6 Deontic Logic

6.1 Permission and obligation

Deontic logic studies formal properties of obligation, permission, prohibition, and related normative concepts. The box in deontic logic is usually written 'O' (for 'obligation' or 'ought'), the diamond 'P' (for 'permission'). So if q translates 'you cook dinner', we might use $\mathsf{O}\,q$ to express that you must cook dinner, in the deontic sense of 'must': it is obligatory that you cook dinner.

We assume that obligation and permission are duals: you are not obligated to do something iff you are allowed to not do it; you are not allowed to do something iff you are obligated to not do it. Before we take a closer look at the logic of these concepts, a few more clarifications are in order.

First, there are many kinds of norms: legal norms, moral norms, prudential norms, social norms, and so on. There may also be overarching norms that combine some or all of the others. Deontic logic is applicable to all kinds of norms, so we do not have to settle whether O expresses legal obligation or moral obligation or some other kind of obligation. However, it is important not to equivocate. If the law requires p and morality $\neg p$, we should not formalize this as $\mathsf{O}\,p \wedge \mathsf{O}\,\neg p$. It would be better to use a multi-modal language with different operators for legal and moral obligation.

Second, we need to clarify how obligations and permissions are related to agents. Intuitively, obligations and permissions vary from agent to agent. If it is your turn to cook dinner, then you are obligated to cook dinner, but I am not. To capture this agent-relativity, we could add agent subscripts to the operators, as we did in epistemic logic. We could then express our different obligations as $\mathsf{O}_1\,p \wedge \neg\,\mathsf{O}_2\,p$. But what does the sentence letter p stand for? When I say that you are obligated to cook dinner, the object of the obligation appears to be a type of act: cooking dinner. In the language of modal propositional logic, however, O and P are sentence operators. So unless we want to say that verb phrases in English (like 'cook dinner') should be translated into sentences of $\mathfrak{L}_M$ – which is actually doable, but non-standard –

we have to transform the acts an agent is obligated or permitted to perform into propositions.

Consider sentence (1), which is arguably equivalent to (2).

(1) You ought to cook dinner.

(2) You ought to see to it that you cook dinner.

In (2), the operator 'you ought to see to it that' attaches to a sentence, 'you cook dinner'. So we can translate (1) via (2) as $\mathsf{O}_1\, p$, where p translates 'you cook dinner', and O_1 corresponds to 'you ought to see to it that'.

The subject (you) is mentioned twice in (2), which may seem redundant. A common assumption in deontic logic is that we can drop the agent subscripts from deontic operators, since the embedded proposition will tell us upon whom the obligation or permission falls. Informally, the idea is that (2) is equivalent to (3), with an impersonal 'ought'.

(3) It ought to be the case that you cook dinner.

The impersonal 'ought' also figures in statements like (4).

(4) Nobody ought to die of hunger.

When I say (4), I don't mean that nobody is obligated to die of hunger, nor do I mean that everybody is obligated to not die of hunger. Rather, I mean that a certain state of affairs – that nobody dies of hunger – ought to be the case. By itself, this does not impose any obligations on anyone.

On closer inspection, the equivalence between personalized 'ought' statements like (2) and impersonal 'ought' statements like (3) is questionable. Suppose Amy has promised to play with Betty. So Amy is obligated to play with Betty. But Betty is not thereby obligated to play with Amy. Betty may even have promised not to play with Amy. It is hard to express these facts in terms of impersonal oughts. If we say that it ought to be the case that Amy plays with Betty, we're missing the fact that the obligation falls on Amy, not on Betty (who might be under a contrary obligation). So perhaps it would be better to keep the agent subscripts after all.

It can also be useful to make the 'see to it that' component in statements like (2) explicit. That Amy ought to play with Betty could then be translated as $\mathsf{O}_a\, \mathsf{STIT}\, p$, where STIT formalizes 'sees to it that'. This allows us to distinguish between the following three claims.

$\mathsf{O}_a\,\mathsf{STIT}\,\neg p$	Amy ought to see to it that she doesn't play with Betty.
$\mathsf{O}_a\,\neg\,\mathsf{STIT}\,p$	Amy ought to not see to it that she plays with Betty.
$\neg\,\mathsf{O}_a\,\mathsf{STIT}\,p$	It is not the case that Amy ought to see to it that she plays with Betty.

The STIT operator has proved useful to represent different concepts of rights and duties. In what follows, we will nonetheless stick to the simplest (and oldest) approach, without a STIT operator and agent subscripts. This approach is sufficient for many applications, but its limitations should be kept in mind.

Exercise 6.1

Let $\mathsf{F}A$ mean that A is forbidden. Can you define F in terms of O or P (or both)?

Exercise 6.2

Translate the following sentences into the standard language of deontic logic.

(a) You must not go into the garden.
(b) You may not go into the garden.
(c) Jones ought to help his neighbours.
(d) If Jones is going to help his neighbours, then he ought to tell them he's coming.
(e) If Jones isn't going to help his neighbours, then he ought to not tell them he's coming.

6.2 Standard deontic logic

Think of a possible world as a history of events. For any such history, and any system of norms, we can ask whether the history conforms to the norms or not. Let's call a world *acceptable* relative to some norms if everything that happens at the world conforms to the norms. That is, a world is acceptable if it contains no violation of any relevant norm.

By definition, whatever happens at an acceptable world is permitted, in the sense that it does not violate any (relevant) norms. The converse is plausible as well:

whenever something is permitted then it is the case at some acceptable world. For example, if it is permitted that Amy plays with Betty, then there should be a complete history of events in which Amy plays with Betty and no norms are violated. If there were no such history, then Amy's playing with Betty would logically entail the violation of some norms; but if an act entails the violation of some norms, then it is hard to see how the act could be permitted relative to these norms.

So we have the following connection between permission and acceptable worlds, which amounts to a "possible-worlds analysis" of permission:

> *A* is permitted relative to some norms iff *A* is the case at some possible world that is acceptable relative to these norms.

Given the duality of permission and obligation, we also get a possible-worlds analysis of obligation:

> *A* is obligatory relative to some norms iff *A* is the case at all worlds that are acceptable relative to these norms.

For example, if it is obligatory that you cook dinner, then every history of events in which all obligations are met is a history in which you cook dinner.

As in earlier chapters, we can turn this analysis into a "model theory" for deontic logic. In logic, we are not interested in who is obligated to do what, but in whether a given deontic statement is logically valid, or whether it logically follows from other statements. Validity means truth in every conceivable scenario under every interpretation of the descriptive vocabulary. In deontic logic, a scenario has to settle not just the descriptive facts, but also the relevant norms, which can be fixed by specifying which worlds are acceptable relative to which other worlds.

We package a scenario of this type and an interpretation of the sentence letters into a Kripke model, interpreting the accessibility relation as specifying which worlds are acceptable from the perspective of any given world.

So a standard model of (propositional) deontic logic consists of a set of "worlds" W, an "accessibility" relation on W, and an interpretation function V that assigns a truth-value to every sentence letter at every world. A world v is accessible from a world w if everything that goes on at v is permitted by the norms at w – equivalently, if everything that ought to be the case at w is the case at v. Worlds that are accessible from w in this sense are also called **ideal** relative to w.

Our possible-worlds analysis of obligation and permission is reflected in definition 3.2, which settles under what conditions a sentence is true at a world in a model. Writing the box as 'O' and the diamond as P', clauses (g) and (h) of the definition state:

$$M, w \models \mathsf{O}\, A \text{ iff } M, v \models A \text{ for all } v \text{ such that } wRv;$$
$$M, w \models \mathsf{P}\, A \text{ iff } M, v \models A \text{ for some } v \text{ such that } wRv.$$

Formally, a sentence is valid iff it is true at every world in every suitable model. If we count all Kripke models as suitable, the logic of obligation and permission will be the minimal normal modal logic K. We can get stronger logics by imposing constraints on the accessibility relation. Let's have a look at a few options.

If we stipulate that the deontic accessibility relation is reflexive, so that every world can see itself, then the **T**-schema for obligation becomes valid:

(**T**) $\quad \mathsf{O}\, A \rightarrow A$

But there are many intuitive counterexamples to **T**. The fact that you are obligated to cook dinner does not logically entail that you cook dinner. Semantically speaking, many worlds are not ideal relative to themselves. So we should not assume reflexivity.

We might, however, impose the weaker condition of seriality – that each world can see some world. This would validate principle **D**:

(**D**) $\quad \mathsf{O}\, A \rightarrow \mathsf{P}\, A$

Intuitively, **D** says that the norms are consistent: if you're obligated to do A, then you can't also be obligated to do not-A. (Remember that $\mathsf{P}\, A$ is equivalent to $\neg\, \mathsf{O} \neg A$.) Semantically, **D** corresponds to the assumption that there is always at least one world at which all the norms are satisfied.

Without seriality, we have to allow for worlds from which no world is accessible. At such a world, all sentences of the form $\mathsf{O}\, A$ are true, and all sentences of the form $\mathsf{P}\, A$ are false. Everything is obligatory, but nothing is allowed. That makes little sense. If we use use Kripke semantics for deontic logic, we should therefore rule out inconsistent norms and accept **D** as valid.

Here it may be important to distinguish *prima facie* obligations from *actual*, or *all-things-considered* obligations. If you've promised to cook dinner, you are under

a *prima facie* obligation to cook dinner. But the obligation can be overridden by intervening circumstances or contrary obligations. If your child has an accident and needs urgent medical care, the right thing to do may well be to not cook dinner and instead bring your child to the hospital. In a sense, you are under conflicting obligations: you ought to cook dinner, and you ought to look after your child (and not cook dinner). There is no world at which you meet both of these obligations. But that is not a counterexample to **D**, if we understand O as all-things-considered obligation. You are *prima facie* obligated to cook dinner, but all things considered, you should not cook dinner.

Let's return to the non-reflexivity of the deontic accessibility relation: many things that are not the case nonetheless ought to be the case. Some have argued that this is only true in non-ideal worlds. In an ideal world, everything that ought to be the case is the case. By this line of thought, if a world v is accessible from some world w – meaning that v is ideal relative to w – then v should be accessible from itself. This condition is sometimes called "shift reflexivity" and corresponds to the following schema **U** (for "utopia")

(**U**) $\quad \mathsf{O}(\mathsf{O}A \rightarrow A)$

In words: it ought to be the case that whatever ought to be the case is the case.

The **U** principle is entailed by an alternative way of formalizing obligation and permission that goes back to Leibniz. Let 'OK' be a propositional constant whose intended meaning is that all norms are satisfied, no obligations violated. Suppose we add this expression to $\mathfrak{L}_M$, and we interpret the box of $\mathfrak{L}_M$ as circumstantial necessity. (So $\Box A$ means that A is entailed by relevant circumstances.) Arguably, $\mathsf{O}A$ is then definable as $\Box(\mathsf{OK} \rightarrow A)$: it ought to be that A iff, necessarily, A is the case whenever all obligations are met. It is not hard to show that if the **T**-schema is valid for the circumstantial box, then the **U**-schema is valid for O (on the present definition).

Exercise 6.3

(a) Translate the **U**-schema into the Leibnizian language just proposed.
(b) Give a tree proof for the translated **U**-schema, using the T-rules for the box.

Exercise 6.4

How could we define P in terms of $\Box$ and OK?

Turning to more familiar schemas and frame conditions, what shall we say about transitivity and euclideanness, and the corresponding schemas **4** and **5**?

(**4**) $\quad \mathsf{O}\,A \rightarrow \mathsf{O}\,\mathsf{O}\,A$

(**5**) $\quad \mathsf{P}\,A \rightarrow \mathsf{O}\,\mathsf{P}\,A$

Here we face a problem. Translated back into English, it is rather unclear what these are supposed to say. If something ought to be the case, ought it to be the case that it ought to be the case? If something is permitted, is it obligatory that it is permitted?

Iterations of deontic operators sound strange in ordinary language. But they have a well-defined meaning in our Kripke semantics. The validity of **4** would mean that whenever something is obligatory at a world, then it is also obligatory at all ideal alternatives to that world. Similarly, **5** would mean that if something is permissible at a world, then it's also permissible at all ideal alternatives to that world. On the background of **D**, these two assumptions would imply that for each world there is a class of ideal worlds all of which are ideal relative to each other.

To get a clearer grip on whether that is plausible, we need to clarify how obligations and permissions can vary from world to world.

One obvious sense in which norms can vary across worlds is that people subscribe to different norms at different worlds. At our world, UK traffic law requires driving on the left, and most people think it is morally wrong to torture animals for fun. At other worlds, the laws and attitudes are different.

Let v be a world at which the traffic laws require driving on the right, and at which everyone thinks it is fine to torture animals. Suppose Norman at v is torturing kittens, while driving on the right (in the UK). Is Norman doing something that's morally wrong? Is he doing something that violates the traffic laws?

The answer depends on whether we evaluate Norman's acts relative to our norms – the norms at our world – or relative to the norms at Norman's world. Both perspectives make sense, and they lead to different deontic logics.

On an **absolutist** conception, the basic norms do not vary from world to world. Whichever world we look at, we always assess it relative to the same set of norms. On this conception, it is natural to assume that the very same worlds are ideal relative

to any world: a world will be accessible from any world just in case it contains no violation of the norms. The resulting logic of obligation and permission is KD45.

Exercise 6.5

Explain why the deontic accessibility relation is transitive and euclidean, on the absolutist conception.

On a **relativist** conception of norms, we evaluate the events at other worlds relative to the norms at these worlds. Transitivity and euclidity become implausible, as does shift reflexivity. For example, consider another world u in which the law says that one should drive on the right but everyone nonetheless drives on the left. Nothing that happens at u, we may assume, violates the traffic laws of our world. So u is deontically accessible from the actual world. But if we evaluate the events at u relative to the laws at u, then much of what happens at u violates the norms, so u is not deontically accessible from itself. Shift reflexivity fails.

Exercise 6.6

Explain why deontic accessibility is neither transitive nor euclidean, on the relativist conception.

The relativist conception is more common in deontic logic. As a consequence, so-called **standard deontic logic** assumes only that the accessibility relation is serial, making the system D the complete logic of obligation and permission.

The proposed logics of absolutism and relativism only disagree on sentences in which a deontic operator occurs in the scope of another deontic operator. Any sentence that does not contain an O or P operator embedded under another O or P operator is D-valid iff it is KD45-valid.

Exercise 6.7

Consider a world in which there are no sentient beings, and nothing else that could introduce norms or laws. Since there are no norms at this world, nothing is obligatory relative to the world's norms, and nothing is permitted. Explain why this casts doubt on the validity of **Dual1** and **Dual2** in the logic

of relativist obligation and permission.

Exercise 6.8

Suppose Amy ought to either promise to help Betty or promise to help Carla. If she were to promise to help Betty, she would be obligated to help her. And if she were to promise to help Carla, she would be obligated to help Carla. So it ought to be the case that Amy is either obligated to help Betty or obligated to help Carla. In fact, Amy makes neither promise, so she is neither obligated to help Betty nor to help Carla. Explain why this casts doubt on the validity of **5**.

Exercise 6.9

Consider the **C4** schema $\mathsf{O}\,\mathsf{O}\,A \to \mathsf{O}\,A$. Show that

(a) if **U** is valid on a frame, then so is **C4**;
(b) it is not the case that if **C4** is valid on a frame, then so is **U**.

Exercise 6.10

Give either a proof or a counterexample for the following sentences, assuming the logic of obligation and permission is D.

(a) $\mathsf{O}\,p \to \mathsf{O}(p \vee q)$
(b) $\mathsf{P}(p \vee q) \to (\mathsf{P}\,p \wedge \mathsf{P}\,q)$
(c) $(\mathsf{O}\,p \wedge \mathsf{O}\,q) \to \mathsf{O}(p \wedge q)$
(d) $\mathsf{P}\,p \to \mathsf{P}(p \vee q)$
(e) $\mathsf{O}(q \wedge \neg q)$
(f) $\neg(\mathsf{O}\,p \wedge \mathsf{O}\,\neg p)$
(g) $\neg\,\mathsf{P}(p \vee q) \to (\mathsf{P}\,\neg p \vee \mathsf{P}\,\neg q)$
(h) $\mathsf{O}\,\mathsf{P}\,q \vee \mathsf{P}\,\mathsf{O}\,q$

6.3 Norms and circumstances

The possible-worlds analysis from the previous section assumes that something ought to be the case iff it is the case at all ideal worlds, where no norms are violated. Many ordinary statements about oughts and obligations do not fit this analysis.

Suppose you are walking past a drowning baby. You ought to save the baby. But are you saving the baby at every world at which no norms are violated? Clearly not. There are worlds at which the baby never fell into the pond, and others at which you are overseas and have no means to rescue the baby. These worlds need not involve any violations of norms.

The general point is that whether something ought to be the case depends not just on the norms but also on the circumstances. If you walk past a drowning baby, you ought to rescue the baby; under other circumstances (if the baby never fell into the pond), no such obligation arises.

We can account for the dependence of obligations on circumstances by changing our interpretation of the accessibility relation. Previously, we assumed that a world v is accessible from w iff all the norms of w are respected at v. On the new interpretation, we also require that relevant circumstances at w are preserved at v. For example, if w is a world at which you're walking past a drowning baby, then any accessible world will also be a world at which you're walking past a drowning baby. Worlds at which the baby hasn't fallen into the pond are ignored.

As a first stab, we might therefore redefine deontic accessibility as follows:

> A world v is deontically accessible from a world w iff (a) the relevant circumstances at w are also the case at v, and (b) no norms from w are violated at v.

I use 'relevant circumstances' as a placeholder for whatever we hold fixed when we consider what ought to be the case. In general, we tend to hold fixed anything that can no longer be changed. If the baby has fallen into the pond at w, then there is nothing anyone can do to undo the falling; so the falling is a "relevant circumstance" that takes place at every world accessible from w. A more informative definition of relevant circumstances would have to address difficult philosophical questions about free will, among other things. Let's set these issues aside.

Clause (b) in the above definition assumes that no norms are violated at any accessible world. But if accessibility is restricted by circumstances, then this is implausible, as the relevant circumstances will often involve violations of norms.

The problem is brought ought by Arthur Prior's "Samaritan Paradox". Suppose someone has been injured in a robbery, and Jones has the opportunity to help. We want to say that Jones ought to help the victim. On the possible-worlds analysis of 'ought', this means that Jones helps the victim at all worlds accessible from the actual world. It follows that the robbery took place at all these worlds. (In a world without a robbery, there is no victim to help.) But then all the accessible worlds contain a violation of norms. In a truly ideal world, nobody would have been robbed and injured.

In the Samaritan Paradox, the robbery cannot be undone; it has happened at all worlds that are compatible with the "relevant circumstances". All of these worlds are therefore non-ideal. But worlds at which Jones doesn't help the victim are even *worse*, in terms of norm violations, than worlds at which he helps the victim. Both kinds of worlds are bad, because the victim got robbed. But our norms don't just divide the possible worlds into good and bad; they allow for finer distinctions between bad worlds and even worse worlds. Jones ought to help the victim because that's what he does in the *best* worlds among those he can bring about, even though none of these worlds are ideal.

So here is a second pass at the revised definition of deontic accessibility.

> A world v is deontically accessible from a world w iff (a) the relevant circumstances at w are also the case at v, and (b) v is one of the best worlds, by the norms at w, among worlds at which the relevant circumstances from w are held fixed.

In the previous section, I argued that if we adopt an absolutist conception of norms (so that we hold fixed the actual norms when evaluating events at other worlds), then the same worlds will be ideal relative to all worlds, giving us KD45 as the logic of obligation and permission. That is no longer true on the present, revised interpretation of deontic accessibility, unless the circumstantial accessibility relation that implicitly figures in clause (a) is an equivalence relation.

It is often useful to divide the revised accessibility relation into its circumstantial and deontic components. To this end, we first add a circumstantial accessibility relation to our models. Informally (and roughly), a world v is circumstantially

accessible from w iff there is something one can do at w that would bring about v. I will sometimes call such worlds *open*, for brevity. Next, we need to settle which worlds in a model are better than others, relative to the norms at any given world (which may be the norms at every world, on an absolutist approach).

Let '$u \prec_w v$' mean that world u is better than world v relative to the norms at w. (Roughly speaking, '$u \prec_w v$' means that w contains *fewer* violations of norms.) We assume that for any world w, the relation $\prec_w$ is asymmetric and transitive, where asymmetry means that if $u \prec_w v$ then it is not the case that $v \prec_w u$. Asymmetric and transitive relations are also known as **partial orders**.

Definition 6.1

An **deontic ordering model** consists of

- a non-empty set W (the worlds),
- a binary relation R on W (the circumstantial accessibility relation),
- for each world $w \in W$, a partial order $\prec_w$ on W (the world-relative ranking of worlds as better or worse), and
- a function V that assigns to each sentence letter of $\mathfrak{L}_M$ and each member of W a truth-value.

Now we need to say under what conditions a sentence of the form $\mathsf{O}\,A$ is true at a world in a model. Informally, $\mathsf{O}\,A$ should be true at w iff A is true at the best worlds among those that are circumstantially accessible (i.e., open). Let's introduce one more piece of notation. For any set of worlds S and any partial order $\prec$, let $Min^{\prec}(S)$ be the set of $\prec$-minimal members of S:

$$Min^{\prec}(S) =_{\text{def}} \{v \in S : \neg\exists u(u \prec v)\}.$$

Intuitively, $Min^{\prec}(S)$ contains all the worlds from S with the fewest norm violations.

Here, then, are the truth-conditions of $\mathsf{O}\,A$ and $\mathsf{P}\,A$ in deontic ordering models M:

$$M, w \models \mathsf{O}\,A \text{ iff } M, v \models A \text{ for all } v \in Min^{\prec_w}(\{u : wRv\})$$
$$M, w \models \mathsf{P}\,A \text{ iff } M, v \models A \text{ for some } v \in Min^{\prec_w}(\{u : wRv\})$$

This is just a formal way of saying that $\mathsf{O}\,A$ is true at w iff A is true at the best worlds (by the norms at w) among the worlds that are open at w.

If we want the **D**-schema to be valid, we have to assume that for any world w, there is always at least one best world among the open worlds, so that $Min^{\prec_w}(\{u : wRv\})$ is never the empty set. Let's make this assumption.

Exercise 6.11

Mary must go to prison because she has robbed a bank. Does Mary go to prison at all legally "ideal" worlds, where no laws are violated? Describe a simple deontic ordering model in which 'Mary must go to prison', translated as $\mathsf{O}\,p$, is true.

Ordering models not only help to clarify how the logic of obligation and permission depends on formal properties of the circumstantial accessibility relation R and the deontic orderings $\prec_w$. They also help with problems that often arises when we try to formalize statements containing modal operators and if-clauses, like (1)–(3).

(1) If you smoke then you must smoke outside.
(2) If you miss the deadline for tax returns then you must pay a fine.
(3) If you have promised to call your parents then you must call them.

When we translate these into $\mathfrak{Q}_M$, we seem to face a choice between (W) and (N).

(W) $\mathsf{O}(p \to q)$
(N) $p \to \mathsf{O}\,q$

In (W), the operator O is said to have **wide scope** because it applies to the entire conditional $p \to q$. In (N), the operator has **narrow scope** because it only applies to the consequent q.

On reflection, neither translation is satisfactory. Starting with (N), remember that a material conditional $A \to B$ is true whenever the antecedent A is false. So if p is false then $p \to \mathsf{O}\,q$ is automatically true. But the mere fact that (say) you don't smoke surely doesn't entail that if you smoke then you must smoke outside; one can easily imagine a scenario in which you don't smoke and there is no such rule.

(W) is not much better. For one, in our Kripke semantics, $\mathsf{O}(p \to q)$ is entailed by $\mathsf{O}(\neg p)$: if p is false at all accessible worlds, then $p \to q$ is true at all these worlds. But it is easy to imagine a scenario in which you must not smoke, or you must submit your tax return before the deadline, and yet (1) and (2) are false.

Another problem with both (N) and (W) is that they would license a problematic form of "strengthening the antecedent". For example, they both suggest that (3) entails (4).

(4) If you have promised to call your parents and you know that someone has attached a bomb to your parents' phone that will go off if you call, then you must call them.

Exercise 6.12

Give a tree proof with the K-rules to show that $p \to \mathsf{O}\, r$ entails $(p \wedge q) \to \mathsf{O}\, r$, and that $\mathsf{O}(p \to r)$ entails $\mathsf{O}((p \wedge q) \to r)$.

Let's think about what is expressed by a statement like (1)–(4). Intuitively, when we ask what must be done if p is the case, we bracket the question whether p is really the case, or whether p ought to be the case. We are limiting our attention to situations in which p is the case, and consider which of these situations best conform to the relevant norms. Thus (1) says – roughly – that among worlds where you smoke, the "best" worlds are worlds where you smoke outside: worlds where you smoke inside are worse than worlds where you smoke outside. Similarly for (2). A world at which you miss the deadline for tax returns and pay the fine contains only one violation of the tax rules; worlds at which you miss the deadline and don't pay the fine contain two. So the "best" worlds among those at which you miss the deadline are worlds at which you pay the fine. And similarly for (3): among worlds at which you have promised to call your parents, the "best" are worlds at which you keep the promise and call them.

So the if-clause in sentences like (1)–(3) seems to restrict the worlds over which the modal operator quantifies. Whereas 'ought q' alone says that q is true at the best of the open worlds, 'if p then ought q' seems to say that q is true at the best of the open worlds *at which p is true*.

There is no way to express these truth-conditions with the resources of $\mathfrak{L}_M$. We will therefore introduce a new, binary operator for **conditional obligation**. The operator is often written '$\mathsf{O}(\cdot/\cdot)$', with a slash separating the two argument places. Intuitively, $\mathsf{O}(B/A)$ means that if condition A is satisfied then B ought to be the case.

Formally, when we evaluate $O(B/A)$, we simply add the assumption A to the circumstances that are held fixed:

$$M, w \models \mathsf{O}(B/A) \text{ iff } M, v \models B \text{ for all } v \in Min^{\prec_w}(\{u : wRv \text{ and } M, v \models A\}).$$

Exercise 6.13

Explain why the answers you gave to exercise 6.1.(d) and 6.1.(e) were inadequate.

Exercise 6.14

The dual of conditional obligation is conditional permission. Outline a semantics for $\mathsf{P}(B/A)$ that parallels the semantics I have outlined for $\mathsf{O}(B/A)$, so that $\mathsf{P}(B/A)$ is equivalent to $\neg\,\mathsf{O}(\neg B/A)$.

6.4 Further challenges

Many apparent problems for standard deontic logic arise from the dependence of obligations on (actual and hypothetical) circumstances. We can avoid these problems by using deontic ordering models and formalizing conditional obligation statements with the binary $\mathsf{O}(\cdot/\cdot)$ operator. But there are other problems for which this move doesn't help. I will mention three.

First, we already saw that standard deontic logic does not allow for conflicting obligations. Suppose you have promised your family to be home for dinner and your friends to join them at the pub. You are under conflicting *prima facie* obligations, and it is not clear that one of them overrides the other. Legal systems can also contain contradictory rules, without any higher-level rules for how to resolve such contradictions.

We can, of course, drop principle **D**. But even in the minimal logic K, $\mathsf{O}\,p$ and $\mathsf{O}\,\neg p$ entail $\mathsf{O}\,A$, for any sentence A. Intuitively, however, the fact that you have given incompatible promises does not entail that you are obligated to, say, kill the Prime Minister.

Another family of problems arises from the fact that in any logic defined in terms of Kripke semantics, O is closed under logical consequence: if $\mathsf{O}(A)$ is true and A entails B, then $\mathsf{O}(B)$ is true. Since logical truths are logically entailed by everything, it follows that all logical truths come out as obligatory: it ought to be that 2+2=4, or

that it either rains or doesn't rain. This is easy to see semantically. Any logical truth is true at all worlds; so it is also true at all deontically accessible worlds.

In response, one might argue that the relevant statements sound wrong not because they are false, but because their utterance would violate a **pragmatic** norm of cooperative communication. A basic norm of pragmatics is that utterances should make a helpful contribution to the relevant conversation. In a normal conversational context, it would be pointless to say that something ought (or ought not) to be the case if it is logically guaranteed to be the case anyway. An utterance of 'it ought to be that p' is pragmatically appropriate only if p could be false. This might explain why it sounds wrong to say that it ought to either rain or not rain.

Note also that by duality, $\neg \mathsf{O}(p \vee \neg p)$ entails $\mathsf{P} \neg(p \vee \neg p)$. So if we deny that it ought to either rain or not rain, and we accept the duality of obligation and permission, we have to say that it is allowed that it neither rains nor doesn't rain. That sounds even worse.

The problem of closure under entailment has special bite when obligation statements are restricted by circumstances. Return to the Samaritan puzzle. Suppose the victim is bleeding, and Jones ought to stop the blood flow. It is logically impossible to stop a blood flow if no blood is flowing. In all the deontic logics we have so far considered, the claim that Jones ought to stop the victim's blood flow therefore entails that the victim ought to be bleeding.

Here, too, one might appeal to a pragmatic explanation. When we say that Jones ought to stop the blood flow, we take for granted that the victim is bleeding. We are interested in what should be done *given* the state in which Jones found the victim. Worlds where the victim isn't injured are set aside; they are not circumstantially accessible. But circumstantial accessibility can shift with conversational context. The claim that the victim ought to be bleeding is pointless if we hold fixed the victim's state of injury. So when we evaluate this claim, we naturally assume that the relevant circumstantial accessibility relation does not hold fixed the injuries. Intuitively, we are no longer considering what should be done given the state in which Jones found the victim, but whether that state itself should have obtained. So worlds in which the state doesn't obtain become circumstantially accessible.

A third family of problems arises from disjunctive statements of permission and obligation. Consider (1).

(1) You ought to either mail the letter or burn it.

Intuitively, (1) suggests that both mailing the letter and burning it are permitted. In standard deontic logic, however, $\mathsf{O}(A \vee B)$ does not entail $\mathsf{P}\,A \wedge \mathsf{P}\,B$. (This puzzle was first noticed by Alf Ross and is known as "Ross's Paradox".)

A similar puzzle arises for permissions. (This one is known as the "Paradox of Free Choice".)

(2) You may have beer or wine.

Intuitively, (2) implies that beer and wine are both permitted. But in standard deontic logic, $\mathsf{P}(A \vee B)$ does not entail $\mathsf{P}\,A \wedge \mathsf{P}\,B$.

Can't we simply add the missing principles?

(**R**) $\quad \mathsf{O}(A \vee B) \rightarrow (\mathsf{P}\,A \wedge \mathsf{P}\,B)$

(**FC**) $\quad \mathsf{P}(A \vee B) \rightarrow (\mathsf{P}\,A \wedge \mathsf{P}\,B)$

No. Both of these have unacceptable consequences when added to the minimal modal logic K. With the help of **R**, we could show that $\mathsf{O}\,A$ entails $\mathsf{P}\,B$: $\mathsf{O}\,A$ entails $\mathsf{O}(A \vee B)$, which by **R** entails $\mathsf{P}\,B$. But clearly 'you ought to mail the letter' does not entail 'you may burn the letter'.

Similarly for **FC**. In K, $\mathsf{P}\,A$ entails $\mathsf{P}(A \vee B)$; by **FC**, $\mathsf{P}(A \vee B)$ entails $\mathsf{P}\,B$. But clearly 'you may have beer' does not entail 'you may have wine'.

Exercise 6.15

Analogous puzzles to those raised by Ross's Paradox and the Paradox of Free Choice arise for the epistemic 'must' and 'might'. Can you give examples?

6.5 Neighbourhood semantics

In reaction to the problems we have reviewed, some have argued that we should not interpret obligation and permission as quantifying over possible worlds. If we give up this core tenet of Kripke semantics, we can define "non-normal" logics weaker than K.

A popular alternative to Kripke semantic is **neighbourhood semantics** (also known as **Scott-Montague semantics**, after its inventors Dana Scott and Richard Montague).

Models in neighbourhood semantics still involve possible worlds. Validity is still defined as truth at all worlds in all (suitable) models. But the box and the diamond are no longer interpreted as quantifiers over accessible worlds. Instead, we simply assume that at every world, some propositions are "necessary" and others are not. $\Box A$ is true at a world if A expresses one of the necessary propositions at that world.

Formally, the accessibility relation in Kripke models is replaced by a **neighbourhood function** N that associates each world in a model with the propositions that are necessary relative to w. Propositions are identified with sets of possible worlds. So $N(w)$ is a set of sets of worlds. Each set of world in $N(w)$ is necessary at w.

Definition 6.2

A **neighbourhood model** consists of

- a non-empty set W,
- a function N that assigns to each member of W a set of subsets of W, and
- a function V that assigns to each sentence letter of $\mathfrak{L}_M$ and each member of W a truth-value.

The interpretation of non-modal sentences at neighbourhood models is just like in Kripke semantics (definition 3.2). To state the semantics for modal sentences, let $[A]^M$ be the set of worlds in model M at which A is true. This is our model proxy for the proposition expressed by A. Then:

$$M, w \models \Box A \text{ iff } [A]^M \text{ is in } N(w).$$
$$M, w \models \Diamond A \text{ iff } [\neg A]^M \text{ is not in } N(w).$$

Intuitively, the clause for the box says that $\Box A$ is true at w iff the proposition expressed by A is one of those that are necessary at w. The clause for the diamond ensures that the box and the diamond are duals.

In neighbourhood semantics, the modal operators are not closed under logical consequence. For example, the neighbourhood function N can easily make p necessary at a world without making $p \vee q$ necessary, even thought p entails $p \vee q$.

Exercise 6.16

What formal condition on the neighbourhood function would ensure that $\Box$ is closed under logical consequence?

If we interpret O and P as the box and the diamond in neighbourhood semantics, we can therefore say that Jones ought to tend to the victim's injuries even thought it is not the case that someone ought to be injured.

We can also allow for conflicting obligations. If the laws at w require both p and $\neg p$, we simply have $[p]^M \in N(w)$ and $[\neg p]^M \in N(w)$. It longer follows that any proposition whatsoever is obligatory.

We may also escape the problems from section 6.3 that led us to introduce a primitive conditional obligation operator. For example, I argued that the wide-scope translation $\mathsf{O}(A \to B)$ of conditional obligation sentences is problematic because $\mathsf{O}(A \to B)$ is entailed by $\mathsf{O}(\neg A)$. In neighbourhood semantics, this entailment fails.

Bare neighbourhood semantics determines a very weak logic. By imposing conditions on the neighbourhood function N, we can get a stronger logic, with more validities.

For example, suppose we want to maintain that if something is logically guaranteed to be true, then it can't be forbidden. Equivalently, any logically necessary truth should be permissible. By the neighbourhood semantics for P, A is permissible at a world w in a model M iff $[\neg A]^M$ is not in $N(w)$. If A is a logical truth, then A is true at all worlds; so $\neg A$ is true at no worlds, and $[\neg A]^M$ is the empty set. So if we want logical truths to be permissible, we have to stipulate that $N(w)$ never contains the empty set.

In Kripke semantics, the assumption that logically necessary truths are permissible is equivalent to the assumption that the **D**-schema $\mathsf{O}\,A \to \mathsf{P}\,A$ is valid. Both assumptions correspond to seriality of the accessibility relation. In neighbourhood semantics, we can distinguish between the two assumptions. While the permissibility of logical truths requires that $N(w)$ never contains the empty set, the validity of $\mathsf{O}\,A \to \mathsf{P}\,A$ requires that $N(w)$ never contains contradictory propositions $[A]^M$ and $[\neg A]^M$.

If we assume that the neighbourhood function is closed under intersection, in the sense that whenever two sets X and Y are in $N(w)$, then so is their intersection $X \cap Y$, then $(\Box A \wedge \Box B) \to \Box(A \wedge B)$ becomes valid. If we also require the converse,

that whenever $X \cap Y \in N(w)$ then $X \in N(w)$ and $Y \in N(w)$, and in addition that $W \in N(w)$, we get back the minimal normal logic K.

Exercise 6.17

Can you find a condition on the neighbourhood function that renders the **T**-schema valid?

For some purposes, even the minimal logic of neighbourhood semantics is too strong. For example, return to the intuitive "Free Choice" principle from the previous section:

(**FC**) $\quad \mathsf{P}(A \vee B) \rightarrow (\mathsf{P}\,A \wedge \mathsf{P}\,B)$

We have seen that this principle is untenable in Kripke semantics. As it turns out, it is still untenable in neighbourhood semantics.

To see why, note first that whenever two sentences A and B are logically equivalent, then in neighbourhood semantics $\mathsf{P}\,A$ and $\mathsf{P}\,B$ are also equivalent. The reason is that the modal operators in neighbourhood semantics operate on the set of worlds at which the embedded sentence is true. If A and B are logically equivalent, then in any model M, the set $[A]^M$ is the same set as $[B]^M$, and so $[A]^M \in N(w)$ iff $[B]^M \in N(w)$. Likewise, $[\neg A]^M \in N(w)$ iff $[\neg B]^M \in N(w)$.

Now any sentence A is logically equivalent to $(A \wedge B) \vee (A \wedge \neg B)$, for any B. In neighbourhood semantics, $\mathsf{P}\,A$ therefore entails $\mathsf{P}((A \wedge B) \vee (A \wedge \neg B))$. By **FC**, $\mathsf{P}((A \wedge B) \vee (A \wedge \neg B))$ entails $\mathsf{P}(A \wedge B)$. So by **FC**, we could still reason from 'you may have a cookie' to 'you may have a cookie and burn down the house'.

Exercise 6.18

Rational beliefs come in degrees, which are often assumed to satisfy the formal rules of probability. Suppose we say that someone believes A iff their degree of belief in A is above a certain threshold – say, 0.9. Explain why we can't give a Kripke semantics for this concept of belief. (Although one can give a neighbourhood semantics.) *Hint*: One rule of probability says that if p and q are independent propositions, then the probability of their conjunction $p \wedge q$ is the product of their individual probabilities.