4.2.

For example, consider the rules asserting that you should not eat with your hands, you should have a napkin on your lap, and you shouldn't insult your host's choice of wine:

$$r_1 : \top \rightarrow \neg h,\ r_2 : \top \rightarrow n,\ r_3 : \top \rightarrow \neg i.$$

Given the weights: $W(r_1) = 1, W(r_2) = 1, W(r_3) = 4$ we obtain the ordering: $\neg hn\neg i \prec \langle hn\neg i, \neg h\neg n\neg i \rangle \prec h\neg n\neg i$
$\prec \neg hni \prec \langle hni, \neg h\neg ni \rangle \prec h\neg ni$

Thus it is always better to not insult your host's wine than to do so; otherwise, the ordering is determined by which of the other rules are violated. If we had $W(r_3) = 2$, then it would be just as bad to insult the wine as it is to eat with the hands and not use a napkin.

**Qualitative Weights** Here ranks are attached to a rule in the spirit of *ordinal conditional rankings* (Spohn 1988). Ranks are given by a function from $\mathbf{R}$ to the natural numbers, $R : \mathbf{R} \mapsto \mathbb{N}$, where conventionally there is a rule with rank 0. Violated rules at a lower rank never outweigh rules at a higher rank.

It is convenient to consider $\mathbf{R}$ as partitioned into disjoint sets of rules $\langle \mathbf{R}_0, \mathbf{R}_1, \dots \rangle$ in which $r \in \mathbf{R}_i$ iff $R(r) = i$. Then we can define $\preceq$ in terms of a *lexicographic order*:

$$w_1 \prec_{\mathbf{R}} w_2 \text{ iff } \exists i \geq 0 \text{ s.t. } F_m(w_1, \mathbf{R}_i) \subset F_m(w_2, \mathbf{R}_i) \text{ and } F_m(w_1, \mathbf{R}_j) = F_m(w_2, \mathbf{R}_j) \text{ for every } j > i$$

$$w_1 \approx_{\mathbf{R}} w_2 \text{ iff } \forall i \geq 0, F_m(w_1, \mathbf{R}_i) = F_m(w_2, \mathbf{R}_i)$$

This is an *extension* of the original approach, since if all rules have rank zero, the above reduces to Definition 4.2. For example, consider the rules asserting that you should not eat with your hands and that you shouldn't kill people:

$$r_1 : \top \to \neg h, r_2 : \top \to \neg m, \text{ with } R(r_1) = 0, R(r_2) = 1$$

This gives the ordering: $\neg h \neg m \prec h \neg m \prec \neg h m \prec h m$.

Finally, these approaches can be combined in the obvious way, where each rule is assigned a rank and then a weight within that rank. The ranking on rules would induce a total preorder on worlds, and then the weights on rules would further "refine" each set of equivalently-ranked worlds. Thus rules involving life-and-death matters would be assigned some non-zero rank, while rules involving social mores would (presumably) be assigned a rank of zero.

## 6 Comparison with Normality Conditionals

We agree with Horty (1993) that conditional deontic reasoning is a form of nonmonotonic reasoning. However, we suggest that deontic conditionals differ from normality conditionals in a couple of key aspects. We have noted that our approach yields a unique ordering over possible worlds, or a single deontic extension. In this way it is more like the *rational closure* (Lehmann and Magidor 1992), which yields a unique ordering, than default logic (Reiter 1980), which may produce multiple extensions.

However, most notably, in conditional reasoning involving normality defaults, specificity information is reflected in the resulting ranking. Consider our pizza-eating example:

$$\top \to \neg h, \; p \to h \tag{2}$$

but consider it under a normality interpretation; that is, interpret the conditionals as stating that *normally* one doesn't eat with the hands; but in eating pizza one *normally* eats with the hands. One obtains (e.g. in (Kraus, Lehmann, and Magidor 1990; Boutilier 1994) and most others) the ordering:

$$\neg p \neg h \prec p h \prec p \neg h \quad \text{and} \quad \neg p \neg h \prec \neg p h$$

Thus *most normally* one isn't eating with the hands. Moreover, in these approaches, the following is a theorem:

$$(\top \to \neg h \wedge p \to h) \supset \top \to \neg p$$

and so most normally pizza isn't eaten. While this seems fine for a normality interpretation, it is clearly too strong for a deontic interpretation: Given that one should in general not eat with the hands, but that if eating pizza one should eat with the hands, it is unreasonable to conclude that in general one should not eat pizza.

Under our treatment of specificity with deontic conditionals (Definition 4.3), things are different. We have the ordering corresponding to the deontic interpretation of (2):

$$\langle p h, \neg p \neg h \rangle \prec \langle p \neg h, \neg p h \rangle$$

Here, in the *most desirable* worlds, either one is eating pizza with the hands, or else is not eating pizza and not using the hands. Neither of the situations $ph$ and $\neg p \neg h$ is better than the other. This is as it should be: a world in which one is eating pizza with the hands is exactly as good as one in which one is not eating pizza and not using the hands.

This difference is a result of the difference in how a conditional $\phi \to \psi$ is interpreted. In a logic of normality we have that (all other things being equal) $\phi \wedge \psi$ worlds are preferred to $\phi \wedge \neg \psi$ worlds. For the deontic interpretation we have that (all other things being equal) $\phi \supset \psi$ worlds are preferred to $\phi \wedge \neg \psi$ worlds.

## 7 Conclusion

In this paper we have presented an approach to defeasible deontic reasoning. The general idea is that a set of rules $\mathbf{R}$ expresses notions of conditional obligation and permission. The task at hand is to suitably "apply" the rules within a given context. To this end, in semantic terms, these rules induce a partial preorder on the set of models, giving the relative desirability of each model. For contingent information $\gamma$, the set of least $\gamma$-models in the ordering then gives the best, or most desirable, states of affairs. A syntactic approach is also given whereby from a set of rules and a formula $\gamma$ one obtains a formula that expresses the best outcome possible, given that $\gamma$ is the case. These approaches coincide, in that for any formula $\gamma$, the least set of models in the ordering induced from the rules exactly characterises the expansion of $\gamma$ according to $\mathbf{R}$.

The approach yields desirable results, both for basic examples and for the various "paradoxes" of deontic reasoning. The second approach is readily implementable and so, for further work, it would be of interest to see how this approach could be implemented in conjunction with (for example) some planner. Complexity is the same as most approaches in nonmonotonic reasoning being (for propositional logic) at the second level of the polynomial hierarchy. Last, we argue that properties of defeasible deontic conditionals differ from those of normality conditionals in the literature.


## Acknowledgements

I thank Nicolas Fillion, John Horty, and Leon van der Torre (who inter alia provided the observation in footnote 6) for their helpful and insightful suggestions and I thank the reviewers for their pertinent comments. I particularly thank two reviewers for this and an earlier version of the paper who gave highly detailed and thoughtful critiques. Financial support was gratefully received from the Natural Sciences and Engineering Research Council of Canada.

## References


Antoniou, G.; Dimaresis, N.; and Governatori, G. 2009. A modal and deontic defeasible reasoning system for modelling policies and multi-agent systems. *Expert Syst. Appl.* 36(2):4125–4134.

Bartha, P. 1999. Moral preference, contrary-to-duty obligation and defeasible oughts. *Norms, Logics and Information Systems: New Studies in Deontic Logic and Computer Science* 93–108.

Boutilier, C. 1994. Conditional logics of normality: A modal approach. *Artificial Intelligence* 68(1):87–154.

Chellas, B. 1980. *Modal Logic*. Cambridge: Cambridge University Press.

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

Forrester, J. 1984. Gentle murder, or the adverbial samaritan. *Journal of Philosophy* 81:193197.

Gärdenfors, P. 1975. Qualitative probability as an intensional logic. *Journal of Philosophical Logic* 4(2):171–185.

Geffner, H., and Pearl, J. 1992. Conditional entailment: Bridging two approaches to default reasoning. *Artificial Intelligence* 53(2-3):209–244.

Goble, L. 1991. Murder most gentle: The paradox deepens. *Philosophical Studies* 64(2):217–227.

Goble, L. 2003. Preference semantics for deontic logic. part i: Simple models. *Logique Et Analyse* 46:383–418.

Hansson, B. 1971. An analysis of some deontic logics. In Hilpinen, R., ed., *Deontic Logic: Introductory and Systematic Readings*. Reidel, Dordrecht. 121147.

Hansson, S. O. 1990. Preference-based deontic logic (PDL). *Journal of Philosophical Logic* 19(1):75–93.

Hansson, S. O. 2004. Semantics for more plausible deontic logics. *Journal of Applied Logic* 2:3–18.

Hilpinen, R., and McNamara, P. 2013. Deontic logic: A historical survey and introduction. In Gabbay, D.; Horty, J.; Parent, X.; van der Meyden, R.; and van der Torre, L., eds., *Handbook of Deontic Logic and Normative Systems*. College Publications. 3–136.

Horty, J. F., and Belnap, N. 1995. The deliberative stit: A study of action, omission, ability, and obligation. *Journal of Philosophical Logic* 24(6):583–644.

Horty, J. 1993. Deontic logic as founded on nonmonotonic logic. *Annals of Mathematics and Artificial Intelligence* 9:69–91.

Horty, J. 2007. Defaults with priorities. *Journal of Philosophical Logic* 36:367–413.

Horty, J. 2014. Deontic modals: Why abandon the classical semantics? *Pacific Philosophical Quarterly* 95:424–460.

Kowalski, R., and Satoh, K. 2018. Obligation as optimal goal satisfaction. *Journal of Philosophical Logic* 47(4):579–609.

Kraus, S.; Lehmann, D.; and Magidor, M. 1990. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence* 44(1-2):167–207.

Lehmann, D., and Magidor, M. 1992. What does a conditional knowledge base entail? *Artificial Intelligence* 55(1):1–60.

Lewis, D. 1973. *Counterfactuals*. Cambridge, Mass.: Harvard University Press.

Makinson, D., and van der Torre, L. 2000. Input/output logics. *Journal of Philosophical Logic* 29(4):383–408.

McCarty, L. T. 1994. Defeasible deontic reasoning. *Fundamenta Informaticae* 21(1,2):125–148.

McNamara, P. 2019. Deontic logic. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2019 edition.

Nebel, B. 1998. How hard is it to revise a belief base? In Dubois, D., and Prade, H., eds., *Belief Change. Handbook of Defeasible Reasoning and Uncertainty Management*, volume 3. Springer, Dordrecht.

Nute, D. 1994. Defeasible logic. In Gabbay, D. M.; Hogger, C. J.; and Robinson, J. A., eds., *Nonmonotonic Reasoning and Uncertain Reasoning*, volume 3 of *Handbook of Logic in Artifical Intelligence and Logic Programming*. Oxford. 353–395.

Reiter, R. 1980. A logic for default reasoning. *Artificial Intelligence* 13(1-2):81–132.

Ryu, Y. U., and Lee, R. M. 1995. Defeasible deontic reasoning and its applications to normative systems. *Decision Support Systems* 14(1):59–73.

Ryu, Y. U. 1995. Conditional deontic logic augmented with defeasible reasoning. *Data & Knowledge Engineering* 16(1):73–91.

Spohn, W. 1988. Ordinal conditional functions: A dynamic theory of epistemic states. In Harper, W., and Skyrms, B., eds., *Causation in Decision, Belief Change, and Statistics*, volume II. Kluwer Academic Publishers. 105–134.

Straßer, C., and Arieli, O. 2015. Normative reasoning by sequent-based argumentation. *Journal of Logic and Computation* 29(3):387–415.

Straßer, C. 2011. A deontic logic framework allowing for factual detachment. *Journal of Applied Logic* 9(1):61 – 80.

van der Torre, L. W. N., and Tan, Y. 1995. Cancelling and overshadowing: Two types of defeasibility in defeasible deontic logic. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 1525–1533. Morgan Kaufmann.

van der Torre, L. W. N., and Tan, Y. 1999. Contrary-to-duty reasoning with preference-based dyadic obligations. *Annals of Mathematics and Artificial Intelligence* 27(1-4):49–78.

van der Torre, L. 1994. Violated obligations in a defeasible deontic logic. In *Proceedings of the European Conference on Artificial Intelligence*, 371–375.

von Wright, G. 1951. Deontic logic. *Mind* 60:1–15.