Semantics for more plausible deontic logics

Invited speech, DEON´02, Imperial College, London, May 22-24, 2002.

Sven Ove Hansson
Royal Institute of Technology
Stockholm, Sweden
soh@infra.kth.se

Abstract: In order to avoid the paradoxes of standard deontic logic, we have to give up the semantic construction that identifies obligatory status with presence in all elements of a subset of the set of possible worlds. It is proposed that deontic logic should instead be based on a preference relation, according to the principle that whatever is better than something permitted is itself permitted. Close connections hold between the logical properties of a preference relation and those of the deontic logics that are derived from it in this way. The paradoxes of SDL can be avoided with this construction, but it is still an open question what type of preference relation is best suited to be used as a basis for deontic logic.

1. Introduction

Modern deontic logic began with a seminal paper by Georg Henrik von Wright in 1951.[1] With a minor modification,[2] his list of postulates has turned out to be characterizable by a simple semantical construction that has since then dominated the subject: It is assumed that there is a subset of the set of possible worlds (the "ideal worlds") such that for any sentence p, Op (meaning that p is obligatory) holds if and only if p holds in all of these worlds. This is *standard deontic logic* (SDL); its basic principle is illustrated in Diagram 1.

The term "possible world" is ambiguous. In a logical sense, a possible world is a maximal consistent subset of the sentences of a given language. In a metaphysical sense, a possible world is a complete description of how the world could be. The holistic alternatives used in the SDL semantics have to be possible worlds in a logical sense, but not

---

1 von Wright 1951. On the origins of deontic logic, see Føllesdal and Hilpinen 1970 and von Wright 1998.

2 Acceptance of the postulate $O(p \vee \neg p)$.

necessarily in a metaphysical sense. They can, for instance, represent the combinations of actions that are open to an individual (rather than the worlds that she might inhabit). In order to avoid confusion with metaphysical possible worlds, I will use the terms "holistic alternative" and "alternative" instead of "possible world".

The valid sentences of SDL coincide with the theorems derivable from the following three axioms:

$Op \rightarrow \neg O \neg p$,
$Op \,\&\, Oq \leftrightarrow O(p \& q)$, and
$O(p \vee \neg p)$.[3]

It is common to assume that the selection of alternatives shown in Diagram 1 is based on an ordering of the alternatives, as in Diagram 2. The ideal alternatives are then identified with the maximal (best) alternatives according to that ordering. This construction makes it possible to extend SDL to conditional norms. In order to determine what obligations hold if *s* is true, we just restrict our attention to alternatives in which *s* is true, and identify the ideal alternatives with the alternatives that are best (maximal) in this restricted set. (Diagram 3)

SDL semantics is admirably simple and elegant. Unfortunately, it forces us to rather implausible conclusions. If we identify obligatory status with presence in all elements of a certain subset of the alternative set, then the following property will invariably hold:

If $\mathtt{k}\ p \rightarrow q$, then $\mathtt{k}\ Op \rightarrow Oq$.

[3] Føllesdal and Hilpinen 1970, p. 13.

This property has many names. I prefer to call it *necessitation* since it says that whatever is necessitated by a moral requirement is itself a moral requirement. As an example, suppose that I am morally required to take a boat without the consent of its owner and use it to rescue a drowning person. Let p denote this composite action that I am required to perform, and let q denote the part of it that consists in taking the boat without leave. Since q follows logically from p, I am logically necessitated to perform q in order to perform p. According to the postulate of necessitation, I then have an obligation to q. This is contestable, since I have no obligation to q in isolation.

Necessitation gives rise to most of the major deontic paradoxes. We may call these the *necessitation paradoxes*. Four of the most prominent are Ross's paradox, the paradox of commitment, the Good Samaritan, and the Knower. Ross's paradox is based on the instance $Op \rightarrow O(p \vee q)$ of necessitation. ("If you ought to mail the letter, then you ought to either mail or burn it.")[4] The paradox of commitment is based on the instance $O\neg p \rightarrow O(p \rightarrow q)$, which is interpreted as saying that if you do what is forbidden, then you are required to do anything whatsoever. ("If it is forbidden for you to steal this car, then if you steal it you ought to run over a pedestrian.")[5] The Good Samaritan operates on two sentences p and q, such that q denotes some atrocity and p some good act that can only take place if q takes place. We then have $\mathtt{k}\ p \rightarrow q$, and it follows by necessitation that if Op then Oq. ("You ought to help the assaulted person. Therefore, there ought to be an assaulted person.")[6] Åqvist's Knower paradox makes use of the epistemic principle that only that which is true can be known. Here, q denotes some wrongful action, and p denotes that q is known by someone who is required

4 Ross 1941, p. 62.
5 Prior 1954.
6 Prior 1958, p. 144.

to know it. Again, we have $\vdash_k p \rightarrow q$ and Op, and it follows by necessitation that Oq. ("If the police officer ought to know that Smith robbed Jones, then Smith ought to rob Jones.")[7]

It is the purpose of the present presentation to show how we can save deontic logic from the necessitation paradoxes. In order to achieve this we will have to give up the basic semantic idea of SDL, and hopefully find some other, more plausible semantic principle for deontic logic.

The subject-matter of deontic logic is quite complex. It includes defeasible norms, counterfactual norms, normative rules, multiagent norms, etc. For the purposes of the present presentation I will leave as much as possible of this complexity aside, and focus on situationist deontic logic, i.e. the deontic logic of that fraction of normative discourse which refers only to one moral appraisal of one situation. No changes in the situation or shifts in the perspective are allowed (which excludes deontic counterfactuals), and no general deontic statements (such as rules) will be represented.[8] Furthermore, I will assume that obligations refer to actions (or omissions), so that p in Op is a sentence that represents an action. (For simple reference I will use "action" and as a short term for "action-representing sentence".)

2. Basing deontic logic more directly on preferences

I propose that we base deontic logic on preferences, but not in the indirect way shown in Diagrams 2–3. Instead of using a preference relation on (holistic) alternatives, we can apply a preference relation directly to the actions that are the actual objects of obligations and permisssions, as in Diagram 4. Clearly, preferences over actions can in their turn be based on preferences over holistic alternatives. To begin with, however, I will make no such assumption but take the preference relation over actions to be

[7] Åqvist 1967.
[8] Hansson 2001, chapter 9.

primitive. (In sections 4–5 I will return to the derivation of preference relations over single actions from preference relations over alternatives.)

Can we insert normative predicates into a preference structure in the same way that we can insert a monadic predicate such as "best" or "bad"? To pursue this possibility, consider the following two classes of value predicates that can be inserted into a preference structure:

A monadic predicate H is $\geq$-*positive* if and only if for all p and q:

$Hp \ \& \ q \geq p \rightarrow Hq$.

It is $\geq$-*negative* if and only if for all p and q:

$Hp \ \& \ p \geq q \rightarrow Hq$.

Among the positive predicates we find such value predicates as "good", "best", "not worst", "very good", "excellent", "not very bad", "acceptable", etc. If one of these predicates holds for p, then it also holds for everything that is better than p or equal in value to p. Among the negative predicates we find "bad", "very bad", "worst", and "not best". If one of these predicates holds for p, then it also holds for everything that is worse than p or equal in value to p.

An obvious option for a semantics of "ought" is to construct it as a positive predicate:

The positivity thesis:

Prescriptive predicates satisfy positivity.

One advantage of this approach is that it allows us to insert several prescriptive predicates of different strengths. This is useful since both our prescriptions and our prescriptive expressions differ in strength; "must" is more stringent than "ought", and "ought" is more stringent than "should".

Unfortunately, however, this simple construction is not at all plausible. There are at least two classes of counterexamples that can be used against it.

The first class of counterexamples follows the recipe: Let p represent something morally required, and q a supererogatory variant of p. Concretely, let p denote that you return a borrowed motorcar in time to its owner and q that you return it in time to its owner after first having washed it and filled the petrol tank. It is quite plausible to value q higher than p but nevertheless maintain that p but not q is morally required.

In the other class of counterexamples, the recipe is as follows: Let p represent something morally required, and q a variant of p that is specified in some morally irrelevant way. Concretely, let p denote that I visit my sick aunt, and q that I do this, entering her flat with my left foot first. Then p and q have equal value, but nevertheless p but not q has obligatory status, contrary to the positivity thesis.

Fortunately, there is an alternative to the positivity thesis. To introduce it we need the following definition:

> A (monadic) predicate H is $\geq$-*contranegative* if and only if for all p and q:
>
> $Hp \,\&\, (\neg p) \geq (\neg q) \rightarrow Hq$.

My proposal is to base the semantics of deontic logic on the semantic principle that prescriptive predicates are contranegative. Hence, if you ought to work hard, and it is worse to be drunk than not to work hard, then you ought not to be drunk:

> *The contranegativity thesis:*[9]
> Prescriptive predicates satisfy contranegativity.

[9] Hansson 1991.

This principle shares with the positivity thesis the advantage of allowing for several prescriptive predicates with different strengths. Neither of the two types of counterexamples that were so easily constructed for the positivity thesis can be transferred to contranegative predicates. Furthermore, the contranegativity thesis yields plausible results for the corresponding permissive and prohibitive predicates, defined in the standard way:

> OBSERVATION 1: Let O, P, and F be predicates with a common domain that is closed under negation, and such that for all p, Op if and only if $\neg P \neg p$, and Op if and only if $F \neg p$. Let $\geq$ be a relation over this domain. Then the following three conditions are equivalent:
> (1) O satisfies $\geq$-contranegativity,
> (2) P satisfies $\geq$-positivity, and
> (3) F satisfies $\geq$-negativity.

PROOF: Left to the reader.

Hence, the contranegativity thesis supports the idea that what is better than something permitted is itself permitted, and that what is worse than something forbidden is itself forbidden. The contranegativity thesis cannot be proved, but it can be corroborated by examples and by the lack of counterexamples. I propose that we accept it on a preliminary basis, or at least as a hypothesis to be tested. The ultimate criterion for its acceptability should of course be whether or not a plausible deontic logic can be based on it. In particular, can the well-known counter-intuitive results in SDL be avoided in a deontic logic that satisfies contranegativity? In order to answer this question, we need to investigate the logical properties of contranegative predicates.

3. General results for contranegative predicates

It turns out that important properties of contranegative predicates correspond closely to properties of the preference relation on which they are based. The following theorem summarizes some major results.

THEOREM 1: A transitive and complete relation $\geq$ satisfies

(a) $(p \geq (p \vee q)) \vee (q \geq (p \vee q))$

(b) $((p \vee q) \geq p) \vee ((p \vee q) \geq q)$

(c) $(p \geq (p \& q)) \vee (p \geq (p \& \neg q))$

(d) If $\mathtt{k}\ q \rightarrow p$, then $p \geq q$.

(e) $(p \geq (p \& q)) \vee (q \geq (p \& q))$

(f) $(p \geq q) \vee ((\neg p \& q) \geq q)$

(g) $p \geq (p \& \neg p)$

if and only if every $\geq$-contranegative predicate O satisfies

(a) $Op \ \& \ Oq \rightarrow O(p \& q)$ (agglomeration)

(b) $O(p \& q) \rightarrow Op \vee Oq$ (disjunctive division)

(c) $P(p \& q) \ \& \ P(p \& \neg q) \rightarrow Pp$ (permissive cancellation[10])

(d) If $\mathtt{k}\ p \rightarrow q$, then $Op \rightarrow Oq$. (necessitation)

(e) $Op \ \& \ Oq \rightarrow O(p \vee q)$ (disjunctive closure)

(f) $Op \ \& \ O(p \rightarrow q) \rightarrow Oq$ (deontic detachment)

(g) $Op \rightarrow O(p \vee \neg p)$

PROOF: See Appendix 1.

The deontic properties listed in (a), (b), and (c) are more plausible than most other deontic postulates, and they are also closely related. To see their

[10] Obviously, for the corresponding permissive predicate P.

relatedness, first note that permissive cancellation follows from disjunctive division.[11]

> OBSERVATION 2: If a prescriptive predicate O satisfies disjunctive division, $(O(p\&q) \rightarrow Op \vee Oq)$, then the corresponding permissive predicate P satisfies permissive cancellation $(P(p\&q)\ \&\ P(p\&\neg q) \rightarrow Pp)$.

PROOF: Substitute $\neg(p\&q)$ for p and $\neg(p\&\neg q)$ for q:

$O(\neg(p\&q)\&\neg\ (p\&\neg q)) \rightarrow O\neg(p\&q) \vee O\neg(p\&\neg q)$

$\neg O\neg(p\&q)\ \&\ \neg O\neg(p\&\neg q) \rightarrow \neg\mathrm{O}(\neg(p\&q)\&\neg(p\&\neg q))$

$P(p\&q)\ \&\ P(p\&\neg q) \rightarrow P((p\&q)\vee\ (p\&\neg q))$

$P(p\&q)\ \&\ P(p\&\neg q) \rightarrow Pp$

The following property of a preference relation:

> $p\geq q \rightarrow p\geq(p\vee q)\ \&\ (p\vee q)\geq q$ (*disjunctive interpolation*)

is quite plausible. It says that $p\vee q$ is intermediate in value between p and q. The following observation provides us with two alternative formulations of this property:

> OBSERVATION 3: Let $\geq$ be a transitive and complete relation. Then it satisfies disjunctive interpolation:
> (1) iff it satisfies $(p\geq(p\vee q)\geq q) \vee\ \ (q\geq(p\vee q)\geq p)$, and

[11] It is also worth noting that if $\geq$ satisifies $(p\vee q\geq p) \vee (p\vee q\geq q)$, then it satisfies $(p\geq(p\&q)) \vee (p\geq(p\&\neg q))$. This can be shown by substituting $p\&q$ for p and $p\&\neg q$ for q.

(2) iff it satisfies both $(p \geq (p \vee q)) \vee (q \geq (p \vee q))$ and $((p \vee q) \geq p) \vee ((p \vee q) \geq q)$

PROOF: Left to the reader.

The two properties of $\geq$ referred to in part (2) of Observation 3 coincide with the two properties used in parts (a) and (b) of Theorem 1. We therefore obtain the following corollary:

COROLLARY TO THEOREM 1, PARTS (a) AND (b): A transitive and complete relation $\geq$ satisfies
(h) $p \geq q \rightarrow p \geq (p \vee q)$ & $(p \vee q) \geq q$ (*disjunctive interpolation*)
if and only if every $\geq$-contranegative predicate O satisfies
(h) both Op & $Oq \rightarrow O(p \& q)$ (agglomeration) and $O(p \& q) \rightarrow Op \vee Oq$ (disjunctive division)

PROOF: From Theorem 1 and Observation 3.

I have already argued that necessitation, the deontic postulate referred to in part (d) of Theorem 1, is an implausible property. Disjunctive closure, the property referred to in part (e), is also quite implausible. To see this, let p denote that I give the oldest of my two children enough to eat, and q that I give my youngest child enough to eat. Then I am subject to the obligations representable as Op and Oq, but it would be strange to claim that $O(p \vee q)$ represents one of the obligations that I have. (Since $p \vee q$ has to be satisfied in order for my obligations to have been fulfilled, it has the same status as the disjunctive statement in Ross's paradox.)

Counterexamples are also available against deontic detachment, the property referred to in part (e).[12] Consider the following equivalent formulation of deontic detachment:

$$O(p \vee q) \,\&\, O\neg p \rightarrow Oq.$$

Suppose that you are for some reason morally required to come to a conference. You are also required not to come unannounced. Let p denote that you stay away from the conference and q that you give notice that you will come. Then $O(p \vee q)$ and $O\neg p$ both hold, but since you should not notify unless you come, Oq does not hold.

The property $Op \rightarrow O(p \vee \neg p)$, that is referred to in part (g) of Theorem 1, is a weakened form of one of the postulates of SDL, namely $O(p \vee \neg p)$. The weakened form is used here for technical reasons.[13] The postulate is implausible both in its original form and in this weakened form. In particular, consider the equivalent version $Oq \rightarrow O(p \vee \neg p)$ of the weakened form, and counterexamples such as "If you are morally required to pay your debts then you are morally required to either commit or not commit mass murder".[14]

The *consistency axiom* of SDL, $Op \rightarrow \neg O\neg p$, is not suitable for being treated in the same way as the deontic postulates listed in Theorem 1. The reason for this is that there is no plausible way to construct a preference relation $\geq$ such that all $\geq$-contranegative predicates satisfy the consistency postulate. This is shown in the following observation:

---

12 McLaughlin 1955. Hansson 1988.

13 There is no non-trivial preference relation $\geq$ such that $O(p \vee \neg p)$ holds for all $\geq$-contranegative predicates. The reason for this is that the empty predicate H, such that $\neg Hr$ holds for all arguments r in its domain, vacuously satisfies contranegativity with respect to any preference relation with an appropriate domain.

14 Cf. Jackson 1985 p. 191 and Lenk 1978 p. 31.

OBSERVATION 4: Let ≥ be a complete and transitive relation, and let there be an element p of its domain such that either $p \geq (\neg p)$ or $(\neg p) \geq p$. Then there is some ≥-contranegative predicate O that does not satisfy the consistency postulate $(Op \rightarrow \neg O \neg p)$.

PROOF: If $p \geq \neg p$, then let O be such that for all q, Oq if and only if $p \geq \neg q$. If $(\neg p) \geq p$, then let O be such that for all q, Oq if and only if $\neg p \geq \neg q$.

4. Reintroducing holistic semantics

In order to obtain a more credible semantic basis for our deontic logic, it will be useful to try to reintroduce holistic alternatives, and use a preference relation (denoted ≥) on them to derive the preference relation on actions that we use as the direct base of the deontic logic. See Diagram 5, and note that no intermediate selection among the holistic alternatives (as in Diagrams 2–3) is used. The idea behind this construction is of course that the normative appraisal of actions should cohere with some reasonable appraisal of the holistic alternatives in which these actions may appear.

I will focus on extremal preference relations. In the present context, this means that the value of an action is completely determined by the values of the best and the worst holistic alternatives that include this action. More precisely, for each action p, let $max(p)$ be the ≥-best alternative in which p is included (or one of them, if there are several of them). Similarly, let $min(p)$ be the ≥-worst alternative in which p is included. We can then define the following extremal preferences:

Maximin preferences:

$p \geq_{\mathbf{i}} q$ iff $min(p) \geq min(q)$

Maximax preferences:

$p \geq_{\mathbf{x}} q$ iff $max(p) \geq max(q)$

Interval maximin preferences:

$p \geq_{\mathbf{ix}} q$ iff either $min(p) > min(q)$ or both $min(p) \equiv min(q)$ and $max(p) \geq max(q)$

Interval maximax preferences:

$p \geq_{\mathbf{xi}} q$ iff either $max(p) > max(q)$ or both $max(p) \equiv max(q)$ and $min(p) \geq min(q)$

Doubly maximizing preferences

$p \geq_{\ddagger} q$ iff $max(p) \geq max(q)$ and $min(p) \geq min(q)$

Two of these preference relations, namely *maximin* and *maximax* preferences, are well-known. They can be said to represent extremely cautious respectively extremely risk-taking decision-making. It has not always been appreciated how extreme the maximin rule is. It requires, for instance, that one be indifferent between owning a valueless piece of paper and owning a ticket in a two-ticket lottery in which the winner will receive ¤1000000 and the loser will receive nothing. The *interval maximin rule* is a modification of the maximin rule that avoids such extreme results. This rule maximizes both worst and best alternatives, but gives maximization of the former absolute priority over maximization of the latter. Similarly, *interval maximax* preference relations maximize both worst and best alternatives,

but give maximization of the latter absolute priority over maximization of the former. The *doubly maximizing* preference relation requires maximization of both maximum and minimum, at the price of not being a complete relation.

Although the interval maximin and interval maximax preference relations mitigate the rather strict principles of maximin and maximax preference relations, respectively, they do so only to a limited degree. It is therefore also of interest to study a wider category of extremal preferences that allows for all assignments of relative priorities to maximization of the best and of the worst alternatives. This can be done as follows:

(1) v is a function that assigns a real number to each (holistic) alternative.
(2) $v_{MAX}(p)$ is the highest value of of any alternative that includes p, and $v_{MIN}(p)$ the lowest value of any alternative that includes p.
(3) δ is a number such that $0<\delta<1$. For all p:
$v_\delta(p) = \delta \cdot v_{MAX}(p) + (1-\delta) \cdot v_{MIN}(p)$
(4) For all p, q :
$p \geq_E q$ iff $v_\delta(p) \geq v_\delta(q)$. (*max-min weighted preferences*)

5. Representation theorems

Elsewhere I have reported a series of representation theorems for contranegative deontic logics that are based on the types of preference relations introduced in the previous section.[15] These theorems make use of a series of background assumptions, primarily that the action-representing sentences can be divided into a finite number of equivalence classes with respect to logical equivalence, and also some conditions relating to the

[15] Hansson 2001, pp. 161–164.

limiting cases of tautologous and contradictory action-representing sentences. Leaving aside these details, the representation theorems are as follows:

1. Maximin preferences:

O is $\geq_{\mathbf{i}}$ -contranegative iff it satisfies

(i) $O(p\&q) \rightarrow Op \vee Oq$

(ii) If Oq and $\mathtt{k}\ p \rightarrow q$, then Op. (*reverse necessitation*), and

(iii) $O^{\perp}$

2. Maximax preferences:

O is $\geq_{\mathbf{x}}$-contranegative iff it satisfies

(i) $Op\ \&\ Oq \rightarrow O(p\&q)$ (*agglomeration*)

(ii) If Op and $\mathtt{k}\ p \rightarrow q$, then Oq. (*necessitation*)

(iii) $\neg O^{\perp}$ (*consistency*)

(iv) There is some p such that Op. (*non-emptiness*)

3. Interval maximin preferences:

O is $\geq_{\mathbf{ix}}$ -contranegative iff it satisfies

(i) $Op\ \&\ Oq \rightarrow O(p\&q)$ (*agglomeration*)

(ii) $O(p\&q) \rightarrow Op \vee Oq$

(iii) If $\mathtt{k}\ r \rightarrow s$, $\mathtt{k}\ s \rightarrow p$, and $\mathtt{k}\ p \rightarrow q$, and $\neg Or$, Os and $\neg Op$, then $\neg Oq$.

(iv) $O^{\perp}$

4. Interval maximax preferences:

O is $\geq_{\mathbf{xi}}$-contranegative iff it satisfies

(i) $Op\ \&\ Oq \rightarrow O(p\&q)$ (*agglomeration*)

(ii) $O(p\&q) \rightarrow Op \vee Oq$

(iii) If $\mathtt{k}\ r \rightarrow s$, $\mathtt{k}\ s \rightarrow p$, and $\mathtt{k}\ p \rightarrow q$, and Or, $\neg Os$ and Op, then Oq.

(iv) $\neg O\bot$ (*consistency*)

(v) There is some p such that Op. (*non-emptiness*)

5. Doubly maximizing preferences:

O is $\geq_{\ddagger}$-contranegative iff it satisfies:

(i) $Op \ \& \ Oq \rightarrow O(p\&q)$

(ii) If $O(p\&q) \rightarrow Op \vee Oq$

(iii) If $\mathtt{k}\ p \rightarrow q$, $\mathtt{k}\ q \rightarrow r$, Op and Or, then Oq.

(iv) $\neg O\bot$

(v) There is some p such that Op.

6. Max-min weighted preferences

O is $\geq_{\mathbf{E}}$-contranegative (with $0<\delta<1$) iff it satisifes:

(i) $Op \ \& \ Oq \rightarrow O(p\&q)$ (*agglomeration*)

(ii) $O(p\&q) \rightarrow Op \vee Oq$

(iii) If $F^{+}p$, $F^{+}q$, $F(p \vee r)$, $\neg F(p \vee s)$, and $\neg F(q \vee r)$, then $\neg F(q \vee s)$.

(iv) If $P^{+}p$, $P^{+}q$, $P(p \vee r)$, $\neg P(p \vee s)$, and $\neg P(q \vee r)$, then $\neg P(q \vee s)$.

where:

$F^{+}p$ iff $F(p\&t)$ for all t such that $p\&t$ is consistent.

$P^{+}p$ iff $P(p\&t)$ for all t such that $p\&t$ is consistent

The proofs can be found in Hansson (2001) except the proof of part 5, that is given in Appendix 2.

The $\geq_{\mathbf{X}}$-, $\geq_{\mathbf{i}}$-, and $\geq_{\mathbf{iX}}$-based operators all satisfy clearly implausible postulates (necessitation, reverse necessitation, and $O\bot$). The other three

types of deontic operators come out better, but unfortunately each of these characterizations makes use of postulates that are unsatisfactorily complex and difficult to grasp. Therefore, although some progress has been made in the construction of a plausible semantics for a deontic logic based on contranegativity, it still remains to develop a plausible semantic structure that has the simplicity of SDL but not its implausible consequences.

Appendix 1

PROOF OF THEOREM 1: *Part a, from LHS to RHS*: Let Op & Oq. We can apply the substitution-instance $((\neg p)\geq(\neg p\vee\neg q)) \vee ((\neg q)\geq\neg p\vee\neg q))$ of LHS. If $((\neg p)\geq(\neg p\vee\neg q))$, then due to contranegativity Op yields $O\neg(\neg p\vee\neg q)$, or equivalently $O(p\&q)$. If $((\neg q)\geq(\neg p\vee\neg q))$, then we can use Oq to obtain the same result.

Part a, from RHS to LHS: Let LHS be violated. We have to show that RHS does not hold either. Since LHS is violated there are p and q such that $p\vee q > p$ and $p\vee q > q$. Since $\geq$ is complete, either $p\geq q$ or $q\geq p$. Without loss of generality we may assume that $p\geq q$. Let O be the predicate such that for all r, Or holds if and only if $p\geq\neg r$. Clearly, O is contranegative. We have $p\geq p$, $p\geq q$, and $p\vee q > p$. Hence, $O(\neg p)$, $O(\neg q)$ and $\neg O(\neg p\&\neg q)$, so that RHS does not hold.

Part b, from LHS to RHS: Let $O(p\&q)$, i.e. $O\neg(\neg p\vee\neg q)$. According to LHS, either $\neg p\vee\neg q\geq\neg p$ or $\neg p\vee\neg q\geq\neg q$. In the first case, Op follows from contranegativity, and in the second case Oq follows in the same way.

Part b, from RHS to LHS: Let LHS be violated. Then there are p and q such that $p>(p\vee q)$ and $q>(p\vee q)$. Let O be such that for all r, Or holds if and only if $p\vee q\geq\neg r$. Then O is contranegative, and $O(\neg p\&\neg q)$, $\neg O\neg p$, and $\neg O\neg q$.

Part c, from LHS to RHS: If $p \geq (p \& q)$, then we can use $P(p \& q)$ and the positivity of P to obtain Pp. If $p \geq (p \& \neg q)$, then we can use $P(p \& \neg q)$ and the positivity of P to obtain Pp.

Part c, from RHS to LHS: Let LHS be violated. Then there are p and q such that $p \& q > p$ and $p \& \neg q > p$. Due to completeness, either $(p \& q) \geq (p \& \neg q)$ or $(p \& \neg q) \geq (p \& q)$. If $(p \& q) \geq (p \& \neg q)$, let P be such that for all r, Pr iff $r \geq (p \& \neg q)$. Then P is $\geq$-positive, and it follows directly that $P(p \& q)$ and $P(p \& \neg q)$. It follows from $p \& \neg q > p$ that $\neg Pp$. The case when $(p \& \neg q) \geq (p \& q)$ is treated analogously.

Part d, from LHS to RHS: Let LHS hold. Let $\mathtt{k}\ p \rightarrow q$ and Op. Then, equivalently: $\mathtt{k}\ \neg q \rightarrow \neg p$. It follows from LHS that $\neg p \geq \neg q$ and from the contranegativity of O that Oq.

Part d, from RHS to LHS: Let LHS be violated. Then there are there are p and q such that $\mathtt{k}\ q \rightarrow p$ and $q > p$.

Let O be the predicate such that for all r , Or holds if and only if $p \geq \neg r$. Then O is $\geq$-contranegative. It follows from $p \geq p$ and $q > p$ that $O(\neg p)$ and $\neg O(\neg q)$. Since $\mathtt{k}\ \neg p \rightarrow \neg q$, RHS is violated.

Part e, from LHS to RHS: Let Op & Oq. From LHS we obtain $\neg p \geq \neg p \& \neg q$ or $\neg q \geq \neg p \& \neg q$, equivalently: $\neg p \geq \neg(p \vee q)$ or $\neg q \geq \neg(p \vee q)$. The rest follows by $\geq$-contranegativity.

Part e, from RHS to LHS: Let LHS be violated. Then there are p and q such that $p \& q > p$ and $p \& q > q$. Due to completeness, either $p \geq q$ or $q \geq p$. If $p \geq q$, let O be the predicate such that for all r, Or if and only if $p \geq \neg r$. Then O is contranegative, and $O \neg p$, $O \neg q$, and $\neg O \neg (p \& q)$, contrary to RHS. The other case is proved analogously.

Part f, from LHS to RHS: Let Op & $O(p \rightarrow q)$. It follows from LHS that $((\neg p) \geq (\neg q)) \vee ((p \& \neg q) \geq (\neg q))$. If $((\neg p) \geq (\neg q))$, then Oq follows from Op due to contranegativity. If $((p \& \neg q) \geq (\neg q))$ or equivalently, $\neg(p \rightarrow q) \geq (\neg q)$, then Oq follows in the same way from $O(p \rightarrow q)$.

Part f, fromRHS to LHS: We are going to assume that LHS does not hold, and prove that then neither does RHS. Since LHS does not hold, there are p and q such that $q>p$ and $q>(\neg p\&q)$. Due to completeness, either $p\geq(\neg p\&q)$ or $(\neg p\&q)\geq p$.

Case 1, $p\geq(\neg p\&q)$: Let O be the such that for all r, Or holds if and only if $p\geq\neg r$. Then O is $\geq$-contranegative. It therefore follows from $p\geq p$ that $O\neg p$, from $p\geq(\neg p\&q)$ that $O(q\rightarrow p)$, and from $q>p$ that $\neg O\neg q$.

Case 2, $(\neg p\&q)\geq p$: Let O be such that for all r, Or holds if and only if $(\neg p\&q)\geq\neg r$. Then O is contranegative. It follows from $(\neg p\&q)\geq(\neg p\&q)$ that $O(q\rightarrow p)$, from $(\neg p\&q)\geq p$ that $O\neg p$, and from $q>(\neg p\&q)$ that $\neg O\neg q$.

Hence, in both cases we have $O\neg p$, $O(\neg p\rightarrow\neg q)$, and $\neg O\neg q$, so that RHS does not hold.

Part g, from LHS to RHS: Let Op. It follows from this and the substitution-instance $\neg p\geq\neg(p\vee\neg p)$ of LHS that $O(p\vee\neg p)$.

Part g, from RHS to LHS: We are going to assume that LHS is violated, and show that then RHS is also violated. Since LHS is violated, there is some p such that $\neg(p\vee\neg p)> p$. Let O be a predicate such that for all r, Or iff $p\geq\neg r$. Then O is contranegative. It follows that $O\neg p$ that $\neg O(p\vee\neg p)$.

Appendix 2

The following is a proof of the result on doubly maximizing preferences referred to in section 5.

> DEFINITION: Let $\geq$ be a relation and O a monadic predicate that both take the elements of a language $\mathcal{L}$ as arguments. Then O is a *sentence-limited $\geq$-contranegative predicate* if and only if there is some sentence $f\in\mathcal{L}$ such that for all $p\in\mathcal{L}$: $Op\leftrightarrow f\geq\neg p$.

LEMMA: Let $\geq$ be a transitive and complete relation over a set $\mathcal{L}$ of sentences that has a finite number of equivalence classes with respect to logical equivalence. Then O is a non-empty $\geq$-contranegative predicate on $\mathcal{L}$ if and only if it is a sentence-limited $\geq$-contranegative predicate on $\mathcal{L}$.

PROOF: For the non-trivial direction, let O be $\geq$-contranegative and non-empty. Since $\geq$ is a finite ordering, there is some r such that Or and that $\neg Op$ for all p such that $\neg p > \neg r$. We need to show that for all q, Oq iff $\neg r \geq \neg q$. It follows from the construction of r that if $\neg q > \neg r$, then $\neg Oq$. It follows from Or and the $\geq$-contranegativity of O that if $\neg r \geq \neg q$ then Oq.

THEOREM 2: The following are equivalent conditions on a predicate O:

(1) O is a sentence-limited contranegative predicate with respect to a doubly maximizing preference relation, and it holds for the sentence limit f that $max(f)$ is non-maximal.

(2) O satisfies the postulates

(i) $Op \ \& \ Oq \rightarrow O(p \& q)$

(ii) If $O(p \& q) \rightarrow Op \vee Oq$

(iii) If $\mathtt{k}\ p \rightarrow q$, $\mathtt{k}\ q \rightarrow r$, Op and Or, then Oq

(iv) $\neg O\bot$

(v) There is some p such that Op.

PROOF: *From 1 to 2*: For (i), let Op and Oq. Then $max(f) \geq max(\neg p)$ and $max(f) \geq max(\neg q)$, hence $max(f) \geq max(\neg p \vee \neg q)$. It follows from $min(f) \geq min(\neg p)$ that $min(f) \geq min(\neg p \vee \neg q)$. Thus $O(p \& q)$.

For (ii), let $O(p\&q)$. Then $max(f) \geq max(\neg p \vee \neg q)$, hence both $max(f) \geq max(\neg p)$ and $max(f) \geq max(\neg q)$. Furthermore, since $min(f) \geq min(\neg p \vee \neg q)$, either $min(f) \geq min(\neg p)$ or $min(f) \geq min(\neg q)$. Hence, either Op or Oq.

For (iii), let $\mathtt{k}\ p \rightarrow q$, $\mathtt{k}\ q \rightarrow r$, Op and Or. It follows from $\mathtt{k}\ \neg q \rightarrow \neg p$ that $max(\neg p) \geq max(\neg q)$ and from Op that $max(f) \geq max(\neg p)$, hence $max(f) \geq max(\neg q)$. Similarly, it follows from $\mathtt{k}\ \neg r \rightarrow \neg q$ that $min(\neg r) \geq min(\neg q)$ and from Or that $min(f) \geq min(\neg r)$, hence $min(f) \geq min(\neg q)$. We may conclude that Oq.

For (iv), suppose to the contrary that $O\bot$, i.e. $f \geq_{\ddagger} \neg\bot$. Then $max(f)$ is maximal, contrary to the condition.

For (v), it follows from $f \geq_{\ddagger} f$ that $O\neg f$.

From 2 to 1: Let $Z = \{p \mid Op\}$, $A = \mathrm{Cn}(Z)$ and $B = A \backslash Z$. We need the following two properties of these sets: (A) A is consistent, and (B) B is logically closed.

For (A), suppose to the contrary that A is inconsistent. Then due to compactness there is a finite and inconsistent subset of Z, but due to (i) and (iv) this is impossible.

For (B), it is sufficient to show (Bα) that if $\mathtt{k}\ p \rightarrow q$ and $p \in B$, then $q \in B$, and (Bβ) if $p, q \in B$, then $p\&q \in B$.

(Bα): Let $p \rightarrow q$ and $p \in B$. It follows from $p \in B \subseteq A$, due to (i) and compactness, that there is some s such that Os and $\mathtt{k}\ s \rightarrow p$. Suppose that Oq. Then it follows from (iii) that Op, contrary to $p \in B$. Hence $\neg Oq$. It follows that $q \in B$.

(Bβ): Let $p, q \in B$. Clearly $p\&q \in A$. Suppose that $p\&q \notin B$. Then $O(p\&q)$, hence due to (ii) either Op or Oq, contrary to $p, q \in B$.

It follows from (v) that $Z \neq \emptyset$, hence $B \neq A$, so that $B \subset A$. Let $a = \&A$ and $b = \&B$. Let $B' = \mathrm{Cn}(\{b\&\neg a\})$. Then $A \cap B = A \cap B'$, hence $A \backslash B = A \backslash B'$.

Let $\mathcal{A}$ be the set of maximal consistent subsets of the language. We are going to construct a preference relation $\geq$ on $\mathcal{A}$ with the strict part $>$ ("better than") and the symmetric part $\equiv$ ("equal in value to"). Let $\mathcal{A}_1 = \{X \in \mathcal{A} \mid A \subseteq X\}$ and $\mathcal{A}_3 = \{X \in \mathcal{A} \mid B' \subseteq X\}$. (Note that they have been constructed to be mutually exclusive.) Let $\mathcal{A}_2 = \mathcal{A} \backslash (\mathcal{A}_1 \cup \mathcal{A}_3)$. If $\mathcal{A}_2$ is empty, then let f be a sentence such that $max(f)$ and $min(f)$ are both in $\mathcal{A}_3$. Let $\mathcal{A}_1$ and $\mathcal{A}_3$ be equivalence classes with respect to $\equiv$, and let $\mathcal{A}_1 > \mathcal{A}_3$. If $\mathcal{A}_2$ is non-empty, then let f be such that $max(f) \in \mathcal{A}_2$ and $min(f) \in \mathcal{A}_3$. Let $\mathcal{A}_1$, $\mathcal{A}_2$, and $\mathcal{A}_3$ be equivalence classes with respect to $\equiv$, and let $\mathcal{A}_1 > \mathcal{A}_2 > \mathcal{A}_3$. Then in both cases:

$Op \leftrightarrow p \in Z$

$\leftrightarrow p \in A \backslash B$

$\leftrightarrow p \in A \backslash B'$

$\leftrightarrow p \in \cap \mathcal{A}_1$ and p $\notin \cap \mathcal{A}_3$

$\leftrightarrow \neg p \notin \cup \mathcal{A}_1$ and $\neg p \in \cup \mathcal{A}_3$

$\leftrightarrow max(f) \geq max(\neg p)$ and $min(f) \geq min(\neg p)$

$\leftrightarrow f \geq_{\ddagger} \neg p$.

References

Åqvist, Lennart (1967) "Good Samaritans, Contrary–to–Duty Imperatives, and Epistemic Obligations", *Noûs*, 1:361–379.

Føllesdal, Dagfinn and Risto Hilpinen (1970) "Deontic Logic: An Introduction", pp. 1–35 in Risto Hilpinen (ed.), *Deontic Logic: Introductory and Systematic Readings*. Reidel, Dordrecht.

Hansson, Sven Ove (1988) "Deontic Logic Without Misleading Alethic Analogies – Part I", *Logique et Analyse* 31:337–353.

Hansson, Sven Ove (1991) "Norms and values", *Crítica* 23:3–13.

Hansson, Sven Ove (2001) The Structure of Values and Norms, Cambridge University Press.

Jackson, Frank (1985) “On the Semantics and Logic of Obligation”, *Mind* 94:177–195.

Lenk, Hans (1978) “Varieties of Commitment”, *Theory and Decision* 9:17–37.

McLaughlin, RN (1955) “Further Problems of Derived Obligation”, *Mind* 64:400–402.

Prior, AN (1954) “The Paradoxes of Derived Obligation”, *Mind* 63:64–65.

Prior, AN (1958) “Escapism”, pp. 135–146 in AI Melden (ed.) *Essays in Moral Philosophy*. Univ. of Washington Press, Seattle.

Ross, Alf (1941) “Imperatives and Logic”, *Theoria* 7:53–71.

von Wright, Georg Henrik (1951) “Deontic Logic”, *Mind* 60:1–15.

von Wright, Georg Henrik (1998) “Deontic Logic – as I see it”, paper presented at the Fourth International Workshop on Deontic Logic in Computer Science (DEON'98), Bologna.

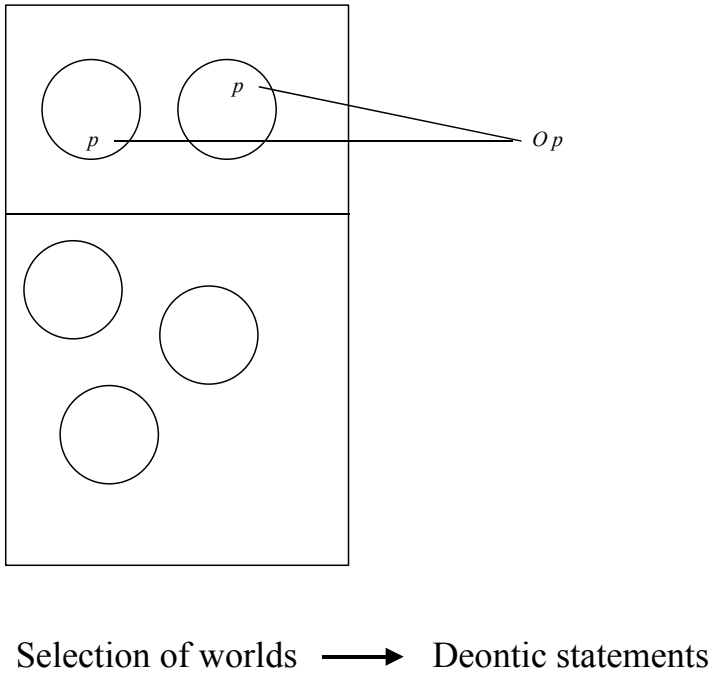

Figure 1. Standard deontic logic.

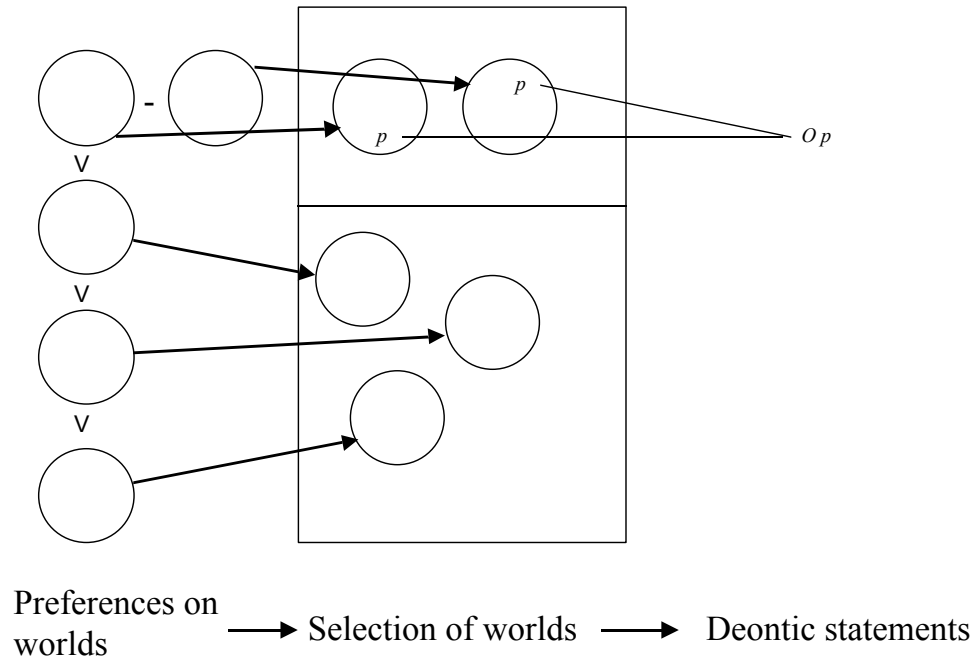

Figure 2. Standard deontic logic, based on a preference relation over worlds.

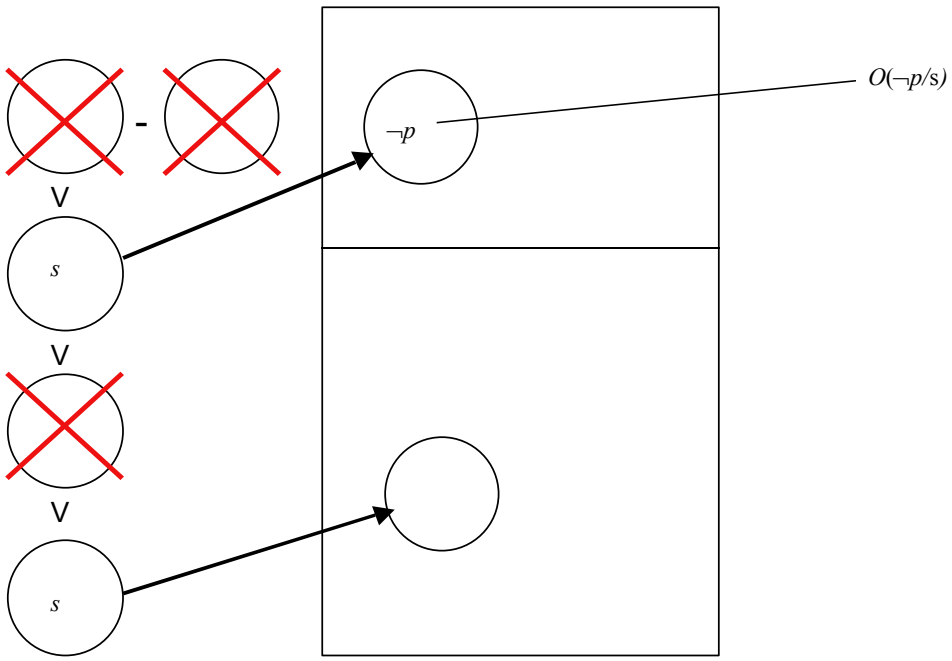

Figure 3. Standard deontic logic, modified to account for conditional obligations.

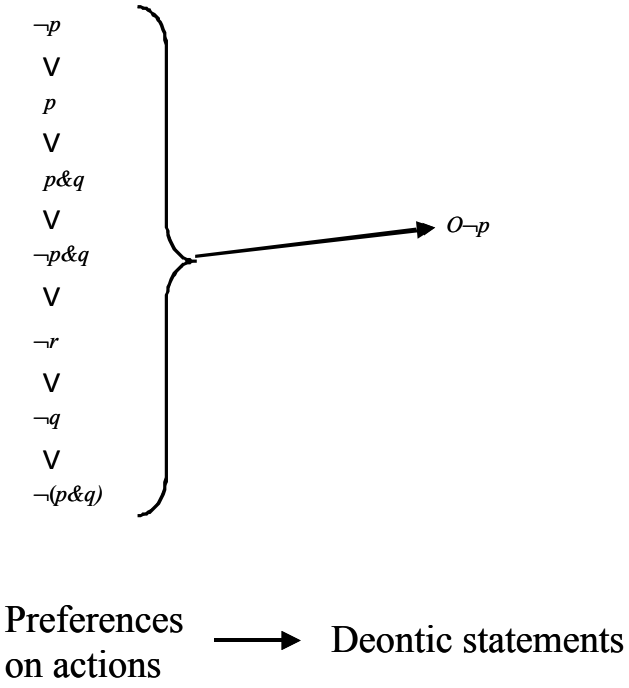

Figure 4. Deontic logic based directly on a preference relation over actions.

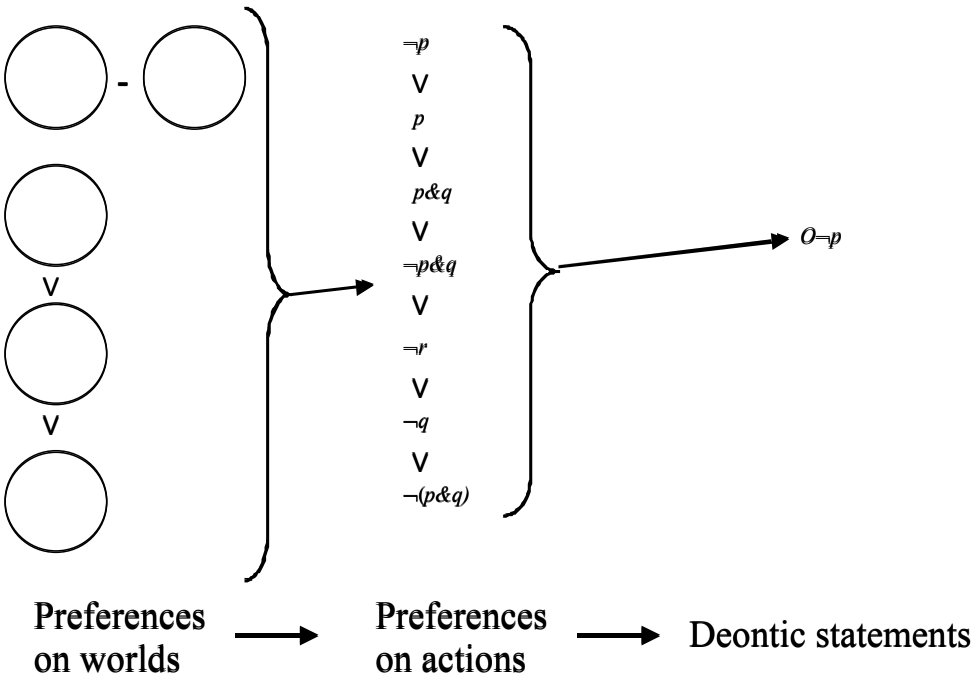

Figure 5. Deontic logic based directly on a preference relation over actions that is in its turn based on a preference relation over holistic alternatives.