# REASONING ABOUT OBLIGATIONS

## DEFEASIBILITY IN PREFERENCE-BASED DEONTIC LOGIC

Cover design: Mirjam Bode

The book is no. 140 of the Tinbergen Institute Research Series. This series is established through cooperation between Thesis Publishers and the Tinbergen Institute. A list of books which already appeared in the series can be found in the back.

# REASONING ABOUT OBLIGATIONS

## DEFEASIBILITY IN PREFERENCE-BASED DEONTIC LOGIC

(REDENEREN OVER VERPLICHTINGEN,
DEFEASIBILITY IN OP PREFERENTIES GEBASEERDE DEONTISCHE LOGICA)

PROEFSCHRIFT

TER VERKRIJGING VAN DE GRAAD VAN DOCTOR
AAN DE ERASMUS UNIVERSITEIT ROTTERDAM
OP GEZAG VAN DE RECTOR MAGNIFICUS
PROF. DR P.W.C. AKKERMANS, M.A.
EN VOLGENS BESLUIT VAN HET COLLEGE VOOR PROMOTIES

DE OPENBARE VERDEDIGING ZAL PLAATSVINDEN OP
27 FEBRUARI 1997 OM 13:30 UUR

DOOR

**LEENDERT WILLEM NICOLAAS VAN DER TORRE**

GEBOREN TE ROTTERDAM

PROMOTIECOMMISSIE

| | |
|---|---|
| PROMOTOR: | Prof. dr R.M. Lee |
| OVERIGE LEDEN: | Prof. dr A. de Bruin<br>Prof. dr J.-J.Ch. Meyer<br>Mr M. Sergot |
| CO-PROMOTOREN: | Dr J.C. Bioch<br>Dr Y.-H. Tan |

# Acknowledgements

Many people have contributed to the realization of this thesis.

First and foremost, this thesis owes much of its shape and content to Yao-Hua Tan. Many of the here presented results were obtained in collaboration with him. I am grateful to him for all the time and effort he put in the development of me and my work. I also thank Patrick van der Laag for contributing to the content of this thesis. He always found time to read what was written. Furthermore, I could always rely on his moral and social support. I thank Pedro Ramos for joint research reported in Section 5.3, John-Jules Meyer, Henry Prakken and Marek Sergot for several discussions on the issues raised in this thesis, and Wiebe van der Hoek for a discussion on modal logic. I thank Mark Polman for proof-reading this thesis.

During my stay at the Erasmus University, I was in the lucky position to have two offices, and the freedom to combine the best of both worlds. The first office was located at the Department of Computer Science of the Economic Faculty of the Erasmus University, and my second office was located at the Erasmus University Research Institute for Decision and Information Systems, EURIDIS. The department is an ideal environment for research on formal methods. and the institute is an ideal environment for application oriented research. I thank Cor Bioch, Arie de Bruin and Ron Lee for making this unique position possible. They always supported my joint research with 'the other side.'

Many people made my stay at Erasmus pleasant, interesting and rewarding. I will not forget the stimulating daily dinner discussions and the enthusiastic monthly drinks at the department. In particular, I thank my room mate Christ Leijtens, who was good company during and after working hours. Finally, I thank my friends and colleagues for several late-night discussions on interesting topics related to the topics of this thesis, such as the logic of legal reasoning and the existence or non-existence of obligatory violations.

Leon van der Torre
Rotterdam, January 1997

# Contents

# Chapter 1

# Deontic Logic

> *Dans la lutte antiterroriste, il y a des choses qui ne doivent pas se faire. Si elles se font, on ne doit pas les dire. Si elles se disent, il faut les nier.*[1] Jose Antonio Saenz de Santa Maria, Le Monde, September 9, 1995, p.3.

> *There should then be an obligation to make the best out of the sad circumstances*. Bengt Hansson [Han71]

In this thesis we investigate reasoning about obligations. In particular, we focus on *defeasible* deontic logics with *preference*-based semantics. In this chapter, we give a general introduction to deontic logic. We identify when deontic logic can be used as a knowledge representation language for computer applications. Moreover, we give a survey of deontic logic, as developed within philosophy. Finally, we give a personal perspective, the objectives of this thesis and its layout.

## 1.1 Defeasible deontic logic

There does not seem to be an agreement in deontic logic literature on the definition of 'defeasible deontic logic.' It is generally accepted that a defeasible deontic logic has to formalize reasoning about the following two issues.

1. **Resolving conflicts.** Defeasibility becomes relevant when there is a (potential) conflict between two obligations. In a defeasible deontic logic a conflict can be *resolved*, because one of the obligations overrides, in some sense, the other one.

2. **Diagnosing violations.** Consider the obligation 'normally, you should do $p$.' Now the problem is what to conclude about somebody who does not do $p$. Is this an exception to the normality claim, or is it a violation of the obligation to do $p$?

However, it is an open question whether defeasibility is related to the following issue.

[1] In fighting terrorism, there are things that should not happen. If they happen, then we must not tell them. If they are told, we must deny them.

3. **? Solving paradoxes.** Deontic logic has been supplied with a wealth of puzzles – usually referred to as *deontic paradoxes*. They are discussed in detail later in this chapter. A paradox is an intuitively consistent set of sentences that derives an inconsistency (or a counterintuitive sentence). Most of the deontic paradoxes can easily be solved, but some of them – most notably the Chisholm and Forrester paradoxes – still trouble deontic logicians. These troublesome paradoxes contain obligations conditional on a violation, which are called contrary-to-duty obligations. An example is 'you should not trade drugs, but if you trade drugs, then you should pay taxes for it.' The second sentence is a contrary-to-duty obligation, because its condition – trading drugs – is a violation (of the first sentence). This kind of sentences often have counterintuitive consequences.

There is a relation between defeasibility and solving paradoxes via an analogy between solving paradoxes and resolving conflicts. For example, in a survey on deontic logic Meyer and Wieringa interpret overriding to resolve conflicts as deleting (declaring out of force) obligations in order to restore consistency. This quote also illustrates that there has been a lively debate in deontic logic literature during the last five years whether deontic reasoning is a kind of defeasible reasoning.

> 'Treating contradictory norms by means of defeasible reasoning means, however, is totally different. It is more pragmatical in nature, ordering the various (contradictory) obligations or norms according to relevance / importance / priority, *deleting* (declaring out of force) least relevant ones to obtain a *consistent* set of norms. So one or more of the normative rules one was confronted with to begin with is not considered as relevant / applicable any more to the case (situation) of concern.
>
> Rather than judging this way of proceeding against other ones we consider this approach a very interesting one, leaving it to future developments as to how successful this will be; the proof of the pudding will be in the eating.' [MW93]

The problem of a deontic paradox is that it is inconsistent, whereas intuitively it is consistent. Hence, a pragmatic solution of the contrary-to-duty paradoxes can make use of restoring consistency techniques in case of a paradox. However, this solution is ad hoc. Restoring consistency is like treating symptoms without treating the disease. The term hack comes to mind! Moreover, defeasible deontic logic based on conflict resolution does not solve all deontic problems related to contrary-to-duty reasoning. This is probably best illustrated by Prakken and Sergot [PS96], when they discuss the so-called pragmatic oddity. We discuss this pragmatic oddity in a modal language in which $Op$ stands for the obligation to do $p$. The details of the modal logic are explained later in this chapter. Prakken and Sergot consider the following three sentences of a modal theory:

1. $Ok$: You should keep your promise.
2. $\neg k \rightarrow Oa$: If you have not kept your promise, you should apologize.
3. $\neg k$: You have not kept your promise.

The second sentence is a contrary-to-duty obligation, because its condition – not keeping your promise – is a violation (of the first sentence). The problem is that $Ok \wedge Oa$ can be derived from

these three sentences, from which $O(k \wedge a)$ can be derived in most deontic logics (as is explained later). Prakken and Sergot remark 'but it is a bit odd to say that in all ideal versions of this world you keep your promise and you apologize for not keeping it. This oddity – we might call it the 'pragmatic oddity' – seems to be absent from the natural language version, which means that the representation is not fully adequate.' In our opinion, the sentence 'you ought to keep your promise and apologize for not keeping it' $O(k \wedge a)$ is paradoxical and should not be derived by a deontic logic. As a consequence, conflict resolution does not seem the right metaphor to approach contrary-to-duty reasoning, because it does not solve the pragmatic oddity.

A second open question about defeasible deontic logic is whether it has to formalize reasoning about the following issue.

4. **? Deontic choice.** According to the concept of deontic choice, an obligation $Op$ compares situations in which $p$ is true (is done) with situations in which $p$ is false (is not done). If an agent can choose between a situation in which $p$ is true and a situation in which $p$ is false, then she has to choose the former.

The relation between defeasibility and deontic choice is as follows. The concept of deontic choice gives rise to a preference relation, i.e. a binary relation that represents different degrees of ideality. Jennings observes that the preference-based semantics is a kind of utilitarian semantics [Jen74] and that preferences are related to a kind of defeasibility (non-monotonicity) [Jen85]. At this moment, we do not have the tools to discuss this issue. It is one of the central issues in this thesis and discussed in detail in Chapter 4. For clarity, in the following we use the term 'preference-based deontic logics' for deontic logics based on deontic choice, and we leave the term 'defeasible deontic logic' for deontic logics with a conflict resolution mechanism.

There is a second relation between defeasibility and solving paradoxes via deontic choice. Deontic choice is a solution for the contrary-to-duty paradoxes, as is explained in detail later in this thesis. Moreover, preference-based deontic logics do not have the pragmatic oddity. In contrast to the conflict resolution approach, the preference-based approach is not ad hoc, but it is based on the intuitive semantic notion of deontic choice. In this thesis we study the relation between obligations, preferences and defeasibility.

### 1.1.1 Obligations, preferences and defeasibility

In the remainder of this introduction, we discuss two relations between obligations, preferences and defeasibility. First, deontic logic and default logic can both be preference-based logics. Second, defeasible deontic logic can combine deontic preferences and default preferences in a multi preference semantics. We start with the use of preferences in deontic logic and default logic.

Deontic logic formalizes reasoning about obligations. It is based on the fundamental distinction between what is ideally the case on the one hand and what is actually the case on the other hand. Agents should try to reach the ideal. If an agent wants to know what she should do, then she only has to inspect the ideal. Unfortunately, sometimes obligations are violated. The ideal is no longer reachable, because the actual starts to deviate from the ideal as a result of a violation. If an agent wants to know what she should do, then it does not make sense anymore to inspect the ideal. Instead, she considers a state that approximates the ideal as much as possible, and that is still possible given the violation. That is, she considers an optimal state. Contrary-to-duty

obligations tell what is obligatory, given the sad circumstances that obligations have been violated. Preference-based deontic logics are developed to formalize contrary-to-duty reasoning. The preferences represent different degrees of ideality. Preferences can be represented explicitly in the language of the logic (the deontic logic is defined in a so-called preference logic), but usually the preference ordering is only part of the semantics. In this thesis we introduce several preference-based deontic logics.

Default logic (also called defeasible logics or logics of defeasible reasoning) formalizes reasoning about default assumptions. It is based on the distinction between what normally is the case on the one hand and what is actually the case on the other hand. In such a logic, conclusions can be defeated. This defeasibility is usually formalized in a non-monotonic logic.[2] For example, consider the default that birds normally fly. If we know that a certain animal is a bird, then we can infer by default that it can fly. However, if we get to know that this bird is a penguin, then the conclusion that it can fly is retracted. That is, the conclusion that it can fly is defeated. In a non-monotonic logic, the conclusion that the bird can fly is no longer derivable. In a preference-based default logic, the preferences represent different degrees of normality. In the most normal case, birds fly. However, penguins are abnormal birds (as far as flying is regarded). If we know that the bird is a penguin, then we know that the situation is exceptional. The most normal of these exceptional cases is that the penguin cannot fly.

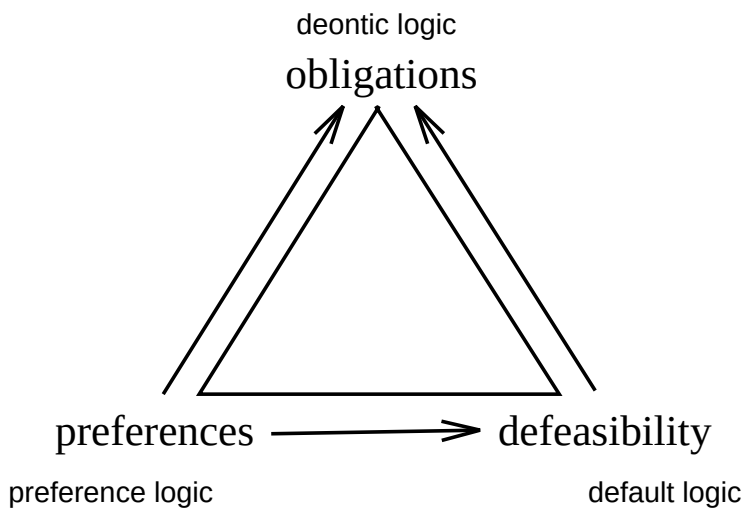



Figure 1.1: Obligations, preferences and defeasibility

The first relation between obligations, preferences and defeasibility is represented in Figure 1.1. Deontic logic and default logic can both be preference-based logics. In particular, there is an analogy between the treatment of violations in preference-based deontic logics and the treatment of exceptions in preference-based default logics. Given this analogy, it is no surprise that most preference-based deontic logics are defeasible. However, *a priori* this is quite remarkable. A violation makes the ideal unreachable, but a violated obligation is still in force. The obligation is not cancelled. It is only no longer a cue for action. In this thesis, we study the

[2] Another possibility to formalize this defeasibility is a conditional logic in which the conditionals do not have strengthening of the antecedent. This is explained later in this thesis.

source of the defeasibility in preference-based deontic logic, and we study the similarities and the distinctions between the defeasibility in deontic logic and default logic.

The second relation between obligations, preferences and defeasibility is defeasible deontic logic, i.e. deontic logic with a conflict resolution mechanism. Defeasible deontic logic can combine deontic preferences and default preferences in a multi preference semantics. There is a substantial overlap between deontic and defeasibility aspects. As a consequence, the diagnosis of violations has to distinguish carefully between exceptions and violations. In this thesis we analyze this overlap, and we also show that this confusion between exceptions and violations can be avoided if one makes the proper distinctions between different types of defeasibility.

The layout of this chapter is as follows. In Section 1.2 we discuss *when* deontic logic can be used as a knowledge representation language, and *which type* of deontic logics should be used. In Section 1.3 we give a survey of deontic logic as developed within philosophy. In Section 1.4 we give a personal perspective on deontic logic. In Section 1.5 we present the objectives of this thesis and in Section 1.6 we present the layout of the rest of this thesis.

## 1.2 Computer applications

Deontic logic might be a useful knowledge representation language if (and only if) the modeler wants to represent violations and contrary-to-duty obligations. In this section we discuss several examples of the use of deontic logic from deontic logic literature, and we discuss an example from the international trade domain.

### 1.2.1 Knowledge representation language

Jones and Sergot [JS92, JS93] argue extensively and convincingly that deontic logic is a useful knowledge representation language when the modeler wants to formalize reasoning about violations and obligations that arise as a result of these violations, the contrary-to-duty obligations. McCarty [McC94b] observes that 'one of the main features of deontic logic is the fact that actors do not always obey the law. Indeed, it is precisely when a forbidden act occurs, or an obligatory action does not occur, that we need the machinery of deontic logic, to detect a violation and to take appropriate action.' In this section we give three examples of deontic logic literature: the United Nations Convention on Contracts, the Imperial College Library Regulations and the Cottage Regulations. These examples are very small – we can consider them as *toy* examples – but they neatly illustrate the contrary-to-duty reasoning in deontic logic applications. Before we consider situations in which deontic logic might be used, we warn against a naive use of deontic logic. In particular, the use of 'ought' in a text is not a reason to use deontic logic as a knowledge representation language. Susskind [Sus87] argues that the fact that typically normative vocabulary such as must, ought, may shall are found in law-formulations necessitates the inclusion of some deontic logic within a legal inference engine. This claim is disputed by Bench-Capon [BC89], who points out that words like must, shall and ought do not always signal deontic modalities. Jones and Sergot [Jon90, JS92, JS93] discuss this issue and observe that in general, natural language is a bad guide for formal representations: 'if we must over-simplify, then our preference is to assert bluntly that surface is *never* a faithful guide to

context.'

Since deontic logic traditionally has been used to analyze the structure of normative law and normative reasoning in law, it is only natural that interest in applications in deontic logic started in the area of legal applications. For example, Smith [Smi94] notices several contrary-to-duty constructions in legal reasoning (which hints at the use of deontic logic to formalize legal reasoning).

> "In the process of leafing through law-books, law-reports and legal literature it becomes apparent that contrary-to-duty constructions are by no means exceptional:[3] the law is not only concerned with describing the deontically perfect world, but also with what should be done in case norms are not complied with (for whatever reason). Very often, breaking obligations makes the agent liable to punishment, damages and the like, and/or gives other persons the right to see to it that the desired situation arises – sometimes at the expense of the person who was responsible for it." [Smi94]

However, later she notices: 'It turned out to be rather difficult to find clearly formulated examples of genuine contrary-to-duty obligations, i.e. things that the agent must do if (or as long as) his primary obligation remains unfulfilled. The examples that I have found come down to: making it known (to the parties concerned) that the primary obligation is not fulfilled.' The apparent contradiction (contrary-to-duty constructions are by no means exceptional, but clearly formulated examples are rather difficult to find) can also be explained by the fact that natural language is a bad guide to context. The surface structure of law texts does not show many contrary-to-duty structures, but the deep structure of law does. Smith's first example is from the United Nations Convention on Contracts for the International Sale of Goods.[4]

**Example 1.1 (Convention on Contracts)** [Smi94, p.127] Section 79 subsection 4 reads as follows: "The party who fails to perform must give notice to the other party of the impediment and its effect on his ability to perform. If the notice is not received by the other party within a reasonable time after the party who fails to perform knew or ought to have known of the impediment, he is liable for damages resulting from such non-receipt." Here we have a double contrary-to-duty construction: first a contrary-to-duty obligation (to give notice), and then a prevision of what the consequences will be if that contrary-to-duty obligation remains unfulfilled (liability for damages). □

A second topic identified – besides legal knowledge-based systems – where deontic logic can be used is the specification of fault tolerant systems. A benchmark example of deontic logic is the following Library Regulations example of Jones and Sergot [JS93]. They argue that deontic logic can be used for (normative) system specification. In particular, Jones and Sergot identify three situations in which it might be useful to formalize contrary-to-duty reasoning.

**Example 1.2 (Library regulations)** [JS93] Consider the following rules of the Imperial College Library Regulations.

[3] See [Jon90]. In every-day cases, the presence of a contrary-to-duty provision is sometimes felt to be essential even for the existence of a primary norm: if nothing happens after a norm violation, that is sometimes taken to mean that apparently there is no norm.

[4] The same convention is quoted by [Jon90].

1. A separate form must be completed by the borrower for each volume borrowed.

2. Books should be returned by the date due.

3. Borrowers must not exceed their allowance of books on loan at any one time. Allowances: undergraduates 6, postgraduates 10, academic staff 20.

4. No book will be issued to borrowers who have books overdue for return to the library.

The rules of the library regulations can be interpreted from several perspectives. For example, they contain obligations for users like 'if you borrow a book, then you ought to complete a separate form.' Moreover, they also contain obligations for the system designer: 'it ought to be that the system does not issue a book to borrowers that do not complete a separate form for each volume borrowed.' It is the latter perspective Jones and Sergot are interested in.

We now consider Jones and Sergot's discussion whether deontic logic can be used for system specification. Imagine the following scenario. The chief librarian wishes to improve the efficiency of the library, and asks us to develop a computer representation of the library regulations. Jones and Sergot distinguish the following two representations of the regulations.

1. He want us to develop a system that advises on the obligations and rights of the various users of the library as it currently exists.

2. He wants us to take the regulations as a specification of how the library ought to function, giving us the task of developing a computer system which automates the library, at least as regards the issueing and returning of books.

In the latter case, the chief librarian wants us (the system engineers) to introduce a system of computers *as a specification of how the system should operate*. We may call this the *system specification* scenario. The task here facing the system designer is to develop a library system which actually behaves in the way the library regulations require. Taking the chief librarian at his word, we might consider how we can *force* actuality and ideality to coincide in this example. However, such a so-called 'regimentation' of the regulation in the system is quite inflexible. Jones and Sergot see the following three roles for deontic logic in system specification.

1. We might require a formal language in which to express precisely the specification of an organization (here the library together with its computer and other administrative procedures). Such a specification must usually make provision for the possibility of violation, where actual behavior deviates from the ideal, and for this a deontic logic is necessary.

2. We might require in addition a formal language in which to specify precisely the intended operation of a computer system. A deontic logic is necessary for specifying computer systems if we want to make provision for violations – whether resulting from faulty components or from extraneous factors. We should like to be able to reason with these specifications, for example to test the internal consistency of the specification, or to determine whether one is a logical consequence of another.

3. We might wish to use an automated theorem prover for a deontic logic as a means of *implementing* some of the software components of a computer system.

The Library Regulations Example illustrates that deontic logic can be used if the formal language has to be able to express violations and the associated contrary-to-duty structures. □

The third example we discuss is the following Cottage Regulations Example of Prakken and Sergot [PS96]. The sentences are, according to Prakken and Sergot, genuine. The example illustrates the occurrence of contrary-to-duty structures in *defeasible* deontic reasoning.

**Example 1.3 (Cottage regulations)** [PS96] The following three sentences come from a set of formal and informal regulations governing the appearance and use of holiday cottages. Both of the regulations (1) and (2) are intended to hold at one and the same time, and (2) is intended as a contrary-to-duty rule for (1) rather than some kind of exception. In contrast, (3) is an exception to (1). If someone has a fence because the cottage is by the sea, then (1) has certainly not been violated.

1. There must be no fence.
2. If there is a fence, then it must be a white fence.
3. If the cottage is by the sea, then there may be a fence.

Two important features of the example are that it concerns states of affairs instead of actions and that all three statements pertain to the same point in time; this makes proposals based on action logics or temporal logics inapplicable. □

In this section we discussed the fact that deontic logic can be used as a knowledge representation language if (and only if) the modeler wants to formalize violations and the obligations that arise as a result of violations, the so-called contrary-to-duty obligations. We gave three examples from deontic logic literature that illustrate the occurrence of contrary-to-duty obligations. Besides legal knowledge based systems and the specification of fault tolerant systems, topics identified in deontic logic literature are the specification of security policies, the automatization of contracting and the specification of normative integrity constraints for databases [WM93]. We discuss some new applications, (robot) planning and diagnosis, in Chapter 5. In the following section we discuss an example of the international trade domain.

### 1.2.2 Trade procedures

> *You cannot trust people, only the deals they make. People are only as good as the deals they make and keep.* The unbelievable truth – Hal Hartley

The research institute EURIDIS is located at Rotterdam, the (at least in some respects) largest port of the world. One of the interests of EURIDIS is modeling inter-organizational trade procedures, with the long term goal to facilitate electronic commerce. For example, in [BLWW95] it is observed that the introduction of EDI can have tremendous benefits for the efficiency of the execution of trade procedures, both among and within organizations. The most obvious benefit is the reduction of time needed for the execution of the transaction. It is now possible to replace many paper documents with electronic equivalents, particularly since standards for the structure

of the messages have matured. Regarding the benefits, it could be expected that many organizations would be eager to start with EDI implementation. However, it is observed in [BLWW95] that successful EDI implementations have mainly been realized in trading relationships with frequent transactions, mostly over a longer period of time. EDI linkages are seldom observed when the partnership is established for a limited period, covering a few transactions only, since the cost of the necessary negotiations cannot be recovered from EDI efficiency gains. These shorter term partnerships are called 'electronic market relationships'. The introduction of electronic market relationships can be facilitated by decreasing the set-up costs for EDI linkages. In particular, they have to agree on the *trade procedure* (also called the trade scenario, business scenario, or business protocol) they are going to follow. A trade procedure is a mutually agreed upon set of rules that governs the activities of all parties involved in a set of related business transactions [BLWW95]. The following example illustrates a simple trade procedure.

**Example 1.4 (Trade procedure)** Consider a simple trade procedure with two agents, a buyer and a seller, represented in Figure 1.2. The seller makes an offer, which the buyer can accept. If the offer is accepted and everything goes well, then the seller delivers some goods, and when the goods are delivered the buyer pays for the goods. This is the ideal behavior of the agents. Obviously, other scenarios are possible, for example in which the buyer makes the offer, or the buyer has to pay before or at the same time as the goods are delivered. Moreover, the protocol can be described in more detail, if necessary for the analysis. □

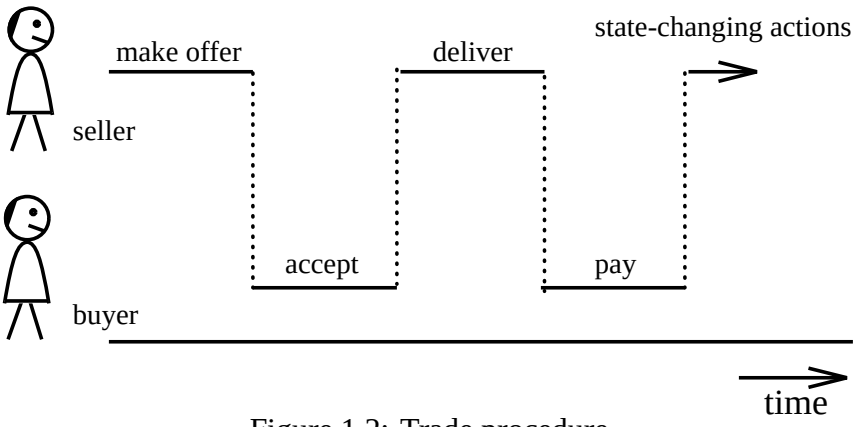



Figure 1.2: Trade procedure

The set-up costs of EDI linkages can be decreased by facilitating the agreement on the trade procedure [BLWW95]. Automated support for setting up the linkages depends crucially on the possibility to reason about these norms. At EURIDIS, trade procedures are modeled in Petri nets. Petri nets [Pet81] are a popular formalism for the modeling and analysis of discrete dynamic (distributed) systems, because they combine the advantages of a graphical representation with the expressive power of parallelism and synchronization. One of the application domains is the modeling of procedures and processes within and between organizations. For example, Van der Aalst [vdA92] developed the ITCPN model to model logistic processes in organizations, and Lee [Lee91, Lee92] developed the CASE/EDI tool to model bureaucratic procedures. The latter tool can be used to dynamically simulate Petri nets (scenario analysis) and check the procedures represented in a Petri net for consistency and dead-locks. It has been applied to model

inter-organizational procedures in international trade, like contract negotiation, the exchange of bill-of-lading for letter-of-credit and custom clearance [BLWW95].

A drawback of the representation of the trade procedure in Figure 1.2, as well as representations in Petri nets [RTvdT96] is that they only model the ideal behavior. It is important that violations of obligations, i.e. sub-ideal states, are represented explicitly in the modeling of procedures, because in most procedures it is described explicitly what is considered as ill-behavior, and how this will be punished (the corresponding sanction). Since this violation behavior is described explicitly, it should also be represented explicitly. Representing ideal behavior does not make sense if there is no way to represent sub-ideal behavior, just as the notion of master does not have a meaning without there being a slave. Hence, both ideal and sub-ideal behaviors must be represented and distinguished from each other for modeling procedures. As a consequence, deontic logic is a candidate to formalize the normative aspects of the procedures, because deontic logic can be used as a knowledge representation language if the modeler wants to model violations and the new obligations that arise as a consequence of violations, the contrary-to-duty obligations, as we we discussed in Section 1.2.1. The following example illustrates that the procedures can be modeled as a set of norms, i.e. as a set of conditional obligations.

**Example 1.5 (Trade procedure, continued)** If the offer is accepted, then there are several conditional obligations, according to the protocol. For example, the seller is obliged to deliver the goods, and the buyer is obliged to pay for the goods, when the seller has delivered them. Moreover, there are (conditional) obligations when no offer has been accepted yet. For example, if the seller makes an offer, then she is obliged to deliver the goods, if a buyer accepts the offer. □

Finally, we discuss some desirable properties of a deontic logic that can be used for the formalization of the trade procedures. In Section 1.2.1 we discussed that the minimal requirement to use deontic logic is that the modeler wants to model violations and the contrary-to-duty obligations. Additional requirements of a deontic logic are that it can model exceptions of norms, that it can discriminate between different agents and that it can model that obligations vary in time.

1. **Violations and contrary-to-duty obligations.** A deontic logic has to be suitable to represent violations and formalize contrary-to-duty reasoning. For example, it is important that violations of obligations are represented explicitly in the modeling of trade procedures, because in most procedures it is described explicitly what is considered as ill-behavior, and how this will be punished.

2. **Exceptions of norms.** Knowledge is usually represented by general rules and exceptions to these general rules. For example, consider the following rule of a trade procedure: *'Normally goods have to be paid on delivery, but regular customers may have some credit.'* If the formalization follows this scheme, then we need defeasibility of a general rule (goods must be paid on delivery), in case of exceptional circumstances (regular customer). It is very important that the language is able to distinguish between exceptions and violations. The problem is what to conclude about somebody who does not pay on delivery. Is this an exception to the normality claim, or is it a violation of the obligation to pay at delivery?

3. **Agents.** Usually, there are several different agents that each have their own set of obligations and rights. For example, in a trade procedure there are, besides the buyer and seller

in Example 1.4, also forwarders, banks, customs, etc. As observed in e.g. [AB81, Alc93, Roy96], obligations expressed by a modal operator $O$ can easily be relativized to particular agents and normative systems by $O^s_a$, where $a$ is the agent and $s$ the normative system (sometimes called the authority). Moreover, an obligation of an agent $a_1$ is often directed towards a specific agent $a_2$, in the sense that if the agent $a_1$ violates the obligation, then the agent $a_2$ has a claim towards her. This can be represented by $O^s_{a1,a2}$. Alternatively, the agents can be formalized as an element of the $\alpha$ of $O\alpha$, see e.g. [HB95, Hor96]. Although the agents are easily introduced, we do not suggest that the formalization of the agents themselves and the formalization of reasoning with obligations of different agents is easy! On the contrary, we think that reasoning with obligations in a multi agent system is one of the most interesting challenges of the formalization of normative reasoning.

4. **Time.** Obligations change in the course of time; thus the relation between time and obligations has to be expressed. Obligations can be relativized to a certain moment (or interval) $t$ in time by $O_t\alpha$. For example, certain obligations have to be fulfilled before a certain moment in time, the so-called dead-line. Moreover, in time obligations are created and destructed. For example, in a trade procedure obligations are created by making an offer, accepting an offer and other acts. Thus, related to the interaction of obligations and time is the representation of actions with their causal structure.

In this thesis we only investigate the first two items. We do not investigate the relation between deontic logic, agents and time. There is some recent research in the area of obligations and agents. For example, directed obligations $O^s_{a1,a2}$ can be modeled in the logic of Krogh and Herrestad [KH96]. Deontic statements invariably suppress explicit references to time, strongly suggesting that temporal information is redundant, namely, it can be reconstructed if required, but glossed over otherwise [Pea93]. This is in contrast to theories of action, that are normally formulated as theories of temporal changes [Sho88, DK89]. Actions and causal structure have been investigated in Pearl's logic of pragmatic obligation [Pea93], see also Section 1.3.7 and 5.2. However, most problems in this area of agents, time and actions are not related to the deontic interpretation of the modal operator. For example, the notorious frame problem of temporal logic also occurs in the deontic setting: when does an obligation $O\alpha$ persist in time?

In this section we discussed desiderata for a deontic logic that can be used as a knowledge representation language to formalize trade procedures. It is an example of the application of logical tools to the task of modeling aspects of commercial activity, including communication about contracts. This is a domain in which a great deal of work remains to be done, and where the rapid development of electronic commerce is creating an urgent need for precise formal methods for modeling the processes involved. In the following section we discuss the philosophical foundations of deontic logic, and we focus on the problems caused by contrary-to-duty obligations.

## 1.3 Philosophical foundations

We start with a few observations from Føllesdal and Hilpinen [FH71], who give an introduction on the area of deontic logic, as developed within philosophical logic. Ernst Mally [Mal26] was the first to use the term *Deontik* to refer to the logical study of the normative use of language.

Normative expressions include the words ‘obligation’, ‘duty’, ‘permission’, ‘right’ and the related expressions. These expressions may be termed *deontic words*, and sentences involving them *deontic sentences*. A deontic sentence is a truth of deontic logic if it is true and remains true for all variations of its non-logical and non-deontic words (that is, expressions which are not logical or deontic words). Deontic logic is closely related to the logic of imperatives (or the logic of commands); in fact, many authors regard these fields as essentially the same. Bengt Hansson [Han71] observes that a deontic statement is not simply an imperative, because ‘one can point out to a person that he ought to do so-and-so without actually telling him to do it.’ What is here called deontic logic has also been referred to as the *logic of obligation* and *logic of norms* (or *logic of normative systems*). One may say that Deontic Logic came into existence in 1951 with the publication of G.H. von Wright’s paper ‘*Deontic Logic*’ [vW51] that appeared in Mind (see [FH71, Knu81] for a discussion on earlier proposals). In that paper, von Wright presented the first viable system of deontic logic. Most of the discussion of deontic logic after 1951 has been inspired – directly or indirectly – by von Wright’s article.

Two characteristic properties of obligations have been identified. The first property can be paraphrased as *what is obligatory is thereby permitted*. The intuition for this principle is that you cannot oblige someone to do something, without giving him at the same time a permission to do it. In a formal language, this property is expressed as $O\alpha \rightarrow P\alpha$, to be read as ‘if $\alpha$ is obligatory, then $\alpha$ is permitted.’

The second property of obligations goes back to Kant’s famous dictum ‘ought implies can’ or in the original German terms as the ‘sollen-können’ principle, see for example [vW71b, p122-p125]. This concept brings us closer to traditional discussions of moral philosophy. As the name indicates, the question here is whether each obligation presupposes a possibility of fulfilling it. The related second deontic axiom is $\neg O\bot$, to be read as ‘the impossible is not obligatory’ ($\bot$ is a symbol that stands for a contradiction like $p \wedge \neg p$). Alternatively, if we express the concept of possibility by the modal operator $M$, then the second deontic axiom can be written as the formula $O\alpha \rightarrow M\alpha$.

Besides these two usually[5] undisputed properties, there is a third property that has been defended by some authors, see e.g. [Con82, Pra96]: the absence of conflicting obligations. We call it the no-dilemma assumption. The related axiom is $\neg(O\alpha \wedge O\neg\alpha)$. Von Wright [vW81, p.5] observes an analogue with Bentham’s logic of the Will. Bentham regarded it as a law of his Logic of the Will that if something is obligatory (Bentham says ‘commanded’) then it is not also prohibited.

Besides these properties, the area of deontic logic is characterized by lack of consensus. The philosophers encountered several problems, which we briefly discuss in the following subsection. In the remainder of this section on the philosophical foundations of deontic logic, we give a crash course in deontic logic by discussing the main formal systems.

[5] Hintikka [Hin71, p.83] argues that the discussions of the ‘ought implies can’ problem one can find in the literature can scarcely to be said to have resulted in any kind of consensus. He therefore concludes that, in the context of a logical discussion, ‘it therefore seems advisable not to try to salvage the ‘ought implies can’ principle by means of additional assumptions.’

### 1.3.1 Problems

In this section we discuss several philosophical problems: whether norms have truth values, whether deontic logic should be developed as a branch of modal logic, what kind of entities the operators operate on, and the deontic paradoxes. In this thesis, in which we study deontic logic from a knowledge representation perspective, the first three problems do not play an important role. The deontic paradoxes, however, are used extensively in this thesis to analyze the properties of the logics that are developed.

The first problem philosophers encountered was the question whether norms have *truth values*. For example, von Wright [vW81] feels a certain hesitation to call deontic formulas 'logical truths' at all, because 'it seems to be a matter of extra-logical decision when we shall say that "there are" or "are not" such and such norms.' We restrict ourselves to quotes from respectively Alchourrón and Bulygin, von Wright and B. Hansson, because this philosophic problem is somewhat hard to follow for non-philosophers. First, Alchourrón and Bulygin [AB81] make a distinction between norms and normative propositions.

> 'For the *hyletic conception* norms are proposition-like entities, i.e. meanings of certain expressions, called normative sentences. [ . . . ] In this conception, norms are not language-dependent; they can only be expressed by linguistic means, but their existence is independent of any linguistic expression. [ . . . ] Norms must be distinguished from *normative propositions*, i.e. descriptive propositions stating that $p$ is obligatory according to some unspecified norm or set of norms. For the *expressive conception*, instead, norms are the result of the prescriptive use of language. [ . . . ] It is only on the pragmatic level of the use of language where the difference between statements, questions, commands etc. arises: [ . . . ] The expression $\vdash p$ indicates that $p$ is asserted and $!p$ indicates that $p$ is commanded, whereas $Op$ expresses a proposition that $p$ ought to be (done). So $Op$ is the symbol for a norm in the hyletic conception whereas '!p' symbolizes a norm in the expressive conception.'

Alchourrón [Alc93] further explains the distinction between norms and normative propositions with a box metaphor.

> 'We may depict the difference between the descriptive meaning (normative propositions) and the prescriptive meaning (norm) of deontic sentences by means of thinking the obligatory sets as well as the permitted sets as different boxes ready to be filled. When the authority $\alpha$ uses a deontic sentence prescriptively to norm an action, his activity belongs to the same category as *putting something into a box*. When $\alpha$, or someone else, uses the deontic sentence descriptively his activity belongs to the same category as *making a picture of* $\alpha$ *putting something into a box*. A proposition is like a picture of reality, so to assert a proposition is like making a picture of reality. On the other hand to issue (enact) a norm is like putting something in a box. It is a way of creating something, of building a part of reality (the normative qualification of an action) with the purpose that the addressees have the option to perform the authorized actions while performing the commanded actions.'

Von Wright [vW81] makes a distinction between norm-formulations and norm-propositions, that is analogous to Alchourrón's distinction:

> 'Normative sentences will be called *norm-formulations*. A characteristic use of them is for giving (issuing, laying down) norms or rules for human agents. When this use is in question, the normative sentences may be said to *express norms*. Normative sentences, however, can also be used for making statements to the effect that there are (have been given or issued) such and such norms or rules. When used in this way, normative sentences express what I propose to call *norm-propositions*.'

Bengt Hansson [Han71] has a pragmatic approach to the problem.

> 'I will take here the view that deontic statements are descriptive, that they describe what is obligatory, forbidden and permitted respectively, according to some (undetermined) system of norms or moral or legal theory. [ . . . ] This descriptive interpretation has one advantage in addition to making formulas into propositions; it makes clear that deontic logic is a tool of meta-ethics and not part of ethics proper.'

The second problem encountered was whether deontic logic should be developed as *a branch of modal logic*. This issue has been discussed extensively in many papers by von Wright, here we repeat the arguments of [vW71a]. Deontic logic was, in origin, an off-shoot of modal logic. It got its decisive impetus from observations of some obvious analogies between the modal notions of normative ideas of necessity, possibility, and impossibility on the one hand and the deontic or normative ideas of obligation, permission and prohibition on the other hand. Von Wright observes that besides the analogies and similarities, however, there is also a number of striking dissimilarities between the two types of modalities, and moreover, that many of the problems which have beset deontic logic since its birth are related to these discrepancies.

1. One difference is the absence in deontic logic of an analogue to the principle $Np \to p$ of modal logic. That what necessarily is the case is also as a matter of fact the case; but that which ought to be the case is far from being always actually the case.

2. The second formal difference between modal and deontic logic observed by von Wright is that, whereas it is obvious that the tautology necessarily is true ($N\top$), it is not intuitively clear that the tautology also ought to be true ($O\top$). The idea of $O\top$ does not seem to make good sense. For example, it excludes the possibility of an empty norm system. Von Wright shows that we can get rid of the unwanted result by adding a so-called contingency clause to the definition of an obligation.[6] If we express the concept of possibility with the modal operator $M$, then the contingency clause of an obligation for $p$ is $Mp \wedge M\neg p$, such that $Op$ implies $Mp$ and $M\neg p$ (hence, we have $Op \to (Mp \wedge M\neg p)$ as a theorem).

3. Finally, von Wright observes a further noteworthy difference between the two logics. In modal logic, the interdefinability of the two ideas of necessity and possibility through the schema $Mp =_{def} \neg N\neg p$ provokes no serious objection. But the corresponding schema or equivalence in deontic logic $Pp =_{def} \neg O\neg p$ is by no means unproblematic. For example, a consequence of the definition is the counterintuitive theorem $Op \vee P\neg p$, which is discussed in the next section. It seems feasible to admit a 'weak' notion of permittedness, according

[6] The consistency clause was introduced in a deontic logic based on necessary and sufficient conditions. Stelzner [Ste92] recently proposed a formalization based on relevance. Unfortunately, these approaches are problematic, because it is difficult to formalize concepts like 'sufficient condition' and 'relevance'.

> to which something *may* be, if and only if it is *not* the case that the contradictory of this thing *ought* to be.[7]

Despite von Wright's observations, the modal approach to deontic logic has become the standard approach to deontic logic. This approach is discussed in detail in the following section.

The third problem the philosophers encountered – related to the second one – is what kind of entities the deontic operators operate on. In von Wright's [vW51], deontic operators are applied to names of *acts*, not to descriptions of states-of-affairs. Thus, the system of propositional logic which constitutes the basis of von Wright's system is not, strictly speaking, a logic of propositions, but a logic of act-names. In this logic, the notion of truth-value is replaced by the notion of performance-value: a proposition can be performed or not performed. Later, von Wright [vW71b, p.106] no longer regards the reading of $O\alpha$ as *one ought to do* $\alpha$ as 'fully correct,' and reads the formula $O\alpha$ as follows: *one ought to see to it that* $\alpha$. Hence, the relation between the operator $O$ and $\alpha$ in $O\alpha$ is either operator and act-type, or operator and proposition. This distinction is analogous to the distinction between the relation between predicate and term in predicate logic and between modal operator and propositional sentence in modal logic. An advantage of the latter reading is that the language contains mixed formula. For example, the mixed formula $\alpha \wedge O\neg\alpha$ can represent a violation.[8]

Since the publication of von Wright's *'Deontic Logic'*, the discussion of the subject and the proliferation of different deontic logics is almost incredible [Alc93]. There is no principle or feature of von Wright's original system which has not been the object of the strongest criticism. Even von Wright himself has advanced serious objections to each of his original statements and has produced many different deontic logical systems in order to overcome the supposed deficiencies of his first ideas. Alchourrón [Alc93] observes that the source of the doubts lay in the difficulties involved in the process of finding intuitive correlates to his principles in the highly ambiguous uses of deontic and related sentences (such as imperatives) in everyday discourse and in more sophisticated (legal and moral) contexts.

The criticism of principles or features of deontic logic has been formulated in terms of deontic puzzles. Consider the following quote of Bertrand Russell in '*On Denoting*'.

> 'A logical theory may be tested by its capacity for dealing with puzzles, and it is a wholesome plan, in thinking about logic, to stock the mind with as many puzzles as possible, since these serve much the same purpose as is served by experiments in physical science.'

---

[7] Alchourrón and Bulygin [AB81] discuss the question whether there are permissive norms. They notice that a great number of philosophers (especially philosophers of law) deny that there are permissive norms, admitting only one type of norms (mandatory norms, imperatives, commands), whereas logicians and lawyers – though probably for different reasons – feel less inclined to such a monistic conception and see no obstacle that would prevent them from speaking of permissive norms (independently of the question whether they are definable in terms of obligations or not).

[8] Moreover, in modal logic the deontic operators can be nested. However, there is no intuitive or useful reading of nested operators (what does 'it ought to be that it ought to be that $\alpha$' tell us?). Nested formula like $O(O\alpha \rightarrow \alpha)$ have been discussed by e.g. Prior [Pri62], see also [FH71, p.15]. A potential use is the nesting of operators in a multi-modal logic, like $P_1P_2p$: 'I permit you to permit someone else to do p.' Nevertheless, it remains to be shown that such sentences can be formalized and reasoned with. The multi agent perspective remains still mainly unexplored.

Tomberlin [Tom81] observes that deontic logicians agree with Russell, because deontic logic has been supplied with a wealth of puzzles – usually referred to as *deontic paradoxes*. Moreover, he observes that the majority of the paradoxes evaporate comfortably enough under close scrutiny and yet, some – most notably the contrary-to-duty imperative paradox and certain strengthened versions of the Good Samaritan paradox – persist. We discuss these troublesome cases in Section 1.3.3. First, we discuss the main types of monadic deontic logics.

### 1.3.2 Monadic obligations

We start with a classification of deontic logics based on three formal properties represented in Table 1.1. For example, consider the obligation to pay for received goods within two weeks. The first property is the derivability of weaker obligations, like the obligation to pay for received goods (at an unspecified time). The second property is the derivability of the conjunction of two obligations, like the obligation to pay for two shipments of received goods. The third property is the possibility to represent violations in the language, like the violation that certain received goods are not paid for within two weeks.

| | | SDL | OS | MDL | PDL |
|---|---|---|---|---|---|
| **W**eakening | $O(\alpha_1 \wedge \alpha_2) \rightarrow O\alpha_1$ | X | X | X | |
| **And** | $(O\alpha_1 \wedge O\alpha_2) \rightarrow O(\alpha_1 \wedge \alpha_2)$ | X | X | | X |
| **V**iolations | wff: $\alpha \wedge O\neg\alpha$ | X | | X | X |

Table 1.1: Three classification properties of deontic logic

Table 1.1 shows four deontic logics, which are discussed later in this section. The most familiar deontic logic, so-called *Standard Deontic Logic* SDL, has all three properties. Three well-known weakenings of this logic each lack one property. Von Wright's *Old System* OS [vW51] lacks the possibility to express violations (because the $\alpha$ in $O\alpha$ is not a proposition but an act-type), Chellas' *Minimal Deontic Logic* MDL [Che74] lacks the derivability of the conjunction of obligations, and S.O. Hansson's *Preference-based Deontic Logic* PDL [Han90b] lacks the property to derive weaker obligations. Since logics can also lack two or all three of the properties, our classification gives eight classes of deontic logics.

Most deontic logics have weakening, because deontic logic has been developed as a branch of modal logic. We therefore first discuss the classes of deontic logics that have weakening, and criticize weakening at the end of this section. We give a proof-theoretic analysis of different types of deontic logics that have weakening by discussing the formulas in Table 1.2. We assume a complete set of principles (axioms and rules of inference) for classical propositional logic and the following two rules of inference, which state that there is substitution of logical equivalents within the scope of the modal operator.

$$\frac{\vdash \alpha_1 \leftrightarrow \alpha_1}{\vdash O\alpha_1 \leftrightarrow O\alpha_2} \qquad \frac{\vdash \alpha_1 \leftrightarrow \alpha_2}{\vdash P\alpha_1 \leftrightarrow P\alpha_2}$$

The two main types of deontic logics are variants of so-called minimal deontic logic (MDL) and standard deontic logic (SDL). All variants of minimal deontic logic have as axioms **OW**, **PW** and **D**. The axioms **OW** and **PW** express that obligations and permissions are closed under

| | | SDL | OS | MDL | PDL |
|---|---|---|---|---|---|
| **OW**eakening | $O(\alpha_1 \wedge \alpha_2) \rightarrow O\alpha_1$ | X | X | X | |
| **PW**eakening | $P(\alpha_1 \wedge \alpha_2) \rightarrow P\alpha_1$ | X | X | X | |
| **D** | $O\alpha \rightarrow P\alpha$ | X | X | X | X |
| **D′** | $\neg O\bot$ | X | X | X | X |
| | $O\top$ | X | | | X |
| **And** | $(O\alpha_1 \wedge O\alpha_2) \rightarrow O(\alpha_1 \wedge \alpha_2)$ | X | X | | X |
| **D*** | $\neg(O\alpha \wedge O\neg\alpha)$ | X | X | | X |
| | $\neg(O\alpha \wedge P\neg\alpha)$ | X | X | | X |
| | $O\alpha \vee P\neg\alpha$ | | | | |
| | $O\alpha \leftrightarrow \neg P\neg\alpha$ | | | | |
| Violations | wff: $\alpha \wedge O\neg\alpha$ | X | | X | X |
| Nested O | wff: $OO\alpha$ | X | | X | X |

Table 1.2: Different types of deontic logics

logical consequence and axiom **D** gives the relation between obligations and permissions: what is obliged is thereby permitted. Chellas' [Che74] MDL is a non-normal modal logic (i.e. it does not have axiom **K**: $O(\alpha \rightarrow \beta) \rightarrow (O\alpha \rightarrow O\beta)$) and has the expressive power to represent violations and nested modal operators. Two extensions of minimal deontic logic which only induce a minor change of the system are the axioms $O\top$ and $\neg O\bot$, where $\top$ stands for any tautology like $p \vee \neg p$ (and $\bot$ for a contradiction like $p \wedge \neg p$). $O\top$ and $\neg O\bot$ refer to extreme cases, and they are not so interesting for the concept of obligation. It is therefore very easy for any system to adapt the definitions such that these extreme cases are excluded or included. An axiom with larger impact is the axiom $O\alpha \vee P\neg\alpha$. The axiom states that every situation is normed, because for *every* proposition $\alpha$ we have that $\alpha$ is either obligatory or its negation is permitted. However, usually not everything is normed, and this axiom should not be accepted.[9]

Standard deontic logic consists of minimal deontic logic plus the axiom **And**.[10] In minimal deontic logic it is possible that there are two obligations for $\alpha_1$ and $\alpha_2$ without an obligation for $\alpha_1 \wedge \alpha_2$. In minimal deontic logic these two obligations refer to two unrelated normative standards. There is not a normative standard that says that $\alpha_1 \wedge \alpha_2$ is obligatory. The axiom **And** of standard deontic logic expresses that such a situation is not possible. It expresses the uniqueness of the normative standard. The theorem **D***: $\neg(O\alpha \wedge O\neg\alpha)$ follows from **And** and **D′**. It says that there cannot be a dilemma. Standard deontic logic has in contrast to minimal deontic logic the no-dilemma assumption. Again, the axioms $O\top$ and $\neg O\bot$ may be added. SDL is usually rep-

[9] When we discussed the representation of deontic logic in modal logic, we observed that the interdefinability of the two ideas of necessity and possibility through the schema $Mp =_{def} \neg N\neg p$ provokes no serious objection. But the corresponding schema or equivalence in deontic logic is by no means unproblematic. It seems feasible to admit a 'weak' notion of permittedness, according to which something *may* be, if and only if it is *not* the case that the contradictory of this thing *ought* to be. A weak permission $P^-$ is introduced by the definition $P^-\alpha =_{def} \neg O\neg\alpha$, see [Alc93]. If the axiom $O\alpha \vee P\neg\alpha$ is added to $\neg(O\alpha \wedge P\neg\alpha)$, then we have $P^-\alpha \leftrightarrow P\alpha$.

[10] A restricted version of the conjunction rule **And**, so-called consistent aggregation, was given by Van Fraassen [vF73, Hor94]. With consistent aggregation, we may derive $O(\alpha_1 \wedge \alpha_2)$ from $O\alpha_1$ and $O\alpha_2$ only when $\alpha_1 \wedge \alpha_2$ is consistent. The extension of minimal deontic logic with consistent aggregation is a system in between MDL and SDL. Restricted **And** does not imply the no-dilemma assumption.

resented by a normal modal system of type **KD** according to the Chellas classification [Che80]. The axiom **K** states that modus ponens holds within the scope of the modal operator, and the axiom **D** states that the impossible cannot be obliged.[11]

**Definition 1.6 (SDL)** The language $\mathcal{L}$ is formed from a denumerable set **P** of propositional variables together with the connectives $\neg$, $\rightarrow$, and $O$. The connectives $\wedge$, $\vee$ and $\leftrightarrow$ are defined in terms of them in the usual way. The logic SDL is the smallest $S \subseteq \mathcal{L}$ such that $S$ contains classical propositional logic and the following axiom schemata, and is closed under the following rules of inference.

| | | | |
|---|---|---|---|
| **K** | $O(\alpha \rightarrow \beta) \rightarrow (O\alpha \rightarrow O\beta)$ | **MP** | from $\alpha$ and $\alpha \rightarrow \beta$ derive $\beta$ |
| **D** | $\neg O\bot$ | **Nes** | from $\alpha$ derive $O\alpha$ |

□

**Definition 1.7 (Kripke semantics)** A possible world (Kripke) model for a deontic theory in SDL is a tuple $M = \langle W, R, V \rangle$ that consists of a nonempty set of worlds $W$, a binary serial[12] accessibility relation $R$ between worlds, and a valuation function $V$ that assigns in each world $w \in W$ a truth value to the atomic propositions. A formula $O\alpha$ is true in world $w$ in $M$, written $M, w \models_{\text{SDL}} O\alpha$, iff, for all accessible worlds $w'$ with $R(w, w')$, it is true that $M, w' \models_{\text{SDL}} \alpha$. As usual, a formula $\alpha$ entails $\beta$, written $\alpha \models_{\text{SDL}} \beta$, iff $M, w \models_{\text{SDL}} \alpha$ implies $M, w \models_{\text{SDL}} \beta$ for all models $M$ and worlds $w$. □

The following example illustrates that the binary accessibility relation associates with each world a set of ideal alternatives. Moreover, there are also ideal alternatives of ideal alternatives, which are used to give nested modal sentences like $OOp$ a truth value.

**Example 1.8 (Kripke semantics)** Consider the Kripke model $M = \langle W, R, V \rangle$ in Figure 1.3. The set of worlds $W$ consists of six worlds, and the accessibility relation is such that the actual world $w$ sees four worlds, which see the sixth world. Notice that the accessibility relation is serial, because all worlds see another world (the right-most world sees itself). We have $M, w \models_{\text{SDL}} Op$, because $p$ is true in all the ideal alternatives of the world $w$. □

Although MDL and SDL have been the prime deontic logics for at least three decades, there are several paradoxes related to weakening. Whereas $O(p \wedge q) \rightarrow Op$ might seem intuitive, the equivalent $Op \rightarrow O(p \vee q)$ looks less convincing. The following four problems related to weakening illustrate the purpose served by these logical puzzles (which, as Russell remarked, is much the same purpose as served by experiments in physical science). In general, it is argued that the following puzzles related to weakening evaporate comfortably enough under close scrutiny.

[11] The inference rules are sometimes written as follows.

$$\textbf{MP}\frac{\alpha, \alpha \rightarrow \beta}{\beta} \qquad \textbf{Nes}\frac{\vdash \alpha}{\vdash O\alpha}$$

In this representation, a distinction between the two inference rules is made explicit. Modus ponens is applicable on every two derivable formulas, whereas necessitation is only applicable on logical theorems. However, in Definition 1.6 we only describe the set of logical theorems, and this distinction is not relevant.

[12] A binary relation $R$ is serial iff for all $w \in W$, there is a $w' \in W$ such that $R(w, w')$. Hence, a binary relation is serial if there are no so-called 'dead ends'.

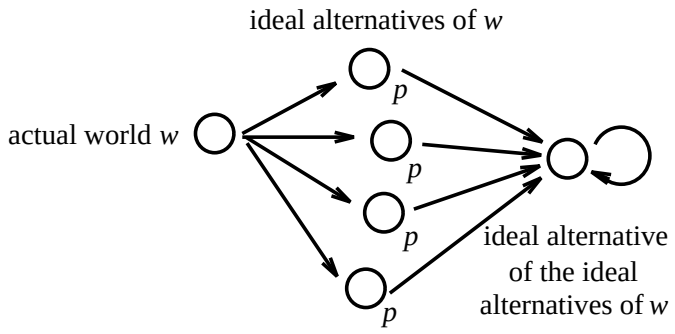



Figure 1.3: A Kripke model

**Example 1.9 (Ross paradox)** [Ros41] Ross gave the following counterintuitive example of the sentence $Op \to O(p \vee q)$, called the Ross paradox: 'If you should mail the letter, then you should mail or burn the letter.' □

Føllesdal and Hilpinen [FH71, p.22] observe that the Ross paradox above 'may perhaps be explained by reference to very general conventions regarding the use of language. For instance, it is generally assumed that a person makes as strong statements as he is in a position to make.' Castañeda [Cas81, p.65] also observes that some writers have observed correctly that few persons would engage in reasoned commanding of the form 'Do A, therefore, do A or B.' However, he argues that this fact about *communication* or *speech* cannot tell against the implications behind the inferences in question, because 'there are reasons pertaining to the transfer of information that explain why those inferences are not drawn,' and 'the very same reasons apply to the corresponding indicative or propositional inferences.' Thus, Castañeda concludes, 'we must simply forget about Ross's paradox.' However, there are more paradoxes associated with weakening.

**Example 1.10 (Good Samaritan paradox)**[Åqv67] The following counterintuitive example of the sentence $O(p \wedge q) \to Op$ is called the Good Samaritan paradox: 'If you ought to help someone who has been robbed, then he ought to be robbed.' □

The usually accepted solution (see e.g. [Cas81, Tom81]) of the Good Samaritan paradox is that the obligation 'you ought to help someone who has been robbed' is a conditional obligation which cannot be represented as a monadic obligation. We consider this solution of the paradox in Section 1.3.3.

**Example 1.11 (Paradox of the knower)** [Åqv67] The following counterintuitive example of the sentence $OKp \to Op$ is called the paradox of the knower: 'If you ought to know that $p$, then it ought to be that $p$.' The paradox follows from the well-known theorem of epistemic logic 'what is known is factually true' $K\alpha \to \alpha$. □

Surprisingly, the paradox of the knower has received very limited attention in deontic logic literature, although it is more complicated than the Good Samaritan paradox. The following example illustrates that weakening is also problematic for the concept of permission.

**Example 1.12 (Free choice paradox)**[Kam74] The following counterintuitive example of the sentence $Pp \to P(p \vee q)$ is called the free-choice paradox: 'If a person is permitted to smoke, then she is also permitted to smoke or to kill.' □

Von Wright [vW71a] suggests 'that a so-called free-choice permission and the permission concept defined by the standard system of deontic logic have different logics; the former concept does not satisfy the distribution principle $P(p \vee q) \leftrightarrow Pp \vee Pq$, but instead the law represented by $P(p \vee q) \leftrightarrow Pp \wedge Pq$.'

Several authors have criticized weakening. Beatty [Bea73] discusses what he calls the descriptive aspect of obligation sentences, and observes that, in terms of description, weakening 'seems much less plausible.' Von Wright [vW81, p.7] observes that 'in a deontic logic which rejects the implication from left to right in the equivalence $O(p \wedge q) \leftrightarrow (Op \wedge Oq)$ while retaining the implication from right to left, the paradoxes would not appear.' Later, Von Wright [vW81] develops a deontic action logic that does not have weakening, because 'from the fact that an agent is under an obligation to perform actions which exhibit two characteristics, it does not follow that he is under an obligation to perform actions which have (only) one of the characteristics.'

Surprisingly, hardly any arguments can be found in deontic literature that defend the property weakening. One argument (which seems to be implicit in some discussions but as far as we know has not been defended explicitly) is based on the analogy between deontic logic and the logic of necessity. The argument runs as follows. If weakening should be rejected in deontic logic, then it should be rejected in the logic of necessity too. For example, the sentence 'if it is necessary that you mail the letter, then it is necessary that you mail or burn the letter' is just as counterintuitive as 'if you ought to mail the letter, then you ought to mail or burn the letter.' However, weakening seems to be intuitive for logics of necessity. We do not reject weakening for the logic of necessity, and as a consequence we should not reject weakening for deontic logic. This argument relies on the analogies and similarities between the two modalities, whereas there are also a number of striking dissimilarities, as discussed in Section 1.3.1. The following (in our opinion also not very convincing) example is given by Sinnot-Armstrong (in response to Forrester [For84]).

> "For example, if I both mow and water your grass, I mow your grass, so, if it is obligatory for me to mow and water your grass, it is obligatory for me to mow your grass. Such arguments cannot be justified if [weakening] is rejected, unless some other rule or principle is substituted. Forrester gives no substitute for [weakening] in his article, and it seems that any substitute would have to be very complex and unintuitive in order to justify all such obviously valid arguments." [SA85]

Finally, the following argument is given by Nute and Yu in the introduction of a book on defeasible deontic logic.

> "[Weakening] is one of the most fundamental principles in SDL and has strong intuitive appeal. The principle states that consequences of what ought to be the case ought to be the case. It hence requires the agent to take the moral responsibility for the possible consequences of what he/she has committed to do. The rejection of the principle, therefore, will not only jeopardize SDL but also seems to be contrary to one of our basic moral reasoning patterns." [NY97]

From a technical point of view, it is easy to construct a non-normal modal logic that does not have weakening, but such a logic does not *explain* why the derivation weakening is not valid. Forrester [For84] observes that most of standard deontic logic must be reconstructed, which is 'a large task.' S.O.Hansson [Han90b] observes that 'this situation seems to depend, to a large degree, on the lack of a credible semantical basis for a weaker deontic logic.' Jennings [Jen85] provides such an explanation. He observes that 'it has been suggested that a unary operator $O$ capable of bearing a deontic interpretation might be defined in a logic of preference by $O\alpha =_{def} \alpha \succ \neg\alpha$', where $\alpha_1 \succ \alpha_2$ stands for a preference of $\alpha_1$ over $\alpha_2$, and that 'if the preference logic has the natural distributive properties as von Wright advocates, the defined deontic necessity will be nonmonotonic' (i.e. does not have weakening). This has been formalized in preference-based deontic logic by Jackson [Jac85] and Goble [Gob90a, Gob90b]. In Chapter 2 we introduce the so-called ordering logic that is defined in a preference logic and we illustrate why such preference-based logics do not have weakening. S.O. Hansson [Han90b] introduces Preference-based Deontic Logic PDL that explains the absence of weakening by preferential semantics, though differently from Jennings. In S.O.Hansson's logic, obligations are defined by the so-called property of negativity: 'what is worse than something wrong is itself wrong.' We discuss this preference-based deontic logic in Chapter 2.

### 1.3.3 The contrary-to-duty paradoxes of SDL

Deontic logic is hampered by many paradoxes, intuitively consistent sentences which are formally inconsistent, or from which counterintuitive sentences can be derived. The most notorious paradoxes are caused by so-called *Contrary-To-Duty* (CTD) obligations, obligations that refer to sub-ideal situations. For example, Lewis describes the following example of the CTD obligation that you ought to be helped when you are robbed.

**Example 1.13 (Good Samaritan paradox)** "It ought not to be that you are robbed. *A fortiori*, it ought not to be that you are robbed and then helped. But you ought to be helped, given that you have been robbed. This robbing excludes the best possibilities that might otherwise have been actualized, and the helping is needed in order to actualize the best of those that remain. Among the best possible worlds marred by the robbing, the best of the bad lot are some of those where the robbing is followed by helping." [Lew74] □

The Forrester [For84] and Chisholm [Chi63] paradoxes are the most notorious CTD paradoxes. In Standard Deontic Logic (SDL), a conditional obligation is usually formalized by the formula $\beta \rightarrow O\alpha$, where $\beta$ is the condition and $\alpha$ the (deontic) conclusion. The conditional obligation $\beta \rightarrow O\alpha$ is a *Contrary-To-Duty* (or *secondary*) obligation of the (*primary*) obligation $O\alpha_1$ when $\beta$ and $\alpha_1$ are contradictory. In analyzing CTD paradoxes, it is useful to make the following distinction. We call an obligation $O\alpha$ that can be derived from a theory $T$ of SDL a fulfilled obligation, violated obligation or deontic cue respectively, depending on whether $\alpha$ is entailed by $T$, $\neg\alpha$ is entailed by $T$ or neither of them. Violated obligations represent what has been done wrong (what is the case but should not be the case) and the deontic cues represent what should be done now (what is not yet realized but should be made the case).[13] In Exam-

[13]This is not the only possible reading of the obligations. For example, assume that $s$ stands for smoking. We can have that $s \wedge O\neg s$ does not represent a violation that you are smoking, but the deontic cue to stop smoking.

ple 1.15 and 1.16 we illustrate that the CTD paradoxes of SDL are conflicts between different types of obligations.

**Definition 1.14 (SDL obligations)** Let $T$ be a theory of SDL. The formula $O\alpha$ is *a fulfilled obligation* of the theory $T$ iff $T \models_{\text{SDL}} O\alpha$ and $T \models_{\text{SDL}} \alpha$. The formula $O\alpha$ is *a violated obligation* of the theory $T$ iff $T \models_{\text{SDL}} O\alpha$ and $T \models_{\text{SDL}} \neg\alpha$. The formula $O\alpha$ is a *deontic cue* of the theory $T$ iff $T \models_{\text{SDL}} O\alpha$, $T \not\models_{\text{SDL}} \alpha$ and $T \not\models_{\text{SDL}} \neg\alpha$. □

The following example is the notorious Forrester paradox [For84], sometimes called the gentle murderer paradox.[14] It is a strengthened version of the Good Samaritan paradox.

**Example 1.15 (Forrester paradox)** Consider the following sentences of an SDL theory $T$:

1. $O\neg k$: Smith should not kill Jones;
2. $k \to O(k \wedge g)$: If Smith kills Jones, then he should do it gently;
3. $k$: Smith kills Jones.

The second obligation is a CTD obligation of the first obligation, because $\neg k$ and $k$ are contradictory. SDL allows so-called factual detachment, i.e.

$$\models_{\text{SDL}} (\beta \wedge (\beta \to O\alpha)) \to O\alpha$$

and therefore we have $T \models_{\text{SDL}} O(k \wedge g)$ from the last two sentences of $T$. Moreover, we have $T \models_{\text{SDL}} O\neg k$ and $T \models_{\text{SDL}} Ok$, where the latter is derived from the CTD obligation $O(k \wedge g)$. The main problem of this paradox is that $O\neg k$ and $Ok$ are inconsistent in SDL, although the set of premises is intuitively consistent. Note that the paradox is a conflict between different types of obligations: $O\neg k$ is a violated obligation and $Ok$ is derived from the deontic cue $O(k \wedge g)$. □

The following, more complicated, CTD paradox was given by Chisholm [Chi63]. It is more complicated, because it also contains an *According-To-Duty* (ATD) obligation. A conditional obligation '$\alpha$ ought to be the case if $\beta$ is the case,' represented by the sentence $\beta \to O\alpha$ or by the sentence $O(\beta \to \alpha)$, is an ATD obligation of $O\alpha_1$ iff $\beta$ logically implies $\alpha_1$. The condition of an ATD obligation is satisfied only if the primary obligation is fulfilled. Hence, the definition of ATD is analogous to the definition of CTD. A CTD obligation is an obligation conditional to a violation and an ATD obligation is an obligation conditional to a fulfillment of an obligation. The paradox can be represented by several different sets of SDL formulas, which are either inconsistent or not logically independent, see e.g. [Chi63, Åqv67, Smi93]. We first give an inconsistent representation.

**Example 1.16 (Chisholm paradox)** Consider the following sentences of an SDL theory $T$:

1. $Oa$: A certain man should go to the assistance of his neighbors;

However, the violation reading we give here of the formula is the reading given in the standard examples, like the Forrester and Chisholm paradoxes below.

[14] Forrester writes $k \to Og$ and $g \to k$, where the latter has the status of a theorem (for the details, see Forrester's paper [For84]). Our simple representation is from [PS96].

2. $O(a \to t)$: It should be that if he goes, then he tells them that he will come;

3. $\neg a \to O\neg t$: If he does not go, then he should not tell them that he will come;

4. $\neg a$: He does not go.

The second obligation is an ATD obligation and the third obligation is a CTD obligation of the first obligation, see Figure 1.4. Since SDL allows a kind of so-called deontic detachment as a result of the **K**-axiom, i.e.

$$\models_{\text{SDL}} (O\beta \wedge O(\beta \to \alpha)) \to O\alpha$$

we have $T \models_{\text{SDL}} Ot$ from the first two sentences. We have $T \models_{\text{SDL}} O\neg t$ from the last two sentences by factual detachment. The paradox is that these two derived obligations are inconsistent, although the set of premises is intuitively consistent. The two conflicting obligations are different types of obligations, because $Ot$ is a consequence of a violated obligation and $O\neg t$ is a deontic cue. □

$$\begin{tikzcd} Oa & Oa \\ O(a \to t) \arrow[u, "\text{implies}"] & \neg a \to O\neg t \arrow[u, leftrightarrow, "\text{inconsistent}"] \end{tikzcd}$$

Figure 1.4: $O(a \to t)$ is an ATD of $Oa$ and $\neg a \to O\neg t$ is a CTD of $Oa$

An alternative representation of the Chisholm set (see e.g. [Chi63, Åqv67, Smi93]) is to represent the second sentence by $a \to Ot$. An advantage of this representation is that the second and third sentence are represented by logical formulas with the same structure. Moreover, it is an attempt to solve the Chisholm paradox, because it makes the Chisholm set consistent. However, the following example illustrates that this 'solution' misses the point of the paradox. The solution does not have deontic detachment whereas deontic detachment is in most cases intuitive. Thus, the representation $O(a \to t)$ derives too much (always deontic detachment) and the representation $a \to Ot$ derives too little (never deontic detachment).

**Example 1.17 (Chisholm paradox, continued)** Consider the following SDL theory

$$T = \{Oa, a \to Ot, \neg a \to O\neg t, \neg a\}$$

Notice that the second formula can be derived from the fourth formula. Chisholm remarked that this logical dependence is counterintuitive, and several logicians (see e.g. [Åqv67]) have demanded that a solution of the Chisholm paradox should represent the sentences such that they are logically independent. However, Tomberlin [Tom81] observes that the criterion is a 'rather glaring theoretical commitment' which 'would be a case of flagrant methodological question-begging.' Moreover, this logical dependence is easily solved by introducing a weaker notion of implication. For example, the two conditional obligations can be represented by $a > Ot$ and $\neg a > O\neg t$ where '$>$' is a so-called strict implication. This solves the logical dependence, because the formula $\alpha \to (\alpha > \beta)$ is in contrast to the formula $\alpha \to (\alpha \to \beta)$ not a theorem, as is explained in Section 1.3.5 below.

The main problem underlying the Chisholm paradox is whether we allow for deontic detachment or not. If a conditional obligation is represented by $O(\beta \to \alpha)$ then deontic detachment is $(O\beta \wedge O(\beta \to \alpha)) \to O\alpha$, and when a conditional obligation is represented by $\beta \to O\alpha$, then deontic detachment in SDL is represented by the SDL formula

$$(O\beta \wedge (\beta \to O\alpha)) \to O\alpha$$

This latter formula is not valid in SDL, hence $Ot$ cannot be deontically detached from the first two sentences in SDL, unless the formula is added as an axiom. However, if it is added as an axiom, then it reinstates the inconsistency of the Chisholm paradox. □

We argue in Chapter 4 that deontic detachment should sometimes hold and sometimes not. Most people have a clear intuition that $O\neg t$ should be preferred over $Ot$. Also we would expect that if the fact $\neg a$ was not the case, the preference would be the other way round. But this cannot be obtained from the SDL representation, because in both cases $Oa$ is true, from which $Ot$ is derived. In the first case $Oa$ is a violated obligation and in the second case a deontic cue. Given the intuitive reading above, $Ot$ should only be inferred from $Oa$ when $Oa$ is a fulfilled obligation or a deontic cue. When it is a deontic cue, $Ot$ is derived on the assumption that $a$ will be true. This reading of the example has a non-monotonic character, i.e. conclusions can be lost by the addition of new information. On the other hand, SDL is a classical modal logic, in which it is impossible to model such non-monotonic reasoning.

### 1.3.4 Solutions of the contrary-to-duty paradoxes of SDL

B.Hansson [Han71] shows that the fundamental problem underlying these paradoxes is that the type of possible world semantics of SDL is not flexible enough. In these semantics only two types of worlds are distinguished in a model; *actual* and *ideal* ones. The ideal worlds have to satisfy all obligations in a deontic theory $T$. Clearly, if these obligations contradict each other, then a problem arises. As Lewis [Lew74] observes, 'a mere division of worlds into the ideal and the less-than-ideal will not meet our needs. We must use more complicated value structures that somehow bear information about comparisons or gradations of value.' For example, in the Chisholm paradox both $Ot$ and $O\neg t$ are implied. No ideal world can satisfy both $t$ and $\neg t$, and this causes the paradox. Hence, ideal worlds are simply not enough. In order to model these paradoxes properly, we need a notion of *sub-ideal* worlds, in which some but not all obligations are satisfied. For example, in the Chisholm paradox we could distinguish between two types of (sub-)ideal worlds: (sub-ideal) worlds in which $\neg t$ is true but not $t$, and (ideal) worlds in which $t$ is true but not $\neg t$. This solves the inconsistency in the ideal worlds. Moreover, having the finer distinction between a hierarchy of (sub-)ideal worlds instead of one type of ideal world, we can define a preference ordering on these sub-ideal worlds. Given that it is a fact that the man does not go to the assistance, it is better not to tell the neighbors that he is coming than not going and telling them that he will come. Hence, although the sub-ideal worlds in which $\neg a$ and $\neg t$ are true are not ideal, they are better than the worlds in which $\neg a$ and $t$ are true. We say that B.Hansson [Han71] introduced a dyadic deontic logic in which the *ideality principle* of SDL is replaced by an *optimality principle*.

> "The problem of conditional obligation is what happens if somebody nevertheless performs a forbidden act. Ideal worlds are excluded. But it may be the case that among the still achievable worlds some are better than others. There should then be an obligation to make the best out of the sad circumstances." Bengt Hansson [Han71]

The optimality principle is a solution of the Forrester paradox phrased in semantic terms. It is formalized by dyadic deontic logic, and discussed in the following section. Four other solutions of the Forrester paradox are based on temporal distinctions, the distinction between settled and non-settled facts, scope distinctions and lack of weakening. All these solutions have their shortcomings. In the remainder of this section we discuss a solution in temporal deontic logic, and we discuss the alternative solutions and their shortcomings in Chapter 2.

Since the late seventies, several temporal deontic logics and deontic action logics were introduced, which formalize satisfactorily a special type of CTD obligations, see for example [Tho81, vE82, LB83, Mak93, Alc93]. We can distinguish two types of temporal deontic logics.

1. The modal operators are relativized with a temporal index. We mentioned this type of temporal deontic logic in Section 1.2.2. By itself, this extension does not solve the paradoxes.

2. The condition of a conditional obligation occurs before the obligatory formula. In this type of temporal deontic logic, the dyadic obligation $O(\alpha \mid \beta)$ is read as 'if $\beta$ is the case, then at later moments $\alpha$ ought to be the case.' The temporal lag between the condition and the obligatory formula can either be formalized by a modal operator [Alc93] or by a temporal connective [Mak93].

In the latter temporal approach, the underlying principle of the formalization of CTD obligations is that facts of the past are not in the 'context of deliberation' [Tho81]. Hence, they can formalize the Good Samaritan paradox in Example 1.13. We call the temporal deontic logics of the second type – condition before conclusion – the temporal solution of the CTD paradoxes. However, they cannot formalize the variant of the paradox described by Forrester in Example 1.15 and the Chisholm paradox in Example 1.16, because in these paradoxes there are CTD obligations of which the consequent occurs at the same time or even before its antecedent.

The distinction between the ideality principle and the optimality principle can be rephrased in terms of this temporal perspective. Thomason [Tho81] makes a distinction between the context of deliberation and the context of justification, the latter is called the 'context of judgment' by Loewer and Belzer [LB83]. He distinguishes between two ways in which the truth values of deontic sentences are time-dependent. First, these values are time-dependent in the same, familiar way that the truth values of all tensed sentences are time-dependent. Second, their truth values are dependent of a set of choices or future options that varies as a function of time. The context of deliberation is the set of choices when you are looking for practical advice, whereas the context of justification is the set of choices for someone who is judging you. The essential contrast between judgment and deliberation is a difference in what we take as settled [Tho81, LB83, AB96]. The crucial distinction between the ideality and the optimality principle is also in what we take as being settled.

### 1.3.5 Dyadic obligations

*Dyadic* deontic logic formalizes conditional obligations. A dyadic obligation $O(\alpha \mid \beta)$ is read as '$\alpha$ ought to be the case if $\beta$ is the case.' The monadic obligations $O\alpha$ we discussed in the previous sections can be seen as a special kind of dyadic obligations $O(\alpha \mid \top)$. However, this is neither the only nor the obvious way to represent absolute obligations in a dyadic deontic logic, as we discuss below. The introduction of the dyadic representation was inspired by the standard way of representing conditional probability, that is, by $Pr(\alpha|\beta)$ which stands for 'the probability for $\alpha$ given $\beta$.' In Table 1.3 we reconsider the three main classification properties of Table 1.1 for *dyadic* deontic logics. In this table $X$ stands for the validity of a theorem in a logic (first two rows) or for the fact that a formula is a well-formed formula (wff) and satisfiable (last two rows). $R$ stands for restricted validity.

| | | Condition | | | Context | |
|---|---|---|---|---|---|---|
| | | Ch-MDL | Ch-SDL | A-W | H-L | ORD |
| **W**eakening | $O(\alpha_1 \wedge \alpha_2\|\beta) \rightarrow O(\alpha_1\|\beta)$ | X | X | X | X | |
| **And** | $(O(\alpha_1\|\beta) \wedge O(\alpha_2\|\beta)) \rightarrow O(\alpha_1 \wedge \alpha_2\|\beta)$ | | X | X | X | R |
| Contextual | wff+sat: $\alpha \wedge O(\neg\alpha\|\top)$ | | | | X | X |
| Conditional | wff+sat: $\alpha \wedge O(\neg\alpha\|\top)$ and $O(\neg\alpha\|\alpha)$ | X | X | | | |

Table 1.3: Three main classification properties for dyadic deontic logic

The two properties weakening and conjunction remain the same, but the possibility to represent violations becomes more complicated. A dyadic deontic logic either has no possibility to represent violations, only the possibility to represent them by the formula $\alpha \wedge O(\neg\alpha|\top)$, or the two possibilities $\alpha \wedge O(\neg\alpha|\top)$ and $O(\neg\alpha|\alpha)$. Hence, in the former case the formula $O(\neg\alpha|\alpha)$ is not satisfiable, but in the latter case it is. We say that the logic gives a contextual interpretation of the antecedent of the dyadic obligations if the logic can represent violations by $\alpha \wedge O(\neg\alpha|\top)$ but not by $O(\neg\alpha|\alpha)$. In the other cases, we say that the logic gives a conditional interpretation of the antecedent. The distinction between the conditional and the contextual interpretation of the antecedent of dyadic deontic logics corresponds to the distinction between the ideality and the optimality principle. The ideality principle corresponds to the conditional interpretation, and the optimality principle is underlying the contextual interpretation of the antecedent of dyadic deontic logics. Hence, the distinction between the conditional and contextual interpretation of the antecedent is motivated by attempts to model contrary-to-duty reasoning. Moreover, the distinction corresponds to the distinction between the context of justification and the context of deliberation. The contextual interpretation of the antecedent refers to the set of options when you are looking for practical advice (to make the best out of the sad circumstances). For example, consider the formula $O(\neg\alpha \mid \alpha)$, that characterizes the distinction between the conditional and the contextual interpretation. With a conditional interpretation of the antecedent, the obligation $O(\neg\alpha|\alpha)$ considers the set of choices of someone who is judging you, and this person observes a violation. With a contextual interpretation of the antecedent, the obligation $O(\neg\alpha|\alpha)$ considers the set of choices when you are looking for practical advice. In the latter case, $\neg\alpha$ cannot be reasonable advice if we know that $\alpha$ is the case.

Whereas the distinction is very clear in the semantics, it is not very clear in the proof the-

ory.[15] In deontic logic literature, the distinction between the two classes is rather confusing. For example, B.Hansson [Han71] introduced the first deontic logic with a contextual interpretation, and remarked that 'circumstances are regarded as something which has actually happened (or will unavoidably happen) and which cannot be changed afterwards.' Lewis [Lew74] discusses four logics with a contextual interpretation of the antecedent and says that some other treatments of dyadic deontic logic 'seem to be based on ideas quite unlike the one I wish to consider,' but he does not give any formal discriminating properties. Moreover, Prakken and Sergot [PS97] call a dyadic obligation a contextual obligation if its antecedent 'stands for a constellation of acts or situations that agents regard as being settled in determining what they should do,' quoting Hilpinen [Hil93], but they also do not give formal properties of how contextual obligations differ from other types of obligations.

Table 1.3 shows five deontic logics, three with a conditional interpretation of the antecedent and two with a contextual interpretation of the antecedent, which are discussed below. Ch-MDL and Ch-SDL are two 'conditional' logics proposed by Chellas [Che74], A-W is a dyadic logic proposed by Alchourrón [Alc93] in the tradition of von Wright's dyadic deontic logics, H-L are deontic logics proposed by B. Hansson [Han71] and further investigated by Lewis [Lew74], and ORD is the ordering logic proposed in Chapter 2 of this thesis.

We give a proof-theoretic analysis of the distinction between the conditional and the contextual interpretation of the antecedent of dyadic deontic logics by discussing the axioms in Table 1.4.[16] In the dyadic case, we assume the following two rules for substitution of logical equivalents within the modal operator $O$ (and similar rules for $P$).

$$\frac{\vdash \alpha_1 \leftrightarrow \alpha_2}{\vdash O(\alpha_1|\beta) \leftrightarrow O(\alpha_2|\beta)} \qquad \frac{\vdash \beta_1 \leftrightarrow \beta_2}{\vdash O(\alpha|\beta_1) \leftrightarrow O(\alpha|\beta_2)}$$

Chellas [Che74] proposed dyadic obligations defined in terms of monadic obligations and a conditional from conditional logic, that is, a strict or relevant implication. For an excellent survey on Chellas-type conditional obligations, see [Alc93, Alc96]. In the table, we discriminate between MDL and SDL for the monadic obligations.

**Definition 1.18 (Chellas-Alchourrón-von Wright)** Assume a multi modal logic where $\Box$ stands for an alethic modal operator that satisfies at least **T**: $\Box\alpha \to \alpha$, and the operators $O$ and $P$ stand for obligation and permission, as before. A conditional obligation '$\alpha$ ought to be the case if $\beta$ is the case' is defined by $O(\alpha\,|\,\beta) =_{def} \beta > O\alpha$, where $\beta > \alpha =_{def} \Box(\beta \to \alpha)$ (thus combined $O(\alpha|\beta) =_{def} \Box(\beta \to O\alpha)$). Similarly, permission is defined by $P(\alpha|\beta) =_{def} \beta > P\alpha$. $\Box$

[15] The distinction between the contextual and conditional interpretation of the antecedent originates in the distinction between the two representations of conditional obligations in monadic deontic logic, $O(\alpha|\beta) =_{def} O(\beta \to \alpha)$ and $O(\alpha|\beta) =_{def} \beta \to O\alpha$. Alchourrón [Alc96] calls the former representation the *insular* representation of conditional norms and the latter representation the *bridge* conception of conditional norms. In the latter representation conditional norms are like bridges which link what *is* (or might be) the case to what *ought* to be (done). They relate the 'Sein-reign' with the 'Sollen-reign'. A second origin is the distinction between 'relative obligation' and 'conditional obligation' discussed by Jackson [Jac85].

[16] In this thesis, we have chosen not to refer to the 'standard' names of axioms and inference rules from conditional logic for our inference patterns, because we want to stress that our inference patterns should be read as semi-formal analysis tools instead of rules of a logic. Moreover, the standard names sometimes refer to left and right (like LLE = left logical equivalence) and left and right have been exchanged for the deontic conditional $O(\alpha|\beta)$ compared to the conditional $\beta > \alpha$ from conditional logic. Finally, we do not use these names, because they are difficult to read for people outside the field of conditional logic. For example, weakening of the consequent is often represented by RCM.

| | | Condition | | | Context | |
|---|---|---|---|---|---|---|
| | | Ch-MDL | Ch-SDL | A-W | H-L | ORD |
| **WC** | $O(\alpha_1\|\beta) \to O(\alpha_1 \vee \alpha_2\|\beta)$ | X | X | X | X | |
| **And** | $(O(\alpha_1\|\beta) \wedge O(\alpha_2\|\beta)) \to O(\alpha_1 \wedge \alpha_2\|\beta)$ | | X | X | X | R |
| | $O(\alpha\|\beta) =_{def} \beta > O\alpha$ | X | X | X | | |
| | $O(\alpha\|\beta) =_{def} \beta >_O \alpha$ | | | | X | R |
| | $O(\alpha\|\beta) \to O(\alpha \wedge \beta\|\beta)$ | | | | X | R |
| | $O(\alpha\|\alpha)$ | | | | X | R |
| **SA** | $O(\alpha\|\beta_1) \to O(\alpha\|\beta_1 \wedge \beta_2)$ | X | X | X | | R |
| **FD** | $(O(\alpha\|\beta) \wedge \beta) \to O\alpha$ | X | X | X | | |
| **DD** | $(O(\alpha\|\beta) \wedge O(\beta\|\gamma)) \to O(\alpha\|\gamma)$ | | | | | |
| **DD**$\top$ | $(O(\alpha\|\beta) \wedge O(\beta\|\top)) \to O(\alpha\|\top)$ | | | | X | |
| **DD**$'$ | $(O(\alpha\|\beta) \wedge O(\beta\|\gamma)) \to O(\alpha \wedge \beta\|\gamma)$ | | | | | R |
| **RBC** | $(O(\alpha\|\beta_1) \wedge O(\alpha\|\beta_2)) \to O(\alpha\|\beta_1 \vee \beta_2)$ | X | X | X | X | |
| | $P(\alpha\|\beta) =_{def} \neg O(\neg\alpha\|\beta)$ | | | | | |
| | $O\alpha =_{def} O(\alpha\|\top)$ | | | | | |

Table 1.4: Different types of dyadic deontic logics

In the discussion of the Chisholm paradox in SDL in Section 1.3.3, we observed that the representation $a > Ot$ solves the logical dependence between $a \to Ot$ and $a$. The formula $\alpha \to (\alpha \to \beta)$ a theorem of SDL, and as a consequence we can derive $a \to Ot$ from $a$. However, with the strict implication we do not have the theorem $\alpha \to (\alpha > \beta)$, i.e. we do not have the theorem $\alpha \to \Box(\alpha \to \beta)$, because we do not have $\alpha \to \Box\alpha$. We now consider the properties in Table 1.4. Given the well-known theorems of **T**,

$$
\begin{aligned}
&(\beta > \alpha) \to (\beta \to \alpha)\\
&(\beta_1 > \alpha) \to ((\beta_1 \wedge \beta_2) \to \alpha)\\
&(\beta > \alpha_1 \wedge \beta > \alpha_2) \leftrightarrow (\beta > (\alpha_1 \wedge \alpha_2))\\
&(\beta_1 > \alpha \wedge \beta_2 > \alpha) \to ((\beta_1 \vee \beta_2) > \alpha)
\end{aligned}
$$

it follows directly that the dyadic obligations have the theorems **FD**, **SA**, **And** and **RBC** of Table 1.4. Factual detachment **FD**, or the derivability of monadic obligations from the dyadic ones, follows from the first theorem, strengthening of the antecedent **SA** follows from the second theorem, the conjunction rule **And** follows from the third theorem and the conjunction rule of the monadic obligations, and reasoning by cases **RBC**, or the sure-thing principle, follows from the fourth theorem. Chellas argues that these dyadic obligations represented by $\beta > O\alpha$ do not 'fuse the notions of obligation and conditionality.' These deontic logics can be contrasted to conditional obligations which are expressed with a primitive dyadic operator (which we represent with $>_O$ in Table 1.4), like the deontic logics with a contextual interpretation of the antecedent. In particular, Chellas argues that for the latter obligations, the connection between the notion of obligation involved in $O(\alpha \mid \beta)$ and that expressed in nonconditional contexts is 'not evident'. Two potential extensions of the Chellas-type logics are the definitions of absolute obligations and permissions in terms of the dyadic obligations. Alchourrón [Alc93] shows that the repre-

sentation of permission $P(\alpha \mid \beta) =_{def} \neg O(\neg\alpha \mid \beta)$ is wrong (the 'first misfortune'), because 'the negation of a conditional statement usually is not a conditional statement.' Alchourrón further remarks that this was also denounced as mistaken by Castañeda. The definition of absolute obligations in terms of dyadic obligations with a tautological antecedent $O\alpha =_{def} O(\alpha \mid \top)$ is the 'second misfortune.' Von Wright [vW68] proposed two ways to represent monadic obligations $O\alpha$ in a dyadic logic: by $O(\alpha \mid \top)$ and by $O(\alpha \mid S)$, where $S$ stands for what von Wright calls the actual circumstance. Alchourrón [Alc93] observes that the former has been followed unanimously by all deontic logicians, although there is a difficulty in the procedure. The obligation $O(\alpha \mid \top)$ represents $\Box O\alpha$, i.e. the obligation to do $\alpha$ in *all possible circumstances*. On the other hand, the simple monadic operators $O$ and $P$ admit the possibility of being used to norm for a single possible circumstance $c$ (which may be the actual circumstance). In this sense, they may be used to express what Alchourrón calls *categorical norms*, i.e. norms that are not general in relation to circumstances. In Alchourrón's words, this misrepresentation is 'the ghost of categorical norms.'

Lewis [Lew74] compares four deontic logics that give a contextual interpretation of the antecedent of the dyadic obligations: dyadic logics proposed by B.Hansson [Han71], by Føllesdal and Hilpinen [FH71], by van Fraassen [vF72] and by himself. All these logics are quite similar, because they do not have strengthening of the antecedent **SA** and factual detachment **FD**. It is exactly for the lack of these two properties that these logics have been criticized. Lack of strengthening of the antecedent is generally felt to be counterintuitive for logics of obligations, see e.g. [Alc93]. Castañeda [Cas81] argues that the denial of strengthening of the antecedent is a 'negative solution that looks like overkill and leaves a converse problem unsolved,' because 'it tempers with the extensionality of ought and it cannot account for the fact that only in *very* few cases we *seem* to have the situation of an obligation cancelled by the addition of another circumstance.' Loewer and Belzer [LB83] distinguish between dyadic obligations that have factual detachment **FD**, and those that have deontic detachment **DD** (terminology introduced by Greenspan [Gre75]). Moreover, they criticize Hansson-Lewis semantics because 'it does not contain the resources to express actual obligations and no way of inferring actual obligations from conditional ones.' Again, two potential extensions of the logics are the definitions of absolute obligations and permissions in terms of the dyadic obligations. Chellas [Che74] remarks that 'another problem with the unanalyzed deontic conditional operator concerns conditional permission, typically taken to be expressed by sentences of the form $\neg O(\neg\alpha \mid \beta)$, by analogy with the rendering $\neg O\neg\alpha$ for unconditional permission. It is not clear that this is a faithful representation of conditional permission; $P(\alpha \mid \beta)$ appears to be more the denial of a conditional prohibition than the conditional affirmation of a permission.'

The contrary-to-duty paradoxes lead to problems for dyadic obligations with a conditional interpretation of the antecedent similar to the problems in standard deontic logic. As a consequence, these logics cannot formalize contrary-to-duty reasoning. In Chapter 2 we show by the dyadic version of the Forrester paradox $\{O(\neg k \mid \top), O(k \wedge g \mid k), k\}$ in Figure 1.5 and 1.6 that the underlying problem in the proof theory is twofold:

1. The combination of unrestricted strengthening of the antecedent **SA** and the conjunction rule **And** derives $O(\neg k \wedge (k \wedge g) \mid k)$, as is represented in Figure 1.5 below. Strengthening of the antecedent has to be restricted such that $O(\neg k \mid k)$ cannot be derived from $O(\neg k \mid \top)$. This can be done with a consistency check on the antecedent and the consequent of the

obligation. If we express a 'consistency check' of $\alpha$ by the formula $\overset{\leftrightarrow}{\Diamond}\alpha$, then we can express a restricted form of strengthening of the antecedent by the following formula **RSA**: $(O(\alpha|\beta_1) \wedge \overset{\leftrightarrow}{\Diamond}(\alpha \wedge \beta_1 \wedge \beta_2)) \rightarrow O(\alpha|\beta_1 \wedge \beta_2)$.

$$\dfrac{\dfrac{O(\neg k|\top)}{O(\neg k|k)}\,\text{SA} \qquad O(k \wedge g|k)}{O(\neg k \wedge (k \wedge g)|k)}\,\text{AND}$$

Figure 1.5: The Forrester paradox in dyadic deontic logic (1)

2. The combination of restricted strengthening of the antecedent **RSA**, weakening of the consequent **WC** and the conjunction rule **And** derives the obligation $O(\neg(k \wedge g) \wedge (k \wedge g)|k)$, as is represented in Figure 1.6 below. Thus, the weakening of **SA** to **RSA** is not sufficient to solve the paradox. A logic that combines (restricted) strengthening of the antecedent with weakening of the consequent also has to block this second counterintuitive derivation.

$$\dfrac{\dfrac{\dfrac{O(\neg k|\top)}{O(\neg(k \wedge g)|\top)}\,\text{WC}}{O(\neg(k \wedge g)|k)}\,\text{RSA} \qquad O(k \wedge g|k)}{O(\neg(k \wedge g) \wedge (k \wedge g)|k)}\,\text{AND}$$

Figure 1.6: The Forrester paradox in dyadic deontic logic (2)

From the discussion of the Forrester paradox in dyadic deontic logic follows that Chellas type of dyadic deontic logics with a conditional interpretation do not solve the paradox, because they have SA and WC, as illustrated in Table 1.4. The Hansson-Lewis dyadic obligations with a contextual interpretation of the antecedent can formalize the contrary-to-duty paradoxes, because they do not have strengthening of the antecedent. Moreover, they cannot have any restricted form of strengthening of the antecedent, because they also have weakening of the consequent. However, as we already remarked the denial of strengthening of the antecedent is a 'negative solution that looks like overkill' [Cas81]. The ordering logic ORD in Table 1.4 we introduce in Chapter 2 has strengthening of the antecedent and therefore does not have weakening of the consequent. In this thesis, we introduce three solutions of the problem of *combining* (restricted) strengthening of the antecedent and weakening of the consequent. In Chapter 2 we introduce two-phase deontic logic 2DL that blocks the second counterintuitive derivation in Figure 1.5 by introducing two phases in the proof theory. The first phase has strengthening of the antecedent and the second phase has weakening of the consequent. Hence, it is not possible to use first weakening of the consequent and afterwards strengthening of the antecedent. This blocks the second derivation. The two other solutions introduced in Chapter 3 extend the language of dyadic deontic logic. Labeled deontic logic LDL is based on the distinction between implicit and explicit obligations, a distinction analogous to the distinction between implicit and explicit knowledge. Contextual deontic logic CDL explicitly represents exceptions of the context of obligations.

In this section we discussed the problem of the formalization of contrary-to-duty paradoxes in monadic and dyadic deontic logic. In the remainder of this discussion of the philosophical foundations of deontic logic, we discuss several other issues. In the following subsection we discuss conflict resolution and violation diagnosis, and in the last subsection we discuss the representation of actions in deontic logic.

### 1.3.6 Defeasible obligations

In the introduction in Section 1.1 we already observed that there is no undisputed definition of 'defeasible deontic logic.' However, it is generally accepted that a defeasible deontic logic should be able to deal with conflict resolution and that it should be able to diagnose violations.

**Conflict resolution.** In a defeasible deontic logic, obligations can be overridden by other obligations. Overridden defeasibility becomes relevant when there is a (potential) conflict[17] between two obligations. For example, there is a conflict between $O(\alpha_1|\beta_1)$ and $O(\alpha_2|\beta_2)$ when $\alpha_1$ and $\alpha_2$ are contradictory, and $\beta_1$ and $\beta_2$ are factually true. There are several different approaches to deal with deontic conflicts. In standard deontic logic SDL a deontic conflict is inconsistent. In minimal deontic logic MDL a conflict is consistent and called a 'deontic dilemma'. In a defeasible deontic logic a conflict can be *resolved*, because one of the obligations overrides the other one. For example, overridden structures can be based on a notion of specificity, like in Horty's well-known example that 'you should not eat with your fingers,' but 'if you are served asparagus, then you should eat with your fingers' [Hor93]. In such cases, we say that an obligation is *cancelled* when it is overridden. The obligation not to eat with your fingers is cancelled by the exceptional circumstances that you are served asparagus. A different kind of overridden structures have been proposed by Ross [Ros30] and formalized, for example, by Morreau [Mor96]. In Ross' ethical theory, an obligation which is overridden has not become a 'proper' or actual duty, but it remains in force as a prima facie obligation. For example, the obligation not to break a promise may be overridden to prevent disaster, but even when it is overridden it remains in force as a prima facie obligation. We say that as actual obligation the overridden obligation is cancelled, but as prima facie obligation it is only overshadowed. We analyze the different types of defeasibility in defeasible deontic logic in Chapter 4.

**Violation diagnosis.** Deontic logic is the logic of obligations, i.e. reasoning about what should be the case. Defeasible logic is the logic of default assumptions, i.e. reasoning about what normally is the case. In defeasible deontic logic these two are combined. An example of this combination is the sentence 'normally, you should do $p$.' Defeasible deontic logic is complex, because there are two interfering notions. Consider the sentence 'normally, you should do $p$.' Now the problem is what to conclude about somebody who does not do $p$? Is this an exception to the normality claim, or is it a violation of the obligation to do $p$? This confusion arises because there is a substantial overlap between deontic and defeasibility aspects. We analyze the distinction between violations and exceptions in Chapter 4.

There are at least two formal definitions of defeasibility. A defeasible expression can be formalized in a non-monotonic logic, or in a conditional logic in which the conditionals do not have

[17] When we say conflicts, we mean conflicts within one normative system. Von Wright [vW71b] observed that 'unconditional duties under different laws or systems of norms may, of course, conflict, in the sense that they impose logically contradictory demands on an agent. Such cases, however, are logically uninteresting and should better not be regarded as genuine 'conflicts of duties' at all.'

strengthening of the antecedent. For example, consider the default expressions 'birds fly' and 'penguins do not fly.' In a non-monotonic logic, flying ($f$) can be derived from bird ($b$), but not from penguin ($b \wedge p$). In contrast, non-flying ($\neg f$) can be derived from penguin ($b \wedge p$). In a conditional logic the two default expressions are represented by $b > f$ and $(b \wedge p) > \neg f$ respectively, where $>$ is a conditional implication from conditional logic. The conditional expression $(b \wedge p) > f$ cannot be derived from $b > f$, which corresponds to the non-derivability of $f$ from $b \wedge p$ in a non-monotonic logic. Hence, in a conditional logic the defeasibility is represented by lack of strengthening of the antecedent of the conditionals. Alchourrón [Alc93] dubbed such conditionals 'defeasible conditionals.' In this thesis we mainly use the latter representation of defeasibility. Strengthening of the antecedent is the most discussed derivation is this thesis.

### 1.3.7 Actions

It is an open question whether the logic of statements like 'it ought to be that $\alpha$ is the case' is analogous or different from the logic of statements like '$\alpha$ ought to be done.' The two logics are sometimes called the logics of ought-to-be and ought-to-do obligations, or the logics of sein-sollen and tun-sollen. Humberstone [Hum71] also makes a distinction between two kinds of ought statements – what he calls 'situational' and 'agent-implicating' oughts. Horty and Belnap [HB95] describe the following example. Let us first imagine a case in which Albert has competed in a gymnastics event. Suppose Albert's performance is clearly superior, but the judge is known to be biased, and it is likely that he will award the medal to someone else. If one then said 'Albert ought to win the medal,' this is a kind of statement that Humberstone would classify as a situational ought. It reflects a judgment about the situation, not about Albert, and can be paraphrased as 'it ought to be that Albert wins the medal.' There is no implication that Albert will be at fault if he fails to win the medal, or that winning the medal is now within his power. By contrast, suppose Albert has not kept up with his training schedule. One might then say, 'Albert ought to practice harder,' and this would be the kind of ought statement that Humberstone classifies as agent-implicating. It implies that Albert is able to practice harder, and places the blame on him if he fails to do so.

In deontic logic literature, deontic statements of the type 'you ought to do $\alpha$' are presumed applicable to any proposition $\alpha$. For example, in STIT theory (see [HB95] for an overview) an obligation $O\alpha$ is read as 'the agent ought to see to it that $\alpha$ is the case.' This $\alpha$ can be any propositional sentence. On the other hand, decision theoretic methods treat actions as distinct, predefined objects. In deontic logic literature, this is observed and investigated by Meyer [Mey88]. As is well-known from the action logic literature, the formalization of actions induces several typical problems. For example, one of the most notorious problems is the interpretation of the negation or complement of an action (the omission of the action). Actions in Meyer's dynamic deontic logic have the following properties.

1. Actions are deterministic. See e.g. [TH96] how nondeterministic actions can be formalized.

2. The actions do not interfere. The global outcome when actions $a_1, \ldots a_n$ are concurrently performed at $m$ – that is, the state associated with the moment that results when the actions are performed concurrently in $m$ – is determined by the local outcomes of the separate actions.

Deontic logic literature has mainly addressed propositional languages, because analogues of most properties (for example weakening and strengthening) also occur in an action logic. As a consequence, the problems discussed in ought-to-be logics also occur in ought-to-do logics. For example, the CTD paradoxes also occur in ought-to-do logics. To illustrate the occurrence of the CTD paradoxes, we first give a semantic intuition and then give the syntactic counterpart. As we discussed, the problem of standard deontic logic is that there is only a distinction between good ideal and bad violation. The Hansson-Lewis dyadic deontic logics replace the distinction between good ideal and bad violation worlds of the monadic deontic logics by a preference ordering on worlds. Analogously, simple deontic action logics only have a distinction between good ideal actions and bad violation actions. Dynamic dyadic deontic logics replace the distinction between good ideal and bad violation actions by a preference ordering on actions. For the introduction of a preference ordering in the STIT approach, see [HB95, p.617]. We now consider the occurrence of weakening and strengthening in the syntax. We discuss the occurrence in Meyer's dynamic deontic logic, without discussing the proof theory or the semantics of the logic. Meyer's logic contains an action language that has among others atomic actions and the connectives '$\cup$' for choice and '$\&$' for concurrency. Moreover, the logic has the following theorems that express weakening. It illustrates that in this calculus choice and concurrency play a similar role as disjunction and conjunction in the propositional languages. For example, the analogy between choice and disjunction is that the introduction of choice $\alpha_1 \rightarrow (\alpha_1 \cup \alpha_2)$ and disjunction $\alpha_1 \rightarrow (\alpha_1 \vee \alpha_2)$ introduce 'weaker' descriptions of actions respectively states.[18]

$$O\alpha_1 \rightarrow O(\alpha_1 \cup \alpha_2)$$
$$O(\alpha_1 \& \alpha_2) \rightarrow O\alpha_1$$

Meyer does not define dyadic deontic operators, but they can easily be introduced (at least in the language). Given the analogy of choice and concurrency with disjunction and conjunction, we can define analogues of the theorems which we discussed when we analyzed propositional dyadic deontic logic. The following formulas express weakening of the consequent and strengthening of the antecedent.

$$\begin{array}{ll}
\mathbf{WC}_{\cup} & O(\alpha_1|\beta) \rightarrow O(\alpha_1 \cup \alpha_2|\beta) \\
\mathbf{WC}_{\&} & O(\alpha_1 \& \alpha_2|\beta) \rightarrow O(\alpha_1|\beta) \\
\mathbf{SA}_{\&} & O(\alpha|\beta_1) \rightarrow O(\alpha|\beta_1 \& \beta_2) \\
\mathbf{SA}_{\cup} & O(\alpha|\beta_1 \cup \beta_2) \rightarrow O(\alpha|\beta_1)
\end{array}$$

The point is that many *deontic* problems, like the formalization of contrary-to-duty structures we study in this thesis, can be analyzed in a propositional language or in an action language. Deontic logic literature has therefore mainly addressed the simpler propositional languages. For example, in a dynamic dyadic deontic logic we can analyze the Forrester paradox as a problem of combining strengthening of the antecedent and weakening of the consequent (the formulas

[18]This weakening has a semantic counterpart. For propositional logic, the set of models of $\alpha_1$ is a subset of the models of $\alpha_1 \vee \alpha_2$. In dynamic logic, the set of worlds referred to by the modal operator $[\alpha_1]$ is a subset of the set of worlds referred to by the modal operator $[\alpha_1 \cup \alpha_2]$. For concurrency $\&$, the semantic analogy follows from property (2), i.e. that the global outcome of actions performed concurrently is determined by the local outcomes of the actions.

above). Makinson [Mak93] also argues that deontic problems should be analyzed in a simple setting. He lists some open problems of deontic logic. Besides the distinction between *sein-sollen* and *tun-sollen* he mentions that the focus on the very best worlds satisfying $\alpha$ (or satisfying some more complex condition) may to be be relaxed somewhat, that the use of a single ordering relation of degrees of goodness of worlds may not be enough, that the representation of human agency by an accessibility relation between worlds may be too simplistic for some purposes, and finally, that an analysis of what should be done does not determine which agents should be bearers of the obligation. Makinson observes that deontic logic is thus a subtle affair even without conditionality. Logics of conditional obligation that take into account factors such as those listed above will tend to become quite intricate. Makinson concludes that such a situation faces the logician with a dilemma: simple structures convey very basic distinctions and insights, but are gross oversimplifications. Complex structures may come closer to the contours of discourse, but can be extremely cumbersome to handle, with insights disappearing in a mass of overheads and book-keeping. At the present state of play, it would not seem advisable to try to cover all complicating factors at once, but rather to get an initial appreciation of them few at a time, only subsequently putting them together and investigating their interactions.

An interesting question of tun-sollen and sein-sollen which can only be analyzed in a dynamic deontic logic is the logical relationship between ought-to-be and ought-to-do obligations. For example, if $\alpha$ is an obligatory action and $\beta$ is a necessary post condition of the action $\alpha$, does this imply that there is an ought-to-be obligation for $\beta$? Or, if an agent has the obligation to see to it that $\alpha$ is the case, does this imply that $\alpha$ ought-to-be the case? This type of questions are beyond the scope of this thesis, see e.g. [HB95].

## 1.4 A personal perspective

Thus far, a survey of philosophical and artificial intelligence literature of deontic logic has been given. In this section I give my personal view on deontic logic.

### 1.4.1 Research vision

Deontic logic has been studied as a modal system, and it is usually presented as a syntactic-axiomatic system. Unfortunately, in my opinion this approach has not been very successful, as is illustrated by the many deontic paradoxes. The semantic approach has had less attention. The modal logics have a Kripke semantics, of course, but this semantics is not very useful or insightful. It is not very clear what an ideal alternative of a world is, and it is even more obscure what an ideal alternative of an ideal alternative is. When we consider the semantics, we can distinguish two main classes of approaches since SDL has been departed. The first approach is based on a semantic concept of time and action [Tho81, vE82, LB83, Mak93, Alc93, HB95, Hor96]. The second approach is based on preferences, either in monadic deontic logic [Jac85, Gob90b, Han90b] or in dyadic deontic logic [Han71, Lew74], and is related to defeasible deontic logic. The relation between preference-based deontic logic and defeasible deontic logic has to be interpreted carefully. A defeasible deontic logic has to be able to deal with conflict resolution and the diagnosis of violations. However, in my opinion there is more to defeasibility. In particular, defeasibility can be used to solve deontic paradoxes, but this does not mean that I argue that deontic

paradoxes may be solved by interpreting them as conflicts to be resolved. Hence, I do not argue that the deontic paradoxes may be solved with restoring consistency techniques. We have to use another type of defeasibility.

Thus far, the time- and preference-based approaches are rivals and have not been merged successfully. Moreover, both have made claims concerning the formalization of contrary-to-duty reasoning. I favor the complex solution of the contrary-to-duty paradoxes based on preferences and defeasibility because, in my opinion, the other solutions are not very satisfactory. This follows from the pragmatic oddity discussed by Prakken and Sergot [PS96], which was already mentioned in the introduction in Section 1.1. From the following three sentences of an SDL theory 'you should keep your promise' $Ok$, 'if you have not kept your promise, you should apologize' $\neg k \rightarrow Oa$ and 'you have not kept your promise' $\neg k$ the formulas $Ok \wedge Oa$ and $O(k \wedge a)$ can be derived. Prakken and Sergot remark 'but it is a bit odd to say that in all ideal versions of this world you keep your promise and you apologize for not keeping it. This oddity – we might call it the'pragmatic oddity' – seems to be absent from the natural language version, which means that the SDL representation is not fully adequate.' This is what I consider a major understatement. In my opinion, the derivation of $O(k \wedge a)$ is not only 'a bit odd' and does not mean that the logic is only 'not fully adequate,' but it means that the logic is not capable of formalizing contrary-to-duty reasoning. For example, just think of what the semantics of the obligation $O(k \wedge a)$ must look like, if it is given a sensible reading. In my opinion, the consequence is that SDL and several other logics have to be rejected as candidates to formalize contrary-to-duty reasoning. In particular, it remains to be shown whether the time-based approach can deal with the pragmatic oddity.

I also followed the semantic approach when I analyzed *defeasible* deontic reasoning, i.e. deontic reasoning combined with conflict resolution. The problem that faced me when I started my research was how to compare the different defeasible deontic logics. There did not seem to be any good tools for the analysis, because defeasible deontic logic was a new and undeveloped (but therefore exciting) area. I developed the following two tools.

1. The first tool I used was multi-preference semantics. The preferences represent the notion of deontic choice: an obligation for $\alpha$ is formalized as a kind of choice between $\alpha$ and $\neg\alpha$. The semantic analysis is based on the distinction between two preference orderings, which reflects the distinction between violations and exceptions.

2. Second, I found inspiration in the Kraus-Lehmann-Magidor approach [KLM90] in artificial intelligence on meta-level analysis of non-monotonic reasoning (applied to default reasoning). I used similar techniques to study defeasible deontic logic, the so-called inference patterns. These patterns focus on structural properties of the logic such that I became able to get a grip on the underlying mechanisms. The two most basic mechanisms are represented in Figure 1.7 below.

   (a) $O(\alpha|\beta)$ is an overriding obligation (based on specificity) of $O(\alpha_1|\beta_1)$ iff $\alpha \wedge \alpha_1$ is inconsistent, and $\beta$ is more specific than $\beta_1$, and

   (b) $O(\alpha|\beta)$ is a contrary-to-duty obligation of $O(\alpha_1|\beta_1)$ iff $\beta \wedge \alpha_1$ is inconsistent.

The multi preference semantics and the inference patterns work together like a tandem. The result is a general analysis of different types of defeasibility in defeasible deontic logics, where the intuitions behind the various distinctions are illustrated with preference-based semantics.


Figure 1.7: Specificity and CTD

Finally, I think that deontic logic only covers a small portion of the formalization of normative reasoning. I started to analyze contrary-to-duty reasoning by formalizing it in a theory of diagnosis. This did not give any problems for contrary-to-duty paradoxes, see Section 5.3. Later, it was shown that contrary-to-duty reasoning can be formalized in a theory of qualitative decision, see Section 5.2. However, in the deontic logicians perspective, this formalization of normative reasoning does not tell anything *about* obligations, because it does not tell which obligations follow from a set of obligations. That is, I did not present a deontic *logic*. This has been pointed out, in particular, by Henry Prakken and Marek Sergot. The point can be illustrated by the following example. In the so-called DIagnostic framework for DEontic reasoning DIODE an obligation for $\alpha$ is represented by the formula $\neg V_i \rightarrow \alpha$, to be read as 'if norm $i$ is not violated then $\alpha$ is the case. A set of such obligations is called a normative system. Consider the normative system with two norms '$a$ should be done' and '$b$ should be done,' represented by $\neg V_1 \rightarrow a$ and $\neg V_2 \rightarrow b$ respectively. Now, a third norm '$a \wedge b$ should be done' can be added without changing any of the conclusions of the normative system. The only distinction is that $V_3$ is part of a diagnosis if and only if $V_1$ and $V_2$ are part of the diagnosis, but the set of diagnoses and the related measurements remain the same (see Section 5.3 for the technical details). This means that the norm '$a \wedge b$ should be done' is *logically implied* by '$a$ should be done' and '$b$ should be done'. This logical relationship explains the meaning of the norms. However, the obligation '$a \wedge b$ should be done' cannot be derived in DIODE. Hence, from $\neg V_1 \rightarrow a$ and $\neg V_2 \rightarrow b$ I could not derive $\neg V_x \rightarrow (a \wedge b)$. So, I understood that I had to reverse the approach. I could not build a deontic logic on top of a theory of diagnosis. Instead, I should use deontic logic in a theory of diagnosis! A theory of diagnosis can use deontic logic to represent system rules and violations of these system rules. Similarly, qualitative decision theory uses preference-based deontic logic to formalize reasoning about context-sensitive goals. Qualitative decision theory can also tell us how norms affect behavior. The behavior of agents depends on the knowledge they have of other agents. In particular, agents behave on the degree of belief they have that other agents have fulfilled their obligations. For example, in the protocol of Example 1.4, the buyer will not pay unless she has sufficient evidence that the seller has actually delivered the goods. I think that theories that formalize reasoning with norms, like diagnosis and decision theory, deserve more attention from researchers that study normative reasoning.

### 1.4.2 Research challenges

When I started my PhD research, I read a few articles on defeasible deontic logics. Unfortunately, I did not find them very satisfactory. For example, the solution of the contrary-to-duty

paradoxes based on conflict resolution is very ad hoc, as I already discussed in the introduction. My first challenge was to describe the relation between defeasibility and the contrary-to-duty paradoxes. Later I studied Horty's logic [Hor93], which seems to represent the paradoxes satisfactorily. However, the logic had some other problems. It is based on Reiter's default logic, a proof-theoretic system which does not have a semantics. More seriously, Horty's logic does not have a possibility to represent violations. As a consequence, the logic did not solve the issue of violation diagnosis, because it confuses the distinction between exceptions and violations. My second challenge was to distinguish these two elements. To distinguish the two elements, I used multi preference semantics. The representation of exceptions by preference-based semantics is well-known from the artificial intelligence literature. My third challenge was to describe the relation between preferences and the contrary-to-duty paradoxes. Summarizing, my research challenge has been the following.

> **Research challenge.** Show the relation between obligations, preferences and defeasibility.

I developed preference-based defeasible deontic logics to study this relation. Many technical problems popped up when I started to develop deontic logics with multi-preference semantics. Some of these problems already appear when there is a single preference ordering. The development of preference-based deontic logic is plagued by the following three (related) problems, which are discussed in detail in Chapter 2.

1. **Strong preference problem.** Strong preferences for $\alpha_1$ and $\alpha_2$ conflict for $\alpha_1 \wedge \neg\alpha_2$ and $\neg\alpha_1 \wedge \alpha_2$.

2. **Contrary-to-duty problem.** Contrary-to-duty reasoning must be formalized without running into the notorious contrary-to-duty paradoxes of deontic logic.

3. **Dilemma problem.** The three combined formulas $O\alpha \wedge O\neg\alpha$, $O(\alpha_1 \wedge \alpha_2) \wedge O\neg\alpha_1$ and $O(\alpha_1 \wedge \alpha_2 | \top) \wedge O(\neg\alpha_1 | \beta)$ represent dilemmas and should therefore be inconsistent.

Two other problems I encountered are related to deontic logics in general, and in particular to dyadic deontic logics. They are also discussed in detail in Chapter 2.

4. **Permissions.** In SDL, (weak) permissions are sometimes defined by $P\alpha =_{def} \neg O\neg\alpha$, but this validates the counterintuitive theorem $O\alpha \vee P\neg\alpha$. I therefore looked for an alternative notion of (strong) permission.

5. **Factual detachment.** Unrestricted factual detachment can be used to derive pragmatic oddities. So I looked for other types of factual detachment.

I encountered more problems when I tried to formalize defeasible deontic logic with multi preference semantics. These problems are related to the distinction between violations and exceptions. The idea of multi-preference semantics can be illustrated by a model of two sentences of the Cottage Housing regulations in Example 1.3 (the second sentence has been adapted).

1. $O(\neg f | \top)$: There should be no fence,

2. $O(f | s)$: If the cottage is by the sea, then there should be a fence.

The typical multi preference model is represented in Figure 1.8. This model is explained in detail in Chapter 4. It can be read as follows. The circles denote equivalence classes of worlds that satisfy the literals inside the circles and the 'horizontal' arrows denote the deontic preference ordering. The boxes denote equivalence classes in the normality ordering and the 'vertical' arrow the normality preference ordering. The two sentences construct two preference orderings on the worlds: one ordering for ideality and one for normality. The idea of the preference ordering on normality is that the worlds with exceptional circumstances (where the cottage is near the sea) are semantically separated from the normal situation (where the cottage is not near the sea). The upper box represents the 'normal' worlds, which is determined by the fact that $s$ is false, i.e. the cottage is not near the sea. Deontically, the $\neg s$ worlds are ordered according to the obligation that, usually, there should be no fence. The lower box contains the worlds where $s$ is true and which are therefore exceptional. These worlds are deontically ordered by the obligation that in this situation, there should be a fence. Because of the exceptional circumstances, the worlds are not subject to the obligation that usually, there should not be a fence. In the ideality ordering, the normal $\neg s \wedge \neg f$ worlds and the exceptional $s \wedge f$ worlds are equivalent.

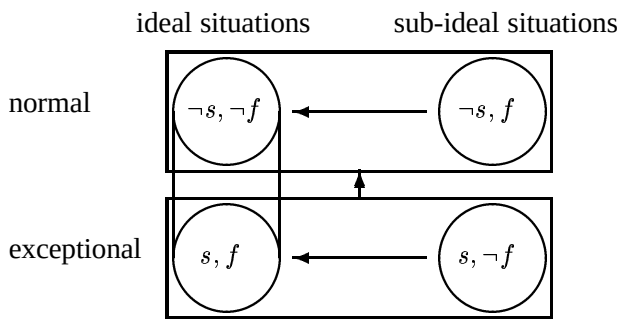



Figure 1.8: Multi-preference relation of the Fence example

This kind of multi preference semantic structures leave us with the following two problems:

6. **Model construction.** How is a multi preference model constructed, given a set of obligations?

7. **Entailment.** Given a multi preference model, which obligations are true in the model? For example, consider the definition $O\alpha$ is true if $\alpha$ is true in the most normal of the best worlds, or in the best of the most normal worlds. It is not very satisfactory, because in that case we have $M \models O\neg s$ for the model $M$ in Model 1.8. This counterintuitive derivation represents that $s$ is a violation, whereas $s$ is not a violation but an exception.

### 1.4.3 Research validation

The validation of the research challenge is based on the development of preference-based defeasible deontic logics. The logics are used to show the relation between obligations, preferences and defeasibility. The logics must have the expressive power to distinguish between exceptions and violations. The success test of any logic is that it derives the 'right' set of sentences from a set of sentences. The problem is, of course, how to distinguish between right and wrong. There are two ways to validate the properties of a logic.

1. **Natural language intuitions**. Intuitions can be used to analyze the derivable theorems. It is the most often used way to analyze a logic, especially in philosophical literature. Unfortunately, it is also the most problematic approach, because the intuitions of people differ on the examples. Alchourrón [Alc93] observes that the source of the problem lays in the difficulties involved in the process of finding intuitive correlates of principles in the highly ambiguous uses of deontic and related sentences (such as imperatives) in everyday discourse and in more sophisticated (legal and moral) contexts.

2. **(Semantic) explanation**. Alternatively, it can be analyzed whether the semantics give a good representation for varying sets of premises. Kripke semantics explain the distinction between actual and ideal. In my opinion, this is a quite limited explanation. Temporal semantics explain that obligations change in time. Preference-based semantics explain the notion of deontic choice.

Moreover, the logics have to solve the seven problems discussed above. In general, a simple solution is preferred to a complex one, because complex solutions are more difficult to formalize. Of most of the seven issues, it can be verified easily whether their solution is successful or not. The most problematic is the formalization of contrary-to-duty reasoning. The success test of the formalization of contrary-to-duty reasoning is a satisfactory formalization of the contrary-to-duty paradoxes.

### 1.4.4 A defense of the paradoxes

I end this personal perspective with a defense of the deontic paradoxes as a success test. First, observe that the examples are typical for a large set of examples that are structurally similar. For instance, Horty's example of table manners 'you should not eat with your fingers, unless you are served asparagus' has the same structure as 'you should not kill, unless the patient is terminally ill and in excruciating pain.' Moreover, there are many sentences with the same structure as the Forrester paradox 'Smith should not kill Jones, but if he kills him he should do it gently.' Consider the following four examples given in deontic logic literature.

1. Smith ought not to perform in South Africa, but if he performs in South Africa he ought to perform in Soweto [Gob91].

2. Jones ought not to wear red to school, but if he wears red to school, then he ought to wear scarlet to school [Gob91].

3. There must be no fence, but if there is a fence it must be a white fence [PS96].

4. The children ought not to be cycling on the street, but if they are cycling on the street they ought to be cycling on the left hand side of the street [PS96].

I think that many examples seem trivial, but they are not. They are the result of several decades of research on deontic logic and have to be taken seriously. The answer to the question whether some sentence should be derivable is often not just yes or no, but something in between. That is, intuitively they are only derivable under certain circumstances or with a certain interpretation of the obligations. To see the point of the puzzles, you have to do some puzzling yourself.

It is like playing cryptograms. You first have to try the puzzle yourself, otherwise you do not see the fun. In the remainder of this section I give four examples of these puzzles, and a possible solution based on the analogy between obligations and imperatives.

**Chisholm paradox.** Remember that the SDL theory

$$\{Oa, O(a \to t), \neg a \to O\neg t, \neg a\}$$

is inconsistent whereas the SDL theory

$$\{Oa, a \to Ot, \neg a \to O\neg t, \neg a\}$$

is consistent. However, the second set does not derive $Ot$, whereas $t$ is true in the ideal state. Even more problematic is the lack of the derivation of $O(a \wedge t)$ from the second set. The set is consistent, but it does not solve the paradox. The problem is whether there is an obligation to tell the neighbors. Consider the case in which $\neg a$ is not a premise. Ask someone who has not been influenced by deontic logic and the odds are that she says that the man should tell his neighbors that he will come. The logician has to explain why.

**Ross paradox.** Many people wonder why this paradox receives attention. Consider the derivation of $O(mail \vee burn)$ from $O(mail)$. The derived obligation $O(mail \vee burn)$ is of course different from $O(mail \wedge burn)$! Now try to give motivation for this derivation, and give motivation against it. In propositional logic, we have the derivation of $p \vee q$ from $p$. Is this an explanation why the derivation should be valid? No, this will not do! Even the analogy between deontic operator and operator for necessity is very questionable, as we discussed in Section 1.3.1. On the other hand, an argument against weakening can be given based on an analogy between obligations and imperatives. Suppose I say:

- you should buy pears,

and later I say:

- you should buy apples or pears.

Now things have changed as a consequence of my second statement. After only the first imperative, no obligations are left if they run out of apples. After the second imperative, if they run out of apples, you should buy pears! Hence, something changed after the second imperative, thus the imperative was not already implied by the first one.

**Apples-and-pears.** A new example is introduced in this thesis. The puzzle contains two premises.

- Premise 1: you should buy apples or pears,
- Premise 2: you should not buy apples.

The question of this puzzle is whether we can derive the following conclusion.

- Conclusion: you should buy pears?

The argument can be based on the analogy between obligations and imperatives. Someone says:

- you should buy apples or pears,
- well, do not buy apples.

This leaves us in a certain deontic state. If she later says:

- you should buy pears,

has anything changed?

**Dominance.** Another example discussed in this thesis contains the following two premises.

- Premise 1: you ought to do $a$ if $b$,
- Premise 2: you ought to do $a$ if not $b$.

The question of this puzzle is whether we can derive the following conclusion.

- Conclusion: you ought to do $a$ without inspecting $b$?

We can again look at the imperatives. Someone says:

- you ought to do $a$ if $b$, and
- you ought to do $a$ if not $b$.

This leaves us in a certain deontic state. If she later says:

- you ought to do $a$,

has anything changed?

The latter two derivations – apples-and-pears and dominance – seem intuitive at first sight, and they are validated by many deontic logics. Arguments against the last two derivations are given in respectively Chapter 3 and 2 of this thesis.

## 1.5 Research objectives

In this thesis we study the relation between obligations, preferences and defeasibility. Obligations are formalized by preference-based dyadic deontic logics, of which the obligations do not have unrestricted strengthening of the antecedent. Hence, preferences are used in the semantics and defeasibility is used in the proof theory of the studied deontic logics. The aims of this study are twofold. First, we study preference-based deontic logics, and we analyze where the defeasibility in these logics comes from. Second, we study defeasible deontic logics and we analyze the many faces of defeasibility in these logics. Our methodology is based on an analysis of a set of deontic puzzles. First, we introduce and discuss a deontic puzzle. Second, we introduce a deontic logic and we test this logic by its capacity of dealing with the puzzle. Thus, the puzzles serve much the same purpose as experiments in physical science.

**Objective-1.** Define a preference-based deontic logic that formalizes contrary-to-duty reasoning.

**Secondary objective.** Explain where the defeasibility of preference-based deontic logics comes from.

Background knowledge of Objective-1.

1. **Present solutions do not suffice.** The simplest (temporal) solution does not suffice, because a-temporal examples like the Cottage Regulations example cannot be represented. Chellas-type of deontic logics do not suffice, because they cannot represent contrary-to-duty obligations, as follows from the contrary-to-duty paradoxes, and semantically from the lack of varying degrees of sub-ideal worlds. Finally, Hansson-Lewis deontic logics do not suffice, because they do not have strengthening of the antecedent.

2. **Use developments in other areas of artificial intelligence.** The problem of the deontic logics based on the optimality principle is how different degrees of sub-ideality can be determined. Recently, it was observed that this aspect of violations can be formalized in non-monotonic logics [McC94a, Hor93], theories of diagnosis (see Section 5.3) or qualitative decision theories (see [Pow67, Jen74, Pea93, Bou94b, TH96, Lan96] and Section 5.2). Common to all these theories is preference-based semantics. The preference-based deontic logics developed in this thesis are inspired by the developments in these areas of artificial intelligence.

In Chapter 2 and 3 of this thesis, we develop logics that can represent violations and can deal with contrary-to-duty reasoning. A preference ordering can be used in two ways to evaluate formulas, which we call *ordering* and *minimizing*. Ordering uses all preference relations between relevant worlds, whereas minimizing uses the most preferred worlds only. B.Hansson's logic is based on the concept of minimization. Our new logics are extensions of our ordering logic ORD represented in Table 1.4. We show that ordering corresponds to strengthening of the antecedent, and minimizing to weakening of the consequent, see respectively the logics ORD and H-L in Table 1.4.

**Objective-2.** Define a preference-based defeasible deontic logic that formalizes both contrary-to-duty reasoning and overridden defeasibility.

**Secondary objective.** Find the proper distinctions between the different types of defeasibility in defeasible deontic logic to avoid confusion between violations and exceptions.

Background knowledge of Objective-2.

1. **Present solutions do not suffice.** Only a few defeasible deontic logics have been proposed. Most of these logics are extensions of dyadic deontic logics with a conditional interpretation of the antecedent [Jon93, Mak93, Pra96] and as a consequence they cannot formalize contrary-to-duty reasoning. An extension of a dyadic deontic logic with a contextual interpretation [Hor93] lacks a semantics and has several other limitations, see the discussions in [vdT94, Pra96]. Concerning the secondary objective, there are no studies of the distinction between the defeasibility in deontic and default logic, nor of the different types of defeasibility in defeasible deontic logic.

2. **No similar developments in other areas.** Defeasible deontic logic is a complex affair that presents us with problems not found in other areas. Makinson [Mak93] also observes the complexity involved in deontic logic. He compares five faces of minimality (defeasible inference, belief revision, counterfactual conditionals, updating and conditional obligation) and concludes that the latter 'face of minimality is the most complex of the five discussed.'

In Chapter 4 we give a general analysis of the problems related to the interaction between defeasibility and violability, and we discuss different ways to extend the introduced deontic logics based on the ordering logic with conflict resolution mechanisms. We argue that (at least) three types of defeasibility must be distinguished in a defeasible deontic logic. First, *factual defeasibility* formalizes overshadowing of an obligation by a violating fact. Second, *strong overridden defeasibility* formalizes cancelling of an obligation by other conditional obligations based on specificity. Third, *weak overridden defeasibility* formalizes the overriding of prima facie obligations.

## 1.6 Layout of this thesis

This thesis consists of three parts. The first part consists of Chapter 2 and 3 and studies the relation between obligations and preferences. In Chapter 2 we introduce the two-phase deontic logic 2DL. The preference-based semantics of 2DL is based on an explicit preference ordering between worlds, representing different degrees of ideality. We argue that an ideality ordering can be used in two ways to evaluate formulas, which we call *ordering* and *minimizing*. Moreover, we show that in the contrary-to-duty paradoxes ordering and minimizing have to be combined to obtain the desirable conclusions, and that in a dyadic deontic logic this can only be done in a so-called *two-phase deontic logic*. In the first phase the preference ordering is constructed, and in the second phase the ordering is used for minimization. If these two phases are not distinguished, then counterintuitive conclusions follow. In Chapter 3 we introduce contextual deontic logic CDL. A contextual obligation is written as $O(\alpha \mid \beta \setminus \gamma)$ and read as '$\alpha$ is the case if $\beta$ is the case unless $\gamma$ is the case.' An ordering obligation '$\alpha$ is the case if $\beta$ is the case' $O(\alpha \mid \beta)$ is logically equivalent to the contextual obligation $O(\alpha|\beta\setminus\bot)$. Hence, CDL is an extension of the dyadic ordering logic developed in Chapter 2. Moreover, the contextual obligations combine the properties strengthening of the antecedent and weakening of the consequent of the dyadic ordering and minimizing obligations of the two-phase deontic logic 2DL developed in Chapter 2.

The second part of this thesis studies the relation between obligations and defeasibility. In Chapter 4 we introduce extensions of contextual deontic logic CDL to formalize obligations that can be overridden by other obligations. This logic is very useful for the analysis, because it explicitly represents exceptions of the obligations. We give a general analysis of different types of defeasibility in defeasible deontic logics. We also show that these distinctions are essential for an adequate analysis of notorious contrary-to-duty paradoxes in a defeasible deontic logic. In particular, these distinctions are essential to avoid confusion between exceptions and violations.

The third part of this thesis discusses future research. This discussion illustrates the limitations of deontic logic to formalize all aspects of normative reasoning. Deontic logic restricts the analysis to reasoning *about* obligations. Notice that the obligations do not tell us anything about the actual behavior of the agents. We give two examples of reasoning *with* obligations

in Chapter 5, and we give desiderata for reasoning with obligations in Chapter 6. The two applications that can use deontic logic are qualitative decision theory and a theory of diagnosis. Qualitative decision theory uses preference-based deontic logic to formalize reasoning about context-sensitive goals. A theory of diagnosis can use deontic logic to represent system rules and violations of these system rules. We discuss reasoning with obligations, but leave the detailed study of this subject for further research. Finally, in Chapter 6 we present the conclusions.

# Chapter 2

# Two-Phase Deontic Logic

In this chapter we study the relation between obligations and preferences. Moreover, we introduce the two-phase deontic logic 2DL. The preference-based semantics of 2DL is based on an explicit preference ordering between worlds, representing different degrees of ideality. We argue that an ideality ordering can be used in two ways to evaluate formulas, which we call *ordering* and *minimizing*. Ordering uses all preference relations between relevant worlds, whereas minimizing uses the most preferred worlds only. We show that ordering corresponds to the inference pattern strengthening of the antecedent, and minimizing to the inference pattern weakening of the consequent. Moreover, we show that in the contrary-to-duty paradoxes ordering and minimizing have to be combined to obtain the desirable conclusions, and that in a dyadic deontic logic this can only be done in a so-called *two-phase deontic logic*. In the first phase the preference ordering is constructed, and in the second phase the ordering is used for minimization. If these two phases are not distinguished, then counterintuitive conclusions follow.

The first three sections of this chapter are modified and extended versions of [TvdT96] and [vdTT97c].

## 2.1 Obligations and preferences

The following example is the formalization of the Forrester paradox [For84] (see Section 1.3.3 and 1.3.5) in a dyadic deontic logic. It illustrates that combining strengthening of the antecedent and weakening of the consequent is problematic. However, both properties are desirable for a dyadic deontic logic. Strengthening of the antecedent is used to derive 'you should not kill in the morning' $O(\neg k \,|\, m)$ from the obligation 'you should not kill' $O(\neg k \,|\, \top)$ and weakening of the consequent is used to derive 'you should not kill' $O(\neg k|\top)$ from the obligation 'you should drive on the right side of the street and not kill' $O(r \wedge \neg k|\top)$.

**Example 2.1 (Forrester paradox)** Assume a dyadic deontic logic that has at least substitution of logical equivalents and the following inference patterns *Weakening of the Consequent* (WC), *Strengthening of the Antecedent* (SA), and *Conjunction* (AND).

$$\text{WC}: \frac{O(\alpha_1|\beta)}{O(\alpha_1 \vee \alpha_2|\beta)} \quad \text{SA}: \frac{O(\alpha|\beta_1)}{O(\alpha|\beta_1 \wedge \beta_2)} \quad \text{AND}: \frac{O(\alpha_1|\beta), O(\alpha_2|\beta)}{O(\alpha_1 \wedge \alpha_2|\beta)}$$

Furthermore, assume the set of dyadic obligations $S = \{O(\neg k \,|\, \top), O(k \wedge g \,|\, k)\}$ as premise set, where $k$ can be read as 'Smith kills Jones' and $k \wedge g$ as 'Smith kills Jones gently.' The

counterintuitive obligation $O(\bot \mid k)$ can be derived from $S$ by SA and AND, where $\bot$ stands for any contradiction. This paradoxical derivation from the set of obligations is represented in Figure 2.1. The derivation is blocked when SA is replaced by the following inference pattern *Restricted Strengthening of the Antecedent* (RSA), in which $\overset{\leftrightarrow}{\Diamond}$ is a modal operator and $\overset{\leftrightarrow}{\Diamond}\alpha$ is true for all consistent propositional formulas $\alpha$.

$$\text{RSA}: \frac{O(\alpha|\beta_1), \overset{\leftrightarrow}{\Diamond}(\alpha \wedge \beta_1 \wedge \beta_2)}{O(\alpha|\beta_1 \wedge \beta_2)}$$

Unfortunately, the counterintuitive obligation $O(\bot|k)$ can still be derived from $S$ by WC, RSA and AND. This paradoxical derivation from the set of obligations is also represented in Figure 2.1. Moreover, in many dyadic deontic logics the obligation $O(\bot|k)$ is inconsistent, whereas the premise set $S$ is intuitively consistent. □

$$\dfrac{\dfrac{O(\neg k|\top)}{O(\neg k|k)}\,\text{SA} \qquad O(k \wedge g|k)}{O(\neg k \wedge (k \wedge g)|k)}\,\text{AND} \qquad \dfrac{\dfrac{\dfrac{O(\neg k|\top)}{O(\neg(k \wedge g)|\top)}\,\text{WC}}{O(\neg(k \wedge g)|k)}\,\text{RSA} \qquad O(k \wedge g|k)}{O(\neg(k \wedge g) \wedge (k \wedge g)|k)}\,\text{AND}$$

Figure 2.1: The Forrester paradox

The Forrester paradox in Example 2.1 shows that combining strengthening of the antecedent and weakening of the consequent is problematic for *any* deontic logic. The underlying problem of the counterintuitive derivation in Example 2.1 is the derivation of $O(\neg(k \wedge g) \mid k)$ from the first premise $O(\neg k \mid \top)$ by WC and RSA. In this chapter we solve the Forrester paradox by a technique, which might look odd at first sight, but which turns out to work well, namely to forbid application of RSA after WC has been applied. We call this the *two-phase approach* in deontic logic. Obviously, this blocks the derivation of the obligation $O(\neg(k \wedge g)|k)$ in Figure 2.1. In the logic, the two phases are represented by two different types of obligations, a phase-1 obligation $O^c$ and a phase-2 obligation $O^c_\exists$. The two phases are linked to each other with the inference pattern REL.

$$\text{REL}: \frac{O^c(\alpha|\beta)}{O^c_\exists(\alpha|\beta)}$$

The blocked derivations are represented in Figure 2.2. Blocked derivation steps are represented by dashed lines. First of all, $O^c(\neg k \mid k)$ is not entailed by $O^c(\neg k \mid \top)$ due to the restriction in RSA. Secondly, $O^c_\exists(\neg(k \wedge g) \mid k)$ is not entailed via the obligation $O^c(\neg(k \wedge g) \mid \top)$, because in the first phase there is no weakening of the consequent. Finally, the obligation $O^c_\exists(\neg(k \wedge g)|k)$ is not entailed via $O^c_\exists(\neg(k \wedge g)|\top)$ either, because in second-phase entailment $O^c_\exists$ does not have strengthening of the antecedent.

Such a sequencing in derivations is rather unnatural and cumbersome from a proof-theoretic point of view. Surprisingly, the two-phase approach can be obtained very intuitively from a semantic point of view. In semantic terms the two-phase approach simply means that first a preference ordering has to be constructed by ordering worlds, and subsequently the constructed ordering can be used for minimization. Preference-based deontic logics are deontic logics of which

$$\dfrac{O^c(\neg k|\top)}{O^c(\neg k|k)}\ (\text{RSA}) \qquad \dfrac{\dfrac{\dfrac{O^c(\neg k|\top)}{O^c(\neg(k\wedge g)|\top)}\ (\text{WC})}{O^c(\neg(k\wedge g)|k)}\ \text{RSA}}{O^c_{\exists}(\neg(k\wedge g)|k)}\ \text{REL} \qquad \dfrac{\dfrac{\dfrac{O^c(\neg k|\top)}{O^c_{\exists}\neg k|\top)}\ \text{REL}}{O^c_{\exists}(\neg(k\wedge g)|\top)}\ \text{WC}}{O^c_{\exists}(\neg(k\wedge g)|k)}\ (\text{RSA})$$

Figure 2.2: Proof-theoretic solution of the Forrester paradox

the semantics contains a preference ordering (usually on worlds of a Kripke style possible world model). This preference ordering reflects different degrees of 'ideality': a world is preferred over another world if it is, in some sense, more ideal than the other world. For example, a numerical value can be associated with each world; in such cases, the ordering is totally connected (for all worlds $w_1$ and $w_2$ we have $w_1 \leq w_2$ or $w_2 \leq w_1$). However, in general the preference ordering can be any partial pre-ordering. Hence, only reflexivity and transitivity are assumed. In such preference orderings there can be incomparable worlds. This represents that a world can be better than another world considering some obligations, but worse when considering other obligations. For example, incomparable worlds can be used to represent dilemmas like $Op \wedge O\neg p$ in a consistent way.

In a preference-based logic, different kinds of preference relations between propositions (sets of worlds) can be derived from the preferences between worlds. The preference relations between propositions are used to formalize different kinds of obligations: $O\alpha$ is some kind of preference of $\alpha$ over $\neg\alpha$, and $O(\alpha|\beta)$ is some kind of preference of $\alpha\wedge\beta$ over $\neg\alpha\wedge\beta$. In this chapter we argue that a preference ordering can be used in two different ways to evaluate formulas. One way, which we call *minimizing*, is to use the ordering to select the minimal elements that satisfy a formula. The other way, which we call *ordering*, is to use the whole ordering to evaluate a formula. One could explain the intuition behind the distinction between ordering and minimizing with the following metaphor. Ordering is what a person does when she envisions the message of the law, issued by the legislator, by determining the preference relations between the possible deontic states. In this envisionment process bad states are as important as good states. Minimizing is what the person does when she also tries to realize the best states. These two things are completely separated. A person might very well know how she should act, without acting accordingly.

The first deontic logic based on a preference ordering was introduced by B. Hansson [Han71]. It is a dyadic logic and it belongs to the first category, because it is based on minimizing. B. Hansson's logic has been criticized because it lacks strengthening of the antecedent. For example, Alchourrón argues [Alc93] that lack of strengthening of the antecedent is acceptable for logics of defeasible reasoning or logics of defeasible obligations (see Section 1.3.6), but not for non-defeasible obligations. Moreover, the semantic concept of minimization is unexplained: whereas in a defeasible logic 'normally $p$' might refer to the most normal worlds only, 'obligatory p' does *not* seem to refer to the most ideal worlds only. Recently, several authors [Jac85, Gob90b, Han90b] introduced a preference ordering in a monadic deontic logic. These logics belongs to the second category of preference-based deontic logics, because the truth of $O\alpha$ depends on the whole ordering. This approach can be traced through a long history of research in preference logics, see e.g. [vW63, Res67, Jen74]. At first sight, it seems that an obligation $O\alpha$ can be

formalized in a preference logic as a preference of every $\alpha$ world over all $\neg\alpha$ worlds. However, it is well-known from preference logics (see e.g. [vW63]) that such a condition is much too strong. For example, consider this strong definition, two unrelated obligations $Op_1$ and $Op_2$ and a model with $p_1 \wedge \neg p_2$ and $\neg p_1 \wedge p_2$ worlds. The obligation $Op_1$ says that the first world is preferred over the second one, and the obligation $Op_2$ implies the opposite. In other words, the model cannot contain $p_1 \wedge \neg p_2$ as well as $\neg p_1 \wedge p_2$ worlds. Jackson [Jac85] and Goble [Gob90b] introduce a second ordering in the semantics as a solution of this problem, and S.O. Hansson [Han90b] introduces complicated ceteris paribus preferences. In this chapter we introduce another solution. An obligation $O\alpha$ is true iff for all $\alpha$ and $\neg\alpha$ worlds, we have that the $\alpha$ world is preferred to the $\neg\alpha$ world, or the two worlds are incomparable. Moreover, we generalize the ordering approach to the dyadic case. An obligation $O(\alpha|\beta)$ is true iff for all $\alpha \wedge \beta$ and $\neg\alpha \wedge \beta$ worlds, we have that the $\alpha \wedge \beta$ world is preferred to the $\neg\alpha \wedge \beta$ world, or the two worlds are incomparable. We show that ordering has strengthening of the antecedent, whereas minimizing has weakening of the consequent. Moreover, we show that in a two-phase deontic logic, ordering can be used as a phase-1 obligation and minimizing as a phase-2 obligation. Thus, they can be used to analyze the Forrester paradox in Example 2.1. The following example illustrates a second problem related to the formalization of $O\alpha$ as a preference of $\alpha$ over $\neg\alpha$. This second problem arises if we formalize the no-dilemma assumption in a dyadic deontic logic.

**Example 2.2 (Cigarettes problem)** Consider the three sets

$$\{O(\neg c|\top), O(c|k)\}, \{O(p_1|\top), O(p_2|\top)\}, \{O(p|q_1), O(\neg p|q_2)\}$$

The first set formalizes Prakken and Sergot's cigarettes example, when $c$ is read as 'offering someone a cigarette' and $k$ as 'killing someone.' The problem is that we want to make the first set inconsistent, but we want to keep the other two sets consistent. The derivations in Figure 2.3 show that a straightforward combination of restricted strengthening of the antecedent and the deontic axiom

$$\textbf{D*} : \neg \overset{\leftrightarrow}{\Diamond} (\alpha_1 \wedge \alpha_2 \wedge \beta) \rightarrow \neg(O(\alpha_1|\beta) \wedge O(\alpha_2|\beta))$$

makes all three sets inconsistent. □

$$\dfrac{\dfrac{O(\neg c|\top)}{O(\neg c|k)}\,\text{RSA} \qquad O(c|k)}{\bot}\,\text{D*}$$

$$\dfrac{\dfrac{O(p_1|\top)}{O(p_1|\neg(p_1 \wedge p_2))}\,\text{RSA} \qquad \dfrac{O(p_2|\top)}{O(p_2|\neg(p_1 \wedge p_2))}\,\text{RSA}}{\bot}\,\text{D*}$$

$$\dfrac{\dfrac{O(p|q_1)}{O(p|q_1 \wedge q_2)}\,\text{RSA} \qquad \dfrac{O(\neg p|q_2)}{O(\neg p|q_1 \wedge q_2)}\,\text{RSA}}{\bot}\,\text{D*}$$

Figure 2.3: Cigarettes problem

The derivations in Figure 2.3 show that there are two possible solutions to the problem in Example 2.2: weakening strengthening of the antecedent or weakening the **D*** axiom. B. Hansson's dyadic deontic logic does not have strengthening of the antecedent, and in monadic deontic logics that do not have weakening [Jac85, Gob90b, Han90b] the **D*** axiom is weakened to $\neg(O\alpha \wedge O\neg\alpha)$, hence $Op \wedge O(\neg p \wedge q)$ is consistent. In Section 2.3 we show how strengthening of the antecedent can be restricted such that this problem is solved. We call B.Hansson's minimizing logic to the rescue. A phase-1 obligation $O_D(\alpha|\beta)$ is true if and only if

1. for all $\alpha \wedge \beta$ and $\neg\alpha \wedge \beta$ worlds, we have that the $\alpha \wedge \beta$ world is preferred to the $\neg\alpha \wedge \beta$ world, or the two worlds are incomparable, *and*

2. the preferred $\beta$ worlds satisfy $\alpha$.

Hence, the formalization of the no-dilemma assumption combines ordering and minimizing in the first phase.

We use a modal preference logic to formalize our dyadic obligations. That is, the binary accessibility relation of the Kripke models of modal logic is interpreted as a preference relation. In modal logic the truth conditions of the modal sentence $\Box\alpha$ is relative to a world, whereas in preference logics a preference statement $\alpha_1 \succ \alpha_2$ is either true or false in a model. We use Boutilier's logic CT4O to represent a single preference ordering, see [Bou92a, Bou94a]. That is, for a world $w$ of a Kripke model $M$ we have $M, w \models \alpha_1 \succ \alpha_2$ if and only if $M \models \alpha_1 \succ \alpha_2$. Boutilier only shows how minimizing conditionals can be formalized in the logic CT4O. He discriminates between what we call existential-minimizing and universal-minimizing conditionals. We also show how ordering conditionals can be formalized in modal logic.

The layout of this chapter is as follows. In Section 2.2 we introduce the dyadic ordering obligation and the dyadic existential-minimizing obligation, and we illustrate the combining of ordering and minimizing. We call the modal preference logic with the definitions of the different types of conditionals our two-phase deontic logic 2DL. Moreover, we analyze the Forrester paradox in 2DL. In Section 2.3 we extend the two-phase deontic logic 2DL by introducing operators that have the no-dilemma assumption, and we analyze the cigarettes problem. In Section 2.4 we introduce different types of dyadic permissions. Finally, in Section 2.5 we discuss the issue of factual detachment, i.e. the detachment of unconditional obligations from the dyadic ones.

## 2.2 The two-phase approach to deontic logic

In this section we introduce the preference-based deontic logic 2DL. We formalize ordering and minimizing obligations in a preference logic. This preference logic is a standard modal system, in which the accessibility relation of the Kripke models is interpreted as a preference 'ideality' ordering on the worlds. Moreover, we analyze the Forrester paradox in 2DL.

### 2.2.1 The logic CT4O

In this chapter, dyadic obligations are formalized in Boutilier's logic CT4O, a bimodal propositional logic of inaccessible worlds. We refer to the modal logic with different types of deontic conditionals as the two-phase deontic logic 2DL. The Kripke models $M = \langle W, \leq, V \rangle$

of CT4O contain a binary accessibility relation $\leq$, that is interpreted as a reflexive and transitive *preference relation*. As is well-known, the standard system S4 is characterized by a partial pre-ordering: the axiom **T**: $\Box\alpha \rightarrow \alpha$ characterizes reflexivity and the axiom **4**: $\Box\alpha \rightarrow \Box\Box\alpha$ characterizes transitivity [HC84, Che80]. The logic CT4 is equivalent to the standard system S4, together with the following definition **C**.[1] The conditional $\beta \Rightarrow \alpha$ is true in a world if $\alpha$ is true in all the $\leq$-minimal $\beta$ worlds (that are accessible from the actual world). Hence, the dyadic conditional is formalized in terms of the monadic operator. We discuss this minimizing conditional in more detail in Section 2.3.

$$\textbf{C}: \quad (\beta \Rightarrow \alpha) \quad =_{def} \quad \Box(\beta \rightarrow \Diamond(\beta \wedge \Box(\beta \rightarrow \alpha)))$$

Boutilier refers to CT4 rather than S4 to emphasize his interest in the conditional aspect of the logic, though he remarks that it should be kept in mind that it is just S4. Boutilier's logic CT4O extends the modal system S4 by introducing a modal operator that refers to inaccessible worlds.[2] The Kripke models of the bimodal logic CT4O can be written as $M = \langle W, \leq, R_2, V\rangle$, i.e. with two accessibility relations for the two modal operators.

$$M, w \models \Box\alpha \text{ iff } \forall w' \in W \text{ if } w' \leq w, \text{ then } M, w' \models \alpha$$
$$M, w \models \overleftarrow{\Box}\alpha \text{ iff } \forall w' \in W \text{ if } w' R_2 w, \text{ then } M, w' \models \alpha$$

Moreover, the relation $R_2$ is the complement of $\leq$, which is expressed by the condition $w_1 R_2 w_2$ iff not $w_1 \leq w_2$. As a consequence of this condition, the truth conditions of the modal operator for inaccessible worlds can also be expressed in terms of the first accessibility relation.

$$M, w \models \Box\alpha \text{ iff } \forall w' \in W \text{ if } w' \leq w, \text{ then } M, w' \models \alpha$$
$$M, w \models \overleftarrow{\Box}\alpha \text{ iff } \forall w' \in W \text{ if } w' \not\leq w, \text{ then } M, w' \models \alpha$$

Boutilier adapts Humberstone's logic [Hum83] of inaccessible worlds for partial pre-orderings and he gives a finite axiomatization. The O of CT4O stands for 'only knowing', which can be expressed in logics of inaccessible worlds. The crucial axiom of the logic CT4O is axiom **H**: $\overleftrightarrow{\Diamond}(\Box\alpha \wedge \overleftarrow{\Box}\beta) \rightarrow \overleftrightarrow{\Box}(\alpha \vee \beta)$, which is an instance of the more general *Humberstone Schemata* $\textbf{H}^*$.

$$\textbf{H}^*: \quad D(\Box\alpha \wedge \overleftarrow{\Box}\beta) \rightarrow B(\alpha \vee \beta)$$

In this schema, $D$ is any sequence of the connectives $\Diamond$ and $\overleftarrow{\Diamond}$ having length $\geq 0$, and $B$ is any such sequence of $\Box$ and $\overleftarrow{\Box}$. In CT4O, the axiom **H** derives all instances of this schema. The axiom **H** axiomatizes the condition $w_1 R_2 w_2$ iff not $w_1 \leq w_2$. For the details and completeness

[1] Boutilier uses the equivalent formula $\Box(\Box\neg\beta \vee \Diamond(\beta \wedge \Box(\beta \rightarrow \alpha)))$. The definition used here is from Makinson, and was independently discovered by Lamarre [Lam91].

[2] We do not need the expressivity of inaccessible worlds to model ordering obligations (although it makes the logic more elegant). We refer to Boutilier's logic CT4O, because we also use the modal system as a preference structure, and we also code our conditionals with monadic modal operators. However, we repeat, the ordering obligations itself are completely different from Boutilier's minimizing conditionals. We use the expressivity of inaccessible worlds when we extend the logic to represent permissions in Section 2.4.

proof of Boutilier's logic CT4O see [Bou92a, Bou94a].

**Definition 2.3 (CT4O)** The bimodal language $\mathcal{L}$ is formed from a denumerable set of propositional variables together with the connectives $\neg$, $\rightarrow$, and the two normal modal connectives $\Box$ and $\overleftarrow{\Box}$. Dual 'possibility' connectives $\Diamond$ and $\overleftarrow{\Diamond}$ are defined as usual and two additional modal connectives $\overset{\leftrightarrow}{\Box}$ and $\overset{\leftrightarrow}{\Diamond}$ are defined as follows.

$$\Diamond\alpha \;=_{def}\; \neg\,\Box\,\neg\alpha \qquad\qquad \overset{\leftrightarrow}{\Box}\alpha \;=_{def}\; \Box\alpha \wedge \overleftarrow{\Box}\alpha$$
$$\overleftarrow{\Diamond}\alpha \;=_{def}\; \neg\,\overleftarrow{\Box}\neg\alpha \qquad\qquad \overset{\leftrightarrow}{\Diamond}\alpha \;=_{def}\; \Diamond\alpha \vee \overleftarrow{\Diamond}\alpha$$

The logic CT4O is the smallest $S \subset \mathcal{L}$ such that $S$ contains classical logic and the following axiom schemata, and is closed under the following rules of inference.

| | | | |
|---|---|---|---|
| **K** | $\Box(\alpha \rightarrow \beta) \rightarrow (\Box\alpha \rightarrow \Box\beta)$ | **Nes** | From $\alpha$ infer $\overset{\leftrightarrow}{\Box}\alpha$ |
| **K′** | $\overleftarrow{\Box}(\alpha \rightarrow \beta) \rightarrow (\overleftarrow{\Box}\alpha \rightarrow \overleftarrow{\Box}\beta)$ | **MP** | From $\alpha \rightarrow \beta$ and $\alpha$ infer $\beta$ |
| **T** | $\Box\alpha \rightarrow \alpha$ | | |
| **4** | $\Box\alpha \rightarrow \Box\Box\alpha$ | | |
| **H** | $\overset{\leftrightarrow}{\Diamond}(\Box\alpha \wedge \overleftarrow{\Box}\beta) \rightarrow \overset{\leftrightarrow}{\Box}(\alpha \vee \beta)$ | | $\Box$ |

**Definition 2.4 (CT4O Semantics)** Kripke models $M = \langle W, \leq, V\rangle$ for CT4O consist of $W$, a set of worlds, $\leq$, a binary transitive and reflexive accessibility relation, and $V$, a valuation of the propositional atoms in the worlds. The partial pre-ordering $\leq$ expresses preferences: $w_1 \leq w_2$ iff $w_1$ is as preferable as $w_2$. The modal connective $\Box$ refers to accessible worlds and the modal connective $\overleftarrow{\Box}$ to inaccessible worlds.

$M, w \models \Box\alpha$ iff $\forall w' \in W$ if $w' \leq w$, then $M, w' \models \alpha$
$M, w \models \overleftarrow{\Box}\alpha$ iff $\forall w' \in W$ if $w' \not\leq w$, then $M, w' \models \alpha$ $\Box$

The following satisfiability conditions for the modal connectives $\overset{\leftrightarrow}{\Box}$ and $\overset{\leftrightarrow}{\Diamond}$ follow immediately from the definitions.

$M, w \models \overset{\leftrightarrow}{\Box}\alpha$ iff $\forall w' \in W$ we have $M, w' \models \alpha$ (i.e. iff $M \models \alpha$)
$M, w \models \overset{\leftrightarrow}{\Diamond}\alpha$ iff $\exists w' \in W$ such that $M, w' \models \alpha$

From the preferences between worlds we derive several preference relations between sets of worlds, i.e. preference relations between *propositions*. Dyadic obligations are defined in terms of these preferences between propositions. We refer to the logic CT4O extended with the new definitions for ordering and minimizing obligations as the logic 2DL.

### 2.2.2 Ordering

In this subsection we only consider the ordering approach to deontic logic. In evaluating formulas, the whole ordering is taken into account. We first give the definitions of the ordering obligations in the modal preference logic, the semantic truth conditions and several properties of the

ordering obligations expressed as theorems of the modal logic. Then we show some adaptations of the definitions to block two counterintuitive theorems.

The ordering obligations are defined in two steps. First, we define a preference ordering on propositions. We write $\alpha_1 \succ \alpha_2$ for $\alpha_1$ is preferred to $\alpha_2$. We use a preference ordering to define obligations $O(\alpha|\beta)$ by a preference of $\alpha \wedge \beta$ over $\neg\alpha \wedge \beta$. We say that a preference ordering $\succ$ formalizes strong preferences when a preference $\alpha_1 \succ \alpha_2$ logically implies $\alpha'_1 \succ \alpha'_2$ when $\alpha'_1$ implies $\alpha_1$ and $\alpha'_2$ implies $\alpha_2$. We call them weak preferences otherwise. Ordering obligations are defined by strong preferences, and minimizing obligations (like the conditionals defined by definition **C** in CT4O) by weak preferences.

For the strong preferences defined in this section we write $\alpha_1 \succ_s \alpha_2$. In the introduction in Section 2.1 we discussed the well-known problem from preference logic that an ordering $\alpha_1 \succ_s \alpha_2$ as a preference of each $\alpha_1$ world over every $\alpha_2$ world is much too strong. For example, consider this strong definition, two unrelated preferences $p_1 \succ_s \neg p_1$ and $p_2 \succ_s \neg p_2$ and a model with $p_1 \wedge \neg p_2$ and $\neg p_1 \wedge p_2$ worlds. The preference $p_1 \succ_s \neg p_1$ says that the first world is preferred to the second one, and the preference $p_2 \succ_s \neg p_2$ implies the opposite. With other words, the model cannot contain $p_1 \wedge \neg p_2$ as well as $\neg p_1 \wedge p_2$ worlds. This property is highly counterintuitive and not very useful. We therefore introduce a weaker notion of strong preference. According to this definition $\alpha_1$ is preferred to $\alpha_2$ if and only if for every $\alpha_1$ world there is not an $\alpha_2$ world that is as preferable. That is, for each pair of $\alpha_1$ and $\alpha_2$ worlds we have either that the $\alpha_1$ world is preferred to the $\alpha_2$ world, or that the two are incomparable.

**Definition 2.5 (Dyadic ordering obligation)** The dyadic ordering obligation '$\alpha$ should be the case if $\beta$ is the case', written as $O(\alpha|\beta)$, is defined as a strong preference of $\alpha \wedge \beta$ over $\neg\alpha \wedge \beta$. A strong preference of $\alpha_1$ over $\alpha_2$, written as $\alpha_1 \succ_s \alpha_2$, is defined as follows.

$$\begin{array}{lll}
\alpha_1 \succ_s \alpha_2 & =_{def} & \overset{\leftrightarrow}{\Box}(\alpha_1 \to \Box\neg\alpha_2) \\
O(\alpha|\beta) & =_{def} & (\alpha \wedge \beta) \succ_s (\neg\alpha \wedge \beta) \\
 & = & \overset{\leftrightarrow}{\Box}((\alpha \wedge \beta) \to \Box\neg(\neg\alpha \wedge \beta)) \\
 & \leftrightarrow & \overset{\leftrightarrow}{\Box}((\alpha \wedge \beta) \to \Box(\beta \to \alpha))
\end{array}$$

□

The preference relation $\succ_s$ is quite weak. For example, it is not anti-symmetric (we cannot derive $\neg(\alpha_2 \succ_s \alpha_1)$ from $\alpha_1 \succ_s \alpha_2$) and it is not transitive (we cannot derive $\alpha_1 \succ_s \alpha_3$ from $\alpha_1 \succ_s \alpha_2$ and $\alpha_2 \succ_s \alpha_3$). The lack of these properties is the result of the fact that we do not have totally connected orderings. In Example 2.40 in Section 2.3.3 we show that this relaxation is crucial for our preference-based deontic logic. In this section we do not further discuss the properties of $\succ_s$ (see Section 2.6.4), but we focus on the properties of the dyadic ordering obligations. Intuitively, an obligation $O(\alpha|\beta)$ expresses a strict preference of all $\alpha \wedge \beta$ over $\neg\alpha \wedge \beta$. The following proposition shows that this preference is represented by the 'negative' condition that $\neg\alpha \wedge \beta$ is not as preferable as $\alpha \wedge \beta$.[3]

[3]We can also define a conditional by $O(\alpha|\beta) =_{def} \Box((\alpha \wedge \beta) \to \Box(\beta \to \alpha))$. The logic has the same properties, in the sense that Proposition 2.8 is still true. The distinction is that the conditional can have different values at distinct worlds, which additional expressivity does not seem very useful. Single preference orderings are also used in standard model preference semantics [Sho88, KLM90].

Alternatively, we can take the dyadic ordering obligation as primitive, defined by the semantic definition in Proposition 2.6, or we can take the preference ordering $\succ_s$ as primitive. In the latter case, we can define the monadic

**Proposition 2.6** *Let $M = \langle W, \leq, V \rangle$ be a CT4O model, $\mid \alpha \mid$ be the set of worlds of $W$ that satisfy $\alpha$, and $|\alpha_1| \not\leq |\alpha_2|$ denote that $\forall w_1 \in |\alpha_1|$ and $\forall w_2 \in |\alpha_2|$, we have $w_1 \not\leq w_2$. For a world $w \in W$, we have $M, w \models O(\alpha|\beta)$ iff ($M \models O(\alpha|\beta)$ iff) $|\neg\alpha \wedge \beta| \not\leq |\alpha \wedge \beta|$.*

**Proof** $\Rightarrow$ *By contraposition. Assume a model $M = \langle W, \leq, V \rangle$ with worlds $w_1, w_2 \in W$ such that $M, w_1 \models \neg\alpha \wedge \beta$, $M, w_2 \models \alpha \wedge \beta$ and $w_1 \leq w_2$. We have $M, w_2 \not\models (\alpha \wedge \beta) \rightarrow \Box(\beta \rightarrow \alpha))$. $M, w \overset{\leftrightarrow}{\models} \Box\ \alpha$ for a world $w \in W$ iff for* all *worlds $w' \in W$ we have $M, w' \models \alpha$. Hence, $M, w \not\models O(\alpha|\beta)$.*

$\Leftarrow$ *Assume $M, w \not\models O(\alpha|\beta)$ for some world $w$. Hence, there is a world $w_2 \in W$ such that $M, w_2 \not\models (\alpha \wedge \beta) \rightarrow \Box(\beta \rightarrow \alpha))$. It follows that $M, w_2 \models \alpha \wedge \beta$ and $M, w_2 \not\models \Box(\beta \rightarrow \alpha)$. Hence, there is a world $w_1 \in W$ such that $M, w_1 \models \neg\alpha \wedge \beta$ and $w_1 \leq w_2$.*

The following example illustrates the definition of ordering obligation as a strong preference.

**Example 2.7** Consider the Kripke model $M$ represented in Figure 2.4 below. The figure should be read as follows. A circle represents a non-empty set of worlds, which satisfy the formulas in the circle. The arrows represent strict preferences for all worlds in the circles (the transitive closure is implicit). We have $M \models O(p|\top)$ and $M \not\models O(q|\top)$. Note that $O(q|\top)$ is not true in $M$, because $|p \wedge \neg q| \leq |\neg p \wedge q|$ and $|\neg p \wedge \neg q| \leq |\neg p \wedge q|$. This shows how in the ordering approach the whole ordering is taken into account in the evaluation of a formula, and not just the most preferred $p \wedge q$ worlds. $\Box$

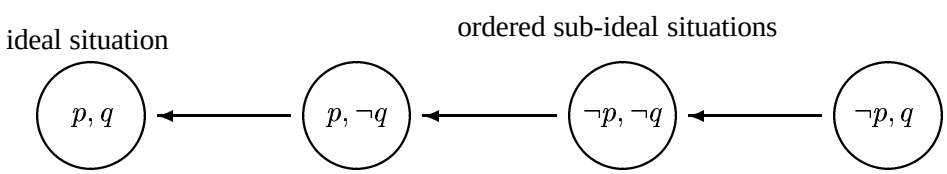



Figure 2.4: Preference-based model

The following proposition shows several properties of the dyadic ordering obligations.

**Proposition 2.8** *The logic* 2DL *has the following theorems of Strengthening of the Antecedent (***SA***), Conjunction (***AND***) and Disjunction (***OR***), and two versions of Deontic Detachment (***DD′*** and* ***DD−****).*

**SA**: $O(\alpha|\beta_1) \rightarrow O(\alpha|\beta_1 \wedge \beta_2)$
**AND**: $(O(\alpha_1|\beta) \wedge O(\alpha_2|\beta)) \rightarrow O(\alpha_1 \wedge \alpha_2|\beta)$
**OR**: $(O(\alpha_1|\beta) \wedge O(\alpha_2|\beta)) \rightarrow O(\alpha_1 \vee \alpha_2|\beta)$
**DD′**: $(O(\alpha|\beta) \wedge O(\beta|\gamma)) \rightarrow O(\alpha \wedge \beta|\gamma)$
**DD−**: $(O(\alpha|\beta) \wedge O(\neg\beta|\gamma)) \rightarrow O((\alpha \wedge \beta) \vee \neg\beta|\gamma)$

operator $\Box$ in terms of $\succ_s$ by $\Box\alpha =_{def} \neg\alpha \succ_s \top$. Analogous definitions of unary modalities in terms of minimizing conditionals $\beta \Rightarrow \alpha$ by $\Box\alpha =_{def} \neg\alpha \Rightarrow \alpha$ are well-known, see e.g. [Sta81, Lew73], and an analogous grounding of the logic CT4 (hence S4) in a minimizing conditional can be found in [Bou92a, p.89].

*The logic* 2DL *does* not *have the following theorems of Weakening of the Consequent (***WC** *and* **WC′***), Deontic Detachment (or deontic transitivity) (***DD***), Reasoning By Cases (***RBC***) and the second Deontic axiom (***D****).*

**WC***:* $O(\alpha_1|\beta) \to O(\alpha_1 \vee \alpha_2|\beta)$
**WC′***:* $O(\alpha_1 \wedge \alpha_2|\beta) \to O(\alpha_1|\beta) \wedge O(\alpha_2|\beta)$
**DD***:* $(O(\alpha|\beta) \wedge O(\beta|\gamma)) \to O(\alpha|\gamma)$
**DD⊤***:* $(O(\alpha|\beta) \wedge O(\beta|\top)) \to O(\alpha|\top)$
**RBC***:* $(O(\alpha|\beta_1) \wedge O(\alpha|\beta_2)) \to O(\alpha|\beta_1 \vee \beta_2)$
**D****:* $\neg(O(\alpha|\beta) \wedge O(\neg\alpha|\beta))$

**Proof** *The (non)theorems can be proven by proving (un)satisfiability in the preference-based semantics. First, consider the validity of strengthening of the antecedent* **SA***. The validity of strengthening $O(\alpha|\beta_1)$ to $O(\alpha|\beta_1 \wedge \beta_2)$ follows directly from the fact that a strong preference of $\alpha \wedge \beta_1$ over $\neg\alpha \wedge \beta_1$ implies a strong preference of $\alpha \wedge \beta_1 \wedge \beta_2$ over $\neg\alpha \wedge \beta_1 \wedge \beta_2$. Secondly, consider the non-theorem* **WC***. $O(\alpha_1|\beta)$ is not weakened to $O(\alpha_1 \vee \alpha_2|\beta)$, because $O(\alpha_1|\beta)$ expresses a preference of* every *$\alpha_1 \wedge \beta$ worlds over any $\neg\alpha_1 \wedge \beta$ world, and from such a preference does not follow that every $(\alpha_1 \vee \alpha_2) \wedge \beta$ world is preferred to any $\neg\alpha_1 \wedge \neg\alpha_2 \wedge \beta$ world. For a counterexample, consider the preference-based model $M$ in Figure 2.4. We have $M \models O(p|\top)$ and $M \not\models O(p \vee q|\top)$, because $|\neg p \wedge \neg q| \leq |\neg p \wedge q|$. Hence, the ordering obligations do not have weakening of the consequent. Verification of the other (non)theorems is left to the reader. Alternatively, the theorems can be proven in the logic* 2DL*.*

The ordering obligations have several remarkable properties. The most remarkable are the non-validity of weakening of the consequent **WC** and reasoning by cases **RBC**. The following example illustrates that the lack of **RBC** is very useful to analyze dominance arguments, see [TH96]. A common sense dominance argument (1) divides possible outcomes into two or more exhaustive, exclusive cases, (2) points out that in each of these alternatives it is better to perform some action than not to perform it, and (3) concludes that this action is best unconditionally. Thomason and Horty observe that, although such arguments are often used, and are convincing when they are used, they are invalid. The following example of [TH96] is a classic illustration of [Jef83].

**Example 2.9 (Cold-war disarmament)** Either there will be a nuclear war or there will not. If there will not be a nuclear war, then it is better for us to disarm because armament is expensive and pointless. If there will be a nuclear war, then we will be dead whether or not we arm, so we are better of saving money in the short term by disarming. So, we should disarm. The fallacy, of course, depends on the assumption that the action of choosing whether to arm or disarm will have no effect on whether there is war or not.

Consider the contextual obligations $O(d|w)$ and $O(d|\neg w)$, which represent that we ought to be disarmed if there will be a nuclear war, and we ought to be disarmed if there will be no war. We cannot derive $O(d|w \vee \neg w)$, because from $(d \wedge w) \succ_s (\neg d \wedge w)$ and $(d \wedge \neg w) \succ_s (\neg d \wedge \neg w)$ we cannot derive $(d \wedge \top) \succ_s (\neg d \wedge \top)$. We might have $(\neg d \wedge \neg w) \succ_s (d \wedge w)$, which represents that we ought to be armed if we have peace if and only if we are armed $O(\neg d|d \leftrightarrow w)$. □

Another remarkable property is the validity of **DD′**. The inference pattern deontic detachment is often considered problematic, in particular as a result of the Chisholm paradox (see Section 1.3.3). The following example illustrates the intuition behind theorem **DD′**.

**Example 2.10** Consider the set of dyadic obligations $S = \{O(a\,|\,\top), O(t\,|\,a)\}$, where $a$ can be read as 'a certain man going to the assistance of his neighbors' and $t$ as 'telling the neighbors that he will come.' Hence, the two obligations can be read as 'a certain man should go to the assistance of his neighbors', and 'he should tell them he is coming, if he goes.' The obligation $O(a \wedge t\,|\,\top)$ can be derived from $S$ with **DD′**, which expresses that ideally, the man goes to the assistance of his neighbors *and* he tells them he is coming. The validity of the inference can be explained by the preference-based semantics. A typical model $M$ of $S$ is given in Figure 2.5 below, and we have $M \models O(a \wedge t\,|\,\top)$. The ideal situation is represented by $a \wedge t$. There are two different ways to deviate from the ideal. The first way is $\neg a$, where the man does not go to the assistance (regardless whether he tells that he will go). The second way is $a \wedge \neg t$, where the man goes to the assistance but he does not tell his neighbors that he is coming. □

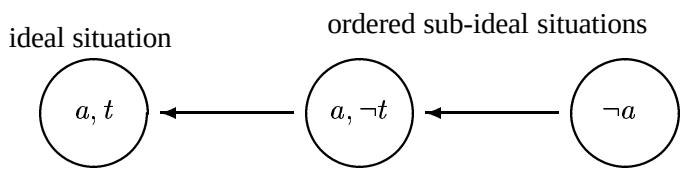



Figure 2.5: Assistance of neighbors

In the remainder of this section we discuss two ways to adapt the ordering logic, with consistency checks and the so-called axiom scheme **LP** respectively. Unfortunately, the logic 2DL has the two counterintuitive theorems $O(\bot \mid \alpha)$ and $O(\alpha \mid \alpha)$, see Proposition 2.12. For this reason, we define various other types of obligations in the preference logic. In the following definition, $O^c(\alpha|\beta)$ has an additional condition that tests whether the obligation can be fulfilled, i.e. whether $\alpha \wedge \beta$ is logically possible ('ought implies can'). The obligation $O^{cc}(\alpha \mid \beta)$ also has another additional condition which tests whether the obligation can be violated, i.e. whether $\neg\alpha \wedge \beta$ is possible.[4] The two conditions formalize von Wright's contingency principle (see Section 1.3). The consistency conditions are based on the concept of choice: if it is not possible to violate or fulfill the obligation, then there is no possibility to choose.

**Definition 2.11 (Dyadic ordering obligation)** Two alternative notions of dyadic ordering obligations '$\alpha$ should be the case if $\beta$ is the case,' written as $O^c(\alpha\,|\,\beta)$ and $O^{cc}(\alpha\,|\,\beta)$ respectively, are defined as a strong preference of $\alpha \wedge \beta$ over $\neg\alpha \wedge \beta$ together with one or two 'consistency checks.'

[4]The conditions only check logical possibility. In an agent environment, the alternatives are to consider stronger conditions which refer to the agent's opportunities or to her abilities. The logical conditions are already stronger than necessary to invalidate the counterintuitive theorems, because the consistency conditions $\overset{\leftrightarrow}{\Diamond}\alpha$ and $\overset{\leftrightarrow}{\Diamond}\neg\alpha$ would (in principle) also do the trick.

$$
\begin{array}{lll}
O^c(\alpha|\beta) & =_{def} & (\alpha \wedge \beta) \succ_s (\neg\alpha \wedge \beta) \wedge \overset{\leftrightarrow}{\Diamond}(\alpha \wedge \beta) \\
 & = & O(\alpha|\beta) \wedge \overset{\leftrightarrow}{\Diamond}(\alpha \wedge \beta) \\
O^{cc}(\alpha|\beta) & =_{def} & (\alpha \wedge \beta) \succ_s (\neg\alpha \wedge \beta) \wedge \overset{\leftrightarrow}{\Diamond}(\alpha \wedge \beta) \wedge \overset{\leftrightarrow}{\Diamond}(\neg\alpha \wedge \beta) \\
 & = & O(\alpha|\beta) \wedge \overset{\leftrightarrow}{\Diamond}(\alpha \wedge \beta) \wedge \overset{\leftrightarrow}{\Diamond}(\neg\alpha \wedge \beta)
\end{array}
$$

□

The following proposition shows that the alternative definitions of dyadic ordering obligations $O^c$ and $O^{cc}$ do not have the counterintuitive theorems of $O$.

**Proposition 2.12** *The logic* 2DL *has the following theorems.*

| | |
|---|---|
| **C***:* | $O(\bot|\alpha)$ |
| **NC**$^c$*:* | $\neg O^c(\bot|\alpha)$ |
| **NC**$^{cc}$*:* | $\neg O^{cc}(\bot|\alpha)$ |
| **Id***:* | $O(\alpha|\alpha)$ |
| **Id**$^c$*:* | $\overset{\leftrightarrow}{\Diamond}\alpha \rightarrow O^c(\alpha|\alpha)$ |
| **NId**$^{cc}$*:* | $\neg O^{cc}(\alpha|\alpha)$ |

**Proof** *Follows directly from Definition 2.5 and 2.11.*

The following proposition shows that the ordering obligations $O^c$ have weaker versions of the theorems of $O$ given in Proposition 2.8. We say that the corresponding inference patterns are restricted. For example, we call the weakened version of **SA** *Restricted Strengthening of the Antecedent* **RSA**. We already saw an example of restricted strengthening of the antecedent in Example 2.1, where a restriction of SA was necessary to block counterintuitive derivations of the Forrester paradox. Obviously, similar results as shown in the following proposition can be obtained for $O^{cc}$.

**Proposition 2.13** *The logic* 2DL *has the following theorems.*

| | |
|---|---|
| **RSA***:* | $(O^c(\alpha|\beta_1) \wedge \overset{\leftrightarrow}{\Diamond}(\alpha \wedge \beta_1 \wedge \beta_2)) \rightarrow O^c(\alpha|\beta_1 \wedge \beta_2)$ |
| **RAND***:* | $(O^c(\alpha_1|\beta) \wedge O^c(\alpha_2|\beta) \wedge \overset{\leftrightarrow}{\Diamond}(\alpha_1 \wedge \alpha_2 \wedge \beta)) \rightarrow O^c(\alpha_1 \wedge \alpha_2|\beta)$ |
| **OR**$^c$*:* | $(O^c(\alpha_1|\beta) \wedge O^c(\alpha_2|\beta)) \rightarrow O^c(\alpha_1 \vee \alpha_2|\beta)$ |
| **RDD**$'$*:* | $(O^c(\alpha|\beta) \wedge O^c(\beta|\gamma) \wedge \overset{\leftrightarrow}{\Diamond}(\alpha \wedge \beta \wedge \gamma)) \rightarrow O^c(\alpha \wedge \beta|\gamma)$ |

**Proof** *Follows directly from Definition 2.11 and Proposition 2.8.*

Proposition 2.13 shows that we cannot derive $O^c(\alpha \mid \beta)$ from $O^c(\alpha \mid \top)$ by **RSA**, unless we have the consistency expression $\overset{\leftrightarrow}{\Diamond}(\alpha \wedge \beta)$ as another premise. Instead of explicitly writing down these consistency expressions in every example, we can consider only models in which all propositionally satisfiable formulas $\alpha$ are true in *some* world. This can be 'axiomatized' with Boutilier's axiom scheme **LP**, see [Bou94a] for a discussion and the completeness proof of the corresponding logic CT4O*. The axiom scheme **LP** states that every formula $\alpha$ without any occurrences of modal operators, which is propositionally satisfiable, is true in some world.

**Definition 2.14 (CT4O*)** The logic CT4O* is CT4O extended with the following axiom scheme **LP**.

**LP**: $\overset{\leftrightarrow}{\Diamond}\alpha$ for all satisfiable propositional $\alpha$ □

**Definition 2.15 (Semantics CT4O*)** Let $\mathcal{P}$ be the set of propositional atoms of the propositional base language $\mathcal{L}$. A CT4O*-model is a CT4O-model $M = \langle W, \leq, V\rangle$ that satisfies the following condition:

$$\{f \mid f \text{ maps } \mathcal{P} \text{ into} \{0,1\}\} \subseteq \{V(w) \mid w \in W\}$$

We write $\models^*$ for logical entailment in CT4O*. □

We write 2DL* for CT4O* with the definitions of our dyadic obligations. The logic 2DL* is illustrated by the following example.

**Example 2.16** Consider the set of ordering obligations $S = \{O^c(p_1 \mid \top), O^c(p_2 \mid \top)\}$. Semantically, the axiom **LP** ensures that the $p_1 \wedge p_2$ worlds exist in all 2DL* models. Hence, we have $\models^* \overset{\leftrightarrow}{\Diamond}(p_1 \wedge p_2)$ whereas we have $\not\models \overset{\leftrightarrow}{\Diamond}(p_1 \wedge p_2)$. Proof-theoretically, in 2DL we can derive $O^c(p_1 \wedge p_2 | \top)$ from $S$ and the premise $\overset{\leftrightarrow}{\Diamond}(p_1 \wedge p_2)$ by **RAND**. In 2DL* the consistency expression $\overset{\leftrightarrow}{\Diamond}(p_1 \wedge p_2)$ can be derived from **LP**, and hence $O^c(p_1 \wedge p_2 | \top)$ can be derived from $S$. This shows that we do not have to write the consistency expressions explicitly in the logic 2DL*. □

In this section we introduced a logic of ordering obligations. In Table 2.1 we compare the new logic with Chellas-type of dyadic deontic logics and the Hansson-Lewis minimizing dyadic deontic logics.

| | | Condition | Context | |
|---|---|---|---|---|
| | | Chellas | H-L | ORD |
| **WC** | $O(\alpha_1\|\beta) \to O(\alpha_1 \vee \alpha_2\|\beta)$ | X | X | |
| **SA** | $O(\alpha\|\beta_1) \to O(\alpha\|\beta_1 \wedge \beta_2)$ | X | | |
| **RSA** | $(O(\alpha\|\beta_1) \wedge \overset{\leftrightarrow}{\Diamond}(\alpha \wedge \beta_1 \wedge \beta_2)) \to O(\alpha\|\beta_1 \wedge \beta_2)$ | | | X |
| **FD** | $(O(\alpha\|\beta) \wedge \beta) \to O\alpha$ | X | | |
| **DD** | $(O(\alpha\|\beta) \wedge O(\beta\|\gamma)) \to O(\alpha\|\gamma)$ | | | |
| **DD**$\top$ | $(O(\alpha\|\beta) \wedge O(\beta\|\top)) \to O(\alpha\|\top)$ | | X | |
| **DD**$'$ | $(O(\alpha\|\beta) \wedge O(\beta\|\gamma)) \to O(\alpha \wedge \beta\|\gamma)$ | | | R |
| **RBC** | $(O(\alpha\|\beta_1) \wedge O(\alpha\|\beta_2)) \to O(\alpha\|\beta_1 \vee \beta_2)$ | X | X | |

Table 2.1: Ordering obligations versus classical dyadic deontic logics

The table shows that the ordering logic has several intuitive properties like strengthening of the antecedent and a version of deontic detachment. This logic combines a contextual interpretation of the antecedent with strengthening of the antecedent. Moreover, we showed that lack of weakening of the consequent and reasoning by cases are sometimes advantageous properties.

Finally, we showed that the logic can easily be adapted to block two counterintuitive properties by adding the consistency checks to the operator. In the next section we discuss 'standard' minimizing obligations and we compare them with the ordering obligations.

### 2.2.3 Minimizing

In this section we introduce a minimizing logic. We first give the definition in modal logic, the semantic truth conditions and several properties expressed as theorems of the modal logic. Then we give three relations between these minimizing obligations and the ordering obligations introduced in the previous section.

In the dyadic deontic logic of Bengt Hansson, an obligation $O(\alpha|\beta)$ is true iff $\alpha$ is true in all minimal (preferred) $\beta$ worlds [Han71]. We therefore say that his logic is based on minimizing. In Section 2.3.2, we discuss Boutilier's reconstruction [Bou94b] of B. Hansson's logic in a modal preference structure. In this section we give a related but weaker logic, in which an obligation $O_\exists(\alpha|\beta)$ is true if and only if $\alpha$ is true in an *equivalence class* of minimal (preferred) $\beta$ worlds.[5] To discriminate between the two types of minimizing conditionals we call the Hansson-Lewis type universal-minimizing and the weaker types discussed in this section existential-minimizing.

The existential-minimizing obligation is defined in a weak preference ordering, written as $\alpha_1 \succ_\exists \alpha_2$. As we discussed in the previous section, we say that a preference ordering $\succ$ formalizes weak preferences when a preference $\alpha_1 \succ \alpha_2$ does not logically imply a preference for $\alpha_1' \succ \alpha_2'$ when $\alpha_1'$ implies $\alpha_1$ and $\alpha_2'$ implies $\alpha_2$. We say that $\alpha_1$ is weakly preferred to $\alpha_2$ iff there is an $\alpha_1$ world such that there is no $\alpha_2$ world which is as preferable. That is, there is a preferred $\alpha_1$ world such that for all preferred $\alpha_2$ worlds we have either that the $\alpha_1$ world is preferred to the $\alpha_2$ world, or that the two worlds are incomparable.

**Definition 2.17 (Dyadic existential-minimizing obligation)** The dyadic existential-minimizing obligation '$\alpha$ should be the case if $\beta$ is the case', written as $O_\exists(\alpha|\beta)$, is defined as a weak preference of $\alpha \wedge \beta$ over $\neg\alpha \wedge \beta$. A weak preference of $\alpha_1$ over $\alpha_2$, written as $\alpha_1 \succ_\exists \alpha_2$, is defined as follows.

$$
\begin{array}{lll}
\alpha_1 \succ_\exists \alpha_2 & =_{def} & \overset{\leftrightarrow}{\Diamond}(\alpha_1 \wedge \Box\neg\alpha_2) \vee \overset{\leftrightarrow}{\Box}\neg\alpha_1 \\
O_\exists(\alpha|\beta) & =_{def} & (\alpha \wedge \beta) \succ_\exists (\neg\alpha \wedge \beta) \\
 & = & \overset{\leftrightarrow}{\Diamond}((\alpha \wedge \beta) \wedge \Box\neg(\neg\alpha \wedge \beta)) \vee \overset{\leftrightarrow}{\Box}\neg(\alpha \wedge \beta) \\
 & \leftrightarrow & \overset{\leftrightarrow}{\Diamond}(\beta \wedge \Box(\beta \rightarrow \alpha)) \vee \overset{\leftrightarrow}{\Box}\neg(\alpha \wedge \beta) \\
O^c_\exists(\alpha|\beta) & =_{def} & (\alpha \wedge \beta) \succ_\exists (\neg\alpha \wedge \beta) \wedge \overset{\leftrightarrow}{\Diamond}(\alpha \wedge \beta) \\
 & \leftrightarrow & \overset{\leftrightarrow}{\Diamond}(\beta \wedge \Box(\beta \rightarrow \alpha)) \\
O^{cc}_\exists(\alpha|\beta) & =_{def} & (\alpha \wedge \beta) \succ_\exists (\neg\alpha \wedge \beta) \wedge \overset{\leftrightarrow}{\Diamond}(\alpha \wedge \beta) \wedge \overset{\leftrightarrow}{\Diamond}(\neg\alpha \wedge \beta)
\end{array}
$$

□

Again, we do not further discuss the preference relation $\succ_\exists$, but we focus on the properties of the dyadic minimizing obligations. Notice that the formula that represents the obligation $O^c_\exists$ is simpler than the formula of $O_\exists$, because the latter has an additional disjunct. We therefore

[5] The definition is adapted from a modal formula of Boutilier. The minor distinction is that Boutilier defines $\overset{\leftrightarrow}{\Diamond}(\beta \wedge \Box(\beta \rightarrow \alpha)) \vee \overset{\leftrightarrow}{\Box}\neg\beta$. We have adapted the definition for our two-phase approach. Boutilier's definition is false if $\overset{\leftrightarrow}{\Box}\neg(\beta \wedge \alpha) \wedge \neg\overset{\leftrightarrow}{\Box}\neg\beta$, and therefore does not validate Proposition 2.21.

usually focus on $O^c_\exists$ in our propositions and examples; the related properties of $O_\exists$ can easily be derived from the properties of $O^c_\exists$. The following proposition shows that the obligation $O_\exists(\alpha|\beta)$ refers to the optimal $\beta$ worlds, and that $O_\exists(\alpha|\top)$ refers to the ideal worlds.

**Proposition 2.18** *Let$M = \langle W, \leq, V\rangle$ be a* 2DL *model and let $|\alpha|$ be the set of worlds that satisfy $\alpha$. For a world $w \in W$, we have $M, w \models O_\exists(\alpha|\beta)$ iff there is no $\alpha \wedge \beta$ world, or there is a world $w_2 \in |\alpha \wedge \beta|$ such that for all worlds $w_1 \in |\neg\alpha \wedge \beta|$ it is true that $w_1 \not\leq w_2$. Hence, we have $M, w \models O_\exists(\alpha|\beta)$ iff there are no $\alpha \wedge \beta$ worlds, or*

1. *$\alpha$ is true in an equivalence class of most preferred $\beta$ worlds of $M$, or*

2. *there is an infinite descending chain in which there is a $\beta$ world $w_2$ such that $\alpha$ is true in all $\beta$ worlds $w_1$ with $w_1 \leq w_2$.*

**Proof** *Analogous to the proof of Proposition 2.6 (see also [Bou94a]). $\Rightarrow$ By contraposition. Assume a model $M = \langle W, \leq, V\rangle$ with a world $w_3 \in W$ such that $M, w_3 \models \alpha \wedge \beta$ and for all worlds $w_2 \in W$ such that $M, w_2 \models \alpha \wedge \beta$ there is a world $w_1 \in W$ such that $M, w_1 \models \neg\alpha \wedge \beta$ and $w_1 \leq w_2$. We have $M, w_2 \not\models (\alpha \wedge \beta) \wedge \Box(\beta \rightarrow \alpha))$. $M, w \models \overset{\leftrightarrow}{\Diamond}\alpha$ for a world $w \in W$ iff there is a world $w' \in W$ such that $M, w' \models \alpha$. Hence, $M, w \not\models O_\exists(\alpha|\beta)$.*

*$\Leftarrow$ By contraposition. Assume $M, w \not\models O_\exists(\alpha|\beta)$ for some world $w$. Hence, for all worlds $w_2 \in W$ we have $M, w_2 \not\models \beta \wedge \Box(\beta \rightarrow \alpha)$ and there is a world $w_3$ such that $M, w_3 \models \alpha \wedge \beta$. It follows that for all worlds $w_2$ such that $M, w_2 \models \alpha \wedge \beta$ we have $M, w_2 \not\models \Box(\beta \rightarrow \alpha)$. Hence, there is a world $w_1 \in W$ such that $M, w_1 \models \neg\alpha \wedge \beta$ and $w_1 \leq w_2$.*

The following example illustrates the dyadic minimizing obligations as weak preferences, and the distinction between ordering and minimizing obligations.

**Example 2.19** Reconsider the Kripke model $M$ of Example 2.7 in Figure 2.6 (repeated from Figure 2.4). We have $M \models O^c_\exists(p|\top)$ and $M \models O^c_\exists(q|\top)$. Since $O^c_\exists(q|\top)$ is equivalent with $\overset{\leftrightarrow}{\Diamond}\Box q$ it is clear that $q$ has to be true in some most preferred world, and that less preferred worlds do not effect the truth of $\overset{\leftrightarrow}{\Diamond}\Box q$. Hence, in the evaluation of $O^c_\exists(q|\top)$ only preferred elements are taken into account and not the whole ordering. In Example 2.7 we showed $M \not\models O^c(q|\top)$, which illustrates the distinction between ordering and minimizing obligations. $\Box$

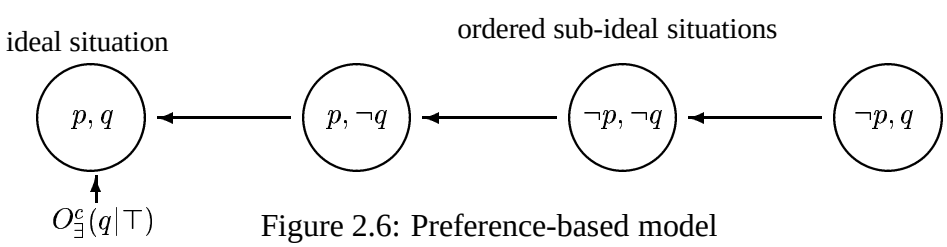



Figure 2.6: Preference-based model

The main properties of the dyadic existential-minimizing obligations are given by the following proposition.

**Proposition 2.20** *The logic* 2DL *has the following theorems.*

**WC**$_\exists$ : $(O_\exists(\alpha_1|\beta) \wedge \overset{\leftrightarrow}{\Diamond}(\alpha_1 \wedge \beta)) \rightarrow O_\exists(\alpha_1 \vee \alpha_2|\beta)$
**WC**$^c_\exists$ : $O^c_\exists(\alpha_1|\beta) \rightarrow O^c_\exists(\alpha_1 \vee \alpha_2|\beta)$
**RBC**$^c_\exists$: $(O^c_\exists(\alpha|\beta_1) \wedge O^c_\exists(\alpha|\beta_2)) \rightarrow O^c_\exists(\alpha|\beta_1 \vee \beta_2)$
**C**$_\exists$: $O_\exists(\bot|\alpha)$
**NC**$^c_\exists$: $\neg O^c_\exists(\bot|\alpha)$
**NC**$^{cc}_\exists$: $\neg O^{cc}_\exists(\bot|\alpha)$
**ID**$_\exists$: $O_\exists(\alpha|\alpha)$
**ID**$^c_\exists$: $\overset{\leftrightarrow}{\Diamond}\alpha \rightarrow O^c_\exists(\alpha|\alpha)$
**NID**$^{cc}_\exists$: $\neg O^{cc}_\exists(\alpha|\alpha)$

*The logic* 2DL *does not have the following theorems.*

**SA**$_\exists$ : $O_\exists(\alpha|\beta_1) \rightarrow O_\exists(\alpha|\beta_1 \wedge \beta_2)$
**AND**$_\exists$ : $O_\exists(\alpha_1|\beta) \wedge O_\exists(\alpha_2|\beta) \rightarrow O_\exists(\alpha_1 \wedge \alpha_2|\beta)$
**DD**$_\exists$ : $O_\exists(\alpha|\beta) \wedge O_\exists(\beta|\gamma) \rightarrow O_\exists(\alpha|\gamma)$
**D***$_\exists$ : $\neg(O_\exists(\alpha|\beta) \wedge O_\exists(\neg\alpha|\beta))$

**Proof** *The (non)theorems can be proven by proving (un)satisfiability in the preference-based semantics. Consider first the validity of weakening of the consequent* **WC**$_\exists$*. The logic has weakening of the consequent of $O^c_\exists(\alpha_1 \mid \beta)$ to $O^c_\exists(\alpha_1 \vee \alpha_2 \mid \beta)$, because the most preferred $\beta$ worlds that satisfy $\alpha_1$ also satisfy $\alpha_1 \vee \alpha_2$. Secondly, consider strengthening of the antecedent* **SA**$_\exists$*. The logic does not have strengthening of the antecedent of $O_\exists(\alpha|\beta_1)$ to $O_\exists(\alpha|\beta_1 \wedge \beta_2)$, because the preferred $\beta_1$ worlds may be different from the preferred $\beta_1 \wedge \beta_2$ worlds. For a counterexample, consider the Kripke model $M$ in Figure 2.6. We have $M \models O_\exists(q|\top)$ and $M \not\models O_\exists(q|\neg p)$. We do not have $M \models O_\exists(q|\neg p)$, because the preferred $\neg p$ worlds are the $\neg p \wedge \neg q$ worlds. Hence, $O_\exists$ does not have strengthening of the antecedent. Verification of the other (non)theorems is left to the reader.*

In the remainder of this section we discuss three relations between ordering and existential-minimizing. The first relation is given by the following proposition.

**Proposition 2.21** *The logic* 2DL *has the following theorems.*

**Rel**$_\exists$: $O(\alpha|\beta) \rightarrow O_\exists(\alpha|\beta)$
**Rel**$^c_\exists$: $O^c(\alpha|\beta) \rightarrow O^c_\exists(\alpha|\beta)$
**Rel**$^{cc}_\exists$: $O^{cc}(\alpha|\beta) \rightarrow O^{cc}_\exists(\alpha|\beta)$

**Proof** *The theorems can easily be proven by proving satisfiability in the preference-based semantics. For example, consider the theorem* **Rel**$^c_\exists$*. $O^c(\alpha \mid \beta)$ is true in a model iff we have $\mid \neg\alpha \wedge \beta \mid \not\leq \mid \alpha \wedge \beta \mid$ and $\mid \alpha \wedge \beta \mid$ is non-empty. Then any world $w \in \mid \alpha \wedge \beta \mid$ is part of a preferred $\beta$ equivalence class (or infinite descending chain) or they can see one. Hence, there is at least one preferred $\beta$ equivalence class (or infinite descending chain) of which the worlds satisfy $\alpha \wedge \beta$. The other theorems follow directly from this result and the definitions of the obliga-*

*tions. Alternatively, the theorems can be proven by proving validity in* 2DL. *For example,* $\mathbf{Rel}^c_\exists$ *is equivalent with the following theorem of* 2DL.

$$\mathbf{Rel}^c_\exists \quad (\overset{\leftrightarrow}{\Box}(\beta \wedge \alpha \rightarrow \Box(\beta \rightarrow \alpha)) \wedge \overset{\leftrightarrow}{\Diamond}(\alpha \wedge \beta)) \rightarrow \overset{\leftrightarrow}{\Diamond}(\beta \wedge \Box(\beta \rightarrow \alpha))$$

*Finally, the theorem is easier to read as an instance of the following formula that relates the preference orderings* $(\alpha_1 \succ_s \alpha_2) \rightarrow (\alpha_1 \succ_\exists \alpha_2)$.

$$\mathbf{Rel}^c_\exists \quad (\overset{\leftrightarrow}{\Box}(\alpha_1 \rightarrow \Box\neg\alpha_2) \wedge \overset{\leftrightarrow}{\Diamond}\alpha_1) \rightarrow \overset{\leftrightarrow}{\Diamond}(\alpha_1 \wedge \Box\neg\alpha_2)$$

The following proposition gives another relation between ordering and existential-minimizing obligations. It shows that an ordering obligation is equivalent to a set of existential-minimizing obligations, when we impose a constraint on the models.

**Proposition 2.22** *Let* $M$ *be a* 2DL *model such that* $M$ *does not contain duplicate worlds, i.e. for all* $w_1, w_2 \in W$ *such that* $w_1 \neq w_2$*, there is a propositional* $\alpha$ *such that* $M, w_1 \models \alpha$ *and* $M, w_2 \not\models \alpha$*. We have* $M, w \models O(\alpha \mid \beta)$ *iff for all* $\beta'$ *such that* $M, w \models \overset{\leftrightarrow}{\Box}(\beta' \rightarrow \beta)$*, we have* $M, w \models O_\exists(\alpha|\beta')$.

**Proof** $\Rightarrow$ *Follows directly from* **SA** *and* $\mathbf{Rel}_\exists$. $\Leftarrow$ *Every world is characterized by a unique propositional sentence. Let* $\overline{w}$ *denote this sentence that characterizes world* $w$*. Proof by contraposition. If* $M, w \not\models O(\alpha \mid \beta)$*, then there are* $w_1, w_2$ *such that* $M, w_1 \models \alpha \wedge \beta$*,* $M, w_2 \models \neg\alpha \wedge \beta$ *and* $w_2 \leq w_1$*. Choose* $\beta' = \overline{w_1} \vee \overline{w_2}$*.* $w_2$ *is one of the preferred* $\beta'$ *worlds, because there are no duplicate worlds. (If duplicate worlds are allowed, then there could be a* $\beta'$ *world* $w_3$ *which is a duplicate of* $w_1$*, and which is strictly preferred to* $w_1$ *and* $w_2$*.) We have* $M, w_2 \not\models \alpha$ *and therefore* $M, w \not\models O_\exists(\alpha|\beta')$.

The latter result is rather surprising for the following reason. When two ordering obligations are represented by two sets of minimizing obligations, then one would expect that the ordering obligations have at least the properties of the minimizing obligations. In particular, at first sight it seems that the ordering obligations have weakening of the consequent. The obligation $O(\alpha_1|\beta)$ is equivalent to the set of obligations $\{O_\exists(\alpha_1|\beta') \mid \overset{\leftrightarrow}{\Box}(\beta' \rightarrow \beta)\}$. Minimizing obligations have weakening, thus the set of obligations implies $\{O_\exists(\alpha_1 \vee \alpha_2 \mid \beta') \mid \overset{\leftrightarrow}{\Box}(\beta' \rightarrow \beta)\}$. The latter set is equivalent to the obligation $O(\alpha_1 \vee \alpha_2 \mid \beta)$. Hence, it seems that the ordering obligation $O(\alpha_1 \mid \beta)$ implies the ordering obligation $O(\alpha_1 \vee \alpha_2 \mid \beta)$. A careful analysis of the definitions reveals that the argument is wrong due to subtle consistency checks. For the operators $O$ and $O_\exists$ the consistency check is in $\mathbf{WC}_\exists$. Hence, the implication from $\{O_\exists(\alpha_1 \mid \beta') \mid \overset{\leftrightarrow}{\Box}(\beta' \rightarrow \beta)\}$ to $\{O_\exists(\alpha_1 \vee \alpha_2|\beta') \mid \overset{\leftrightarrow}{\Box}(\beta' \rightarrow \beta)\}$ is not valid. For the operators $O^c$ and $O^c_\exists$ the consistency check is part of **RSA**. Hence, the implication of $O^c(\alpha_1 \mid \beta)$ to $\{O^c_\exists(\alpha_1|\beta') \mid \overset{\leftrightarrow}{\Box}(\beta' \rightarrow \beta)\}$ is not valid. This is illustrated by the following example.

**Example 2.23** Consider the ordering obligations $O(p \wedge q \mid \top)$ and $O(q \mid \top)$, and the model $M$ in Figure 2.4 we discussed in Example 2.7. We have $M, w \models O(p \wedge q \mid \top)$ but we also have $M, w \not\models O(q|\top)$. The obligations correspond respectively to the sets of minimizing obligations

(for any $\beta$) $O_\exists(p \wedge q \,|\, \beta)$ and $O_\exists(q \,|\, \beta)$. However, $O_\exists(p \wedge q \,|\, \beta)$ does not imply $O_\exists(q \,|\, \beta)$ when $\beta$ implies $\neg p$, because of the consistency check $\overset{\leftrightarrow}{\Diamond}(p \wedge q \wedge \beta)$ of $\mathbf{WC}_\exists$. Moreover, consider the obligations $O^c(p \wedge q \,|\, \top)$ and $O^c(q \,|\, \top)$. The obligations correspond to minimizing obligations $O_\exists(p \wedge q|\beta)$ and $O_\exists(q|\beta')$, for any $\beta$ such that $\overset{\leftrightarrow}{\Diamond}(p \wedge q \wedge \beta)$ and for any $\beta'$ such that $\overset{\leftrightarrow}{\Diamond}(q \wedge \beta')$, respectively. We have $O_\exists(p \wedge q \,|\, \beta)$ implies $O_\exists(q \,|\, \beta)$. However, the first set of $\beta$ is a subset of the second set of $\beta'$. Hence $O^c(p \wedge q|\top)$ does not imply $O^c(q|\top)$. $\square$

In this section we introduced existential-minimizing obligations. They are weak variants of Bengt Hansson's universal-minimizing obligations. We showed three relations between the existential-minimizing obligations and the ordering obligations. First, we showed that ordering and existential-minimizing obligations are duals when we consider the inference patterns strengthening of the antecedent and weakening of the consequent, because the former only validates the first inference pattern whereas the latter only validates the second one. Second, we showed that an ordering obligation is stronger than existential-minimizing obligations in the sense that the former logically implies the latter. Third, we showed that an ordering obligation does not only derive a set of existential-minimizing obligations, but even corresponds to them if we add the additional condition that there are no duplicate worlds. Finally we showed that in the proof theory this surprising property corresponds to some consistency checks. In the next section we show how minimizing and ordering can be combined in a two-phase deontic logic. The two-phase approach combines strengthening of the antecedent and weakening of the consequent.

### 2.2.4 Combining ordering and minimizing

In this section we analyze the Forrester paradox in Example 2.1. The problem of the paradox is the combination of strengthening of the antecedent and weakening of the consequent. Thus far, we have discussed the ordering logic $O$ that has strengthening of the antecedent but not weakening of the consequent, and the minimizing logic $O_\exists$ that has weakening of the consequent but not strengthening of the antecedent. The two phases in a deontic logic correspond to the two different kinds of obligations $O^c$ and $O^c_\exists$ (or $O$ and $O_\exists$, or $O^{cc}$ and $O^{cc}_\exists$). From a proof-theoretic point of view, the first phase corresponds to applying valid inferences of $O^c$ like RSA, RAND etc, and the second phase corresponds to applying valid inferences of $O^c_\exists$ like WC and RBC. The following example shows how the two-phase approach solves the Forrester paradox of Example 2.1.

**Example 2.24 (Forrester paradox, continued)** Consider the set of premises

$$S = \{O^c(\neg k|\top), O^c(k \wedge g|k)\}$$

The crucial observation is that $O^c_\exists(\neg(k \wedge g)\,|\,k)$ is not entailed by $S$. A typical countermodel $M$ is represented in Figure 2.7. We have $M \models O^c(\neg k|\top)$ and $M \models O^c(k \wedge g|k)$, because $|k| \not\leq |\neg k|$ and $|k \wedge \neg g| \not\leq |k \wedge g|$ respectively. We have $M \not\models O^c_\exists(\neg(k \wedge g)|k)$, because $|k \wedge g| \leq |k \wedge \neg g|$.

For the proof-theoretic analysis of the non-derivability of $O^c_\exists(\neg(k \wedge g) \,|\, k)$, see the possible derivations in Figure 2.8 (a copy of Figure 2.2). In this figure, a dashed line represents an inference which is no longer valid compared to the derivation in Figure 2.1. First of all, the obligation $O^c_\exists(\neg(k \wedge g) \,|\, k)$ is not entailed by $S$ via $O^c(\neg k \,|\, k)$, because $O^c(\neg k \,|\, k)$ is not entailed by $O^c(\neg k|\top)$ due to the restriction in **RSA**. Secondly, $O^c_\exists(\neg(k \wedge g)|k)$ is not entailed by $S$ via

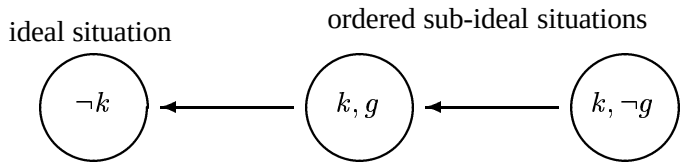



Figure 2.7: Semantic solution of the Forrester paradox

$O^c(\neg(k \wedge g) \mid \top)$, because $O^c(\neg(k \wedge g) \mid \top)$ is not entailed by $S$. Finally, $O^c_\exists(\neg(k \wedge g) \mid k)$ is not entailed by $S$ via $O^c_\exists(\neg(k \wedge g) \mid \top)$ either, because $O^c_\exists$ does not have strengthening of the antecedent at all. □

$$\frac{O^c(\neg k|\top)}{O^c(\neg k|k)}\ (\text{RSA}) \qquad \frac{\dfrac{\dfrac{O^c(\neg k|\top)}{O^c(\neg(k\wedge g)|\top)}\ (\text{WC})}{O^c(\neg(k\wedge g)|k)}\ \text{RSA}}{O^c_\exists(\neg(k\wedge g)|k)}\ \text{REL} \qquad \frac{\dfrac{\dfrac{O^c(\neg k|\top)}{O^c_\exists\neg k|\top)}\ \text{REL}}{O^c_\exists(\neg(k\wedge g)|\top)}\ \text{WC}}{O^c_\exists(\neg(k\wedge g)|k)}\ (\text{RSA})$$

Figure 2.8: Proof-theoretic solution of the Forrester paradox

Semantically, the first phase corresponds to ordering ($O^c$) and the second phase to minimizing ($O^c_\exists$). An intuition behind the two-phase approach is the distinction between 'dynamic' and 'static' processes. The first phase 'dynamically' orders all worlds, and the second phase 'statically' tests the minimal worlds. The distinction is analogous to the distinction between actions and tests in programming languages or dynamic logic. However, the intuition behind the distinction between dynamic and static processes is not represented in the semantics. It is a consequence of the way we use the two types of obligations. As can be verified in Figure 2.8 above, we use the ordering obligations $O^c$ as premises and the minimizing obligations $O^c_\exists$ as conclusions. The standard entailment relation $S \models \alpha$ means that for all models $M$ such that $M \models S$, we have $M \models \alpha$. Hence, all models 'dynamically' ordered by $S$ have the 'static' property $\alpha$. We further discuss the distinction between the dynamic first phase and the static second phase when we discuss the two-phase approach with the no-dilemma assumption in Section 2.3.4.

In this section we have introduced a new dyadic deontic logic 2DL. Any fully-fledged deontic logic has to be able to represent the no-dilemma assumption, permissions and factual detachment. We turn to these issues for our two-phase deontic logic 2DL in the remainder of this chapter.

## 2.3 The no-dilemma assumption in the two-phase approach

In this section we show how the no-dilemma assumption can be incorporated in the two-phase deontic logic. Most deontic logics, for example standard deontic logic SDL, have a no-dilemma assumption. For a philosophical motivation of this assumption, see e.g. [Con82, Hor94, Pra96].

The formalization of this assumption makes dilemmas inconsistent (see Section 1.3). The following example shows how the two-phase deontic logic developed in the previous section can represent dilemmas in a consistent way.

**Example 2.25 (Dilemma)** Consider the set of obligations $S = \{O(p \mid \top), O(\neg p \mid \top)\}$ that represents a dilemma, because the consequents of the obligations are contradictory. The set $S$ is consistent, and a typical model $M$ of $S$ is given in Figure 2.9. We have $M \models O(p \mid \top)$ and $M \models O(\neg p \mid \top)$, because $|\neg p| \not\leq |p|$ and $|p| \not\leq |\neg p|$, respectively. The model illustrates that the dilemma $S$ corresponds semantically to incomparable $p$ and $\neg p$ worlds. □

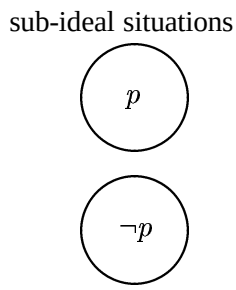



Figure 2.9: Dilemma

In some applications of deontic logic, the consistency of dilemmas is not a desirable property. There are (at least) two reasons to accept the no-dilemma assumption. First, with the assumption we can derive more conclusions than without it. For example, consider the premise $O(p|\top)$. With the assumption, we can derive (the otherwise non-derivable) $\neg O(\neg p|\top)$, $\neg O(\neg p \wedge q|\top)$, and probably even $\neg O(\neg p|q)$ for any $q$. Secondly, the assumption is necessary if we have a conflict resolution mechanism that is based on the the idea of 'restoring consistency.' A dilemma can be considered as a kind of conflict which should be inconsistent. Conflict resolution mechanisms are discussed in Chapter 4.

Thus, it must be possible to add the no-dilemma assumption to a deontic logic. In particular, we have to show how the assumption can be formalized in the two-phase deontic logic, to show it is a fully-fledged deontic logic. Usually, the no-dilemma assumption is formalized by a variant of the **D*** axiom which can be added to the logic to make dilemmas inconsistent. However, there is a problem (called the cigarettes problem) which shows that the addition of the **D*** axiom is unsatisfactory for *any* dyadic deontic logic. For example, the axiom is too weak for the existential-minimizing logic $O_{\exists}(\alpha|\beta)$ and too strong for the dyadic ordering logic $O(\alpha|\beta)$. Instead of adding an axiom to the logic developed in the previous section, we introduce new ordering and minimizing obligations.

### 2.3.1 The cigarettes problem

The following example (adapted from [PS96]) illustrates that strengthening of the antecedent SA is necessary to make dilemmas inconsistent. However, it also illustrates that only a restricted form of SA may be accepted, or counterintuitive conclusions follow.

**Example 2.26 (Cigarettes problem)** Assume a dyadic deontic logic that has at least substitution of logical equivalents, no strengthening of the antecedent and the following axiom which makes dilemmas inconsistent.

**D***: $\neg \overset{\leftrightarrow}{\Diamond}(\alpha_1 \wedge \alpha_2 \wedge \beta) \rightarrow \neg(O(\alpha_1|\beta) \wedge O(\alpha_2|\beta))$

Consider the set of obligations $S = \{O(\neg c \mid \top), O(c \mid k)\}$, where $k$ can be read as 'killing someone (the witness)' and $c$ can be read as 'offering someone a cigarette.' Prakken and Sergot argue in [PS96] that $S$ represents a dilemma. Hence, $S$ should be inconsistent, even when $S$ is extended with the obligation $O(\neg k|\top)$.

> "Is this acceptable? In our opinion it is: what is crucial is that $O(c|k)$ is not a CTD rule of $O(\neg c \mid \top)$ but of $O(\neg k \mid \top)$, for which reason $O(c \mid k)$ and $O(\neg c \mid \top)$ are unrelated obligations. Now one may ask how this conflict should be resolved and, of course, one plausible option is to regard $O(c|k)$ as an exception to $O(\neg c|\top)$ and to formalize this with a suitable nonmonotonic defeat mechanism. However, it is important to note that this is a separate issue, which has nothing to do with the CTD aspects of the example." [PS96]

To obtain the inconsistency, we can consider the inference pattern RSA. However, assume RSA and consider the set of obligations $S' = \{O(p_1|\top), O(p_2|\top)\}$. Obviously, $S'$ does not represent a dilemma when $p_1$ and $p_2$ are unrelated propositions, and it should be consistent. With RSA (and $\overset{\leftrightarrow}{\Diamond}(p_1 \wedge \neg p_2)$ and $\overset{\leftrightarrow}{\Diamond}(\neg p_1 \wedge p_2)$), we can derive the obligations $O(p_1 \mid \neg(p_1 \wedge p_2))$ and $O(p_2|\neg(p_1 \wedge p_2))$ from $S'$. However, the two derived obligations are inconsistent with **D***. This inconsistency can be avoided when the **D*** axiom is replaced by the following axiom **D*′**.

**D*′**: $\neg \overset{\leftrightarrow}{\Diamond}(\alpha_1 \wedge \alpha_2) \rightarrow \neg(O(\alpha_1|\beta) \wedge O(\alpha_2|\beta))$

However, assume **D*′** and consider the set of obligations $S'' = \{O(p \mid q_1), O(\neg p \mid q_2)\}$. It is argued by von Wright [vW71b] that $S''$ does not represent a dilemma and that it should therefore be consistent.[6]

> "Herewith it has been proven that, if there is a duty to see to it that $\alpha$ under circumstances $\beta$, then there is no duty to see to it that not-$\alpha$ under circumstances $\gamma$. For example: It has been proven that, if there is a duty to see to it that a certain window is closed should it start raining, then there cannot be a duty to see to it that the window is open should the sun be shining. This is manifestly absurd. Generally speaking: From a duty to see to a certain thing under certain circumstances nothing can follow logically concerning a duty or not-duty under entirely different, logically unrelated, circumstances. Least of all should one be able to prove that there is under those unrelated circumstances a duty of contradictory content."[vW71b, p.116].

With RSA (and $\overset{\leftrightarrow}{\Diamond}(p \wedge q_1 \wedge q_2)$ and $\overset{\leftrightarrow}{\Diamond}(\neg p \wedge q_1 \wedge q_2)$), we can derive the obligations $O(p|q_1 \wedge q_2)$ and $O(\neg p|q_1 \wedge q_2)$ from $S''$. The two derived obligations are inconsistent with **D*′**. □

[6] To be precise, the set $S''$ conflicts with $\overset{\leftrightarrow}{\Diamond}(q_1 \wedge q_2)$. Alchourrón [Alc93] argues that the derivation of $\overset{\leftrightarrow}{\Box}\neg(q_1 \wedge q_2)$ is intuitive (with the conditional interpretation of the antecedent).

$$\dfrac{\dfrac{O(\neg c|\top)}{O(\neg c|k)}\,\text{RSA} \qquad O(c|k)}{\bot}\,\text{D*}\ /\ \text{D*}'$$

$$\dfrac{\dfrac{O(p_1|\top)}{O(p_1|\neg(p_1\wedge p_2))}\,\text{RSA} \qquad \dfrac{O(p_2|\top)}{O(p_2|\neg(p_1\wedge p_2))}\,\text{RSA}}{\bot}\,\text{D*}$$

$$\dfrac{\dfrac{O(p|q_1)}{O(p|q_1\wedge q_2)}\,\text{RSA} \qquad \dfrac{O(\neg p|q_2)}{O(\neg p|q_1\wedge q_2)}\,\text{RSA}}{\bot}\,\text{D*}\ /\ \text{D*}'$$

Figure 2.10: Cigarettes problem

The derivations in Figure 2.10, a copy of Figure 2.3, illustrate that there are two solutions for the cigarettes problem above: weakening RSA and weakening **D*′**. The first solution of the Cigarettes problem is that $O(p_1 \mid \neg(p_1 \wedge p_2))$ cannot be derived from the obligation $O(p_1 \mid \top)$ when there is another premise $O(p_2 \mid \top)$ (set $S'$), and such that $O(p \mid q_1 \wedge q_2)$ cannot be derived from the obligation $O(p \mid q_1)$ when is there another premise $O(\neg p \mid q_2)$ (set $S''$). However, RSA may not be weakened too far, because the set $S$ has to remain inconsistent. In this section we incorporate this solution in our two-phase deontic logic. We first give minimizing obligations in Section 2.3.2 and then ordering obligations in Section 2.3.3, because the latter uses results from the former.

### 2.3.2 Minimizing

In this section we introduce the second minimizing logic. We first give the definition in modal logic, the semantic truth conditions and several properties expressed as theorems of the modal logic. Then we give a relation between these universal-minimizing obligations and the existential-ordering obligations introduced before. Finally, we investigate whether the universal-minimizing obligations can solve the cigarettes problem by adding strengthening of the antecedent.

The following universal-minimizing obligations $O_\forall$ consider all preferred worlds, not only an equivalence class of preferred worlds like the existential-minimizing obligations $O_\exists$.[7] It is similar to Boutilier's reconstruction [Bou94b] of Bengt Hansson's dyadic deontic logic [Han71].[8] The distinction between the universal-minimizing obligations and the existential-minimizing obligations is analogous to the distinction between standard and minimal deontic logic, see Section 1.3.2. In standard deontic logic an obligation $O\alpha$ is true if $\alpha$ is true in all accessible worlds,

[7] Spohn [Spo75] argues that Hansson's minimizing logic is axiomatized by a standard system, extended with $\neg O_\forall(\neg\beta \mid \alpha) \rightarrow (O_\forall(\gamma \mid \alpha \wedge \beta) \leftrightarrow O_\forall(\beta \rightarrow \gamma \mid \alpha)$. However, **RBC** cannot be derived from this theorem, see e.g. [KLM90, Alc96].

[8] B.Hansson's orderings are totally connected and Boutilier's reconstruction is in the logic CO instead of CT4O, see Definition 2.37. Moreover, Boutilier defines minimizing conditionals by $O_\forall(\alpha|\beta) =_{def} \overset{\leftrightarrow}{\Box}(\beta \rightarrow \Diamond(\beta \wedge \Box(\beta \rightarrow \alpha)))$. We have adapted the definition for the two-phase approach in a similar way as we have adapted the weak minimizing obligations in Section 2.2.3.

and in minimal deontic logic $O\alpha$ is true if $\alpha$ is true in an equivalence class of accessible worlds. Moreover, the distinction is analogous to the distinction between sceptical and credulous inference relations in logics of defeasible reasoning. For example, in circumscription [McC80] a sceptical inference relations considers all minimal worlds and a credulous inference relation an equivalence class of minimal worlds. Moreover, in Reiter's default logic [Rei80] the distinction between sceptical and credulous is related to the distinction between truth in all extensions and truth in at least one extension.

The universal-minimizing obligation is defined in a weak preference ordering, which we write as $\alpha_1 \succ_\forall \alpha_2$. A preference $\alpha_1 \succ_\forall \alpha_2$ does not logically imply a preference for $\alpha_1' \succ \alpha_2'$ when $\alpha_1'$ implies $\alpha_1$ and $\alpha_2'$ implies $\alpha_2$. Thus $\succ_\forall$ formalizes weak preferences. We say that $\alpha_1$ is weakly preferred to $\alpha_2$ if and only if for all $\alpha_2$ worlds there is an $\alpha_1$ world $w$ such that for all $\alpha_2$ worlds we have that they are not as preferable as $w$. That is, for all *preferred* $\alpha_1$ worlds and all *preferred* $\alpha_2$ worlds we have either that the $\alpha_1$ world is preferred to the $\alpha_2$ world, or that the two worlds are incomparable.

**Definition 2.27 (Dyadic minimizing obligation)** The dyadic minimizing obligation '$\alpha$ should be the case if $\beta$ is the case', written as $O_\forall(\alpha \mid \beta)$, is defined as a weak preference of $\alpha \wedge \beta$ over $\neg\alpha \wedge \beta$. A weak preference of $\alpha_1$ over $\alpha_2$, written as $\alpha_1 \succ_\forall \alpha_2$, is defined as follows.

$$
\begin{array}{lll}
\alpha_1 \succ_\forall \alpha_2 & =_{def} & \overset{\leftrightarrow}{\Box}(\alpha_2 \rightarrow \Diamond(\alpha_1 \wedge \Box\neg\alpha_2)) \vee \overset{\leftrightarrow}{\Box}\neg\alpha_1 \\
O_\forall(\alpha|\beta) & =_{def} & (\alpha \wedge \beta) \succ_\forall (\neg\alpha \wedge \beta) \\
 & = & \overset{\leftrightarrow}{\Box}((\neg\alpha \wedge \beta) \rightarrow \Diamond((\alpha \wedge \beta) \wedge \Box\neg(\neg\alpha \wedge \beta))) \vee \overset{\leftrightarrow}{\Box}\neg(\alpha \wedge \beta) \\
 & \leftrightarrow & \overset{\leftrightarrow}{\Box}(\beta \rightarrow \Diamond(\beta \wedge \Box(\beta \rightarrow \alpha))) \vee \overset{\leftrightarrow}{\Box}\neg(\alpha \wedge \beta) \\
O^c_\forall(\alpha|\beta) & =_{def} & (\alpha \wedge \beta) \succ_\forall (\neg\alpha \wedge \beta) \wedge \overset{\leftrightarrow}{\Diamond}(\alpha \wedge \beta) \\
 & \leftrightarrow & \overset{\leftrightarrow}{\Box}(\beta \rightarrow \Diamond(\beta \wedge \Box(\beta \rightarrow \alpha))) \wedge \overset{\leftrightarrow}{\Diamond}\beta \\
O^{cc}_\forall(\alpha|\beta) & =_{def} & (\alpha \wedge \beta) \succ_\forall (\neg\alpha \wedge \beta) \wedge \overset{\leftrightarrow}{\Diamond}(\alpha \wedge \beta) \wedge \overset{\leftrightarrow}{\Diamond}(\neg\alpha \wedge \beta) \\
 & \leftrightarrow & \overset{\leftrightarrow}{\Box}(\beta \rightarrow \Diamond(\beta \wedge \Box(\beta \rightarrow \alpha))) \wedge \overset{\leftrightarrow}{\Diamond}(\neg\alpha \wedge \beta)
\end{array}
$$

□

The following proposition shows that the minimizing obligations $O_\forall$ consider all preferred worlds.

**Proposition 2.28** *Let $M = \langle W, \leq, V\rangle$ be a* 2DL *model. For a world $w \in W$, we have $M, w \models O_\forall(\alpha \mid \beta)$ iff there are no $\alpha \wedge \beta$ worlds or for all $\beta$ worlds $w_3$ there is a world $w_2 \in W$ such that $w_3 \leq w_2$ and and for all $\beta$ worlds $w_1 \in W$ with $w_1 \leq w_2$ we have $M, w_1 \models \alpha$. Hence, $M, w \models O_\forall(\alpha|\beta)$ iff there are no $\alpha \wedge \beta$ worlds, or:*

1. *$\alpha$ is true in all most preferred $\beta$ worlds, and*
2. *in every infinite descending chain that contains $\beta$ worlds there is a $\beta$ world $w_2$ such that $\alpha$ is true in all $\beta$ worlds $w_1$ such that $w_1 \leq w_2$.*

**Proof** *Analogous to the proof of Proposition 2.6 and 2.18 (see also [Bou94a]). $\Rightarrow$ By contraposition. Assume a model $M = \langle W, \leq, V\rangle$ such that there is a world $w_4$ such that $M, w_4 \models \alpha \wedge \beta$ and there is a world $w_3$ such that there is not a $\beta$ world $w_2$ such that $w_2 \leq w_3$ and $\alpha$ is true in all $\beta$ worlds $w_1$ such that $w_1 \leq w_2$. We have $M, w_2 \models \beta \wedge \Box(\beta \rightarrow \alpha)$ and therefore $M, w_3 \models \beta \wedge \neg\Diamond(\beta \wedge \Box(\beta \rightarrow \alpha))$. Hence, $M, w \not\models O_\forall(\alpha|\beta)$.*

$\Leftarrow$ *By contraposition. Assume* $M, w \not\models O_\forall(\alpha \mid \beta)$ *for some world* $w$*. Hence, there is a world* $w_3 \in W$ *such that* $M, w_3 \not\models \beta \to (\beta \wedge \Box(\beta \to \alpha))$ *and there is a world* $w_4$ *such that* $M, w_4 \models \alpha \wedge \beta$*. It follows that* $M, w_3 \models \beta$ *and* $M, w_3 \not\models \Diamond(\beta \wedge \Box(\beta \to \alpha))$*. It follows that for all worlds* $w_2$ *with* $w_2 \leq w_3$ *we have* $M, w_2 \models \neg(\beta \wedge \Box(\beta \to \alpha))$*. Hence, for every* $w_2$ *such that* $M, w_2 \models \alpha \wedge \beta$ *there is a world* $w_1 \in W$ *such that* $M, w_1 \models \neg\alpha \wedge \beta$ *and* $w_1 \leq w_2$*. The set of* $w_1$ *either contains most preferred worlds or an infinite descending chain of* $\beta$ *worlds.*

The following proposition shows various properties of the universal-minimizing obligations as weak preferences.

**Proposition 2.29** *The logic* 2DL *has the following theorems.*

| | |
|---|---|
| **WC**$_\forall$ : | $O^c_\forall(\alpha_1 \mid \beta) \to O^c_\forall(\alpha_1 \vee \alpha_2 \mid \beta)$ |
| **AND**$_\forall$ : | $(O^c_\forall(\alpha_1 \mid \beta) \wedge O^c_\forall(\alpha_2 \mid \beta)) \to O^c_\forall(\alpha_1 \wedge \alpha_2 \mid \beta)$ |
| | $O^c_\forall(\alpha \mid \beta) \to O^c_\forall(\beta \to \alpha \mid \top)$ |
| **DD**$\top_\forall$ : | $O_\forall(\alpha \mid \beta) \wedge O_\forall(\beta \mid \top) \to O_\forall(\alpha \mid \top)$ |
| **RBC**$_\forall$*:* | $(O_\forall(\alpha \mid \beta_1) \wedge O_\forall(\alpha \mid \beta_2)) \to O_\forall(\alpha \mid \beta_1 \vee \beta_2)$ |
| **D***$_\forall$ : | $\neg(O^c_\forall(\alpha \mid \beta) \wedge O^c_\forall(\neg\alpha \mid \beta))$ |
| **Id**$_\forall$ | $O_\forall(\alpha \mid \alpha)$ |
| **Id**$^c_\forall$ | $\overset{\leftrightarrow}{\Diamond}\alpha \to O^c_\forall(\alpha \mid \alpha)$ |
| **NId**$^{cc}_\forall$ | $\neg O^{cc}_\forall(\alpha \mid \alpha)$ |
| **C**$_\forall$ | $O_\forall(\bot \mid \alpha)$ |
| **NC**$^c_\forall$ | $\neg O^c_\forall(\bot \mid \alpha)$ |
| **NC**$^{cc}_\forall$ | $\neg O^{cc}_\forall(\bot \mid \alpha)$ |

*The logic* 2DL *does not have the following theorems.*

| | |
|---|---|
| **SA**$_\forall$ : | $O_\forall(\alpha \mid \beta_1) \to O_\forall(\alpha \mid \beta_1 \wedge \beta_2)$ |
| **DD**$_\forall$ : | $O_\forall(\alpha \mid \beta) \wedge O_\forall(\beta \mid \gamma) \to O_\forall(\alpha \mid \gamma)$ |

**Proof** *The (non)theorems can easily be verified in the preference-based semantics. The proofs are analogous to the proofs of Proposition 2.8, 2.12 and 2.20.*

The following proposition gives a relation between the two types of minimizing obligations $O_\exists$ and $O_\forall$.

**Proposition 2.30** *The logic* 2DL *has the following theorems.*

| | |
|---|---|
| **Rel**$_{\forall\exists}$ | $O_\forall(\alpha \mid \beta) \to O_\exists(\alpha \mid \beta)$ |
| **Rel**$^c_{\forall\exists}$ | $O^c_\forall(\alpha \mid \beta) \to O^c_\exists(\alpha \mid \beta)$ |
| **Rel**$^{cc}_{\forall\exists}$ | $O^{cc}_\forall(\alpha \mid \beta) \to O^{cc}_\exists(\alpha \mid \beta)$ |

**Proof** *The theorems can easily be verified in the preference-based semantics. Truth in all worlds implies truth in an equivalence class of preferred worlds (if such worlds exist). Alternatively, the theorems follow from the following relation* $(\alpha_1 \succ_\forall \alpha_2) \to (\alpha_1 \succ_\exists \alpha_2)$*, i.e. from*

$$(\overset{\leftrightarrow}{\Box}(\alpha_2 \to \Diamond(\alpha_1 \wedge \Box\neg\alpha_2)) \vee \overset{\leftrightarrow}{\Box}\neg\alpha_1) \to (\overset{\leftrightarrow}{\Diamond}(\alpha_1 \wedge \Box\neg\alpha_2) \vee \overset{\leftrightarrow}{\Box}\neg\alpha_1)$$

In the remainder of this section we consider the cigarettes problem for universal-minimizing obligations. The following example shows that the logic of the minimizing obligations $O_\forall$ does not solve the cigarettes problem.

**Example 2.31 (Cigarettes problem, continued)** Consider the set of dyadic obligations $S = \{O_\forall(\neg c|\top), O_\forall(c|k)\}$. $S$ is consistent, as is shown by the model $M$ of $S$ in Figure 2.11 below. We have $M \models O_\forall(\neg c|\top)$, because the minimal worlds satisfy $\neg c$, and we have $M \models O_\forall(c|k)$, because the minimal $k$ worlds satisfy $c$. The set $S$ is consistent, thus the logic $O_\forall$ does not solve the cigarettes problem. □

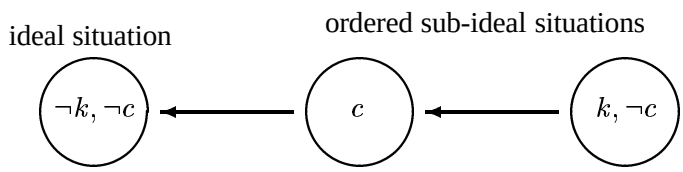



Figure 2.11: The cigarettes problem

In Section 2.3.1 we argued that the cigarettes problem may be solved when the obligations have some strengthening of the antecedent. There are several well-known ways to add some strengthening of the antecedent to the minimizing logic. The most popular is without doubt Pearl's so-called System Z [Pea90] which is equivalent to Lehmann's Rational Closure [LM92] and the so-called minimal specificity principle of possibilistic logic [BDP92]. System Z adds strengthening of the antecedent by assuming that 'worlds gravitate towards most preferred.' Here we give the reconstruction of gravitating towards most preferred of Boutilier [Bou92a].[9] The basic idea of worlds gravitating towards most preferred is that worlds are more preferred in preference relation $\leq_1$ than in $\leq_2$ when they are equivalent to a more preferred world.[10]

**Definition 2.32 (More preferred)** [Bou92a, Definition 5.20] Let $M_1 = \langle W, \leq_1, V\rangle$ and $M_2 = \langle W, \leq_2, V\rangle$ be two CT4O* models with the same $W$ and $V$. $w \in W$ is more preferred in $M_1$ than in $M_2$, written as $N(w, M_1, M_2)$, iff

1. there is some $v \in W$ such that $v \leq_1 w$, $w \leq_1 v$, and not $v \leq_2 w$, or
2. there is no $v$ (with $w \neq v$) such that $w \leq_2 v$ and $v \leq_2 w$.

□

Given the definition of more preferred worlds, Boutilier defines a preference ordering on models. The ordering on models ($\sqsubseteq$) should not be confused with the ordering on worlds ($\leq$). The ordering on models is a technical trick to ensure that the worlds within a model are maximally preferred, whereas the ordering on worlds expresses the ideality ordering.

**Definition 2.33 (Preferred to)** [Bou92a, Definition 5.21] Let $M_1 = \langle W, \leq_1, V\rangle$ and $M_2 = \langle W, \leq_2, V\rangle$ be CT4O* models. The model $M_1$ is as preferable as $M_2$, written as $M_1 \sqsubseteq M_2$, iff

[9] Boutilier's reconstruction of gravitating towards preferred, i.e. System Z, is in the logic CO, see Definition 2.37. The reconstruction in CT4O is analogous.

[10] The definitions are slightly complicated because worlds can be not equivalent to any other world: condition (2.) of Definition 2.32 and to allow that $N(w, M_1, M_2)$ is false in Definition 2.33.

for all $w \in W$, $N(w, M_1, M_2)$ is false only if $\{v \mid w \leq_2 v\} \subseteq \{v \mid w \leq_1 v\}$. $M_1$ is preferred to $M_2$, written as $M_1 \sqsubset M_2$, iff $M_1 \sqsubseteq M_2$ and $M_2 \not\sqsubseteq M_1$. □

The preference ordering on models is used to determine the preferred models. Definition 2.33 compares only models that agree on possible worlds ($W$ and $V$ must agree). Boutilier [Bou92a] notices that if we are considering only CT4O* models, this makes little difference (as long as we 'rename' worlds appropriately), because duplicate worlds (having the same induced valuation) have no effect on preferentially entailed conclusions.

**Definition 2.34 (Preferred model)** [Bou92a, Definition 5.22] Let $M$ be a CT4O* model and let $T \subseteq \mathcal{L}$ be a set of dyadic obligations $O_\forall(\alpha|\beta)$. $M$ is a preferred model of $T$ iff $M \models T$ and for all $M'$ such that $M' \models T$, we have $M' \not\sqsubset M$. □

The preferred models are used for preferential entailment [Sho88, KLM90].

**Definition 2.35 (Preferential entailment)** [Bou92a, Definition 5.23] Let $T \subseteq \mathcal{L}$ be a set of dyadic obligations $O_\forall(\alpha\,|\,\beta)$. $\alpha$ is preferentially entailed by $T$, written as $T \models_\sqsubset \alpha$, iff $M \models \alpha$ for all preferred models $M$ of $T$. □

The following example illustrates that System Z does not solve the cigarettes problem, although it has some strengthening of the antecedent. System Z does not solve the cigarettes problem, because it maintains consistency (if $S$ has a model then $S$ has a preferred model). Methods that maintain consistency do no solve the cigarettes problem, because $\{O(\neg k\,|\,\top), O(k\,|\,c)\}$ is consistent (it has a model), but should be inconsistent. There are many defeasible logics with different methods to add strengthening of the antecedent. For example, a method related to System Z is Brewka's prioritization [Bou92a, BB95]. Another popular irrelevance principle is adding a maximal consistent set of material counterparts of conditionals [Del88, Mor95]. However, as far as we know all these methods maintain consistency. Obviously, these other methods do not solve the cigarettes problem either.

**Example 2.36 (Cigarettes problem, continued)** Reconsider the CT4O* model $M$ in Figure 2.11 of the set of dyadic obligations $S = \{O_\forall(\neg c|\top), O_\forall(c|k)\}$ of Example 2.31. $M$ is a most preferred model of $S$. For any propositional $a$ with $\overset{\leftrightarrow}{\Diamond}(a \wedge \neg c \wedge \neg k)$ we have $S \models_\sqsubset O_\forall(\neg c|a)$. Hence, $O_\forall(\neg c|\top)$ is strengthened to $O_\forall(\neg c|a)$. However, $S$ is consistent and we have $S \not\models_\sqsubset O_\forall(\neg c|k)$. Hence, $O_\forall(\neg c|\top)$ is not strengthened to $O_\forall(\neg c|k)$ and System Z is not a solution of the cigarettes problem. □

Bengt Hansson proposed minimizing obligations for *totally connected* orderings, i.e. for all worlds $w_1$ and $w_2$ we have either $w_1 \leq w_2$ or $w_2 \leq w_1$.[11] If the ordering is connected, then the modal operator $\Box$ satisfies the S4.3 axioms instead of the S4 axioms. The additional axiom is

[11] If the set of premises $S$ contains only formulas of the form $O_\forall(\alpha \mid \beta)$, then we have the well-known result that $S \models_P \alpha$ iff $S \models_R \alpha$, where $\models_P$ is entailment based on S4 (preferential) models and $\models_R$ is based on S4.3 (rational) models. This follows from Lehmann's observation [LM92] about preferential and rational models, see also [Bou92a, p.63].

$\Box(\Box(\alpha \to \beta) \vee \Box(\beta \to \alpha))$. Totally connected is axiomatized by Humberstone's axiom **S** given in the following definition. Boutilier calls the extension of CT4O* with **S** the logic CO*.[12]

**Definition 2.37 (CO*)** [Bou94b] The logic CO* is the smallest $S \subseteq \mathcal{L}$, such that $S$ contains the axioms of CT4O* and the following axiom **S**, and is closed under the rules of inference of CT4O*.

**S** $\alpha \to \overleftarrow{\Box}\Diamond\alpha$ $\Box$

**Definition 2.38 (CO* semantics)** [Bou94b] A CO* model is a CT4O* model $M = \langle W, \leq, V \rangle$ of which the accessibility relation is totally connected. $\Box$

The following proposition shows that we can formalize the no-dilemma assumption for $O_\exists$ by demanding that the partial ordering $\leq$ is totally connected. An important consequence of this proposition is that in the minimizing logics, incomparable worlds are *only* used to model dilemmas.

**Proposition 2.39** *The logic CO* has the following theorem.*

$$O_\exists(\alpha|\beta) \leftrightarrow O_\forall(\alpha|\beta)$$

**Proof** *Follows directly from the semantic definitions. Truth in an equivalence class of preferred worlds and truth in all preferred worlds has become equivalent.*

In this section we discussed Bengt Hansson's universal-minimizing obligations in the framework of modal preference logics. We showed that the logic makes some dilemmas inconsistent, but with the cigarettes problem we showed that it does not make *all* dilemmas inconsistent. The latter problem can be solved by adding some strengthening of the antecedent to the minimizing obligations. However, we also showed that existing methods to add strengthening of the antecedent to minimizing obligations are not sufficient to solve the cigarettes problem, because these methods maintain consistency. In the following section we consider extensions of the ordering obligations, which have strengthening of the antecedent, to solve the cigarettes problem.

### 2.3.3 Ordering

In this section we define new ordering obligations in the modal logic, and we give several properties of the obligations expressed as theorems of the modal logic. Moreover, we introduce a notion of preferential entailment for the obligations, such that the obligations have restricted strengthening of the antecedent. Finally, we analyze the cigarettes problem with the new ordering obligations. The following example illustrates the cigarettes problem of Example 2.26 for the ordering obligations $O^c$.

[12] The preferred System Z model in CO* given $W$ and $V$ is unique. The uniqueness is obviously not true for more complex sentences of the language. Consider for example $O_\forall(p|\top) \vee O_\forall(\neg p|\top)$. There are two preferred models, one with $p$ preferred to $\neg p$ and one vice versa. From the uniqueness of the preferred models follows that it can be axiomatized with the concept of 'conditional only knowing', see [Bou92a]. This axiomatization is Boutilier's motivation to use the logic of inaccessible worlds.

**Example 2.40 (Cigarettes problem, continued)** Assume the extension of 2DL with the following axiom **D***.

$$\textbf{D*}: \quad \neg \overset{\leftrightarrow}{\Diamond}(\alpha_1 \wedge \alpha_2 \wedge \beta) \rightarrow \neg(O^c(\alpha_1|\beta) \wedge O^c(\alpha_2|\beta))$$

Consider the set of dyadic obligations $S = \{O^c(p_1|\top), O^c(p_2|\top)\}$. From the proof-theoretic analysis in Example 2.26 follows that $S$ is inconsistent with **D***.[13] Moreover, the set $S$ is even inconsistent when the axiom **D*** is replaced by the following axiom **D*′**.

$$\textbf{D*}': \quad \neg \overset{\leftrightarrow}{\Diamond}(\alpha_1 \wedge \alpha_2) \rightarrow \neg(O^c(\alpha_1|\beta) \wedge O^c(\alpha_2|\beta))$$

We can derive $O^c(p_1 \wedge \neg(p_1 \wedge p_2) \,|\, \neg(p_1 \wedge p_2))$ and $O^c(p_2 \wedge \neg(p_1 \wedge p_2) \,|\, \neg(p_1 \wedge p_2))$ from $S$ with $\overset{\leftrightarrow}{\Diamond}\alpha \rightarrow O^c(\alpha|\alpha)$ and the **RAND** theorem. They are logically equivalent to the obligations $O^c(p_1 \wedge \neg p_2 \,|\, \neg(p_1 \wedge p_2))$ and $O^c(p_2 \wedge \neg p_1 \,|\, \neg(p_1 \wedge p_2))$, and the latter two obligations are inconsistent with axiom **D*′**. □

For the two-phase approach, we define a new kind of phase-1 obligation $O_D$ as a combination of ordering obligation $O$ and universal-minimizing obligation $O_\forall$. This new obligation combines ordering and minimizing in the same phase. In particular, it combines the strengthening of the antecedent of the ordering obligations with the no-dilemma assumption of the universal-minimizing obligations to solve the Cigarettes problem. The details of this solution are shown at the end of this section.

**Definition 2.41 (Dyadic ordering obligation)** Dyadic ordering obligation '$\alpha$ should be the case if $\beta$ is the case', written as $O_D(\alpha|\beta)$, is defined as a strong and a weak preference of $\alpha \wedge \beta$ over $\neg\alpha \wedge \beta$.

$$\begin{array}{lll}
O_D(\alpha|\beta) & =_{def} & (\alpha \wedge \beta) \succ_s (\neg\alpha \wedge \beta) \wedge (\alpha \wedge \beta) \succ_\forall (\neg\alpha \wedge \beta) \\
O_D^c(\alpha|\beta) & =_{def} & O_D(\alpha|\beta) \wedge \overset{\leftrightarrow}{\Diamond}(\alpha \wedge \beta) \\
O_D^{cc}(\alpha|\beta) & =_{def} & O_D(\alpha|\beta) \wedge \overset{\leftrightarrow}{\Diamond}(\alpha \wedge \beta) \wedge \overset{\leftrightarrow}{\Diamond}(\neg\alpha \wedge \beta)
\end{array}$$ □

The following proposition shows that the dyadic ordering obligations $O_D$ have weaker versions of the theorems of Proposition 2.8.

**Proposition 2.42** *The logic* 2DL *has the following theorems.*

[13] Semantically, the inconsistency is a result of the fact that the axiom **D*** makes the ordering $\leq$ connected. In the previous section we observed that in the minimizing approach, incomparable worlds are *only* used to model dilemmas. However, in the ordering approach (and the two-phase logic) this is not the only usage of incomparable worlds. Let $M$ be a 2DL model of $S$. The $p_1 \wedge \neg p_2$ and $\neg p_1 \wedge p_2$ worlds of $M$ are incomparable, because the first is better than the latter with respect to obligation $O^c(p_1|\top)$, but worse with respect to obligation $O^c(p_2|\top)$.

**RSA**$_D$: $(O_D(\alpha|\beta_1) \wedge O_\forall(\alpha|\beta_1 \wedge \beta_2)) \rightarrow O_D(\alpha|\beta_1 \wedge \beta_2)$
**AND**$_D$: $(O^c_D(\alpha_1|\beta) \wedge O^c_D(\alpha_2|\beta) \wedge \overset{\leftrightarrow}{\Diamond}(\alpha_1 \wedge \alpha_2 \wedge \beta)) \rightarrow O^c_D(\alpha_1 \wedge \alpha_2|\beta)$
**OR**$_D$: $(O^c_D(\alpha_1|\beta) \wedge O^c_D(\alpha_2|\beta)) \rightarrow O^c_D(\alpha_1 \vee \alpha_2|\beta)$
**RDD**$_D$: $(O^c_D(\alpha|\beta) \wedge O^c_D(\beta|\gamma) \wedge O_\forall(\alpha \wedge \beta|\gamma)) \rightarrow O^c_D(\alpha \wedge \beta|\gamma)$
**D***$_D$: $\neg \overset{\leftrightarrow}{\Diamond}(\alpha_1 \wedge \alpha_2 \wedge \beta) \rightarrow \neg(O^c_D(\alpha_1|\beta) \wedge O^c_D(\alpha_2|\beta))$

**Proof** *The properties follow directly from the properties of $O$ and $O_\forall$, see Proposition 2.8 and 2.29.* **RSA**$_D$ *and* **RDD**$'$ *follow from $O$,* **D***$_D$ *follows from $O_\forall$, and* **AND**$_D$ *and* **OR**$_D$ *follow from both.*

The following proposition shows the relation between the ordering and minimizing obligations.

**Proposition 2.43** *The logic* 2DL *has the following theorems.*

**Rel**$_\forall$: $O_D(\alpha|\beta) \rightarrow O_\forall(\alpha|\beta)$
**Rel**$^c_\forall$: $O^c_D(\alpha|\beta) \rightarrow O^c_\forall(\alpha|\beta)$
**Rel**$^{cc}_\forall$: $O^{cc}_D(\alpha|\beta) \rightarrow O^{cc}_\forall(\alpha|\beta)$

**Proof** *Follows directly from Definition 2.41.*

In the remainder of this section on the ordering obligations, we consider the property strengthening of the antecedent and we analyze the cigarettes problem in the logic. In Section 2.2.2 we discussed the following problem of **RSA**. To obtain strengthening of the antecedent of $O(\alpha|\beta_1)$ to $O(\alpha|\beta_1 \wedge \beta_2)$ we have to know the consistency expression $\overset{\leftrightarrow}{\Diamond}(\alpha \wedge \beta_1 \wedge \beta_2)$. The axiom **LP** and the related logic 2DL* obtain strengthening of the antecedent without having to specify the consistency expression for every example. A similar problem occurs for **RSA**$_D$, because now we we can only derive $O_D(\alpha|\beta_1 \wedge \beta_2)$ from $O_D(\alpha|\beta_1)$ when we have $O_\forall(\alpha|\beta_1 \wedge \beta_2)$ as another premise. As a solution, we use preferential entailment (hence, we only use preferred models), but a different one than System Z. The following preference ordering on models prefers models which are *maximally connected* with respect to the partial pre-ordering $\leq$. To distinguish this new notion of preferential entailment from the definitions of preferred models and $\models_\sqsubset$ in Definition 2.33 and 2.35 we write c-preferred and $\models_c$.

**Definition 2.44 (C-preferential entailment)** Let $M_1 = \langle W, \leq_1, V\rangle$ and $M_2 = \langle W, \leq_2, V\rangle$ be two 2DL* models. $M_1$ is as c-preferable as $M_2$, written as $M_1 \sqsubseteq_c M_2$, iff for all $w_1, w_2 \in W$ if $w_1 \leq_2 w_2$ then $w_1 \leq_1 w_2$. $M_1$ is c-preferred to $M_2$, written as $M_1 \sqsubset_c M_2$, iff $M_1 \sqsubseteq_c M_2$ and $M_2 \not\sqsubseteq_c M_1$. $\alpha$ is c-preferentially entailed by $T$, written as $T \models_c \alpha$, iff $M \models \alpha$ for all c-preferred models $M$ of $T$. □

In the ordering logic, c-preferred models are not unique. Consider the models of $O_D(p|q)$: the $\neg q$ worlds can be equivalent to $p \wedge q$ or to $\neg p \wedge q$ worlds.[14] The following example compares $\sqsubseteq$ with $\sqsubseteq_c$.

[14] Hence, we cannot use conditional only knowing. The System Z preferred models of a set of ordering obligations are not unique either.

**Example 2.45** Consider the sets $S = \{O_D(p \mid \top)\}$ and $S' = \{O_\forall(p \mid \top)\}$. The unique maximally connected model $M$ of $S$ consists of two equivalence classes, one consists of all $p$ worlds and one consists of $\neg p$ worlds, such that all $p$ worlds are strictly preferred over $\neg p$ worlds, see Figure 2.12.a. For any propositional $q$ with $\overset{\leftrightarrow}{\Diamond}(p \wedge q)$ we have $M \models O_D(p \mid q)$ and therefore $S \models_c O_D(p|q)$. Hence, $O_D$ with $\models_c$ has strengthening of the antecedent. The model $M$ is the unique System Z model of $S'$ and therefore also $S' \models_\sqsubset O_\forall(p|q)$. Finally, $M$ is also a maximally connected model of $S'$. However, consider a model $M'$ with two equivalence classes of $p \wedge \neg q$ and $\neg p \vee q$ worlds such that $p \wedge \neg q$ worlds are preferred over all $\neg p \vee q$ worlds, see Figure 2.12.b. The model $M'$ is also a maximally connected model of $S'$. We have $M' \not\models O_\forall(p \mid q)$, and thus $S' \not\models_c O_\forall(p|q)$. Hence, $O_\forall$ with $\models_c$ does not have strengthening of the antecedent. □

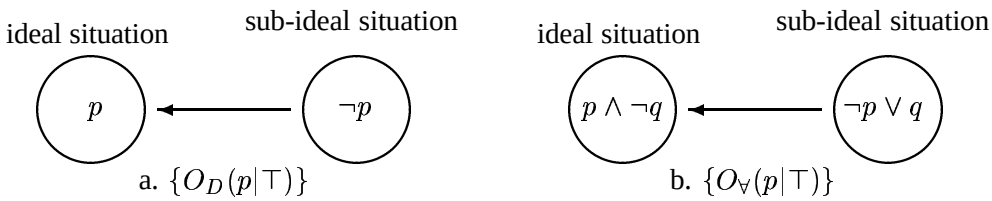



Figure 2.12: C-preferred models

A consequence of Example 2.45 is that preferential entailment $\sqsubseteq_c$ cannot be used if we only have minimizing obligations. This is surprising for the following reason. The reasoning scheme 'maximally connected' can be used for the logic $O_D$ to obtain strengthening of the antecedent, because it is used to derive $O_D(\alpha|\beta_1 \wedge \beta_2)$ from the obligation $O_D(\alpha|\beta_1)$. Thus, the obligation $O_\forall(\alpha \mid \beta_1 \wedge \beta_2)$ is derived from $O_D(\alpha \mid \beta_1) = O(\alpha \mid \beta_1) \wedge O_\forall(\alpha \mid \beta_1)$. One may (wrongly) think that this inference can be decomposed as follows.

$$\frac{O_D(\alpha|\beta_1)}{O_D(\alpha|\beta_1 \wedge \beta_2)} = \frac{O(\alpha|\beta_1) \wedge O_\forall(\alpha|\beta_1)}{O(\alpha|\beta_1 \wedge \beta_2) \wedge O_\forall(\alpha|\beta_1 \wedge \beta_2)} = \frac{O(\alpha|\beta_1)}{O(\alpha|\beta_1 \wedge \beta_2)} + \frac{O_\forall(\alpha|\beta_1)}{O_\forall(\alpha|\beta_1 \wedge \beta_2)}$$

The surprising result is that in the new scheme 'maximally connected,' $O_\forall(\alpha \mid \beta_1 \wedge \beta_2)$ is not derived from $O_\forall(\alpha \mid \beta_1)$, like in System Z, but from $O(\alpha \mid \beta_1)$. Hence, strengthening of the antecedent can be decomposed as follows. The obligation $O_\forall$ in $O_D$ is only used to make some moral dilemmas inconsistent, not for strengthening of the antecedent.

$$\frac{O_D(\alpha|\beta_1)}{O_D(\alpha|\beta_1 \wedge \beta_2)} = \frac{O(\alpha|\beta_1) \wedge O_\forall(\alpha|\beta_1)}{O(\alpha|\beta_1 \wedge \beta_2) \wedge O_\forall(\alpha|\beta_1 \wedge \beta_2)} = \frac{O(\alpha|\beta_1)}{O(\alpha|\beta_1 \wedge \beta_2)} + \frac{O(\alpha|\beta_1)}{O_\forall(\alpha|\beta_1 \wedge \beta_2)}$$

The following example shows that $O_D$ solves the cigarettes problem of Example 2.26 by weakening RSA.

**Example 2.46 (Cigarettes problem, continued)** Consider the set of obligations

$$S = \{O_D^c(\neg c \mid \top), O_D^c(c|k)\}$$

where $\neg c$ does not entail the negation of $k$ ($\overset{\leftrightarrow}{\Diamond}(k \wedge \neg c)$). $S$ is inconsistent, and we can derive an inconsistency as follows. The premise $O_D^c(\neg c \mid \top)$ entails the obligation $O^c(\neg c \mid \top)$, which entails $O^c(\neg c|k)$. The premise $O_D^c(c|k)$ entails $O_\forall^c(c|k)$, which is inconsistent with $O^c(\neg c|k)$.

Moreover, consider the set of obligations $S' = \{O^c_D(p_1 \mid \top), O^c_D(p_2 \mid \top)\}$. $S'$ is consistent, and the unique c-preferred model $M$ of $S'$ is given in Figure 2.13. We have $M \not\models O^c_D(p_1 \mid \neg(p_1 \wedge p_2))$, and thus $S' \not\models_c O^c_D(p_1 \mid \neg(p_1 \wedge p_2))$. Hence, the problematic inconsistency in Example 2.26 is blocked by weakening RSA.

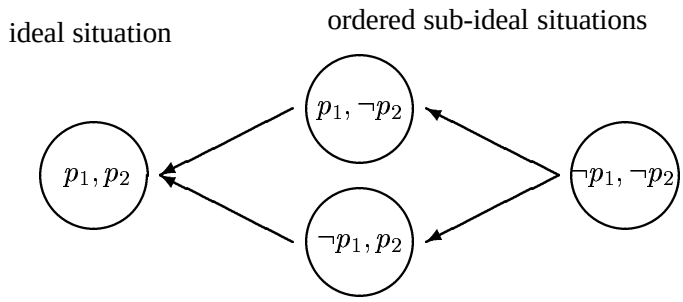



Figure 2.13: Semantic solution of the cigarettes problem (1)

Finally, consider the set of obligations $S'' = \{O^c_D(p \mid q_1), O^c_D(\neg p \mid q_2)\}$. $S''$ is consistent, and a c-preferred model $M$ of $S''$ is given in Figure 2.14. The ideal worlds satisfy $q_1 \rightarrow p$ and $q_2 \rightarrow \neg p$, and the subideal worlds either $\neg p \wedge q_1$ or $p \wedge q_2$. We have $M \not\models O(p \mid q_1 \wedge q_2)$ and thus $S \not\models_c O(p \mid q_1 \wedge q_2)$. Hence, the problematic inconsistency in Example 2.26 is blocked by weakening RSA. □

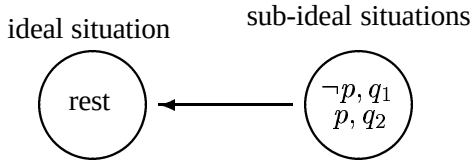



Figure 2.14: Semantic solution of the cigarettes problem (2)

In this section we showed that the cigarettes problem can be solved by the ordering obligation $O_D$, which combines ordering $O$ and minimizing $O_\forall$ obligations. Moreover, we introduced a new notion of preferential entailment based on 'maximally connected' models. At first sight, it seems that we need preferential entailment for universal-minimizing obligations $O_\forall$ like gravitating towards the ideal (System Z), because $\mathbf{RSA}_D$ is restricted by universal-minimizing obligations $O_\forall$. However, this is not the case, because our new scheme is a weaker scheme than gravitating towards the ideal. It is weaker in the sense that 'maximally connected' cannot be used with only universal-minimizing obligations $O_\forall$. In the next section we show how ordering and minimizing are combined in the two-phase approach with the no-dilemma assumption.

### 2.3.4 Combining ordering and minimizing

The two-phase approach with $O^c_D(\alpha|\beta)$ and $O^c_\forall(\alpha|\beta)$ works similar to the two-phase approach with $O^c(\alpha \mid \beta)$ and $O^c_\exists(\alpha \mid \beta)$. Proposition 2.43 is the counterpart of Proposition 2.21 for the second example of the two-phase approach. In this case, we have to use preferential entailment

$\models_c$ instead of $\models$. Preferential entailment is a typical mechanism from non-monotonic reasoning. The following example illustrates why the combination of ordering and minimizing is non-monotonic.

**Example 2.47 (Cigarettes problem, continued)** Consider the sets of dyadic obligations

$$S = \emptyset$$

$$S' = \{O^c_D(p_1|\top)\}$$

$$S'' = \{O^c_D(p_1|\top), O^c_D(p_2|\top)\}$$

The three unique c-preferred models of $S$, $S'$ and $S''$ are represented in Figure 2.15. We have $S' \models_{\sqsubset} O^c_\forall(p_1|\neg(p_1 \wedge p_2))$ and $S'' \not\models_{\sqsubset} O^c_\forall(p_1|\neg(p_1 \wedge p_2))$. Hence, by addition of a formula we loose conclusions. Moreover, it shows that the cigarettes problem in Example 2.26 is solved by weakening RSA, because with $S''$ the obligation $O^c_\forall(p_1|\top)$ is not strengthened to the obligation $O^c_\forall(p_1 \mid \neg(p_1 \wedge p_2))$. RSA is not valid, because there is not a unique most preferred obligation for the antecedent $\neg(p_1 \wedge p_2)$. However, we still have $S'' \models_{\sqsubset} O^c_\exists(p_2 \mid \neg(p_1 \wedge p_2))$ as well as $S'' \models_{\sqsubset} O^c_\exists(\neg p_2|\neg(p_1 \wedge p_2))$. □

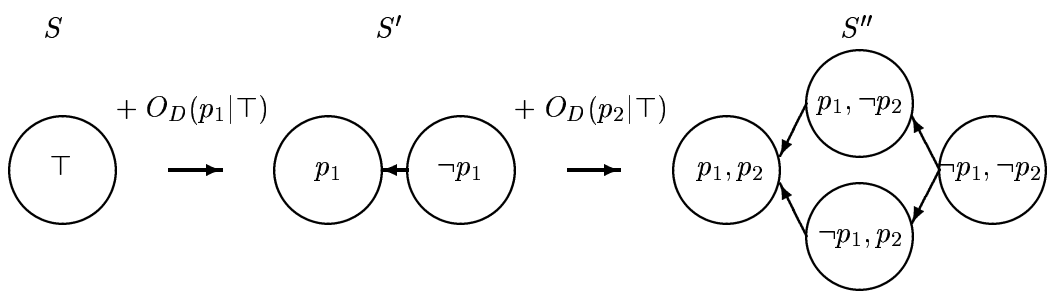



Figure 2.15: Dynamics of the two-phase approach

The models in Figure 2.15 illustrate the dynamics of preferential entailment. At the end of Section 2.2.4 we already remarked that we can distinguish two types of processes: the dynamic ordering of worlds and the static testing of minimal worlds. The dynamics of the ordering process are explicit in Figure 2.15. With no premises, all worlds are equally ideal. By addition of premise $O_D(p_1|\top)$, the $p_1$ worlds are strictly preferred over $\neg p_1$ worlds. By addition of the second premise $O_D(p_2 \mid \top)$, the $p_2$ worlds are strictly preferred over $\neg p_2$ worlds, and the $p_1 \wedge \neg p_2$ and $\neg p_1 \wedge p_2$ worlds become incomparable.

In this section we showed that strengthening of the antecedent and weakening of the consequent can be combined by combining the two usages of the preference ordering in a preference-based semantics of a deontic logic. The combination is the two-phase approach to deontic logic. The first phase corresponds to ordering, and the second phase corresponds to minimizing. The two phases are combined by the theorems $O(\alpha|\beta) \rightarrow O_\exists(\alpha|\beta)$ and $O_D(\alpha|\beta) \rightarrow O_\forall(\alpha|\beta)$. The combination of ordering and minimizing can be illustrated by Figure 2.16. The figure should be read as follows. The corners represent different types of deontic operators. The arrows represent logical implication. The figure also illustrates that the operators with the no-dilemma assumption imply the operators without it, i.e. the theorems $O_D(\alpha|\beta) \rightarrow O(\alpha|\beta)$ and $O_\forall(\alpha|\beta) \rightarrow O_\exists(\alpha|\beta)$. In the following section we extend our study with permission operators.

$$\begin{tikzcd} O \arrow[r] & O_\exists \\ O_D \arrow[u] \arrow[r] & O_\forall \arrow[u] \end{tikzcd}$$

Figure 2.16: Operators for combining ordering and minimizing

## 2.4 Permissions in the two-phase approach

Deontic logic is the logic of obligations, prohibitions and permissions. Different types of prohibitions can easily be formalized in 2DL by defining them in terms of obligations.

**Definition 2.48 (Prohibition)** The prohibition '$\alpha$ is forbidden to be the case if $\beta$ is the case,' written as $F(\alpha|\beta)$, is defined as follows.

$$F(\alpha|\beta) \quad =_{def} \quad O(\neg\alpha|\beta) \qquad \square$$

Dyadic (strong) permissions cannot be defined satisfactorily in terms of dyadic obligations (see Section 1.3.5). The main problem of a definition for dyadic permissions analogous to the definition of weak monadic permissions $P\alpha =_{def} \neg O\neg\alpha$ is that this definition implies the formula $O\alpha \vee P\neg\alpha$. Hence, it implies that every proposition is normed in the sense that every proposition $\alpha$ is either obligatory or its negation is permitted. In this section we extend the logic CT4O such that permissions can be represented satisfactorily. We show that they have two desirable properties. The first desirable property is the standard relation between obligations and permissions $O\alpha \rightarrow P\alpha$. The second desirable property is that we do *not* have that every proposition is normed $O\alpha \vee P\neg\alpha$.

The basic idea is that if obligations are *strict* preferences $O(\alpha|\beta) =_{def} (\alpha \wedge \beta) \succ (\neg\alpha \wedge \beta)$, then permissions are preferences defined by $P(\alpha \mid \beta) =_{def} (\alpha \wedge \beta) \succeq (\neg\alpha \wedge \beta)$. From the definition follows directly the first deontic axiom $O(\alpha|\beta) \rightarrow P(\alpha|\beta)$. Before we can define the permissions in the modal preference logic, we first we have to extend the logic CT4O to be able to refer to strictly preferred worlds.

### 2.4.1 The logic 2DL

In the logic CT4O, we have $\square$ for equivalent and strictly preferred worlds, and the operator $\overleftarrow{\square}$ for inaccessible worlds. For the permission operators in this section we need a modal operator $\overrightarrow{\square}$ for only the strictly preferred worlds. That is, for the preference ordering on worlds $w_1 \leq w_2$ we define the a-symmetric reduct $w_1 < w_2$ by the standard definition: $w_1 < w_2$ iff $w_1 \leq w_2$ and $w_2 \not\leq w_1$. Thus, we have the following truth conditions.

$M, w \models \square\alpha$ iff $\forall w' \in W$ if $w' \leq w$, then $M, w' \models \alpha$
$M, w \models \overleftarrow{\square}\alpha$ iff $\forall w' \in W$ if $w' \not\leq w$, then $M, w' \models \alpha$
$M, w \models \overrightarrow{\square}\alpha$ iff $\forall w' \in W$ if $w' \leq w$ and $w \not\leq w'$, then $M, w' \models \alpha$

The semantics for the trimodal logics will be based on the same Kripke structures used for

(mono-) modal logics, with additional truth conditions for $\overrightarrow{\Box}$ defined on these models, just like the logic of inaccessible worlds CT4O is based on Kripke models with additional truth conditions for $\overleftarrow{\Box}$. Thus, this additional operator adds no 'ontological baggage' to our semantic conception of ideality (see [Bou92a, p.83]). The following axiomatization has been suggested by Wiebe van der Hoek.

**Definition 2.49 (**2DL**)** The trimodal language $\mathcal{L}$ is formed from a denumerable set of propositional variables together with the connectives $\neg$, $\rightarrow$, and the three normal modal connectives $\Box$, $\overleftarrow{\Box}$ and $\overrightarrow{\Box}$. Dual 'possibility' connectives $\Diamond$, $\overleftarrow{\Diamond}$ and $\overrightarrow{\Diamond}$ are defined as usual and two additional modal connectives $\overleftrightarrow{\Box}$ and $\overleftrightarrow{\Diamond}$ are defined as follows.

$$\begin{array}{lll} \Diamond\alpha & =_{def} & \neg\,\Box\,\neg\alpha \\ \overleftarrow{\Diamond}\,\alpha & =_{def} & \neg\,\overleftarrow{\Box}\neg\alpha \\ \overrightarrow{\Diamond}\,\alpha & =_{def} & \neg\,\overrightarrow{\Box}\neg\alpha \end{array} \qquad \begin{array}{lll} \overleftrightarrow{\Box}\alpha & =_{def} & \Box\alpha\wedge\overleftarrow{\Box}\alpha \\ \overleftrightarrow{\Diamond}\,\alpha & =_{def} & \Diamond\alpha\vee\overleftarrow{\Diamond}\,\alpha \end{array}$$

The logic 2DL is the smallest $S \subset \mathcal{L}$ such that $S$ contains classical logic and the following axiom schemata, and is closed under the following rules of inference.

| | | | |
|---|---|---|---|
| **K** | $\Box(\alpha \rightarrow \beta) \rightarrow (\Box\alpha \rightarrow \Box\beta)$ | **Nes** | From $\alpha$ infer $\overleftrightarrow{\Box}\alpha$ |
| **K′** | $\overleftarrow{\Box}(\alpha \rightarrow \beta) \rightarrow (\overleftarrow{\Box}\alpha \rightarrow\overleftarrow{\Box}\beta)$ | **MP** | From $\alpha \rightarrow \beta$ and $\alpha$ infer $\beta$ |
| **K″** | $\overrightarrow{\Box}(\alpha \rightarrow \beta) \rightarrow (\overrightarrow{\Box}\alpha \rightarrow\overrightarrow{\Box}\beta)$ | | |
| **T** | $\Box\alpha \rightarrow \alpha$ | | |
| **4** | $\Box\alpha \rightarrow \Box\Box\alpha$ | | |
| **H** | $\overleftrightarrow{\Diamond}(\Box\alpha\wedge \overleftarrow{\Box}\beta) \rightarrow\overleftrightarrow{\Box}(\alpha \vee \beta)$ | | |
| **R1** | $(\alpha\wedge \overrightarrow{\Box}\beta) \rightarrow \Box(\beta \vee \Diamond\alpha)$ | | |
| **R2** | $\Box\alpha \rightarrow\overrightarrow{\Box}\alpha$ | | |
| **R3** | $\alpha \rightarrow\overrightarrow{\Box}\overleftarrow{\Diamond}\,\alpha$ | | |

We write $\vdash_{2dl}$ for derivability in 2DL. □

The modal connective $\Box$ refers to accessible worlds, the modal connective $\overleftarrow{\Box}$ to inaccessible worlds and the modal connective $\overrightarrow{\Box}$ to one-way accessible worlds.

**Definition 2.50 (**2DL **models)** A 2DL model $M = \langle W, \leq, V\rangle$ is a CT4O model. We have:

$M, w \models \Box\alpha$ iff $\forall w' \in W$ if $w' \leq w$, then $M, w' \models \alpha$
$M, w \models \overleftarrow{\Box}\alpha$ iff $\forall w' \in W$ if $w' \not\leq w$, then $M, w' \models \alpha$
$M, w \models \overrightarrow{\Box}\alpha$ iff $\forall w' \in W$ if $w' \leq w$ and $w \not\leq w'$, then $M, w' \models \alpha$

We write $\models_{2dl}$ for logical entailment in 2DL. □

We can write 2DL-models as follows. A 2DL model $M = \langle W, R_1, R_2, R_3, V\rangle$ is a Kripke model with three accessibility relations such that $R_1$ is reflexive and transitive and the following

two additional conditions hold $R_2(x,y) \leftrightarrow \neg R_1(x,y)$ and $R_3(x,y) \leftrightarrow (R_1(x,y) \wedge \neg R_1(y,x))$. The two conditions are equivalent to the following set of conditions.

$$
\begin{array}{l}
R_1(x,y) \rightarrow \neg R_2(x,y) \\
\neg R_1(x,y) \rightarrow R_2(x,y) \\
R_3(x,y) \rightarrow R_1(x,y) \\
R_3(x,y) \rightarrow R_2(y,x) \\
R_1(x,y) \rightarrow (R_3(x,y) \vee R_1(y,x))
\end{array}
$$

We have:

$$
\begin{array}{l}
M,w \models \Box\alpha \text{ iff } \forall w' \in W \text{ if } w'R_1w, \text{ then } M,w' \models \alpha \\
M,w \models \overleftarrow{\Box}\alpha \text{ iff } \forall w' \in W \text{ if } w'R_2w, \text{ then } M,w' \models \alpha \\
M,w \models \overrightarrow{\Box}\alpha \text{ iff } \forall w' \in W \text{ if } w'R_3w, \text{ then } M,w' \models \alpha
\end{array}
$$

Before we prove soundness and completeness of the logic, we first show that the axioms with which we extended the logic CT4O characterize the desired properties of $R_3$. This, however, is not enough for a completeness proof of the logic 2DL, because the axiomatization of the logic CT4O is not cumulative. That is, completeness of the logic CT4O is not assured when axioms are added to it. We therefore also give a (standard) completeness proof based on canonical models. On first reading, the reader is advised to skip the remainder of this rather complex technical section. It is not used later in this thesis.

**Proposition 2.51** *The class of frames with $R_3(x,y) \rightarrow R_1(x,y)$ is characterized by $\Box\alpha \rightarrow \overrightarrow{\Box}\alpha$, that is, the set of frames $F = \langle W, R_1, R_2, R_3\rangle$ with $R_3(x,y) \rightarrow R_1(x,y)$ is the set of frames such that $F \models \Box\alpha \rightarrow \overrightarrow{\Box}\alpha$ (i.e. for all models $M = \langle W, R_1, R_2, R_3, V\rangle$ of the frame, we have $M \models \Box\alpha \rightarrow \overrightarrow{\Box}\alpha$).*

**Proof** $\Rightarrow$ *Consider a model that satisfies the frame condition $R_3(x,y) \rightarrow R_1(x,y)$. Proof by contraposition. Assume a world $w$ such that $M,w \not\models \Box\alpha \rightarrow \overrightarrow{\Box}\alpha$. We have $M,w \models \Box\alpha$ and $M,w \not\models \overrightarrow{\Box}\alpha$. From $M,w \not\models \overrightarrow{\Box}\alpha$ follows that there is a world $w'$ such that $R_3(w,w')$ and $M,w' \models \neg\alpha$. From the condition on frames follows $R_1(w,w')$. This contradicts $M,w \models \Box\alpha$.*

$\Leftarrow$ *Consider a frame $F = \langle W, R_1, R_2, .R_3\rangle$ that does not satisfy the condition $R_3(x,y) \rightarrow R_1(x,y)$. Hence, there are worlds $w$ and $w'$ such that $R_3(w,w')$ and $\neg R_1(w,w')$. Choose a model $M = \langle W, R_1, R_2, R_3, V\rangle$ on $F$ such that $p$ is false at $w'$ and true at all other worlds. We have $M,w \models \Box p$ and $M,w \not\models \overrightarrow{\Box}p$. Hence, we have $M,w \not\models \Box p \rightarrow \overrightarrow{\Box}p$.*

The following result is due to Lemmon [Seg70], which is of special significance for tense logic (see also [Hum83]).

**Proposition 2.52** *The class of frames with $R_3(x,y) \rightarrow R_2(y,x)$ is characterized by $\alpha \rightarrow \overrightarrow{\Box}\overleftarrow{\Diamond}\alpha$, that is, the set of frames $F = \langle W, R_1, R_2, R_3\rangle$ with $R_3(x,y) \rightarrow R_2(y,x)$ is the set of frames such that $F \models \alpha \rightarrow \overrightarrow{\Box}\overleftarrow{\Diamond}\alpha$ (i.e. for all models $M = \langle W, R_1, R_2, R_3, V\rangle$ of the frame, we have $M \models \alpha \rightarrow \overrightarrow{\Box}\overleftarrow{\Diamond}\alpha$).*

**Proof** *Analogous to the proof of Proposition 2.51.*

$\Rightarrow$ *Consider a model that satisfies the frame condition* $R_3(x,y) \rightarrow R_2(y,x)$. *Proof by contraposition. Assume a world* $w$ *such that* $M,w \not\models \alpha \rightarrow \overrightarrow{\Box}\overleftarrow{\Diamond}\alpha$. *We have* $M,w \models \alpha$ *and* $M,w \not\models \overrightarrow{\Box}\overleftarrow{\Diamond}\alpha$. *From* $M,w \not\models \overrightarrow{\Box}\overleftarrow{\Diamond}\alpha$ *follows that there is a world* $w'$ *such that* $R_3(w,w')$ *and* $M,w' \models \neg\overleftarrow{\Diamond}\alpha$, *i.e.* $M,w' \models \overleftarrow{\Box}\neg\alpha$. *From the frame condition follows* $R_2(w',w)$, *and therefore* $M,w \models \neg\alpha$. *Contradiction.*

$\Leftarrow$ *Consider a frame* $F$ *that does not satisfy the frame condition* $R_3(x,y) \rightarrow R_2(y,x)$. *Hence, there are worlds* $w$ *and* $w'$ *such that* $R_3(w,w')$ *and* $\neg R_2(w',w)$. *Choose a model* $M$ *with interpretation* $V$ *such that* $p$ *is true at* $w$ *and false at all other worlds. We have* $M,w \models p$ *and* $M,w' \not\models \overleftarrow{\Diamond}p$. *From* $M,w' \not\models \overleftarrow{\Diamond}p$ *and* $R_3(w,w')$ *follows* $M,w \not\models \overrightarrow{\Box}\overleftarrow{\Diamond}p$. *Hence, we have* $M,w \not\models p \rightarrow \overrightarrow{\Box}\overleftarrow{\Diamond}p$.

**Proposition 2.53** *The class of frames* $R_1(x,y) \rightarrow (R_3(x,y) \vee R_1(y,x))$ *is characterized by* $(\alpha \wedge \overrightarrow{\Box}\beta) \rightarrow \Box(\beta \vee \Diamond\alpha)$, *that is, the set of frames* $F = \langle W, R_1, R_2, R_3\rangle$ *with* $R_1(x,y) \rightarrow (R_3(x,y) \vee R_1(y,x))$ *is the set of frames such that we have* $F \models (\alpha \wedge \overrightarrow{\Box}\beta) \rightarrow \Box(\beta \vee \Diamond\alpha)$ *(i.e. for all models* $M = \langle W, R_1, R_2, R_3, V\rangle$ *of the frame, we have* $M \models (\alpha \wedge \overrightarrow{\Box}\beta) \rightarrow \Box(\beta \vee \Diamond\alpha)$*).*

**Proof** *Analogous to the proof of Proposition 2.51 and 2.52.*

$\Rightarrow$ *Consider a model that satisfies the condition* $R_1(x,y) \rightarrow (R_3(x,y) \vee R_1(y,x))$. *Proof by contraposition. Assume a* $w$ *such that* $M,w \not\models (\alpha \wedge \overrightarrow{\Box}\beta) \rightarrow \Box(\beta \vee \Diamond\alpha)$. *We have* $M,w \models \alpha$, $M,w \models \overrightarrow{\Box}\beta$ *and* $M,w \not\models \Box(\beta \vee \Diamond\alpha)$. *From* $M,w \not\models \Box(\beta \vee \Diamond\alpha)$ *follows that there is a world* $w'$ *such that* $R_1(w,w')$ *and* $M,w' \not\models \beta$ *and* $M,w' \not\models \Diamond\alpha$. *From the condition on the frames follows* $R_3(w,w')$ *or* $R_1(w',w)$. *The first conflicts with* $M,w \models \overrightarrow{\Box}\beta$ *and* $M,w' \not\models \beta$, *the latter with* $M,w \models \alpha$ *and* $M,w' \not\models \Diamond\alpha$. *Contradiction.*

$\Leftarrow$ *Consider a frame that does not satisfy the condition* $R_1(x,y) \rightarrow (R_3(x,y) \vee R_1(y,x))$. *Hence, there are worlds* $w$ *and* $w'$ *such that* $R_1(w,w')$, $\neg R_3(w,w')$ *and* $\neg R_1(w',w)$. *Choose a model* $M$ *with an interpretation* $V$ *such that* $p$ *is true at* $w$ *and false at all other worlds, and* $q$ *is false at* $w'$ *and true at all other worlds. We have* $M,w \models p$, $M,w \models \overrightarrow{\Box}q$, $M,w' \models \neg q$, $M,w' \models \Box\neg p$. *From the latter two follows* $M,w \models \neg\Box(q \vee \Diamond p)$. *Hence,* $M \not\models (p \wedge \overrightarrow{\Box}q) \rightarrow \Box(q \vee \Diamond p)$.

**Proposition 2.54** *(Soundness) If* $\vdash_{2dl} \alpha$ *, then* $\models_{2dl} \alpha$.

**Proof** *Soundness from* 2DL *follows from the soundness of CT4O and Proposition 2.51, 2.52, and 2.53.*

Proposition 2.51, 2.52, and 2.53 are not sufficient for a completeness proof of 2DL, because the completeness proof of Humberstone and Boutilier of the logic of inaccessible worlds is non-cumulative (or nonadditive). The propositions above belong to correspondence theory (they give the class of all those frames on which every theorem of the system is valid), whereas Humberstone's theorem belongs to theory of completeness, see [Hum83] for a discussion. Hence, for completeness of 2DL we have to show that the canonical model construction of CT4O can be used for the a-symmetric reduct. To show completeness it is sufficient to show that $\alpha$ is falsifiable for any non-theorem $\alpha$. Let $\Gamma$ be some maximal 2DL-consistent set (MCS) which contains

$\neg\alpha$, we will construct a model $M = (W, \leq, V)$ which falsifies $\alpha$. This technique is employed in [Hum83, Bou94a].

Humberstone observes that the argument is modeled after Creswell's adaptation to modal logic of the method Henkin used to prove the completeness of first-order logic, rather than the more widely known adaptation of that method due to Scott and Makinson [Cre67]. In the former case maximal consistent sets of formulas are correlated with elements of the falsifying model but the correlation is not required to be one-one, so that there is more freedom in constructing the required accessibility relation than on the latter approach – as generally implemented – in which the maximal consistent sets are identified with the point of the model serving to falsify any given nontheorem. Humberstone observes that the 'canonical models' of the latter approach are not, as they stand, very helpful in delivering the completeness result.

**Definition 2.55 (Canonical model)** A canonical model $M^c = \langle W, R_1, R_2, R_3, V\rangle$ of $\Gamma$ is constructed with $W$ a set of MCSs and three relations $R_1$, $R_2$ and $R_3$, where $R_2$ is intended to represent the complement of $R_1$, and $R_3$ is intended to be the a-symmetric reduct of $R_1$. The construction proceeds as follows. We start at stage 0 by adding $\Gamma$ to $W$, so that $W = \{\Gamma\}$ and $R_1 = R_2 = R_3 = \emptyset$. At each following stage $i$, for each set $\Lambda$ added to $W$ at stage $i-1$ we do the following:

1. for each formula $\Diamond\beta \in \Lambda$ add a MCS $\Lambda'$ to $W$ where $\{\beta\} \cup \{\gamma \mid \Box\gamma \in \Lambda\} \subseteq \Lambda'$, and add $\langle \Lambda, \Lambda'\rangle$ to $R_1$,

2. similarly for each formula $\overleftarrow{\Diamond}\beta \in \Lambda$ add a MCS $\Lambda'$ to $W$ where $\{\beta\} \cup \{\gamma \mid \overleftarrow{\Box}\gamma \in \Lambda\} \subseteq \Lambda'$, and add $\langle \Lambda, \Lambda'\rangle$ to $R_2$,

3. similarly for each formula $\overrightarrow{\Diamond}\beta \in \Lambda$ add a MCS $\Lambda'$ to $W$ where $\{\beta\} \cup \{\gamma \mid \overrightarrow{\Box}\gamma \in \Lambda\} \subseteq \Lambda'$, and add $\langle \Lambda, \Lambda'\rangle$ to $R_3$.

The canonical model $M^c$ is the totality of this (typically) infinite construction with the interpretation $V$ such that $V_w(\alpha) = \alpha \in w$ for atomic $\alpha$. $\Box$

From the construction follows that we do not have $R_i(w_1, w_2)$ and $R_j(w_1, w_2)$ for $i \neq j$ (and therefore the correlation of MCSs and worlds is not one-one). Evaluating the truth conditions of $\overleftarrow{\Box}$and $\overrightarrow{\Box}$with respect to $R_2$ and $R_3$ (as if $R_2$ and $R_3$ were the complement and the a-symmetric reduct of $R_1$) we can prove the following proposition.

**Proposition 2.56** *Let $M^c$ be a canonical model (for $\Gamma$). $M^c, w \models \beta$ iff $\beta \in w$.*

**Proof** *By induction on the structure of $\beta$. For atomic $\beta$, this follows by the definition of $V$. Assuming this for $\alpha$ and $\beta$, clearly it holds for both $\neg\alpha$ and $\alpha \rightarrow \beta$ by standard properties of maximal consistent sets. Now suppose $\Box\beta \in w$. By the construction of $M^c$, for all $R_1(w, v)$, $M^c, v \models \beta$, therefore $M^c, w \models \Box\beta$. Now suppose $\Box\beta \notin w$, and therefore $\Diamond\neg\beta \in w$. By the construction of $M^c$, there is some $R_1(w, v)$ such that $M^c, v \not\models \beta$, therefore $M^c, w \not\models \Box\beta$. The same arguments hold for $\overleftarrow{\Box}$and $\overrightarrow{\Box}$.*

Now we have a canonical model $M^c$ that falsifies $\alpha$, as $\neg\alpha \in \Gamma$ and by Proposition 2.56, $M^c, \Gamma \models \neg\alpha$. However, $M^c$ is not a 2DL-model, since $R_1$ is neither reflexive nor transitive, $R_2$

is not the complement of $R_1$ and $R_3$ is not the a-symmetric reduct of $R_1$. We now show that $R_1$, $R_2$ and $R_3$ can be extended such that $R_1$ does possess the desired properties and $R_2$ and $R_3$ have the desired relation with $R_1$, while not changing the fact that $M, w \models \beta$ iff $\beta \in w$.

**Definition 2.57 (Extended model)** Let $M^c = \langle W, R_1^c, R_2^c, R_3^c, V\rangle$ be a canonical model. The extended model $M^e = \langle W, R_1, R_2, R_3, V\rangle$ of the canonical model $M^c$ is defined as follows. We add relations such that $R_2$ is the complement of $R_1$, and $R_3$ is the a-symmetric reduct of $R_1$. We use the following order:

1. For each $R_3(x, y)$, we add $R_1(x, y)$ and $R_2(y, x)$.
2. We add $R_1(x, y)$ and $R_2(x, y)$ such that $R_1$ is reflexive and transitive, and for all $x$ and $y$ we have either $R_1(x, y)$ or $R_2(x, y)$. We insist that $R_1$ is completed maximally before we complete $R_2$. For example, at the step where we decide to add each pair of worlds to $R_1$ or $R_2$, we can consider the union of the family of all possible relations $R_1$ on $W \times W$ that respect on restrictions on accessibility; we take this set to be $R_1$ and let $R_2$ then be $W \times W - R_1$.
3. For each $R_1(x, y)$ and not $R_1(y, x)$, we add $R_3(x, y)$.

□

The following proposition shows that the extension of the model does not influence Proposition 2.56.

**Proposition 2.58** *Let $M^e$ be an extended model of a canonical model $M^c$ (for $\Gamma$). We have $M^e, w \models \beta$ iff $\beta \in w$.*

**Proof** *We have $M^c, w \models \beta \Leftrightarrow M^e, w \models \beta$. Hence, the additions do not affect the set of satisfiable formulas.*

1. *The addition of $R_1(x, y)$ and $R_2(y, x)$ for each $R_3(x, y)$ does not affect the set of satisfiable formulas as a result of the axioms $\Box\gamma \rightarrow \overrightarrow{\Box}\gamma$ and $\gamma \rightarrow \overrightarrow{\Box}\overleftarrow{\Diamond}\gamma$. First, suppose that there are $x$ and $y$ with $R_3(x, y)$ such that $R_1(x, y)$ affects the truth of formula. Then there must be $\Box\beta \in x$ and $\beta \notin y$. If $\Box\beta \in x$ then $\overrightarrow{\Box}\beta \in x$ (by the axiom) and $\beta \in y$ (because $R_3(x, y)$). Contradiction. Second, suppose that there are $x$ and $y$ with $R_3(x, y)$ such that $R_2(y, x)$ affects the truth of a formula. Then there must be $\overleftarrow{\Box}\beta \in y$ and $\beta \notin x$. If $\overleftarrow{\Box}\beta \in y$ then $\overrightarrow{\Diamond}\overleftarrow{\Box}\beta \in x$ (by $R_3(x, y)$) and $\beta \in x$ (by contraposition of the axiom). Contradiction.*
2. *We add $R_1(x, y)$ and $R_2(x, y)$ such that $R_1$ is reflexive and transitive, and for all $x$ and $y$ we have either $R_1(x, y)$ or $R_2(x, y)$. Axiom* **H***: $\overleftrightarrow{\Diamond}(\Box\gamma \wedge \overleftarrow{\Box}\beta) \rightarrow \overleftrightarrow{\Box}(\gamma \vee \beta)$ implies the Humberstone schemata, see [Bou92a].*

   **H**$^*$*:* $\quad D(\Box\alpha \wedge \overleftarrow{\Box}\beta) \rightarrow B(\alpha \vee \beta)$

   *In this schema, $D$ is any sequence of the connectives $\Diamond$ and $\overleftarrow{\Diamond}$ having length $\geq 0$, and $B$ is any such sequence of $\Box$ and $\overleftarrow{\Box}$. Suppose that we do not have $R_1(w_1, w_2)$ nor $R_2(w_1, w_2)$,*

*and that they cannot be added without affecting the truth of a sentence. Then there must be some $\Box\beta \in w_1$ such that $\beta \notin w_2$, and some $\overleftarrow{\Box}\gamma \in w_1$ such that $\gamma \notin w_2$. Both $w_1$ and $w_2$ must be some finite 'distance' away from starting point $\Gamma$, say $m$ and $n$ 'steps' respectively. Following the 'path' which lead to the addition of $w$ to $W$, we have $M, \Gamma \models D_1(\Box\beta \wedge \overleftarrow{\Box}\gamma)$ where $D_1$ is a string of $m$ $\Diamond$'s and $\overleftarrow{\Diamond}$'s corresponding to how $w_1$ was added. Similarly, $M, \Gamma \models D_2(\neg\beta \wedge \neg\gamma)$ where $D_2$ is a string of $n$ $\Diamond$'s and $\overleftarrow{\Diamond}$'s corresponding to how $w_2$ was added. But this sentence is equivalent to $\neg B_2(\beta \vee \gamma)$, where $B_2$ is formed by replacing $\Diamond$ and $\overleftarrow{\Diamond}$ with $\Box$ and $\overleftarrow{\Box}$ (respectively) in $D_2$. This means both $D_1(\Box\beta \wedge \overleftarrow{\Box}\gamma) \in \Gamma$ and $\neg B_2(\beta \vee \gamma) \in \Gamma$, contradicting the Humberstone schema. Since $\Gamma$ is consistent, $(w_1, w_2)$ can be added to either $R_1$ or $R_2$ without affecting the truth of formulas at any world in $W$, and hence $R_1$ and $R_2$ can be extended to complement one another, making valuation of $\overleftarrow{\Box}$ with respect to $R_2$ the same as valuation with respect to the standard truth conditions.*

*We can ensure that $R_1$ is reflexive, as well. Adding $R_1(w, w)$ affects the truth of some sentence only if there is some $\beta$ such that $\Box\beta \in w$ and $\beta \notin w$; but this contradicts the axiom* **T** *and the fact that $w$ is a MCS.*

*For transitivity, assume $R_1(w_1, w_2)$ and $R_1(w_2, w_3)$. Adding $R_1(w_1, w_3)$ can only affect truth if there is some $\beta$ such that $\Box\beta \in w_1$ and $\beta \notin w_3$. Since $\Box\beta \in w_1$, by axiom* **4**, *$\Box\Box\beta \in w_1$. This means $\Box\beta \in w_2$ and $\beta \in w_3$, contradicting the assumption.*

*Boutilier observes that there may be some interaction during these 'steps' whereby certain pairs of worlds are moved from the set $R_2$ to $R_1$; but clearly nothing in principle stops one from constructing a suitable model with the appropriate constraints being fulfilled by the relations. He further observes that if we insist that $R_1$ is completed maximally before we complete $R_2$, there need not to be any interaction.*

3. *For each $R_1(x, y)$ and not $R_1(y, x)$, we add $R_3(x, y)$. Given $R_1(x, y)$, the addition of either $R_1(y, x)$ or $R_3(x, y)$ does not influence the set of sentences because of the axiom $(\gamma \wedge \overrightarrow{\Box}\beta) \rightarrow \Box(\beta \vee \Diamond\gamma))$. For suppose that there are $x$ and $y$ with $R_1(x, y)$ such that neither $R_1(y, x)$ nor $R_3(x, y)$ can be added. Then there must be $\Box\neg\gamma \in y$ and $\neg\gamma \notin x$, and there must be $\overrightarrow{\Box}\beta \in x$ and $\beta \notin y$. This conflicts with the axiom. We cannot add $R_1(y, x)$ due to the maximality of the construction under (2), thus we can add $R_3(x, y)$.*

After the second step, $R_2$ is the complement of $R_1$ and after the third step $R_3$ is the a-symmetric reduct of $R_1$. As a consequence, the extended model is a 2DL-model.

**Proposition 2.59** *(Completeness) If $\models_{2dl} \alpha$ , then $\vdash_{2dl} \alpha$.*

**Proof** *By contraposition. For every non-theorem $\alpha$, we can construct a canonical model $M^c$ and extend this model to a 2DL model $M^e$. Hence, we can construct a 2DL-model which falsifies the non-theorem $\alpha$.*

**Theorem 2.60** *The system 2DL is characterized by the class of 2DL-models; that is, $\vdash_{2dl} \alpha$ iff $\models_{2dl} \alpha$.*

**Proof** *Follows directly from Proposition 2.54 (soundness) and Proposition 2.59 (completeness).*

In the following section we use the logic 2DL to model different types of permissions.

### 2.4.2 Permissions

Permissions are defined as a type of non-strict preferences. The definitions are analogous to the definitions of the different types of obligations. For readability we do not give the permission operators with a contingency clause (i.e. with $c$ and $cc$ conditions).

**Definition 2.61 (Dyadic permission)** Dyadic permission '$\alpha$ is permitted to be the case if $\beta$ is the case', written as $P(\alpha \mid \beta)$, is defined as some type of non-strict preference of $\alpha \wedge \beta$ over $\neg\alpha \wedge \beta$. Several types of non-strict preferences of $\alpha_1$ over $\alpha_2$, written as $\alpha_1 \succeq \alpha_2$, $\alpha_1 \succeq_\exists \alpha_2$ and $\alpha_1 \succeq_\forall \alpha_2$, are defined as follows.

$$
\begin{array}{lll}
\alpha_1 \succeq_\exists \alpha_2 & =_{def} & \overset{\leftrightarrow}{\Diamond}(\alpha_1 \wedge \overset{\rightarrow}{\Box}\neg\alpha_2) \vee \overset{\leftrightarrow}{\Box}\neg\alpha_1 \\
P_\exists(\alpha|\beta) & =_{def} & (\alpha \wedge \beta) \succeq_\exists (\neg\alpha \wedge \beta) \\
 & \leftrightarrow & \overset{\leftrightarrow}{\Diamond}(\alpha \wedge \beta \wedge \overset{\rightarrow}{\Box}(\beta \rightarrow \alpha)) \vee \overset{\leftrightarrow}{\Box}\neg(\alpha \wedge \beta) \\
\alpha_1 \succeq_\forall \alpha_2 & =_{def} & \overset{\leftrightarrow}{\Box}(\alpha_2 \rightarrow \Diamond(\alpha_1 \wedge \overset{\rightarrow}{\Box}\neg\alpha_2)) \vee \overset{\leftrightarrow}{\Box}\neg\alpha_1 \\
P_\forall(\alpha|\beta) & =_{def} & (\alpha \wedge \beta) \succeq_\exists (\neg\alpha \wedge \beta) \\
 & \leftrightarrow & \overset{\leftrightarrow}{\Box}(\beta \rightarrow \Diamond(\alpha \wedge \beta \wedge \overset{\rightarrow}{\Box}(\beta \rightarrow \alpha))) \vee \overset{\leftrightarrow}{\Box}\neg(\alpha \wedge \beta) \\
\alpha_1 \succeq_s \alpha_2 & =_{def} & \overset{\leftrightarrow}{\Box}(\alpha_1 \rightarrow \overset{\rightarrow}{\Box}\neg\alpha_2) \\
P(\alpha|\beta) & =_{def} & (\alpha \wedge \beta) \succeq_s (\neg\alpha \wedge \beta) \\
 & \leftrightarrow & \overset{\leftrightarrow}{\Box}((\alpha \wedge \beta) \rightarrow \overset{\rightarrow}{\Box}(\beta \rightarrow \alpha)) \\
P_D(\alpha|\beta) & =_{def} & (\alpha \wedge \beta) \succeq_s (\neg\alpha \wedge \beta) \wedge (\alpha \wedge \beta) \succeq_\forall (\neg\alpha \wedge \beta)
\end{array}
$$

□

The following proposition shows that ordering and minimizing permissions consider all respectively only preferred worlds.

**Proposition 2.62** *Let $M$ be a CT4O model. For a world $w \in W$, we have*

- *$M, w \models P(\alpha|\beta)$ iff for all $w_1, w_2 \in W$ such that $M, w_1 \models \alpha \wedge \beta$ and $M, w_2 \models \neg\alpha \wedge \beta$, it is true that $w_2 \not< w_1$.*
- *$M, w \models P_\exists(\alpha|\beta)$ iff there are no $\alpha \wedge \beta$ worlds, or*
  1. *there is a preferred $\beta$ world that satisfies $\alpha$, or*
  2. *there is an infinite descending chain that does have a $\alpha \wedge \beta$ world strictly below which $\alpha$ is true in all $\beta$ worlds.*
- *$M, w \models P_\forall(\alpha|\beta)$ iff there are no $\alpha \wedge \beta$ worlds, or*
  1. *there is not an equivalence class of preferred $\beta$ worlds which satisfy $\alpha$, and*
  2. *every infinite descending chain that contains $\beta$ worlds has a $\alpha \wedge \beta$ world strictly below which $\alpha$ is true in all $\beta$ worlds.*

**Proof** *Analogous to the proof of Proposition 2.6, 2.18 and 2.28.*

The following proposition shows the relation between the different types of permission.

**Proposition 2.63** *The logic* 2DL *has the following theorems.*

$$P_D(\alpha|\beta) \rightarrow P_\forall(\alpha|\beta)$$
$$P(\alpha|\beta) \rightarrow P_\exists(\alpha|\beta)$$
$$P_D(\alpha|\beta) \rightarrow P(\alpha|\beta)$$
$$P_\forall(\alpha|\beta) \rightarrow P_\exists(\alpha|\beta)$$

**Proof** *Analogous to the proofs of the relations between the different types of obligations (Proposition 2.21, 2.30 and 2.43). It follows directly from the semantics (Proposition 2.62).*

The following proposition shows the relation between obligations and permissions.

**Proposition 2.64** *The logic* 2DL *has the following theorems.*

| | |
|---|---|
| **D** | $O(\alpha\|\beta) \rightarrow P(\alpha\|\beta)$ |
| $\mathbf{D}_D$ | $O_D(\alpha\|\beta) \rightarrow P_D(\alpha\|\beta)$ |
| $\mathbf{D}_\exists$ | $O_\exists(\alpha\|\beta) \rightarrow P_\exists(\alpha\|\beta)$ |
| $\mathbf{D}_\forall$ | $O_\forall(\alpha\|\beta) \rightarrow P_\forall(\alpha\|\beta)$ |

**Proof** *Follows directly from axiom* $\Box\alpha \rightarrow \overrightarrow{\Box}\alpha$.

In Section 2.2.3 we considered dilemmas, i.e. conflicts between two obligations. Obviously, there cannot be conflicts between two permissions. The following example illustrates conflicts between obligations and permissions.

**Proposition 2.65** *The logic* 2DL *has the following theorems.*

$$(\overset{\leftrightarrow}{\Diamond}(\alpha \wedge \beta) \wedge \overset{\leftrightarrow}{\Diamond}(\neg\alpha \wedge \beta)) \rightarrow \neg(P_\exists(\alpha|\beta) \wedge O_\forall(\neg\alpha|\beta))$$
$$(\overset{\leftrightarrow}{\Diamond}(\alpha \wedge \beta) \wedge \overset{\leftrightarrow}{\Diamond}(\neg\alpha \wedge \beta)) \rightarrow \neg(P_\forall(\alpha|\beta) \wedge O_\exists(\neg\alpha|\beta))$$
$$(\overset{\leftrightarrow}{\Diamond}(\alpha \wedge \beta) \wedge \overset{\leftrightarrow}{\Diamond}(\neg\alpha \wedge \beta)) \rightarrow \neg(P_\forall(\alpha|\beta) \wedge O_\forall(\neg\alpha|\beta))$$

*The logic* 2DL *does not have the following theorems.*

$$(\overset{\leftrightarrow}{\Diamond}(\alpha \wedge \beta) \wedge \overset{\leftrightarrow}{\Diamond}(\neg\alpha \wedge \beta)) \rightarrow \neg(P_\exists(\alpha|\beta) \wedge O_\exists(\neg\alpha|\beta))$$
$$P_\exists(\alpha|\beta) \vee O_\forall(\neg\alpha|\beta)$$
$$P_\forall(\alpha|\beta) \vee O_\exists(\neg\alpha|\beta)$$
$$P_\exists(\alpha|\beta) \vee O_\exists(\neg\alpha|\beta)$$

**Proof** *For the latter three non-theorems, consider a model* $M$ *that consists of an infinite descending chain with alternating* $p$ *and* $\neg p$ *worlds, see Figure 2.17. For any world* $w$ *we have* $M, w \models \neg P_\exists(p|\beta)$ *and* $M, w \models \neg O_\exists(p|\beta)$.

We defined two desirable properties for (strong) permissions. The first desirable property is that obligations derive permissions and the second desirable property is that we do not have that

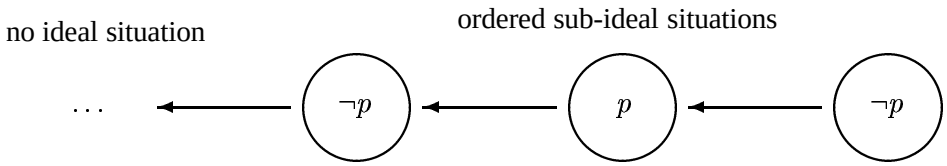



Figure 2.17: Infinite descending chain

every proposition is normed. The operators $P$ introduced in this section have the two desirable properties. The first property is shown in Proposition 2.64 and the second property is shown in Proposition 2.65. From the latter proposition follows that we do not have for any combination of obligations and permissions that all propositions are normed, not even in the weakest logic $P_\exists$ and $O_\exists$. Finally, there is a second conclusion from Proposition 2.65. If we want to represent a dilemma between an obligation and a permission in a consistent way, then we have to use $O_\exists$ and $P_\exists$ operators, because $P_\exists(\alpha\,|\,\top) \wedge O_\exists(\neg\alpha\,|\,\top)$ is the only consistent obligation-permission dilemma.

## 2.5 Factual detachment in the two-phase approach

No dyadic deontic logic can be introduced without a discussion on factual detachment. Minimizing logics have been criticized [Che74, LB83, Alc93], because they do not have strengthening of the antecedent and factual detachment, see Section 1.3.5. In this chapter, we showed by the ordering obligations that strengthening of the antecedent can be accepted if we do not have weakening of the consequent. Moreover, we showed that strengthening of the antecedent can be combined with weakening of the consequent in a two-phase deontic logic. In this final section we investigate the second lack of the minimizing obligations, the lack of factual detachment.

Factual detachment is the inference pattern that derives monadic obligations from dyadic (conditional) obligations. We assume monadic modal obligations $O\alpha$ to represent the detached obligations. No further properties of the monadic operator are assumed. The simplest definition of factual detachment is the following rule FD, alias deontic modes ponens.

$$\text{FD}: \frac{O(\alpha|\beta), \beta}{O\alpha}$$

The following example illustrates that FD detaches counterintuitive obligations. It is the so-called pragmatic oddity of Prakken and Sergot [PS96] in a dyadic deontic logic.

**Example 2.66 (Pragmatic oddity)** Consider the following three sentences:

1. $O(k|\top)$: You should keep your promise.
2. $O(a|\neg k)$: If you have not kept your promise, you should apologize.
3. $\neg k$: You have not kept your promise.

The derivation represented in Figure 2.18 shows that $Ok \wedge Oa$ can be derived, from which the obligation $O(k \wedge a)$ can be derived. Prakken and Sergot remark ‘but it is a bit odd to say that in all ideal versions of this world you keep your promise and you apologize for not keeping it.’ □

$$\dfrac{\dfrac{O(k|\top)}{Ok}\,\text{FD} \qquad \dfrac{O(a|\neg k) \quad \neg k}{Oa}\,\text{FD}}{O(k \wedge a)}\,\text{AND}$$

Figure 2.18: Pragmatic oddity

The following example illustrates a similar problem. We say that defeasible detachment only holds as a defeasible rule, as is explained in Chapter 4 when we discuss the example in more detail.

**Example 2.67** Consider the two obligations $O(t \mid a)$ and $O(a \mid \top)$ of Example 2.10, where $a$ can be read as 'a certain man going to his neighbor's assistance' and $t$ as 'telling the neighbors that that he will come.' We can derive $O(a \wedge t \mid \top)$ and $O_{\exists}(t \mid \top)$ but not $O_{\exists}(t \mid \neg a)$, because the minimizing obligations do not have strengthening of the antecedent. We analyze this type of defeasibility in Chapter 4. The intuition is that if the man does not go to the assistance of his neighbors, then he does not have an obligation to tell that he will come. Factual detachment FD detaches the monadic obligation that the man should tell the neighbors that he will come $Ot$ from the obligation $O(t|\top)$, even when he does not go to their assistance (facts: $\neg a$). This derivation is represented in Figure 2.19. The derivation of $Ot$ when the facts are $\neg a$ is counterintuitive for the same reasons as the derivation of $O_{\exists}(t|\neg a)$ is counterintuitive. □

$$\dfrac{\dfrac{\dfrac{\dfrac{O(t|a) \quad O(a|\top)}{O(t \wedge a|\top)}\,\text{DD'}}{O_{\exists}(t \wedge a|\top)}\,\text{REL}}{O_{\exists}(t|\top)}\,\text{WC} \qquad \neg a}{O_{\exists}t}\,\text{FD}$$

Figure 2.19: Factual detachment

The following inference pattern *Exact Factual Detachment* EFD does not derive these counterintuitive obligations. Exact factual detachment can be represented by the inference pattern

$$\text{EFD}: \frac{O(\alpha|\beta), \mathcal{A}\beta}{O(\alpha)}$$

in which $O\alpha$ is a new, monadic modal oparator, and $\mathcal{A}$ is Levesque's All-I-Know (alias only knowing) operator $\mathcal{A}$ (see [Lev90]). $\mathcal{A}\alpha$ is true iff $\alpha$ is logically equivalent with all factual premises given. The inference pattern EFD is based on the intuition that the antecedent of a dyadic obligation restricts the focus to possible situations in which the antecedent is *assumed* to be factually true, and the consequent represents what is obligatory, given that *only* these facts are assumed. If the facts are equivalent to the antecedent, then the consequent can be considered as an absolute obligation.

A problem with EFD is that it does not derive violated obligations. In fact, we can consider the following *definition* of the monadic modal operator as an extension of the inference pattern

EFD: we have $O\alpha$ if *and only if* we have $O(\alpha \mid \beta)$ and $\mathcal{A}\beta$ (for some $\beta$). In that case, we have as a theorem $\neg(\alpha \wedge O\neg\alpha)$, as a result of the theorem $\neg O(\neg\alpha \mid \alpha)$. If EFD is accepted then the relation between facts and absolute obligations is identical to the relation between antecedent and consequent of the conditional obligations. To formalize a notion of factual detachment that derives violated obligations, we introduce the following so-called *retraction test* (R-test). The test says that if we consider whether $\alpha$ is obligatory, we have to consider possibilities in which $\alpha$ is true and possibilities in which $\alpha$ is false.

> R-test: $\alpha$ is obligatory ($O\alpha$ is an absolute obligation) iff $\alpha$ ought to be the case *on the assumption that* $\neg\alpha$ *and* $\alpha$ *are not the case*, i.e. on the assumption that $\alpha$ is contingent.

In order to evaluate the normative force of factual sentences, we require that we first (hypothetically) give up the belief in $\alpha$ and $\neg\alpha$ and then consider whether the optimal extensions of the beliefs entail $\alpha$. In other words, we contract the facts by $\alpha$ and $\neg\alpha$ and then evaluate the obligation $O(\alpha \mid \beta)$ with respect to the contracted facts. The R-test can be considered as a version of the Kantian principle for factual detachment. In this interpretation of 'ought implies can', *ought* refers to the absolute obligations and *can* means that neither $\neg\alpha$ nor $\alpha$ is factually the case. The R-test is formalized in the following *Retraction Factual Detachment* RFD, where '$-$' is a retraction operator satisfying the Gärdenfors postulates [AGM85, Gär88]. For simplicity we write retraction as $\alpha = \beta - \{\gamma_i\}$, where $\alpha$, $\beta$ and $\gamma_i$ are sentences of the propositional base language $\mathcal{L}$. $\alpha$ is the result of the retraction of the $\gamma_i$ from $\beta$, and therefore $\alpha$ does not derive any of these $\gamma_i$.

$$\text{RFD}: \frac{O(\alpha \mid \beta - \{\alpha, \neg\alpha\}), \mathcal{A}(\beta)}{O\alpha}$$

Notice that this formalization inherits problems of retraction, i.e. that it is not unique and computationally complex. The relation between EFD and RFD is given by the following proposition.

**Proposition 2.68** *Let $\mathcal{F}$ be the conjunction of the factual premises. If $\alpha$ and $\neg\alpha$ are not in $Cn[\mathcal{F}]$, where $Cn$ stands for consequence set, then $O\alpha$ is derived by* EFD *iff it is derived by* RFD.

**Proof** *From the Gärdenfors postulates follows that $Cn[\mathcal{F} - \{\alpha\}] = Cn[\mathcal{F}]$ when $F \wedge \neg\alpha$ is consistent.*

In this section we proposed two alternatives for unrestricted factual detachment. These alternatives were introduced, because unrestricted factual detachment is counterintuitive. First, we proposed the conservative exact factual detachment. A drawback of exact factual detachment is that it does not derive violated obligations. Second, we proposed an extension of exact factual detachment that derives violated obligations, so-called retraction factual detachment. With this analysis we discussed the last drawback of dyadic obligations with a contextual interpretation of the antecedent. We showed that the two-phase deontic logic 2DL has strengthening of the antecedent and that a kind of factual detachment can be accepted. As a consequence, it seems a good candidate to formalize deontic reasoning.

## 2.6 Related research

There has been a lot of research on preference-based logics and the deontic puzzles discussed in this chapter. We discriminate between three areas: deontic logic, preference logic and default logic.

### 2.6.1 Deontic logic: the Forrester paradox

Dyadic deontic logics like 2DL were developed to solve the Forrester paradox. In this section we discuss four other solutions that have been proposed in deontic logic literature to solve the paradox, based on temporal distinctions, a distinction between settled and non-settled facts, scope distinctions and rejection of weakening. The simplest solution of the Forrester problem is to reduce the expressivity of the deontic logic such that a-temporal obligations cannot be represented. For example, some deontic logics make a temporal distinction between antecedent and consequent (see Section 1.3.4). In such logics, the premise $O(k \wedge g \mid k)$ and the counterintuitive obligation $O(\neg(k \wedge g) \mid k)$ derived in Example 2.1 is meaningless, because both antecedent and consequent refer to the same time point. This distinction is made explicit in deontic logics (like [Mey88, Alc93]) that define two types of propositions, one for the antecedent and one for the consequent, following Castañeda's distinction between assertions and actions [Cas81]. Hence, they do not allow that a proposition occurs in one formula in the antecedent and in another formula in the consequent. The drawback of the temporal solution to the Forrester paradox is that the expressivity of the temporal solution is limited (see Section 1.3.4). For example, temporal deontic logics that make a distinction between antecedent and consequent cannot represent the set of premises of the Forrester paradox in Example 2.1.

The second solution of the Forrester paradox is to make a distinction between settled and non-settled facts. A fact can be settled to become true, without factually being true. Loewer and Belzer [LB86] solve the Forrester paradox in their temporal deontic logic 'Dyadic Deontic Detachment' (3D) [LB83]. In 3D a dyadic obligation $O(\alpha \mid \beta)$ is read as 'if it is settled that $\beta$ will the case, then $\alpha$ ought to be the case.' This reading is related to B. Hansson's [Han71] interpretation of circumstances, see also Spohn's comments on B.Hansson's logic [Spo75]. In this reading, the antecedent always refers to an earlier time point than the consequent. Moreover, there is an operator $S\alpha$ in 3D that represents that a proposition $\alpha$ is settled. This operator is related to Greenspan's operator $U\alpha$ for unalterable $\alpha$ [Gre75]. Loewer and Belzer [LB86] also discuss the relation between their solution and Castañeda's approach to the contrary-to-duty paradoxes [Cas81]. The drawback of the settled-unsettled solution in 3D is that it introduces rather complicated mechanisms (in the meta-theory).

The third solution of the Forrester paradox is based on scope distinctions. Scope distinctions have been proposed (see e.g. [Cas81]) to solve the Good Samaritan paradox, the predecessor of the Forrester paradox. Scope confusions seem to be absent from the Forrester paradox. However, Sinnot-Armstrong [SA85] argues that also Forrester's paradox rests on scope confusions. He invokes Davidson's account of the logical form of action statements [Dav67], according to which adverbial modifiers like gently in the consequent of $O(k \wedge g \mid k)$ are represented as predicates of action-events. Hence, the obligation is translated to 'there is an event $e$, which is a murdering event, and it, $e$, is gentle' – $\exists e(Me \wedge Ge)$. Because of the conjunction, we can distinguish between wide scope $O\exists e(Me \wedge Ge)$ and narrow scope $\exists e(Me \wedge OGe)$. The narrow

scope representation solves the paradox, because we cannot derive 'Smith ought to kill Jones' from 'the event $e$ ought to be gentle.' A drawback of this solution [LB86, Gob91] is that not every adverb of action is amenable to treatment as a predicate. For example, Goble [Gob91] gives the example 'Jones ought not to wear red to school' and 'if Jones wears red to school, then Jones ought to wear scarlet to school.' Goble observes that the relation between scarlet and red is not such that we can say scarlet is 'red *and* . . . ,' which might allow us to pull the term red away from the deontic operator in the manner of Sinnot-Armstrong and Castañeda, leaving the operator to apply only to whatever fills the blank. Scarlet is just a determinate shade of red; that is all we can say. Finally, as far as we know no solution based on scope distinctions has been proposed for the paradox of the knower, see Example 1.11. In dyadic deontic logic, this paradox can be formalized by the two obligations '$p$ should not be the case' $O(\neg p \mid \top)$, but 'if $p$ is the case then you ought to know it' $O(Kp|p)$.

The fourth solution of the Forrester paradox is based on rejection of the property weakening. In the remainder of this section on alternative solutions of the Forrester paradox, we discuss four monadic logics that do not have weakening. Goble [Gob91] argues that Forrester's paradox is caused by weakening, following a suggestion of Forrester [For84, p.196]. His monadic deontic logic does not have weakening and represents the paradox by $\{O\neg k, k \to O(k \wedge g), k\}$, which is consistent. Following Jackson [Jac85], Goble defines an obligation $O\alpha$ as a preference of the *closest* $\alpha$ over the *closest* $\neg\alpha$. Such a second ordering of closeness can represent the notion of best world 'from this or that perspective.' The idea can be covered that in certain contexts the way things are in some worlds can be ignored – perhaps they are too remote from the actual world, or outside any agent's control. Alternatively, we can interpret 'closeness' as 'the most normal' as used in logics of defeasible reasoning. As a consequence of this definition, $O\neg k \wedge Ok$ is a dilemma and inconsistent, whereas $O\neg k \wedge O(k \wedge g)$ is not a dilemma and consistent. This does not justice do the fact that only in *very* few cases we *seem* to have that $O\neg\alpha_1 \wedge O(\alpha_1 \wedge \alpha_2)$ is not a dilemma. This solution seems like overkill, see also the discussion in [LB86].

A third monadic logic without weakening, which can be considered to represent the Forrester paradox, is Sven Ove Hansson's so-called Preference-based Deontic Logic (PDL) [Han90b]. The basic idea of PDL is that prohibitions are defined by the property of negativity. This property states that what is worse than something wrong is itself wrong. Obligations are defined in terms of prohibitions in the usual way: $O\alpha =_{def} F\neg\alpha$. The properties of PDL are similar to the properties of the logics of Jackson and Goble discussed above. Hence, the logic has the same drawback. Compared to 2DL, PDL only has monadic obligations and the (ceteris paribus) preferences are not axiomatized, they are only in the semantics. Brown and Mantha [BM91] criticize PDL, because Hansson has to prove the existence of representation functions. Brown and Mantha argue that, because their obligations (see below) are defined on preferences expressed as a modality, they can prove axioms of the logic by proving the existence of models (i.e. derivability in the modal logic), see the discussion in [BM91].

The fourth monadic logic without weakening is proposed by Humberstone [Hum83], who observes that obligations can be defined by $O\alpha =_{def} \overleftarrow{\Box}\neg\alpha$, i.e. $O\alpha$ is true if $\alpha$ is true *only* in accessible worlds, and obviously these obligations lack weakening (although they have strengthening, see below).[15] Brown and Mantha [BM91] further investigate Humberstone's observation.

[15] Humberstone further remarks that permissions defined by $P\alpha =_{def} \overleftarrow{\Box}\neg\alpha$ are so-called free-choice permissions, characterized by the theorem $P(\alpha_1 \vee \alpha_2) \leftrightarrow P\alpha_1 \wedge P\alpha_2$.

They do not accept the axioms **T** and **4**, so the preference ordering does not have to be reflexive and transitive. Their logic is defined as follows.

**Definition 2.69 (Obligation and admissibility)** Obligation and admissibility are defined in Humberstone's logic of inaccessible worlds as follows.

$$\begin{array}{lll} O_{BM}\alpha & =_{def} & \overleftarrow{\Box}\neg\alpha \\ P_{BM}\alpha & =_{def} & \neg\Box\neg\alpha \\ O^c_{BM}\alpha & =_{def} & \overleftarrow{\Box}\neg\alpha \wedge \neg\Box\neg\alpha \end{array}$$ $\Box$

**Proposition 2.70** *(See [BM91]) The modal logic has the following theorems:*

| | |
|---|---|
| **AND** | $O_{BM}\alpha \wedge O_{BM}\beta \rightarrow O_{BM}(\alpha \wedge \beta)$ |
| **RAND** | $O^c_{BM}\alpha \wedge O^c_{BM}\beta \wedge P_{BM}(\alpha \wedge \beta) \rightarrow O^c_{BM}(\alpha \wedge \beta)$ |
| **OR** | $O_{BM}\alpha \wedge O_{BM}\beta \rightarrow O_{BM}(\alpha \vee \beta)$ |
| **C** | $O_{BM}\bot$ |
| **NC**$^c$ | $\neg O^c_{BM}\bot$ |
| **D** | $O^c_{BM}\alpha \rightarrow P_{BM}\alpha$ |
| **FCO** | $O_{BM}(\alpha \vee \beta) \rightarrow O_{BM}\alpha \wedge O_{BM}\beta$ |
| **RFCO** | $O^c_{BM}(\alpha \vee \beta) \wedge P_{BM}\alpha \wedge P_{BM}\beta \rightarrow O^c_{BM}\alpha \wedge O^c_{BM}\beta$ |

*The logic CT4O does not have the following theorems:*

| | |
|---|---|
| **W** | $O_{BM}\alpha \rightarrow O_{BM}(\alpha_1 \vee \alpha_2)$ |
| **W′** | $O_{BM}(\alpha_1 \wedge \alpha_2) \rightarrow O_{BM}\alpha_1 \wedge O_{BM}\alpha_2$ |

Brown and Mantha consider $O_{BM}\bot$ counterintuitive, because 'it flies in the face of ought implies can.' Furthermore, they notice that, as a result of this theorem, 'there are no worlds free of obligations.' As a solution, they propose the new definition of obligation $O^c_{BM}$. The logic has the theorem $\neg O^c_{BM}\bot$ and solves therefore the counterintuitive theorem. Unfortunately, it does not have **AND**, it only has the weaker **RAND**. They further notice that **AND** can be derived again when $P_{BM}\alpha \wedge P_{BM}\beta \rightarrow P_{BM}(\alpha \wedge \beta)$ is accepted. However, $O_{BM}\bot$ is not the only counterintuitive theorem of Brown and Mantha's logic. A more serious counterintuitive theorem of the first definition of obligation is the free-choice theorem **FCO**. As they remark, 'this is somewhat unreasonable, since the premise is weaker than the conclusion.' The second definition of obligation has the related counterintuitive theorem **RFCO**. Brown and Mantha consider the second definition of obligation to be 'quite satisfactorily,' but the validity of **RFCO** contradicts in our opinion this claim. Our ordering logic does not have the counterintuitive theorems **FCO** or **RFCO**.

### 2.6.2 Deontic logic: dyadic deontic logic

In this section we consider alternative dyadic deontic logics that combine strengthening of the antecedent and weakening of the consequent. To solve the Forrester paradox in a *dyadic* deontic logic, we have to block the following two derivations.

1. The derivation of $O(\neg k \,|\, k)$ from $O(\neg k \,|\, \top)$. This derivation is blocked when the dyadic obligations have a consistency check on the antecedent and consequent.

2. The derivation of $O(\neg(k \wedge g) \,|\, k)$ from $O(\neg k \,|\, \top)$. This derivation is blocked when the dyadic obligations do not have weakening of the consequent.

A dyadic deontic logic with a conditional interpretation, defined by $O(\alpha \,|\, \beta) =_{def} \beta > O\alpha$, can represent the Forrester paradox if strengthening of the antecedent is restricted and the monadic obligations do not have weakening of the consequent. As far as we know, such logics have not been proposed in deontic logic literature. Here we give a direction for a possible formalization, that is based on Proposition 2.22. This proposition shows for models $M$ without duplicate worlds that $M, w \models O(\alpha \,|\, \beta)$ iff for all $\beta'$ such that $M, w \models \overset{\leftrightarrow}{\Box}(\beta' \rightarrow \beta)$, we have $M, w \models O_{\exists}(\alpha \,|\, \beta')$. Analogously, we can define a phase-1 operator as a set of dyadic obligations with a conditional interpretation, i.e. $M, w \models O(\alpha \mid \beta)$ iff for all $\beta'$ such that $M, w \models \overset{\leftrightarrow}{\Box}(\beta' \rightarrow \beta)$, we have $M, w \models \beta > O\alpha$. This relaxation of our ordering logic is analogous to the relaxations of B. Hansson's minimizing logic discussed in [Lew74], see also [Mak93]. We leave the technical details of this idea (which type of monadic obligation, which type of implication '$>$', and most importantly the formalization of the quantification) to the reader. Such a logic would be similar to our ordering logic. Two minor distinctions between our ordering logic are represented by the following two theorems of the ordering logic.

1. $O(\alpha|\beta) \rightarrow O(\alpha \wedge \beta|\beta)$.

2. $(O(\alpha|\beta) \wedge O(\beta|\gamma)) \rightarrow O(\alpha \wedge \beta|\gamma)$. The ordering obligations have deontic detachment. We consider deontic detachment an intuitive inference, as illustrated in Example 2.10.

Prakken and Sergot [PS96] propose to formalize CTD obligations by $\beta > O_{\beta}\alpha$, where '$>$' is a conditional implication and $O_{\beta}\alpha$ is called a contextual obligation with context $\beta$. They define absolute obligations $O\alpha =_{def} O_{\top}\alpha$ and absolute permissions $P\alpha =_{def} \neg O \neg \alpha$. An important axiom is the so-called Down axiom, which derives obligations with a more specific context, i.e. a kind of strengthening of the context. The logic does have the no-dilemma assumption. Thus the logic should be compared to $O_D$ and $O_{\forall}$.

1. The most important distinction between Prakken and Sergot's logic and 2DL is represented by the so-called Up axiom: $P\beta \rightarrow (O_{\beta}\alpha \rightarrow O\alpha)$.[16] This axiom performs a kind of factual detachment. Everything that refers to a permitted context may be detached from a contextual obligation.

2. In 2DL all possible states are ordered in the semantics, whereas the semantic ordering in the logic of Prakken and Sergot only considers states with obligations of which the antecedent can be derived from the facts. This is shown by the following example.

   **Example 2.71** In the logic of Prakken and Sergot, the Chisholm paradox (see Example 1.16) is represented by the theory $\{Oa, a > O_a t, \neg a > O_{\neg a}\neg t, \neg a\}$. From this theory,

[16] The logic satisfies a more complicated version of the Up axiom, which performs a kind of weakening of the context. This sentence can be derived from it. However, this simpler formula suffices here for our comparison.

only the contextual obligations $Oa$ and $O_{\neg a}\neg t$ are derived. Hence, the contextual obligation $O_a t$ is not used to build the partial ordering in the models of this theory. Moreover, neither the absolute obligation $Ot$ nor $O\neg t$ is derived from this theory. □

3. The logic 2DL has deontic detachment **DD′**, but there is no deontic detachment in the logic of Prakken and Sergot. It will be difficult to implement this, because the contextual obligations are derived only when the antecedent is factually true (see 2). We consider deontic detachment an intuitive inference, as illustrated in Example 2.10.

4. The notion of absolute obligations in 2DL (defined with RFD) is quite different from the notion of absolute obligations in the logic of Prakken and Sergot. In 2DL, absolute obligations also represent deontic cues from a sub-ideal context, like the absolute obligation $\neg t$ in the Chisholm paradox. In the logic of Prakken and Sergot, only a contextual obligation $O_{\neg a}\neg t$ can be derived.

5. The following variant of the Forrester paradox causes problems in Prakken and Sergot's deontic logic, as discussed in [PS96], but not in the logic 2DL. To facilitate comparison with Example 2.1 we write $O(\alpha|\beta)$ instead of $O_\beta\alpha$.

   **Example 2.72 (Forrester paradox, continued)** Assume a dyadic deontic logic with substitution of logical equivalents and the inference patterns WC, AND, and the following version of *Restricted Strengthening of the Antecedent* RSA′.

$$\text{RSA}' : \frac{O(\alpha|\beta_1), \overset{\leftrightarrow}{\Diamond}(\alpha \wedge \beta_1 \wedge \beta_2), \overset{\leftrightarrow}{\Diamond}(\neg\alpha \wedge \neg(\beta_1 \wedge \beta_2))}{O(\alpha|\beta_1 \wedge \beta_2)}$$

   Consider the set of obligations $S = \{O(\neg k \mid \top), O(k \wedge g \mid k)\}$. The counterintuitive derivation of $O(\neg(k \wedge g)\,|\,k)$ from $O(\neg(k \wedge g)\,|\,\top)$ in Figure 2.1 is blocked, because the second condition $\overset{\leftrightarrow}{\Diamond}(\neg\alpha \wedge \neg(\beta_1 \wedge \beta_2)) = \overset{\leftrightarrow}{\Diamond}(\neg(k \wedge g) \wedge k)$ of RSA′ is false. Unfortunately, the counterintuitive $O(\neg(k \wedge g)\,|\,k)$ can be derived from an extension of $S$. Consider the set of obligations $S' = \{O(\neg k|\top), O(k \wedge g|k), O(p|\top)\}$, where $p$ can be read as 'buying pears.' The counterintuitive derivation of the obligation $O(\neg(k \wedge g)\,|\,k)$ is represented in Figure 2.20 below. □

$$\dfrac{\dfrac{\dfrac{\dfrac{O(\neg k|\top)}{O(\neg(k \wedge g)|\top)}\,\text{WC} \qquad O(p|\top)}{O(\neg(k \wedge g) \wedge p|\top)}\,\text{AND}}{O(\neg(k \wedge g) \wedge p|k)}\,\text{RSA/RSA}'}{O(\neg(k \wedge g)|k)}\,\text{WC}$$

Figure 2.20: Forrester paradox, continued

To solve the problems in Example 2.72, Prakken and Sergot [PS97] propose a system which is very similar to 2DL. Unfortunately, at the moment of writing we do not have a final version to compare the logics.

### 2.6.3 Deontic logic: another two-phase problem

In this chapter, we showed that two phases are necessary to combine strengthening of the antecedent and weakening of the consequent in a dyadic logic. Finally, we observe a related problem in deontic logic literature. The two-phase approach is also necessary for combining restricted conjunction RAND, also called consistent aggregation, and weakening.

$$\text{RAND}\,\frac{O^c\alpha_1, O^c\alpha_2, \overset{\leftrightarrow}{\Diamond}(\alpha_1 \wedge \alpha_2)}{O^c(\alpha_1 \wedge \alpha_2)}$$

Van Fraassen [vF73] discusses a problem, which we reconstruct with inference patterns in Figure 2.21.

$$\frac{Op \quad O\neg p}{O(p \wedge \neg p)}\ (\text{RAND}) \qquad \frac{\dfrac{O(f \vee m) \quad O\neg m}{O(f \wedge \neg m)}\ \text{RAND}}{Of}\ \text{W} \qquad \frac{\dfrac{\dfrac{Op}{O(f \vee p)}\ \text{W} \quad O\neg p}{O(f \wedge \neg p)}\ \text{RAND}}{Of}\ \text{W}$$

Figure 2.21: Consistent aggregation

His monadic deontic logic does not have the no-dilemma assumption, so he does not want to derive $O\bot$ from $Op$ and $O\neg p$. Unrestricted conjunction is too strong. However, he wants to derive from the two premises 'Honor thy father or thy mother!' $O(f \vee m)$ and 'Honor not thy mother!' $O\neg m$ the conclusion 'thou shalt honor thy father' $Of$. This derivation is also represented in Figure 2.21. Now van Fraassen asks himself whether restricted conjunction can be formalized.

> "But can this happy circumstance be reflected in the logic of the ought-statements alone? Or can it be expressed only in a language in which we can talk directly about the imperatives as well? This is an important question because it is the question whether the inferential structure of the 'ought' language game can be stated in so simple a manner that it can be grasped in and by itself. Intuitively, we want to say: there are simple cases, and in the simple cases the axiologist's logic is substantially correct even if it is not in general – but can we state precisely when we find ourselves in such a simple case? These are essentially technical questions for deontic logic, and I shall not persue them here. In conclusion, it seems to me that the problem of possibly irresolvable moral conflict reveals serious flaws in the philosophical and semantic foundations of 'orthodox' deontic logic, but also suggests a rich set of new problems and methods for such logic." [vF73]

The third derivation of Figure 2.21 illustrates that in a monadic deontic logic we cannot accept restricted conjunction and weakening. If we accept these inference patterns, then we can

derive $Of$ from a deontic dilemma $Op \wedge O\neg p$, which is obviously counterintuitive. We can combine restricted conjunction and weakening only if there are two phases, for similar reasons as two phases are necessary to combine strengthening of the antecedent and weakening of the consequent in a dyadic deontic logic. The first phase does not have weakening but it has restricted conjunction, and the second phase vice versa. Obviously, this blocks the counterintuitive third derivation in Figure 2.21. The distinction between phase-1 and phase-2 obligations is analogous to van Fraassen's distinction between 'imperatives' and 'ought-statements'. In Horty's reconstruction [Hor93] of van Fraassen's theory in Reiter's default logic [Rei80] the two phases are *not* explicit. In our terminology, the distinct operators $O_D$ and $O_\forall$ are represented by the same modal operator $O$. As a consequence, it is very difficult if not impossible to construct a semantics for this logic.

### 2.6.4 Preference logic

It has been suggested [Jen85, Jac85, Gob90b, Han90b] that a unary operator $O$ capable of bearing a deontic interpretation might be defined in a logic of preference by

$$O\alpha =_{def} \alpha \succ \neg\alpha$$

However, we cannot define the preference $\alpha \succ \neg\alpha$ by a preference of all instances of $\alpha$ over $\neg\alpha$, because two obligations 'be polite' $Op$ and 'be helpful' $Oh$ would conflict when considering 'being polite and unhelpful' $p \wedge \neg h$ and being impolite and helpful' $\neg p \wedge h$ [vW63]. Jackson [Jac85] and Goble [Gob90b] therefore propose to define the preference $\alpha \succ \neg\alpha$ by a preference of *the closest* $\alpha$ over *the closest* $\neg\alpha$. We already discussed these logics when we discussed related work concerning the Forrester paradox in the previous section. The obligation 'be polite' $Op$ prefers the closest $p$ worlds to the closest $\neg p$ worlds. Hence, the problem is solved by the derivation that 'polite and unhelpful' $p \wedge \neg h$ and 'impolite and helpful' $\neg p \wedge h$ are not among the closest $p$, $\neg p$, $h$ or $\neg h$ worlds. S.O.Hansson [Han90b] introduces complicated ceteris paribus preferences, i.e. all other things are considered to be equal (see also [DW91b, TP94]). The obligation 'be polite' $Op$ prefers 'polite and helpful' $p \wedge h$ to 'impolite and helpful' $\neg p \wedge h$, and 'polite and unhelpful' $p \wedge \neg h$ to 'impolite and unhelpful' $\neg p \wedge \neg h$, but it does not say anything about $p \wedge h$ and $\neg p \wedge \neg h$. We propose a third solution. The preference $\alpha \succ \neg\alpha$ is defined by $\neg\alpha$ is not as preferable as $\alpha$. The two obligations $Op$ and $Oh$ do not conflict when considering $p \wedge \neg h$ and $\neg p \wedge h$ when the underlying preference ordering is not strongly connected. This solution is simpler than the first two solutions, because we do not need the additional semantic baggage of the second ordering or the ceteris paribus preferences. Tan and Pearl [TP94, p.531] argue for a ceteris paribus reading of preferences because 'absolute preferences are not very useful,' given von Wright's problem. However, we show that this motivation is not very convincing, because our logic 2DL does not have this problem. On the other hand, the formalization of ceteris paribus circumstances is problematic, because 'similar circumstances' is difficult to formalize.

The various binary relations $\alpha_1 \succ \alpha_2$ defined in this chapter can be considered as preference relations, and compared with preference logics as proposed in the literature. The preference relation $\succ_s$ is quite weak, because it is not anti-symmetric (we cannot derive $\neg(\alpha_2 \succ_s \alpha_1)$ from $\alpha_1 \succ_s \alpha_2$) and not transitive (we cannot derive $\alpha_1 \succ_s \alpha_3$ from $\alpha_1 \succ_s \alpha_2$ and $\alpha_2 \succ_s \alpha_3$). A distinction is that the relations $\alpha_1 \succ \alpha_2$ of preference logics only compare $\alpha_1 \wedge \neg\alpha_2$ worlds with

$\neg\alpha_1 \wedge \alpha_2$ worlds. For our specific use of the preference relations, this restriction is superfluous because for the dyadic obligations we compare $\alpha \wedge \beta$ with $\neg\alpha \wedge \beta$. A second distinction is that in the classical preference logics (e.g. [vW63, Res67]) a value is associated with each world. Hence, the orderings are totally connected. In [Han89] it is shown that connected orderings are problematic. If we add the condition that $\succ_s$ is (totally) connected, then the relation is anti-symmetric and transitive.

An interesting perspective on preference-based logics is that they formalize the 'combining of preference relations.' That is, each premise represents a preference relation, and the notion of preferential entailment combines them. A dyadic ordering obligation $O(\alpha|\beta)$ is best considered as a preference relation with $w_1 \leq w_2$ iff $w_1 \in|\alpha \wedge \beta|$ and $w_2 \in|\neg\alpha \wedge \beta|$, or $w_1 \in|\neg\beta|$ or $w_2 \in|\neg\beta|$. Hence, the relation is not transitive. Much research in economic theory is based on the perspective of combining preference relations, most notably Arrow's social choice [Arr63], but the lack of transitivity also invalidates Arrow's famous theorem [Arr50]. In fact, our logic is better analyzed in Andreka *et al*'s framework [ARS95]. However, this framework considers *prioritized* mechanisms to combine preference relations. Priorities will be considered in Chapter 4 of this thesis, when we consider *defeasible* deontic logic.

Preference logics now gain popularity as logics for qualitative decision theory to formalize reasoning about context-sensitive goals, which we discuss in Chapter 5. Boutilier [Bou94b] proposes an extension of Hansson's minimizing logic with System Z to represent preferences. Moreover, an idea related to exact factual detachment EFD is proposed by Boutilier, because to determine preferences based on certain actual facts, he considers only the *most ideal* worlds satisfying those facts, rather than *all* worlds satisfying those facts. In Boutilier's logic, this means that the antecedent of his conditional is logically equivalent with the premises, i.e. he considers $\vdash O(\alpha|KB)$, where $KB$ is the set of premises. Another preference logic is introduced by Tan and Pearl [TP94, p.533] and has a what they call 'principle of maximal indifference' which is better called gravitating towards the center. A remarkable property of the logic is that it makes a specificity set like $\{O(p|\top), O(\neg p|q)\}$ inconsistent [TP94, p.537]. Thus it is also a solution for the cigarettes problem. Boutilier's logic as well as Tan and Pearl's logic do not deal satisfactorily with contrary-to-duty preferences. Consider Forrester like graded preferences related to driving speed. For example, 'you should not drive faster than 100 km an hour,' 'if you do drive faster than 100 km an hour, then you should not drive faster than 110 km an hour,' 'if you do drive faster than 110 km an hour, then you should not drive faster than 120 km an hour,' etc. To be more precise, our setting is the following.

**Example 2.73 (Speed limits)** Consider preferences about driving speed. The preferences are divided in two (unrelated) groups: preferences for when a person is in the United States ($u$), which are expressed in miles ($m$), and preferences for when she is in Europe ($\neg u$, not in the United States), which are expressed in kilometers ($k$). The preferences are represented in Figure 2.22. This figure should be read as follows. The circles are equivalence classes of worlds which satisfy the propositions in the circle, and which do not satisfy the propositions in strictly preferred circles. In the United States, there is a preference for less than 80 miles ($m80$) over more than 80 miles. In Europe, the ideal is less than 100 km an hour ($k100$), sub-ideal is less than 110 km an hour ($k110$), sub-sub-ideal is less than 120 km an hour ($k120$) and the worst behavior is going faster than 120 km an hour. Obviously, from the propositions $k100$, $k110$ and $k120$, each logically implies all latter ones. For example, we have $\overset{\leftrightarrow}{\Box}(k100 \rightarrow k110)$, $\overset{\leftrightarrow}{\Box}(k100 \rightarrow k120)$,

etc. We leave this implicit, otherwise the figures become very difficult to read. We assume all propositions are controllable, in the terminology of [Bou94b], thus the agent can control whether she goes to the United States ($u$) or to Europe ($\neg u$), and she can control her speed. Notice that we do not assume any preferences for either the United States or Europe.

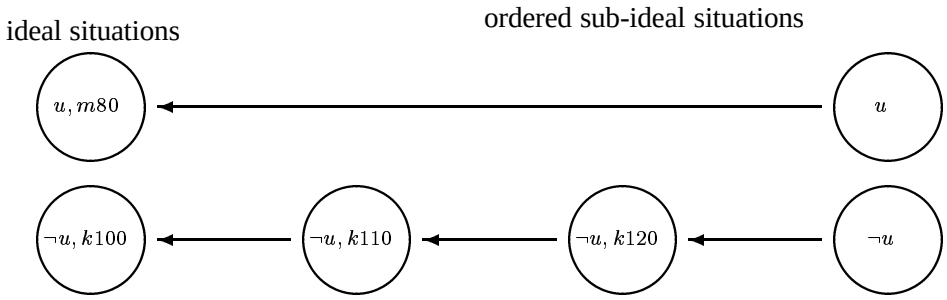



Figure 2.22: Speed limits

Now consider the situation in which we specify the United States preferences more precisely, as represented in Figure 2.23. Instead of a binary distinction between less than 80 miles (good) and more than 80 miles (bad) we distinguish six different categories. One expects that this revised specification does not influence the derivable preferences, except for the preferences that concern driving speeds less than 80 miles. For example, if the first specification says that it is desired to drive less than 120 km an hour, then the revised specification should derive the same. In particular, it is highly counterintuitive if the first specification implies a preference for the United States, and the second specification a preference for Europe (or vice versa). □

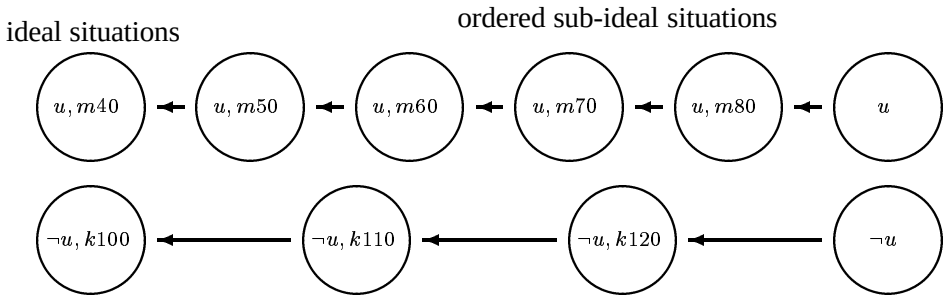



Figure 2.23: Revised speed limits

Boutilier [Bou94b] formalizes the preferences of his logic for qualitative decision theory QDT in his modal preference logic CO with the defeasible reasoning mechanism System Z, see Section 2.3. Tan and Pearl [TP94] criticize this approach of gravitating towards ideality, because 'while it is intuitive to assume that worlds gravitate towards normality because abnormality is a monopolar scale, it is not at all clear that worlds ought to be as preferred as possible since preference is a bipolar scale.' In a preference logic, there are good ideal and bad violation poles. The following example illustrates that gravitating towards the ideal derives counterintuitive consequences for our speed limits example.

**Example 2.74 (Speed limits, continued)** The preferences of Example 2.73 are formalized by the following two sets $S$ and $S'$.

$$\begin{array}{rcl} S & = & \{ \quad O_\forall(m80|u), \\ & & \quad O_\forall(k100|\neg u), O_\forall(k110|\neg u \wedge \neg k100), O_\forall(k120|\neg u \wedge \neg k110)\} \\ S' & = & \{ \quad O_\forall(m40|u), O_\forall(m50|u \wedge \neg m40), O_\forall(m60|u \wedge \neg m50), \\ & & \quad O_\forall(m70|u \wedge \neg m60), O_\forall(m80|u \wedge \neg m70), \\ & & \quad O_\forall(k100|\neg u), O_\forall(k110|\neg u \wedge \neg k100), O_\forall(k120|\neg u \wedge \neg k110)\} \end{array}$$

The unique System Z models of $S$ and $S'$ are represented in Figure 2.24. This figure should be read as follows. The circles are equivalence classes of worlds, that satisfy at least one of the rows of formulas written within them, and that do not satisfy one of the rows of formulas of a preferred circle. For example, in the System Z model of $S$, the $u \wedge m80$ and $\neg u \wedge k100$ worlds are equivalent, as well as the $u \wedge \neg m80$ and $\neg u \wedge k110 \wedge \neg k100$ worlds. Notice that the System Z model might be usable for minimization (the ideal worlds are $u \wedge m80$ and $\neg u \wedge k100$) but not for maximization (intuitively, the worst worlds are $u \wedge \neg m80$ and $\neg u \wedge \neg k120$, whereas only the latter are worst in the System Z model).

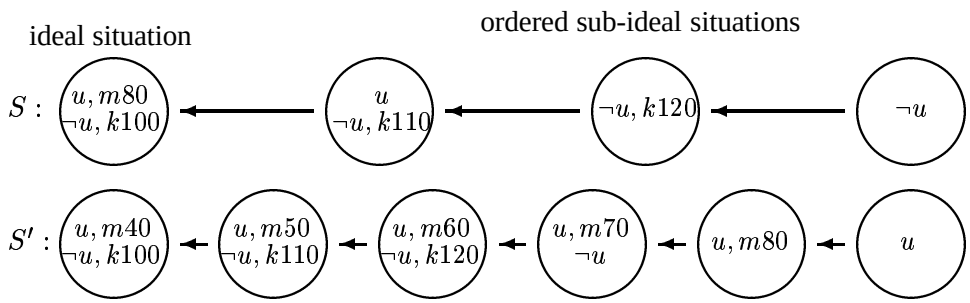



Figure 2.24: Gravitating towards the ideal (System Z)

Unfortunately, we have $S \models_\sqsubset O_\forall(u|(u \wedge \neg m80) \vee (\neg u \wedge \neg k110))$, because the preferred $(u \wedge \neg m80) \vee (\neg u \wedge \neg k110)$ worlds are only the $u \wedge \neg m80$ worlds (remember that $\neg k110$ logically implies $\neg k100$). This is counterintuitive, because intuitively the most ideal states for the United States are the $u \wedge \neg m80$ worlds, and the most preferred worlds for Europe are the $\neg u \wedge k120$ worlds, and these worlds are incomparable. In particular, it is counterintuitive to prefer the United States over Europe. However, gravitating towards ideality prefers the first, because it contains only one violation, whereas the latter contains two violations. Hence, System Z works as a violation counter. Moreover, consider the revised speed limits. We have $S' \models_\sqsubset O_\forall(\neg u|(u \wedge \neg m80) \vee (\neg u \wedge \neg k110))$. Hence, because we have further specified the speed limits in the United States, now we have the desire to go to Europe instead of going to the United States. Obviously, this is highly counterintuitive. □

Tan and Pearl [TP94] propose a gravitation mechanism in which there is no preference for either end of the bipolar preference scale. Instead of a preference for the ideal, there is a preference for the center, which they call gravitating towards indifference: 'In the case of preferences the principle we adopt is the principle of maximal indifference. We want to assume that there

is no preference between two worlds unless compelled to be so by preferences that are explicated by the reasoning agent. From the set of admissible preference rankings we want to select a distinguished ranking which best captures the essence of the principle of maximal indifference. This ranking, the $\pi^+$ ranking, will minimize the difference in the preference ranks.' To formalize this idea, we need rankings, which associate numerical values with worlds (thus the ordering is connected). Boutilier [Bou94b] defends his closure rule gravitating towards the ideal: 'Is the assumption that worlds are preferred unless stated otherwise reasonable? For instance, Tan and Pearl [TP94] argue that worlds should gravitate towards "indifference" rather than preference. We cannot, of course, make sense of such a suggestion in our framework, since we do not have a bipolar scale (where outcomes can be good, bad or neutral).[17] However, even if an "assumption of indifference" were technically feasible, we claim that the "assumption of preference" is the right one in our setting.'

Boutilier [Bou94b] defends this claim as follows: 'Recall that we wish to use preferences to determine the set of goal states for a given context $C$. These are simply the most preferred $C$-worlds according to our ranking; call this set $Pref(C)$. If the agent brings about *any* of these situations, it will have behaved correctly. A conditional preference $I(A \mid C)$ constrains the set $Pref(C)$ to contain only $A$-worlds. Thus an agent will attempt to bring about some $A \wedge C$-world when $C$ holds. But which $A \wedge C$-world is the right one? With no further information, System Z will set $Pref(C) = \mid A \wedge C \mid$; all $A \wedge C$-worlds will be assumed to be equally acceptable. This seems to be appropriate: with no further information, any course of action that makes $A$ true should be judged to be as good as any other. Any other assumption, such as gravitation of worlds toward indifference, must make the set $Pref(C)$ smaller than $|A \wedge C|$. For example, if we rule out worlds satisfying $\alpha$ from $Pref(C)$, then $Pref(C) = |A \wedge C \wedge \neg\alpha|$. This requires that an agent striving for $Pref(C)$ make $\neg\alpha$ true as well as $A$. This imposes unnecessary and unjustified restrictions on the agent's goals, or on the manner in which it decides to achieve them.' The following example illustrates Boutilier's problem for Tan and Pearl's compactness rule.

**Example 2.75 (Speed limits, continued)** We write $\models_{TP}$ for Tan and Pearl's preferential entailment based on the most compact models. Reconsider the set of preferences of Example 2.74. Tan and Pearl's unique most compact model is given in Figure 2.25. We have $S \models_{TP} O_\forall(\neg u \mid \top)$: there is a preference for Europe. Moreover, with revised speed limits we have $S' \models_{TP} O_\forall(u \mid \top)$: there is a preference for the United States. Again, this is highly counterintuitive. □

The following example illustrates that our two phase logic 2DL can deal with the contrary-to-duty preferences of the speed limits example.

**Example 2.76 (Speed limits, continued)** The driving speed preferences in Example 2.73 are formalized by the following two sets $S''$ and $S'''$.

[17] Note that in classical decision theory, such distinctions do not exist. An outcome cannot be good or bad, nor can an agent be indifferent toward an outcome, in isolation; it can only be judged *relative* to other outcomes. An agent can adopt an attitude towards a *proposition*, as we explain below.

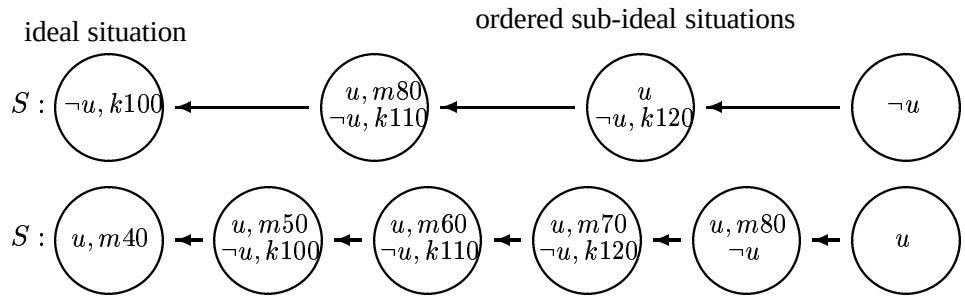



Figure 2.25: Gravitating towards the center (compactness)

$$
\begin{array}{rcl}
S'' & = & \{ \quad O_D(m80|u), \\
 & & \quad O_D(k100|\neg u), O_D(k110|\neg u \wedge \neg k100), O_D(k120|\neg u \wedge \neg k110)\} \\
S''' & = & \{ \quad O_D(m40|u), O_D(m50|u \wedge \neg m40), O_D(m60|u \wedge \neg m50), \\
 & & \quad O_D(m70|u \wedge \neg m60), O_D(m80|u \wedge \neg m70), \\
 & & \quad O_D(k100|\neg u), O_D(k110|\neg u \wedge \neg k100), O_D(k120|\neg u \wedge \neg k110)\}
\end{array}
$$

We do not derive $O_\forall(u \mid \beta)$, unless $\beta$ implies $u$. This property follows from the fact that $u$ worlds and $\neg u$ worlds do not have any constraint in common, and thus for every pair of $u$ and $\neg u$ worlds there is a c-preferred model in which they are equivalent. As a result, we do not have the counterintuitive derivations of Example 2.74 or 2.75.[18] □

The problem of the logics of Boutilier and Tan and Pearl is the combination of a unique model and a totally connected ordering. Consider the unique preferred models in Figure 2.24 and 2.25. Some $u$ and $\neg u$ worlds are forced to be equivalent, whereas intuitively they are incomparable. As a consequence, the contrary-to-duty paradoxes arise.

### 2.6.5 Default logic

Logics of defeasible reasoning formalize reasoning about default assumptions, i.e. what normally is the case. Obviously, the distinction between preference-based default logics and deontic logics is that the former introduces preferences to deal with exceptions, whereas the latter uses preferences to deal with violations. However, the way these preference-based logics deal with respectively exceptions and violations is quite similar. In particular, we were inspired by the logics of defeasible reasoning when we developed the new deontic logics. There are a lot of preference-based default logics, whereas there are only a few preference-based deontic logics. In this section we compare the preference-based default logics and the preference-based deontic logics developed in this chapter.

The first distinction between logics of defeasible reasoning and deontic logic is the fact that an obligation is not cancelled when it is violated, it is only no longer in force as a cue for action. We discuss this distinction in more detail in Chapter 4.

[18] We do not even have $S'' \models_c O_\forall(u|\neg(u \wedge m80))$, although $u$ worlds are the only ideal $\neg(u \wedge m80)$ worlds. These inferences can be added to 2DL by adding the constraint that the equivalence class of ideal worlds is as large as possible.

The second distinction between logics of defeasible reasoning and deontic logic is represented by the cigarettes example. It is an alphabetic variant of the famous Tweety example: birds normally fly, but penguins normally do not fly. The Tweety example should be consistent in a logic of defeasible reasoning, whereas the cigarettes example should be inconsistent in a deontic logic.

The first similarity between deontic logic and default logic is that both can be considered as faces of minimality [Mak93]. In particular, the minimizing approach is commonly taken in preferential semantics for non-monotonic logics, see for example [Sho88, KLM90, Mak93, Bou94a].

The second similarity between preference-based deontic logic and preference-based default logic is that both have an ordering and a minimizing phase. The logics based on circumscription [McC80] have an ordering and minimizing phase. The distinction between ordering and minimizing is especially clear in Veltman's normally-presumably logic based on update semantics [Vel96]. Veltman's logic has like 2DL the distinction between dynamic and static processes, which in contrast to 2DL is represented in the semantics. 'Normally $\alpha$' is a first-phase default and 'presumably $\alpha$' is a second-phase default. The normally defaults do not have weakening of the consequent and they do not have reasoning by cases. Thus, they are comparable to the ordering obligations defined in this chapter. However, update semantics is sensitive for the sequence in which the premises are presented. In our two-phase deontic logic 2DL, we have as premises a set of ordering obligations and as conclusion a minimizing obligation. Hence, the presentation of premises is not important. We therefore do not incorporate the rather complex update semantics.

The distinction between ordering and minimizing in preference-based default reasoning corresponds to the two-phase approach in default logic. Two phases are necessary to solve the inheritance problem, the derivation of 'penguins have wings' from 'birds have wings' given that penguins are exceptional birds. This is illustrated in Figure 2.26. Inheritance is based on the principle of independence (see e.g. [Vel96]): the property of having wings is independent from flying (which makes penguins exceptional). We want to derive 'penguins have wings' $(b \wedge p) > w$ from birds have wings $b > w$, but we do not want to derive 'penguins fly' $(b \wedge p) > f$ from 'birds fly' $b > f$. The latter derivation is blocked, because we have 'penguins do not fly' $(b \wedge p) > \neg f$. Hence, strengthening of the antecedent is blocked by a default with a contradictory consequent. However, we also want to block strengthening of the antecedent for obligations derived from 'birds fly' like 'birds fly or $q$' $b > (f \vee q)$. The third derivation in Figure 2.26 is an example of a counterintuitive derivation. Such inferences are blocked by not allowing strengthening of the antecedent after weakening of the consequent has been used. If 'birds fly or $q$' $b > (f \vee q)$ would have been a premise, then strengthening of the antecedent would have been allowed. Hence, there are two phases. See [Vel96] for an example why phase-1 defaults (called normally defaults) do not have weakening of the consequent.

$$\frac{b > w}{(b \wedge p) > w}\ \text{RSA} \qquad \frac{b > f}{(b \wedge p) > f}\ (\text{RSA}) \qquad \frac{\dfrac{b > f}{b > (f \vee q)}\ \text{WC}}{(b \wedge p) > (f \vee q)}\ \text{RSA}$$

Figure 2.26: Inheritance

In the proof theory of logics of defeasible reasoning, the two phases correspond to the 'appli-

cability of rules' and 'consequences of rules.' In general, conditional expressions can be represented (and analyzed) in the object language of a logic like the expression $\beta > \alpha$ of conditional logic, or alternatively they can be represented in the meta-theory like $\beta \vdash \alpha$ (see e.g. [KLM90]). Makinson [Mak93] compares different types of reasoning and calls it 'part of the folklore' which of these possibilities is used. The first is more expressive, because nested operators cannot be expressed at the meta-level. We illustrated the two-phase approach by two conditional expressions in the object language, for example the dyadic obligations $O^c$ and $O^c_\exists$. An alternative formalization of the two-phase approach is to model the first phase in dyadic obligations and the second phase in the meta-theory. The following definition shows this idea.

**Definition 2.77 ($\vdash_{fd}$)** Assume a deontic logic that contains at least dyadic ordering obligations $O$, monadic obligations and factual sentences. Consider the set $S = \{O(\alpha_i \,|\, \beta_i) \mid i = 1 \ldots n\}$ and the factual sentence $\beta$. The inference relation $\vdash_{fd}$ is defined as follows.

$$S \cup \{\beta\} \vdash_{fd} O\alpha \text{ iff } \{O^c(\alpha_i|\beta_i) \mid i = 1 \ldots n\} \models O^c_\exists(\alpha|\beta)$$

□

The inference relation $\vdash_{fd}$ does not validate a satisfactory notion of factual detachment, because if $S \cup \beta \vdash_{fd} \alpha$ then we cannot derive $O\neg\alpha$. Hence, violated obligations $\alpha \wedge O\neg\alpha$ cannot be derived by $\vdash_{fd}$, for similar reasons as EFD does not derive violated obligations. This can easily be adapted when we consider retraction-based mechanism like RFD. The inference relation $\vdash_{fd}$ gives a proof-theoretic interpretation of the two phases as follows. The first (ordering) phase considers whether a dyadic obligation is *applicable*. The second (minimizing) phase considers the consequences of the applicable obligations. The inference relation $\vdash_{fd}$ is comparable to Reiter's default logic. In particular, it is similar to Horty's reconstruction of van Fraassen's logic [vF73] in Reiter's default logic.

## 2.7 Conclusions

In this chapter we studied the relation between obligations and preferences. The main puzzle of this chapter is a formalization of the Forrester paradox that has the following properties.

1. The language is dyadic deontic logic;
2. The time index of the consequent is not necessarily later than the time index of the antecedent;
3. The obligations have strengthening of the antecedent;
4. The obligations have weakening of the consequent.

The analysis of the Forrester paradox reveals that any solution of this puzzle has to block the following two derivations.

1. The derivation of $O(\neg k \,|\, k)$ from $O(\neg k \,|\, \top)$. This derivation is blocked when the dyadic obligations have a consistency check on the antecedent and consequent. As a consequence, the obligations only have *restricted* strengthening of the antecedent.

2. The derivation of $O(\neg(k \wedge g) \,|\, k)$ from $O(\neg k \,|\, \top)$. This derivation is blocked when RSA cannot be used after WC has been used.

We had to introduce the new two-phase deontic logic 2DL, because other deontic logics do not satisfy these conditions. Any deontic logic has to be able to be extended with the no-dilemma assumption, permissions and factual detachment. Thus, the introduction of the new deontic logic 2DL introduced several new problems. Summarizing, we have established the following results in this chapter.

1. We have proposed a *new ordering logic*, the logic $O$, with two extensions (the logics $O^c$ and $O^{cc}$). The ordering logic is defined in the same preference logic as existing minimizing logic of B. Hansson $O_\forall$. Our deontic logic $O$ gives a contextual interpretation for the antecedent, and like the Hansson-Lewis logics it does not have factual detachment. However, it also differs in several respects from these logics. It does not have weakening of the consequent, but it has strengthening of the antecedent. In the latter respect, it resembles the Chellas-type logics with a conditional interpretation of the antecedent.

2. We have introduced the *new two-phase approach* to deontic reasoning. We have shown that only this approach combines strengthening of the antecedent and weakening of the consequent and thus solves the Forrester paradox.

3. We have given a *new solution for the cigarettes problem*, a problem related to dilemmas. This solution uses the new notion of preferential entailment based on 'maximally connected' models.

4. We have proposed *new preference-based permissions*. The new operators validate the standard relation between obligations and permissions (explained by the semantics in a non-standard way).

5. We have proposed the *new relation between dyadic and monadic obligations*, Retraction Factual Detachment RFD.

The preference-based deontic logic 2DL consists of twenty-four different dyadic obligation and permission operators, which are classified in Figure 2.27. The relations between the operators are represented by arrows, which represent strict entailment. We can make a distinction between three cubes, which represent the number of consistency checks.

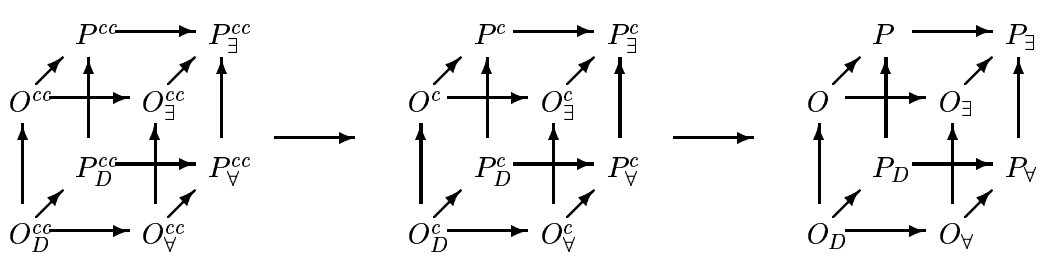



Figure 2.27: 2DL cubes

The logic 2DL is used to analyze the relation between obligations and preferences. The main conclusions are the following.

1. **Bipolarity and deontic choice.** The semantics of 2DL formalizes the semantic notion of deontic choice. A crucial property of deontic choice is that it is bipolar. The notion of deontic choice for an obligation $O(\alpha|\beta)$ compares two options $\alpha \wedge \beta$ and $\neg\alpha \wedge \beta$. Many preference-based semantics are monopolar, because they interpret the dyadic obligation $O(\alpha|\beta)$ as $\alpha \wedge \beta$ is the ideal or optimal of all $\beta$. In Chapter 4 we discuss obligations that can be overridden by other obligations, and we show that the bipolar concept is crucial.

2. **Distinction between ordering and minimizing.** The semantic distinction between ordering and minimizing corresponds to the properties strengthening of the antecedent and weakening of the consequent of the dyadic obligations.

3. **Deontic reasoning as a two-phase process.** The two-phase process consists semantically of an ordering and a minimizing phase. Two phases are necessary to combine:

   (a) strengthening of the antecedent and weakening of the consequent,

   (b) strengthening of the antecedent and reasoning by cases, and

   (c) restricted conjunction and weakening.

   Semantically, the two phases are related, because an ordering obligation $O$ corresponds to a set of existential-minimizing obligations $O_\exists$, see Proposition 2.22.

4. **Another perspective on deontic detachment.** In the preference semantics, the inference pattern deontic detachment is intuitive (for ordering obligations), as illustrated in Example 2.10. Moreover, the semantics explains that the inference patterns factual detachment, reasoning by cases and weakening of the consequent are not valid.

5. **Relation between obligations and permissions.** If obligations are *strict* preferences defined by $O(\alpha\,|\,\beta) =_{def} (\alpha \wedge \beta) \succ (\neg\alpha \wedge \beta)$, then permissions are preferences defined by $P(\alpha\,|\,\beta) =_{def} (\alpha \wedge \beta) \succeq (\neg\alpha \wedge \beta)$. From the definition follows directly the first deontic axiom $O(\alpha|\beta) \rightarrow P(\alpha|\beta)$.

In the following chapter we further investigate the relation between obligations and preferences by introducing a more expressive preference-based deontic logic. This new deontic logic questions whether deontic reasoning is a two-phase process.

The analysis of the relation between obligations and preferences in this chapter shows several relations between obligations and defeasibility. We introduced preference-based deontic logics which are structurally quite similar to preference-based default logics. Moreover, we introduced a new notion of preferential entailment, 'maximally connected', which makes the deontic logic non-monotonic. The preferences and the notion of preferential entailment are necessary to (1) formalize contrary-to-duty reasoning satisfactorily and (2) make the right set of dilemmas inconsistent. However, the introduction of preferences and preferential entailment is quite remarkable, because obligations are in some respects not defeasible. The violation of an obligation does not cancel the obligation in the same sense as an exception cancels a default assumption. The result of a violation is that the obligation is no longer a cue for action. The defeasibility involved in violations is in this sense weaker than the defeasibility involved in exceptions. We investigate the different types of defeasibility in Chapter 4. There we also show that the distinction is crucial in logics that formalize obligations that can be overridden by other obligations.

# Chapter 3

# Contextual deontic logic

In this chapter we further study the relation between obligations and preferences. Moreover, we introduce labeled deontic logic LDL and contextual deontic logic CDL, two extensions of dyadic deontic logic. The labeled and contextual obligations combine the properties strengthening of the antecedent and weakening of the consequent of the dyadic ordering and minimizing obligations of the two-phase deontic logic 2DL developed in Chapter 2. Labeled deontic logic is based on the distinction between implicit and explicit obligations, a distinction analogous to the distinction between implicit and explicit knowledge. Contextual deontic logic explicitly represents exceptions of the context of obligations.

This chapter is a modified and extended version of [vdTT97a].

## 3.1 The apples-and-pears example

In this section we discuss the apples-and-pears example, a new deontic example introduced in this thesis. This example neatly illustrates the distinction between the conditional and contextual interpretations of dyadic obligations, see Section 1.3.5. We give a proof-theoretic analysis of the apples-and-pears example in a 'classical' dyadic deontic logic based on the conditional interpretation, and in the two-phase deontic logic 2DL based on the contextual interpretation.

**Example 3.1 (Apples-and-pears)** Assume a dyadic deontic logic that has at least substitution of logical equivalents and the following inference patterns *Strengthening of the Antecedent* (SA), *Weakening of the Consequent* (WC) and *Conjunction* (AND).

$$\text{SA}: \frac{O(\alpha|\beta_1)}{O(\alpha|\beta_1 \wedge \beta_2)} \qquad \text{WC}: \frac{O(\alpha_1|\beta)}{O(\alpha_1 \vee \alpha_2|\beta)} \qquad \text{AND}: \frac{O(\alpha_1|\beta), O(\alpha_2|\beta)}{O(\alpha_1 \wedge \alpha_2|\beta)}$$

Furthermore, assume the set of dyadic obligations $S = \{O(a \vee p \mid \top), O(\neg a \mid \top)\}$ as premise set, where $a$ can be read as 'buying apples,' $p$ as 'buying pears', and $\top$ stands for any tautology. The intuitive obligation $O(\neg a \wedge p \mid \top)$ can be derived from $S$ by AND, and from this derived obligation, the obligation $O(\neg a \wedge p \mid a)$ can be derived by SA. From the latter obligation, the obligation $O(p \mid a)$ can be derived by WC. The derivability or underivability of $O(p \mid a)$ from $\{O(a \vee p|\top), O(\neg a|\top)\}$ illustrates the distinction between the conditional and contextual interpretation of the antecedent of dyadic obligations. If the antecedent $\top$ of $O(p|\top)$ is interpreted as a conditional, then it means that buying pears is always obligatory, also if $a$ happens to be the

case. On the other hand, if the antecedent of the dyadic obligation is interpreted as a context, then the obligation $O(p|a)$ should not be derivable. In the context $a$, the first premise $O(a \vee p|\top)$ is fulfilled and the second premise $O(\neg a|\top)$ is violated. Since the first premise is already fulfilled, there is intuitively no reason why $p$ should be obliged given the fact that $a$. Hence, within this context of buying apples, there is no reason to buy pears.

This derivation of $O(p\,|\,a)$ can be blocked by replacing SA by the following version of *Restricted Strengthening of the Antecedent* RSA, in which $\overset{\leftrightarrow}{\Diamond}$ is a modal operator and $\overset{\leftrightarrow}{\Diamond}\alpha$ is true for all consistent propositional formulas $\alpha$.

$$\text{RSA}: \frac{O(\alpha|\beta_1), \overset{\leftrightarrow}{\Diamond}(\alpha \wedge \beta_1 \wedge \beta_2)}{O(\alpha|\beta_1 \wedge \beta_2)}$$

The derivation of the obligation $O(p \mid a)$ discussed above is blocked, because the obligation $O(\neg a \wedge p \mid a)$ cannot be derived from the obligation $O(\neg a \wedge p \mid \top)$ by RSA. However, the obligation $O(p\,|\,a)$ can still be derived in another way. From the obligation $O(\neg a \wedge p\,|\,\top)$ the obligation $O(p|\top)$ can be derived by WC. From this latter obligation, the obligation $O(p|a)$ can be derived by RSA. The derivations are represented in Figure 3.1 below. □

$$\dfrac{\dfrac{\dfrac{O(a \vee p|\top) \quad O(\neg a|\top)}{O(\neg a \wedge p|\top)}\,\text{AND}}{O(\neg a \wedge p|a)}\,\text{SA}}{O(p|a)}\,\text{WC} \qquad \dfrac{\dfrac{\dfrac{O(a \vee p|\top) \quad O(\neg a|\top)}{O(\neg a \wedge p|\top)}\,\text{AND}}{O(p|\top)}\,\text{WC}}{O(p|a)}\,\text{SA/RSA}$$

Figure 3.1: The apples-and-pears example

Notice that the obligation $O(p \mid a)$ in Example 3.1 is a contrary-to-duty obligation derived from its primary obligation $O(\neg a|\top)$. The following example shows that the obligation $O(p|a)$ of the apples-and-pears example of Example 3.1 is not derived in the two-phase approach of 2DL.

**Example 3.2 (Apples-and-pears, continued)** Consider the set of dyadic obligations

$$S = \{O^c(a \vee p|\top), O^c(\neg a|\top)\}$$

where $\neg a$ does not entail the negation of $p$. We have $S \models \overset{\leftrightarrow}{\Diamond}(\neg a \wedge p)$, $S \models O^c(\neg a \wedge p\,|\top)$ and $S \models O^c_\exists(\neg a \wedge p|\top)$, $S \not\models O^c(p|\top)$ and $S \models O^c_\exists(p|\top)$. The crucial observation is that $O^c_\exists(p|a)$ is not entailed by $S$. Consider a typical countermodel $M$ of $O^c_\exists(p \mid \top)$ in Figure 3.2 below. We have $M \models O^c(a \vee p\,|\top)$ and $M \models O^c(\neg a\,|\top)$ because $|\neg a \wedge \neg p|\not\leq|a \vee p|$ and $|a|\not\leq|\neg a|$. Moreover, we have $M \not\models O^c_\exists(p|a)$, because all $a$ worlds are equivalent: for all $w_1, w_2 \in |a|$ we have $w_1 \leq w_2$.

For the proof-theoretic analysis of the underivability of $O^c_\exists(p|a)$, see the derivations in Figure 3.3. First of all, $O^c_\exists(p \mid a)$ is not entailed by $S$ via a derivation of $O^c(\neg a \wedge p \mid a)$, because $O^c(\neg a \wedge p|a)$ is not entailed by $O^c(\neg a \wedge p|\top)$ due to the restriction in **RSA**. Secondly, $O^c_\exists(p|a)$

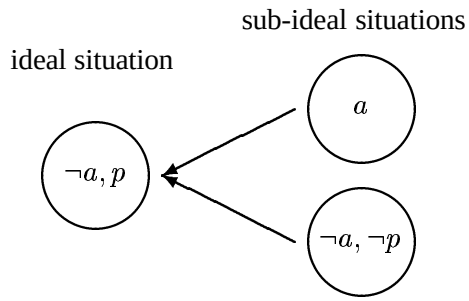



Figure 3.2: The apples-and-pears example in 2DL (semantics)

is not entailed by $S$ via $O^c(p|\top)$, because $O^c(p|\top)$ is not entailed by $S$. Finally, $O^c_\exists(p|a)$ is not entailed by $S$ via $O^c_\exists(p|\top)$ either, because $O^c_\exists$ does not have strengthening of the antecedent at all. □

$$\dfrac{\dfrac{\dfrac{\dfrac{O^c(a \vee p|\top) \quad O^c(\neg a|\top)}{O^c(\neg a \wedge p|\top)}\,\text{RAND}}{O^c(\neg a \wedge p|a)}\,(\text{RSA})}{O^c_\exists(\neg a \wedge p|a)}\,\text{REL}}{O^c_\exists(p|a)}\,\text{WC} \qquad \dfrac{\dfrac{\dfrac{\dfrac{O^c(a \vee p|\top) \quad O^c(\neg a|\top)}{O^c(\neg a \wedge p|\top)}\,\text{RAND}}{O^c_\exists(\neg a \wedge p|\top)}\,\text{REL}}{O^c_\exists(p|\top)}\,\text{WC}}{O^c_\exists(p|a)}\,(\text{RSA})$$

Figure 3.3: The apples-and-pears example in 2DL (proof theory)

When we compare the derivations of the apples-and-pears example with the counterintuitive derivations of the Forrester paradox in a dyadic deontic logic, see Example 2.1, we find that both are instances of the derivation of the obligation $O(\alpha_1|\neg\alpha_2)$ from the obligation $O(\alpha_1 \wedge \alpha_2|\top)$. Despite this similarity between our apples-and-pears example and the classic benchmark problem of deontic logic, we think that there are good reasons to introduce the new apples-and-pears example. Most importantly, the contrary-to-duty reasoning in the Forrester paradox is already visible in the premises, whereas the contrary-to-duty reasoning in the apples-and-pears example only manifests itself in the derivations. Moreover, the derived obligation $O(\neg(g \wedge k)|k)$ in the Forrester paradox does not have any intuitive reading, whereas the derived obligation in the apples-and-pears example still has a possible reading. Finally, the apples-and-pears example is simpler than the Forrester paradox, because (proof-theoretically) it only contains premises with tautological antecedents and (semantically) the preference orderings of the apples-and-pears example only has a binary distinction between ideal and sub-ideal, see Figure 3.2, whereas the preference orderings of the Forrester paradox have varying sub-ideal worlds, see Figure 2.7.

From a semantic point of view, the two-phase approach simply means that first an ordering has to be constructed, before it can be used for minimization. However, proof-theoretically such sequencing of derivations is rather complicated. This can be shown by analyzing the properties of the inference relation, when we consider $O^c(\alpha \mid \beta)$ as a premise and $O^c_\exists(\alpha \mid \beta)$ as a conclusion (as in Figure 3.3). For example, the inference relation does not support lemma handling,

because an intermediate conclusion cannot be used as a lemma for another proof. Obviously, this property follows directly from the two-phase approach, because a conclusion $O^c_{\exists}(\alpha \mid \beta)$ is simply something different than a premise $O^c(\alpha|\beta)$. According to the Kraus-Lehmann-Magidor classification [KLM90], the inference relation is not cumulative.

In this chapter, we discuss two one-phase logics that give a contextual interpretation to the antecedent. We illustrate the two new logics by showing that the obligation to buy pears when buying apples is not derivable in the apples-and-pears example. The first logic (Section 3.2) is based on a label of an obligation. The label is used to discriminate between implicit and explicit obligations, a distinction analogous to the distinction between implicit and explicit knowledge. The second logic (Section 3.3) extends the notion of a context of an obligation. Contextual deontic logic combines the intuitive preferential semantics of 2DL with a one-phase proof theory. This is accomplished by making the obligations more complex: we use a ternary instead of the dyadic representation in Example 3.1. It might seem that derivability of $O(p \mid \top)$ and underivability of $O(p|a)$ are in conflict with each other. However, in contextual deontic logic it simply means that pears should be bought *unless apples are bought*. The unless clause is formalized with contextual obligations $O(\alpha|\beta\backslash\gamma)$, to be read as '$\alpha$ should be the case if $\beta$ is the case unless $\gamma$ is the case'. We say that the obligation to buy pears is only valid in the context in which no apples are bought.

## 3.2 Labeled obligations

In this section we introduce labeled deontic logic LDL. Labeled obligations $O(\alpha|\beta)_L$ can *roughly* be read as '$\alpha$ ought to be the case if $\beta$ is the case, against the background of $L$.' They are based on the distinction of what we call implicit and explicit obligations.

To illustrate the distinction between implicit and explicit obligations, we recall the well-known distinction between implicit and explicit knowledge. The latter distinction originates in the logical omniscience problem: in principle, an agent cannot know all logical consequences of his knowledge. The benchmark example is that knowledge of the laws of mathematics does not imply knowledge of the theorem of Fermat. That is, an agent does not explicitly know the theorem of Fermat, she only implicitly knows it. Analogously, explicit obligations are not deductively closed, in contrast to implicit obligations. The two-phase deontic logic 2DL can be understood as follows: the ordering obligations are explicit obligations and the minimizing obligations are implicit obligations. The idea behind labeled obligations is to represent the explicit obligation, of which the implicit obligation is derived, in the label. The label is the reason for the obligation. This explains our reading of the label obligation $O(\alpha \mid \beta)_L$: '$\alpha$ ought to be the case if $\beta$ is the case, against the background of $L$.' We can use labeled deontic logic to solve the contrary-to-duty paradoxes, if we use the label to check that a derived obligation is not a contrary-to-duty obligation of its premises. An obligation $O(\alpha \mid \beta)$ is a *contrary-to-duty* obligation of the *primary* obligation $O(\alpha_1 \mid \beta_1)$ if and only if $\beta \wedge \alpha_1$ is inconsistent, as represented in Figure 3.4. The label of an obligation represents the consequents of the premises from which the obligation is derived. In labeled deontic logic we use a consistency check of the label of the obligation with its antecedent. If the label and the antecedent are consistent, then the derived obligation is not a contrary-to-duty of its premises.

In this section we introduce a deontic version of a labeled deductive system as it was in-

$$\begin{tikzcd} O(\alpha_1|\beta_1) \arrow[d, leftrightarrow, "\text{inconsistent}"'] \\ O(\alpha|\beta) \end{tikzcd}$$

Figure 3.4: $O(\alpha|\beta)$ is a contrary-to-duty obligation of $O(\alpha_1|\beta_1)$

troduced by Gabbay in [Gab91]. The language of dyadic deontic logic is enriched by allowing labels in the dyadic obligations. Roughly speaking, the label $L$ is a record of the consequents of all the premises that are used in the derivation of $O(\alpha|\beta)$.

**Definition 3.3 (Language of** LDL**)** The language of labeled deontic logic is a propositional base logic $\mathcal{L}$ and labeled dyadic conditional obligations $O(\alpha|\beta)_L$, with $\alpha$ and $\beta$ sentences of $\mathcal{L}$, and $L$ a set of sentences of $\mathcal{L}$. □

Each formula occurring as a premise has its own consequent in its label. The intuition is that the premises are explicit obligations.

**Definition 3.4 (Premises of** LDL**)** A formula which has its own consequent as its label is called a premise. □

We assume that the antecedent and the label of an obligation are always consistent. The label of an obligation derived by an inference rule is the union of the labels of the premises used in this inference rule. Below are some labeled versions of inference schemes. We write $\overset{\leftrightarrow}{\Diamond} L$ for a consistency check of a *set* of formulas.

$$\text{RSA}_V : \frac{O(\alpha \mid \beta_1)_L, \overset{\leftrightarrow}{\Diamond}(L \cup \{\beta_1 \wedge \beta_2\})}{O(\alpha \mid \beta_1 \wedge \beta_2)_L}$$

$$\text{WC}_V : \frac{O(\alpha_1 \mid \beta)_L}{O(\alpha_1 \vee \alpha_2 \mid \beta)_L}$$

$$\text{DD}'_V : \frac{O(\alpha|\beta)_{L_1}, O(\beta|\gamma)_{L_2}, \overset{\leftrightarrow}{\Diamond}(L_1 \cup L_2 \cup \{\gamma\})}{O(\alpha \wedge \beta \mid \gamma)_{L_1 \cup L_2}}$$

$$\text{AND}_V : \frac{O(\alpha_1 \mid \beta)_{L_1}, O(\alpha_2 \mid \beta)_{L_2}, \overset{\leftrightarrow}{\Diamond}(L_1 \cup L_2 \cup \{\beta\})}{O(\alpha_1 \wedge \alpha_2 \mid \beta)_{L_1 \cup L_2}}$$

Informally, the premises used in the derivation tree using hese $V$-rules are not violated by the antecedent of the derived obligation, or, alternatively, the derived obligation is not a CTD obligation of these premises. If the label and the antecedent are consistent, then the derived obligation is not a contrary-to-duty of its premises, see Figure 3.4. We say that the labels formalize the assumptions from which an obligation is derived, and the consistency check $\overset{\leftrightarrow}{\Diamond}$ checks whether the assumptions are violated. The following example illustrates that the labeled deductive system gives the same reading to the apples-and-pears example in Example 3.1 as the two-phase deontic logic 2DL.

**Example 3.5 (Apples-and-pears, continued)** Assume LDL that has at least the inference patterns $\text{RSA}_V$, $\text{RAND}_V$ and $\text{WC}_V$. Consider the set

$$S = \{O(a \vee p|\top)_{\{a\vee p\}}, O(\neg a|\top)_{\{\neg a\}}\}$$

as premise set, where $a$ can be read as 'buying apples' and $p$ as 'buying pears.' In Figure 3.5 below it is shown how the derivations of Example 3.1 in Figure 3.1 are blocked. □

$$\dfrac{\dfrac{\dfrac{O(a \vee p|\top)_{\{a\vee p\}} \quad O(\neg a|\top)_{\{\neg a\}}}{O(\neg a \wedge p|\top)_{\{a\vee p,\neg a\}}}\ \text{AND}}{\dfrac{O(\neg a \wedge p|a)_{\{a\vee p,\neg a\}}}{O(p|a)_{\{a\vee p,\neg a\}}}\ \text{WC}}\ (\text{SA/RSA})}\qquad \dfrac{\dfrac{\dfrac{O(a \vee p|\top)_{\{a\vee p\}} \quad O(\neg a|\top)_{\{\neg a\}}}{O(\neg a \wedge p|\top)_{\{a\vee p,\neg a\}}}\ \text{AND}}{O(p|\top)_{\{a\vee p,\neg a\}}}\ \text{WC}}{O(p|a)_{\{a\vee p,\neg a\}}}\ (\text{SA/RSA})$$

Figure 3.5: The apples-and-pears example in LDL

The apples-and-pears example in labeled deontic logic shows an important property of dyadic deontic logics with a contextual interpretation of the antecedent, namely that the context is restricted to non-violations of premises. If the antecedent is a violation, i.e. if the derived obligation would be a CTD, then the derivation is blocked. Obviously, as a logic the labeled deductive system is quite limited, if only because it lacks a semantics. In the following section, we consider contextual deontic logic, which combines the advantage of labeled deontic logic (only one phase) with the intuitive preference-based semantics of 2DL.

## 3.3 Contextual obligations

In this section, we introduce contextual obligations and discuss the relation between contextual, ordering and minimizing obligations. In contextual deontic logic CDL the obligation 'pears should be bought unless apples are bought' can be derived in the apples-and-pears example.

We start with a semantic intuition for the contextual obligations. Recall the relation between ordering and minimizing obligations in Proposition 2.22. We showed that a model $M$ (without duplicate worlds) satisfies an ordering obligation $O(\alpha \mid \beta)$ iff for all propositional $\beta'$ such that $M \models \overset{\leftrightarrow}{\Box}(\beta' \rightarrow \beta)$, we have $M \models O_{\exists}(\alpha \mid \beta')$. Contextual obligations are in a sense weaker, because only for some propositions $\beta'$ such that $M \models \overset{\leftrightarrow}{\Box}(\beta' \rightarrow \beta)$, we have $M \models O_{\exists}(\alpha \mid \beta')$. We say that the lower bound of $\beta'$ is $\beta$, i.e. $M \models \overset{\leftrightarrow}{\Box}(\beta' \rightarrow \beta)$, and the upper bound is given by the context $\gamma$ of the contextual obligation $O(\alpha|\beta\backslash\gamma)$, i.e.$M \not\models \overset{\leftrightarrow}{\Box}(\beta' \rightarrow \gamma)$, see Proposition 3.12. Contextual obligations are defined as follows.

**Definition 3.6 (Contextual obligation)** The contextual obligation '$\alpha$ should be the case if $\beta$ is the case unless $\gamma$ is the case', written as $O(\alpha|\beta\backslash\gamma)$, is defined as a strong preference of $\alpha \wedge \beta \wedge \neg\gamma$ over $\neg\alpha \wedge \beta$.

$$
\begin{array}{lll}
O(\alpha|\beta\backslash\gamma) & =_{def} & (\alpha \wedge \beta \wedge \neg\gamma) \succ_s (\neg\alpha \wedge \beta) \\
 & = & \overset{\leftrightarrow}{\Box}((\alpha \wedge \beta \wedge \neg\gamma) \rightarrow \Box(\beta \rightarrow \alpha)) \\
O^c(\alpha|\beta\backslash\gamma) & =_{def} & (\alpha \wedge \beta \wedge \neg\gamma) \succ_s (\neg\alpha \wedge \beta) \wedge \overset{\leftrightarrow}{\Diamond}(\alpha \wedge \beta \wedge \neg\gamma) \\
O^{cc}(\alpha|\beta\backslash\gamma) & =_{def} & (\alpha \wedge \beta \wedge \neg\gamma) \succ_s (\neg\alpha \wedge \beta) \wedge \overset{\leftrightarrow}{\Diamond}(\alpha \wedge \beta \wedge \neg\gamma) \wedge \overset{\leftrightarrow}{\Diamond}(\neg\alpha \wedge \beta)
\end{array}
$$

□

We write CDL for the logic CT4O extended with the definition of contextual obligations. The following proposition shows that contextual obligations are in some respects similar to ordering obligations.

**Proposition 3.7** *Let* $M = \langle W, \leq, V\rangle$ *be a* CDL *model and let* $|\alpha|$ *be the set of worlds that satisfy the formula* $\alpha$*. For a world* $w \in W$*, we have* $M, w \models O(\alpha|\beta\backslash\gamma)$ *iff for all* $w_1 \in |\alpha \wedge \beta \wedge \neg\gamma|$ *and all* $w_2 \in |\neg\alpha \wedge \beta|$ *we have* $w_2 \not\leq w_1$*.*

**Proof** *Analogous to the proof of Proposition 2.6, because both are defined in the preference relation* $\succ_s$*.*

The following example illustrates contextual obligations.

**Example 3.8** Consider the Kripke model $M$ represented in Figure 3.6 (repeated from Figure 2.4). We have $M \models O(q|\top\backslash\neg p)$, because $|\neg q| \not\leq |q \wedge p|$. The condition is stronger than $M \models O(q|p)$, because the latter means $|\neg q \wedge p| \not\leq |q \wedge p|$. The condition is weaker than $M \models O(q \wedge p|\top)$, because the latter means $|\neg(q \wedge p)| \not\leq |q \wedge p|$. □

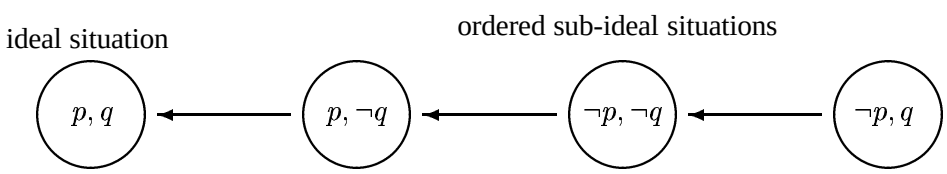



Figure 3.6: Preferential model

The following proposition shows several propositions of contextual obligations.

**Proposition 3.9** *The logic* CDL *has the following theorems.*

**SA**: $O(\alpha|\beta_1\backslash\gamma) \rightarrow O(\alpha|\beta_1 \wedge \beta_2\backslash\gamma)$
**WC**: $O(\alpha_1 \wedge \alpha_2|\beta\backslash\gamma) \rightarrow O(\alpha_1|\beta\backslash\gamma \vee \neg\alpha_2)$
**WT**: $O(\alpha|\beta\backslash\gamma_1) \rightarrow O(\alpha|\beta\backslash\gamma_1 \vee \gamma_2)$
**AND**: $(O(\alpha_1|\beta\backslash\gamma) \wedge O(\alpha_2|\beta\backslash\gamma)) \rightarrow O(\alpha_1 \wedge \alpha_2|\beta\backslash\gamma)$
**RSA**: $(O^c(\alpha|\beta_1\backslash\gamma) \wedge \overset{\leftrightarrow}{\Diamond}(\alpha \wedge \beta_1 \wedge \beta_2 \wedge \neg\gamma)) \rightarrow O^c(\alpha|\beta_1 \wedge \beta_2\backslash\gamma)$
**RAND**: $(O^c(\alpha_1|\beta\backslash\gamma) \wedge O^c(\alpha_2|\beta\backslash\gamma) \wedge \overset{\leftrightarrow}{\Diamond}(\alpha_1 \wedge \alpha_2 \wedge \beta \wedge \neg\gamma)) \rightarrow O^c(\alpha_1 \wedge \alpha_2|\beta\backslash\gamma)$

**Proof** *The theorems can easily be proven in the preference-based semantics. Consider* **WC**. *Let* $W_1 = |\alpha_1 \wedge \alpha_2 \wedge \beta \wedge \neg\gamma|$ *and* $W_2 = |\neg(\alpha_1 \wedge \alpha_2) \wedge \beta|$. $M \models O(\alpha_1 \wedge \alpha_2|\beta\backslash\gamma)$ *implies* $w_2 \not\leq w_1$ *for all* $w_1 \in W_1$ *and* $w_2 \in W_2$*. Let* $W_1' = |\alpha_1 \wedge \beta \wedge \neg(\gamma \vee \neg\alpha_2)|$ *and* $W_2' = |\neg\alpha_1 \wedge \beta|$*. We have*

*$w_2 \not\leq w_1$ for all $w_1 \in W_1'$ and $w_2 \in W_2'$, because $W_1' = W_1$ and $W_2' \subseteq W_2$. Hence we have $M \models O(\alpha_1 | \beta \backslash \gamma \vee \neg\alpha_2)$. Verification of the other theorems is left to the reader.*

The following example illustrates the notion of weakening of the consequent.

**Example 3.10** Reconsider the model $M$ in Example 3.6. We have $M \models O(p \wedge q \mid \top \backslash \bot)$, $M \not\models O(q | \top \backslash \bot)$ and $M \models O(q | \top \backslash \neg p)$. Hence, the context records when the consequent is weakened. □

The following proposition shows several properties of the relation between ordering, minimizing and contextual obligations. In particular, it shows that ordering and minimizing obligations are, in a sense, specific types of contextual obligations. Moreover, from the non-theorems follows that the logic is a non-trivial extension of 2DL, it increases its expressiveness significantly.[1]

**Proposition 3.11** *The logic* CDL *has the following theorems.*

$$\begin{array}{lcl} O(\alpha|\beta\backslash\bot) & \leftrightarrow & O(\alpha|\beta) \\ O^c(\alpha|\beta\backslash\gamma) & \rightarrow & O^c_{\exists}(\alpha|\beta\wedge\neg\gamma) \end{array}$$

*The logic* CDL *does not have the following theorems.*

$$\begin{array}{lcl} O(\alpha|\beta\backslash\gamma) & \leftrightarrow & O(\alpha\wedge\neg\gamma|\beta) \\ O(\alpha|\beta\backslash\gamma) & \leftrightarrow & O(\alpha|\beta\wedge\neg\gamma) \end{array}$$

**Proof** *The theorems follow directly from the semantic definitions. Counterexamples of the non-theorems have been given in Example 3.8.*

The following proposition shows the relation between contextual and minimizing obligations, just like Proposition 2.22 shows the relation between ordering obligations and minimizing obligations.

**Proposition 3.12** *Let $M = \langle W, \leq, V\rangle$ be a Kripke model, such that there are no worlds that satisfy the same propositional sentences. Hence, we identify the set of worlds with a set of propositional interpretations, such that there are no duplicate worlds. We have $M, w \models O^{cc}(\alpha|\beta\backslash\gamma)$ iff there are $\alpha\wedge\beta\wedge\neg\gamma$ and $\neg\alpha\wedge\beta$ worlds, and for all propositional $\beta'$ such that $M, w \models \overset{\leftrightarrow}{\Box}(\beta' \rightarrow \beta)$ and $M, w \not\models \overset{\leftrightarrow}{\Box}(\beta' \rightarrow \gamma)$, we have $M, w \models O^c_{\exists}(\alpha|\beta')$.*

**Proof** *(Analogous to the proof of Proposition 2.22.) $\Rightarrow$ Follows directly from the semantic definitions. $\Leftarrow$ Every world is characterized by a unique propositional sentence. Let $\overline{w}$ denote this sentence that characterizes world $w$. Proof by contraposition. If $M, w \not\models O^{cc}(\alpha|\beta\backslash\gamma)$, then there are $w_1, w_2$ such that $M, w_1 \models \alpha\wedge\beta\wedge\neg\gamma$, $M, w_2 \models \neg\alpha\wedge\beta$ and $w_2 \leq w_1$. Choose $\beta' = \overline{w_1}\vee\overline{w_2}$.*

[1] An inspection of the definitions reveals that we can define contextual obligations in terms of ordering obligations by $O(\alpha|\beta\backslash\gamma) =_{def} O(\alpha\wedge\beta\wedge\neg\gamma|(\alpha\wedge\beta\wedge\neg\gamma)\vee(\neg\alpha\wedge\beta))$.

*$w_2$ is an element of the preferred $\beta'$ worlds, because there are no duplicate worlds. (If duplicate worlds are allowed, then there could be a $\beta'$ world $w_3$ which is a duplicate of $w_1$, and which is strictly preferred to $w_1$ and $w_2$.) We have $M, w_2 \not\models \alpha$ and therefore $M, w \not\models O^c_{\exists}(\alpha|\beta')$,*

The following example illustrates that contextual deontic logic has the same reading of the apples-and-pears problem as our two-phase deontic logic 2DL. In particular, it illustrates that the semantic representation is similar to the representation in 2DL, but the proof-theoretic representation does not depend on two distinct phases.

**Example 3.13 (Apples-and-Pears, continued)** Consider the set of contextual obligations $S = \{O^c(a \vee p|\top\backslash\bot), O^c(\neg a|\top\backslash\bot)\}$. We have $S \models O^c(p|\top\backslash a)$, as is shown in Figure 3.7, which expresses that our little sister should buy pears, unless she buys apples. The crucial observation is that we do not have $O^{cc}(p|a\backslash\gamma)$ for any $\gamma$, and a typical countermodel is again the model in Figure 3.2. □

$$\dfrac{\dfrac{\dfrac{O^c(a \vee p|\top\backslash\bot) \quad O^c(\neg a|\top\backslash\bot)}{O^c(\neg a \wedge p|\top\backslash\bot)}\ \text{AND}}{O^c(p|\top\backslash a)}\ \text{WC}}{O^c(p|a\backslash a)}\ (\text{RSA})$$

Figure 3.7: Apples-and-pears in contextual deontic logic

In this section we introduced contextual obligations and showed that they combine strengthening of the antecedent and weakening of the consequent. A drawback of two-phase logics is that they do not support lemma handling. Compared to the two-phase deontic logic 2DL, in CDL the combining of strengthening of the antecedent and weakening of the consequent is established in one phase. Thus, the logic CDL supports lemma-handling.

## 3.4 Related research

The *apples-and-pears example* is a new example introduced in this thesis, although there are examples of sentences with the same logical structure as the apples-and-pears example. Examples are given by Van Fraassen [vF73] (quoting Stalnaker): (1) you should obey your father or your mother, and (2) you should not obey your mother, and by Horty [Hor93]: (1) you should serve in the army or attend alternative service, and (2) you should not serve in the army. The distinction is that they give the examples in a monadic modal logic to illustrate intuitions behind the restricted conjunction rule RAND, see the discussion in Section 2.6.3.

There are several similarities between our deontic logics and logics of defeasible reasoning. Most importantly, the contextual obligations $O(\alpha|\beta\backslash\gamma)$, read as '$\alpha$ ought to be the case if $\beta$ is the case unless $\gamma$ is the case,' can be compared with Reiter default rules $\frac{\beta:\neg\gamma}{\alpha}$, where $\neg\gamma$ is the justification of the default rule [Rei80]. The main distinction between CDL and Reiter's default logic is that contextual obligations are not used as inference rules. In CDL, we derive contextual

obligations from contextual obligations, which can be compared to the derivation of defaults from defaults. In Reiter's default logic, defaults are used to generate extensions. A similarity between CDL and default logic is that contextual obligations as well as defaults express preferences. Reiter's defaults express preferences on assumptions. In the default $\frac{\beta:\neg\gamma}{\alpha}$ we have that $\alpha \wedge \beta$ are preferred to $\neg\alpha \wedge \beta$, and this preference is cancelled when $\gamma$ is the case.

## 3.5 Conclusions

In this chapter, we further studied the relation between obligations and preferences. We have established the following results.

1. We have identified a *new deontic example*, the apples-and pears example (Example 3.1). We have shown two interpretations of this example.

2. We have proposed the *new labeled deontic logic* LDL. We have shown that labeled obligations give the same interpretation of the apples-and-pears example as the two-phase deontic logic 2DL. Labeled deontic logic explains that the blocked derivations are derivations of contrary-to-duty obligations from their primary obligations.

3. We have proposed the *new contextual deontic logic* CDL. We have shown that contextual obligations give the same interpretation of the apples-and-pears example as the two-phase deontic logic 2DL. Contextual deontic logic combines the preference-based semantics of 2DL with a one-phase proof theory.

CDL has been used to further analyze the relation between obligations and preferences. The contextual obligations $O(\alpha\,|\,\beta\backslash\gamma)$ are a generalization of the ordering obligations developed in Chapter 2, because an ordering obligation $O(\alpha\,|\,\beta)$ is logically equivalent to a contextual obligation $O(\alpha\,|\,\beta\backslash\bot)$. Like 2DL, the contextual obligations can be extended with the no-dilemma assumption, permission operators and factual detachment. The main conclusions of this chapter concerning the relation between obligations and preferences are the following.

1. **Deontic reasoning is a two-phase process?** In the previous chapter we argued that the preference-based semantics indicate that deontic reasoning is a two-phase process of ordering and minimizing. Two phases are necessary in a dyadic deontic logic to combine several inference patterns, for example strengthening of the antecedent and weakening of the consequent. However, in this chapter we showed that they can also be combined in a one-phase proof theory, if the language is made more expressive.

2. **Weakening of the consequent introduces exceptions.** The explicit exceptions of CDL reveal that weakening of the consequent corresponds to introducing exceptions.

The latter relation is surprising, because exceptions are an issue normally formalized in default logic. In the next chapter we use (extensions of) the logic CDL to analyze the relation between obligations and defeasibility.

# Chapter 4

# Defeasible deontic logic

In this chapter we study the relation between obligations and defeasibility. We give a general analysis of different types of defeasibility in defeasible deontic logics. We argue that (at least) three types of defeasibility must be distinguished in a defeasible deontic logic. First, we make a distinction between *factual defeasibility*, that formalizes overshadowing of an obligation by a violating fact, and *overridden defeasibility*, that formalizes cancelling of an obligation by other conditional obligations. Second, we show that overridden defeasibility can be further divided into *strong overridden defeasibility*, that formalizes specificity, and *weak overridden defeasibility*, that formalizes the overriding of prima facie obligations. Our general analysis can be applied to any defeasible deontic logic, because we use inference patterns to analyze the different types of defeasibility. Moreover, we illustrate the intuitions behind the various distinctions with preference-based semantics. We also show that these distinctions are essential for an adequate analysis of notorious contrary-to-duty paradoxes such as the Chisholm and Forrester paradox in a defeasible deontic logic. In particular, they are essential to distinguish between violations and exceptions.

The interference between violability and specificity was first discussed from a proof-theoretic perspective in [vdT94] and from a semantic perspective in [TvdT94b, TvdT95]. This chapter is a modified and extended version of [vdTT95a, vdTT97b].

## 4.1 Obligations and defeasibility

Dyadic modal logics were introduced to formalize deontic reasoning about contrary-to-duty obligations in, for example, the Chisholm paradox that we will discuss later. An example of a conditional obligation in a dyadic modal logic is $O(h \mid r)$, which expresses that "you ought to be helped ($h$) when you are robbed ($r$)." Similarly, $O(\neg r \mid \top)$ expresses that "you ought not to be robbed," where $\top$ stands for any tautology. If both $O(\neg r \mid \top)$ and $r$ are true, then we say that the obligation is *violated* by the fact $r$. In recent years it was argued by several authors that these dyadic obligations can be formalized in non-monotonic logics [McC94a, Hor93, RL93].

In this chapter we argue that contrary-to-duty obligations do have a defeasible aspect, but a different one than is usually thought. The first part of this claim follows directly from Alchourrón's [Alc93] definition of a defeasible conditional as a conditional that lacks strengthening

of the antecedent, represented by the inference pattern

$$\text{SA} : \frac{O(\alpha|\beta_1)}{O(\alpha|\beta_1 \wedge \beta_2)}$$

Alchourrón's definition is based on the idea that lack of strengthening of the antecedent is a kind of implicit non-monotonicity. The relation between strengthening of the antecedent and non-monotonicity can be made explicit with the following inference pattern *Exact Factual Detachment* EFD, see Section 2.5. Exact factual detachment can be represented by the inference pattern

$$\text{EFD} : \frac{O(\alpha|\beta), \mathcal{A}\beta}{O(\alpha)}$$

in which $O\alpha$ is a new, monadic modal operator, and $\mathcal{A}$ is an all-that-is-known operator [Lev90]: $\mathcal{A}\phi$ is true if and only if $\phi$ is logically equivalent with all factual premises given. The inference pattern EFD is based on the intuition that the antecedent of a dyadic obligation restricts the focus to possible situations in which the antecedent is *assumed* to be factually true, and the consequent represent what is obligatory, given that *only* these facts are assumed. If the facts are equivalent to the antecedent, then the consequent can be considered as an absolute obligation. From the properties of $\mathcal{A}$ follows immediately that EFD is monotonic iff the dyadic obligations have strengthening of the antecedent. Dyadic deontic logics that can represent contrary-to-duty reasoning are defeasible deontic logics, because the dyadic obligations typically lack strengthening of the antecedent, see Section 1.3.5. In this sense, contrary-to-duty obligations do have a defeasible aspect.

However, we argue that this defeasible aspect of contrary-to-duty obligations is a different one than is usually proposed. In this chapter we analyze defeasibility in defeasible deontic logic by analyzing different conditions on strengthening of the antecedent. In particular, we analyze the inference relation of defeasible deontic logics with inference patterns, in a similar way as Kraus *et al* [KLM90] analyze logics of defeasible reasoning. Moreover, we give preference-based semantic intuitions for the inference patterns. The advantage of our approach is that (1) it is applicable to any defeasible deontic logic, because of the generality of the inference patterns, and (2) it gives also an explanation of the intuitions behind the inference patterns by the preference semantics.

### 4.1.1 Cancelling and overshadowing

The main claim of this chapter is that the defeasible aspect of contrary-to-duty obligations is different from the defeasible aspect of, for example, Reiter's default rules [Rei80]. Different types of defeasibility in a logic of defeasible reasoning formalize a single notion, whereas defeasible deontic logics formalize two different notions. Consider first the logics of defeasible reasoning and the famous Tweety example. In the case of factual defeasibility, we say that the 'birds fly' default is *cancelled* by the fact $\neg f$, and in the case of overridden defeasibility by the 'penguins do not fly' default. By cancellation we mean, for example, that if $\neg f$ is true, then the default assumption that $f$ is true is null and void. The truth of $\neg f$ implies that the default assumption about $f$ is completely falsified.

The fundamental difference between deontic logic and logics for defeasible reasoning is that $O(\neg r|\top) \wedge r$ is *not* inconsistent. That is the reason why the deontic operator $O$ had to be represented as a modal operator with a possible worlds semantics, to make sure that *both* the obligation

and its violation could be true at the same time. Although the obligation $O(\neg r \mid \top)$ is violated by the fact $r$, the obligation still has its force, so to say. This still being in force of an obligation is reflected, for example, by the fact that someone has to pay a fine even if she does $r$. Even if you are robbed, you should not have been robbed. But if penguins cannot fly, it makes no sense to state that normally they can fly. We will refer to this relation between the obligation and its violation as *overshadowing* to distinguish it from *cancellation* in the case of defeasible logics. By the overshadowing of an obligation we mean that it is still in force, but it is no longer to be acted upon.

The conceptual difference between cancelling and overshadowing is analogous to the distinction between 'defeasibility' and 'violability' made by Smith in [Smi93] and by Prakken and Sergot in [PS96]. An essential difference between those papers and this one is that in this chapter we argue that violability has to be considered as a type of defeasibility too, because it also induces a constraint on strengthening of the antecedent. The main advantage of the violability-as-defeasibility perspective is that it explains the distinctions *and the similarities* between cancelling and overshadowing. Moreover, it can be used to analyze complicated phenomena like prima facie obligations, which have cancelling as well as overriding aspects.

### 4.1.2 Different types of defeasibility

In defeasible reasoning one can distinguish at least three types of defeasibility, based on different semantic intuitions. To illustrate the difference between the different types we discuss the penguin example in Geffner and Pearl's assumption-based default theories [GP92]. In such theories, the 'birds fly' default rule is expressed by a factual sentence $\delta_1 \rightarrow f$ and a default sentence $\top \Rightarrow \delta_1$, and the 'penguins do not fly' default by $p \wedge \delta_2 \rightarrow \neg f$ and $p \Rightarrow \delta_2$. Here, '$\rightarrow$' is the classical material implication and '$\Rightarrow$' some kind of default implication. The $\delta_i$ constants are the so-called assumptions; for each default in the set of premises a distinct constant is introduced. Geffner and Pearl's so-called conditional entailment maximizes these assumptions, given certain constraints. In conditional entailment, the 'birds fly' default can be defeated by the fact $\neg f$, or it can be overridden by the more specific 'penguins do not fly' default. The first follows directly from $\neg f \rightarrow \neg\delta_1$, i.e. the contraposition of the factual sentence $\delta_1 \rightarrow f$, and the second follows from the fact that $p \rightarrow \neg\delta_1$ can be derived from the constraints of conditional entailment (we do not give the complicated proof; see [GP92] for these details). We call the first case *factual defeasibility* and the last case *overridden defeasibility*. The distinction between factual and overridden defeasibility is only the start of a classification of different types of defeasibility. To illustrate the distinction between different types of overridden defeasibility, we consider the adapted 'penguins do not fly and live on the Southern Hemisphere' default $p \wedge \delta_2 \rightarrow (\neg f \wedge s)$. In some logics of defeasible reasoning, the 'birds fly' default is overridden whenever $p$ is true. In other logics it is overridden when $p$ is true *but only as long as $s$ is not false*. If $s$ is false, then the penguin default is no longer applicable. In the first logics the 'birds fly' default is not reinstated, whereas in the second logics it is, because it was only suspended. In other words, in the latter case the penguin default overrides the bird default only when it is applicable itself. We call the first case *strong overridden defeasibility* and the second case *weak overridden defeasibility*. The different types of overridden defeasibility are based on different semantic intuitions. Strong overridden defeasibility is usually based on a probabilistic interpretation of defaults (most birds fly, but penguins are exceptional), like in Pearl's $\epsilon$-semantics [Pea88]. Weak overridden defeasi-

bility is usually based on an argument-based conflict resolution interpretation (there is a conflict between the two rules, and the second one has highest priority), for example in conditional entailment and Brewka's prioritized default logic [Bre94].

The distinction between different types of defeasibility is crucial in logics that formalize reasoning about obligations which can be overridden by other obligations. Overridden defeasibility becomes relevant when there is a (potential) conflict between two obligations. For example, there is a conflict between $O(\alpha_1|\beta_1)$ and $O(\alpha_2|\beta_2)$ when $\alpha_1$ and $\alpha_2$ are contradictory, and $\beta_1$ and $\beta_2$ are factually true. There are several different approaches to deal with deontic conflicts, see Section 1.3.6. In von Wright's so-called standard deontic logic SDL [vW51] a deontic conflict is inconsistent. In weaker deontic logics, like Chellas minimal deontic logic MDL [Che74], a conflict is consistent and called a 'deontic dilemma.' In a defeasible deontic logic a conflict can be *resolved*, because one of the obligations overrides the other one. For example, overridden structures can be based on a notion of specificity, like in Horty's well-known example that 'you should not eat with your fingers,' but 'if you are served asparagus, then you should eat with your fingers' [Hor93]. In such cases, we say that an obligation is *cancelled* when it is overridden, because it is analogous to cancelling in logics of defeasible reasoning. The obligation not to eat with your fingers is cancelled by the exceptional circumstances that you are served asparagus. A different kind of overridden structures have been proposed by Ross [Ros30] and formalized, for example, by Morreau in [Mor96]. In Ross' ethical theory, an obligation which is overridden has not become a 'proper' or actual duty, but it remains in force as a prima facie obligation. For example, the obligation not to break a promise may be overridden to prevent a disaster, but even when it is overridden it remains in force as a prima facie obligation. As actual obligation the overridden obligation is cancelled, but as prima facie obligation it is only overshadowed. Because of this difference between cancellation and overshadowing, it becomes essential not to confuse the types of defeasibility in analyzing the deontic paradoxes. We show that if they are confused, counterintuitive conclusions follow for the Chisholm and Forrester paradoxes.

In the figure below the three different faces of defeasibility in defeasible deontic logic are represented with their corresponding character (cancelling or overshadowing). In non-deontic defeasible logic the different types of defeasibility all have a cancelling character.

| | overshadowing | cancelling |
|---|---|---|
| Factual defeasibility | X | |
| Strong overridden defeasibility | | X |
| Weak overridden defeasibility | X | X |

Table 4.1: Matrix

This chapter is organized as follows. In Section 4.2 we give a detailed comparison of factual and overridden defeasibility in deontic reasoning, and we show that the Chisholm paradox can be analyzed as a case of factual defeasibility rather than overridden defeasibility. In Section 4.3 we focus on the overshadowing aspect of factual defeasibility and the cancellation aspect of overridden defeasibility by analyzing specificity, and we show that in an adequate analysis of an extension of the Forrester paradox both these aspects have to be combined. In Section 4.4 we focus

on the cancelling aspect and the overshadowing aspect of overridden defeasibility by analyzing prima facie obligations.

## 4.2 Overridden versus factual defeasibility

In this section we analyze the fundamental difference between overridden and factual defeasibility in a defeasible deontic logic by formalizing contrary-to-duty reasoning as a kind of overridden defeasibility as well as a kind of factual defeasibility. Moreover, we show that contrary-to-duty reasoning is best formalized by the latter one.

### 4.2.1 Chisholm paradox

The following example describes the notorious Chisholm paradox [Chi63], also called the CTD paradox, or the paradox of deontic detachment. The original paradox was given in a monadic modal logic, see Section 1.3.3. Here we give the obvious formalization in a non-defeasible dyadic logic. See [Tom81] for a discussion of the Chisholm paradox in several dyadic deontic logics. To make our analysis as general as possible, we assume as little as possible about the deontic logic we use. The analyses given in this chapter in terms of inference patterns are, in principle, applicable to any deontic logic.

**Example 4.1 (Chisholm paradox)** Assume a dyadic deontic logic that has at least substitution of logical equivalents and the inference patterns (unrestricted) *Strengthening of the Antecedent* SA, *Weakening of the Consequent* WC and a version of *Deontic Detachment* DD$'$.

$$\text{SA}: \frac{O(\alpha|\beta_1)}{O(\alpha|\beta_1 \wedge \beta_2)} \qquad \text{WC}: \frac{O(\alpha_1|\beta)}{O(\alpha_1 \vee \alpha_2|\beta)} \qquad \text{DD}': \frac{O(\alpha|\beta), O(\beta|\gamma)}{O(\alpha \wedge \beta|\gamma)}$$

Notice that the following inference pattern *Deontic Detachment* (or transitivity) DD can be derived from WC and DD$'$.

$$\text{DD}: \frac{O(\alpha|\beta), O(\beta|\gamma)}{O(\alpha|\gamma)}$$

Furthermore, assume the premises $O(a|\top)$, $O(t|a)$ and $O(\neg t|\neg a)$, where $\top$ stands for any tautology, $a$ can be read as the fact that a certain man goes to the assistance of his neighbors and $t$ as the fact that he tells them he is coming. Notice that the third premise $O(\neg t|\neg a)$ is a CTD obligation of the (primary) obligation $O(a|\top)$, because its antecedent is inconsistent with the consequent of the latter.

The paradoxical derivation of $O(t|\neg a)$ from the Chisholm paradox is represented in Figure 4.1. The intuitive obligation $O(a \wedge t|\top)$ can be derived by DD$'$ from the first two obligations. It seems intuitive, because in the ideal situation the man goes to the assistance of his neighbors and he tells them he is coming. The obligation $O(t|\top)$ can be derived from $O(a \wedge t|\top)$ by WC (or from the premises by DD). The derived obligation $O(t|\top)$ expresses that if the man does not tell his neighbors, then the ideal situation is no longer reachable. However, from $O(t|\top)$ the counterintuitive $O(t|\neg a)$ can be derived by SA. This is counterintuitive, because there is no reason to tell the neighbors he is coming when the man does not go. In contrast, in this violation context the man should do the opposite! Moreover, in several deontic logics the set of obligations $\{O(\neg t|\neg a), O(t|\neg a)\}$ is inconsistent. □

$$\dfrac{\dfrac{\dfrac{O(t|a) \quad O(a|\top)}{O(a \wedge t|\top)}\,\text{DD}'}{O(t|\top)}\,\text{WC}}{O(t|\neg a)}\,\text{SA}$$

Figure 4.1: Chisholm paradox

In this example the Chisholm paradox is presented in a normal dyadic deontic logic, to show its paradoxical character. In the next section, we analyze the paradox in a defeasible deontic logic that has only overridden defeasibility. This analysis solves the paradox, but for the wrong reasons. Finally, in Section 4.2.2 we give an analysis of the Chisholm paradox in terms of factual defeasibility, which is more satisfactory. In Section 4.2.3 we analyze factual defeasibility with a preference semantics.

### 4.2.2 Overridden defeasibility

In recent years several authors proposed to solve the Chisholm paradox by analyzing its problematic CTD obligation as a type of overridden defeasibility (see e.g. [McC94a, RL93]).[1] The underlying idea is that a CTD obligation can be considered as a conflicting obligation that overrides a primary obligation. Although this idea seems to be very intuitive at first sight, we claim that the perspective of CTD obligations as a kind of overridden defeasibility is misleading. It is misleading, because although this perspective yields most (but not all!) of the correct conclusions for the Chisholm paradox, it does so for the wrong reasons. We show that it is more appropriate to consider the CTD obligation as a kind of factual defeasibility. This does not mean that there is no place for overridden defeasibility in deontic logic. By a careful analysis of an extended version of another notorious paradox of deontic logic, the Forrester paradox, we show that sometimes combinations of factual and overridden defeasibility are needed to represent defeasible deontic reasoning. But first we give our analysis of the Chisholm paradox. The following example shows that the counterintuitive obligation of Example 4.1 cannot be derived in a defeasible deontic logic with overridden defeasibility. For our argument we use a notion of overridden based on specificity.

**Example 4.2 (Chisholm paradox, continued)** Assume that SA is replaced by the following Restricted Strengthening of the Antecedent rule $\text{RSA}_O$. $\text{RSA}_O$ contains the so-called non-overridden condition $C_O$, which requires that $O(\alpha|\beta_1)$ is not overridden for $\beta_1 \wedge \beta_2$ by some more specific $O(\alpha'|\beta')$.[2]

$$\text{RSA}_O : \frac{O(\alpha|\beta_1), C_O}{O(\alpha|\beta_1 \wedge \beta_2)}$$

[1] McCarty [McC94a] does not analyze the Chisholm paradox but the so-called Reykjavic paradox, which he considers to contain 'two instances of the Chisholm paradox, each one interacting with the other.'

[2] The overridden condition $C_O$ is based on a simplified notion of specificity, because background knowledge is not taken into account and an obligation cannot be overridden by more than one obligation. A more sophisticated definition of overridden can be found in the literature of logics of defeasible reasoning. For our purposes this simple definition is enough, because it is a weak definition (most definitions of specificity are extensions of this definition). For a discussion on the distinction between background and factual knowledge, see [vdT94].

where condition $C_O$ is defined as follows:

> $C_O$: there is no premise $O(\alpha'|\beta')$ such that $\beta_1 \wedge \beta_2$ logically implies $\beta'$, $\beta'$ logically implies $\beta_1$ and not vice versa and $\alpha$ and $\alpha'$ are contradictory.

The 'solution' for the paradox is represented in Figure 4.2. This figure should be read as follows. The horizontal lines represent *possible* derivation steps. Blocked derivation steps are represented by dashed lines. For example, the last derivation step is blocked, and the cause of the blocking is represented by the obligation $O(\neg t|\neg a)$ above the blocked inference rule. We compare the blocked derivation in Figure 4.2 with the derivation in Figure 4.1. The intuitive obligation $O(t\,|\,\top)$ can still be derived by DD (hence, by DD$'$ and WC) from the first two obligations. From $O(t\,|\,\top)$ the counterintuitive $O(t\,|\,\neg a)$ cannot be derived by RSA$_O$, because $O(t\,|\,\top)$ is overridden for $\neg a$ by the CTD obligation $O(\neg t\,|\,\neg a)$, i.e. $C_O$ is false. Hence, the counterintuitive obligation is cancelled by the exceptional circumstances that the man does not go to the assistance. □

$$\begin{array}{ccc}
\dfrac{\dfrac{O(t|a) \quad O(a|\top)}{O(a \wedge t|\top)}\,\text{DD}'}{O(t|\top)}\,\text{WC} & & O(\neg t|\neg a) \\
 & & \downarrow \\
- - - - - - - - - & & (\text{RSA}_O) \\
O(t|\neg a) & &
\end{array}$$

Figure 4.2: Chisholm paradox solved by overridden defeasibility

Overridden defeasibility yields intuitive results from the Chisholm paradox, but for the wrong reasons. A simple counterargument against the solution of the paradox in Example 4.2 is that overriding based on specificity does not solve the paradox anymore when the premise $O(a\,|\,\top)$ is replaced by a premise with a non-tautological antecedent. For example, if it is replaced by $O(a|i)$, where $i$ can be read as the fact that the man is personally invited to assist. Another counterargument against the solution of the paradox for *any* definition of overridden is that the derivation of $O(t\,|\,\neg a)$ is also counterintuitive when the set of premises contains only the first two obligations, as is the case in the following example.

**Example 4.3 (Chisholm paradox, continued)** Assume only the premises $O(a|\top)$ and $O(t|a)$. Again the intuitive obligation $O(t\,|\,\top)$ can be derived by DD. From this derived obligation the counterintuitive $O(t\,|\,\neg a)$ can be derived by RSA$_O$, because there is no CTD obligation which cancels the counterintuitive obligation. □

If the obligation $O(t\,|\,\top)$ can be derived but not the obligation $O(t\,|\,\neg a)$, then we say that 'deontic detachment holds as a defeasible rule.' Unrestricted strengthening of the antecedent cannot be applied to the obligation $O(t|\top)$, derived by deontic detachment DD. This restriction is the characteristic property of defeasible conditionals [Alc93]. The underlying intuition is that the inference of the obligation of the man to tell his neighbors that he is coming is made *on the assumption that he goes to their assistance*. If he does not go, then this assumption is violated

$$\dfrac{\dfrac{\dfrac{O(t|a) \quad O(a|\top)}{O(a \wedge t|\top)}\,\text{DD}'}{O(t|\top)}\,\text{WC}}{O(t|\neg a)}\,\text{RSA}_O$$

Figure 4.3: Chisholm paradox, continued

and the obligation based on this assumption is factually defeated. We say that the man should tell his neighbors, unless he does not go to their assistance.

The problematic character of DD is well-known from the Chisholm paradox. A popular 'solution' of the paradox is not to accept DD′ for a deontic logic. However, this rejection of DD′ causes serious semantic problems for these logics. For example, Tomberlin [Tom81] showed that there are semantic problems related to the rejection of DD′ for Mott's solution of the Chisholm paradox [Mot73], which does not accept DD′. Moreover, the apples-and-pears example of Section 3.1 shows that similar problems occur when $\text{RSA}_O$, WC and the conjunction rule AND are accepted. This last rule is accepted by many deontic logics.

**Example 4.4 (Apples-and-Pears)** Assume a dyadic deontic logic that has at least substitution of logical equivalents and the inference patterns $\text{RSA}_O$, WC and the following conjunction rule AND.

$$\text{AND}: \frac{O(\alpha_1|\beta), O(\alpha_2|\beta)}{O(\alpha_1 \wedge \alpha_2|\beta)}$$

Notice that the following inference pattern *Consequential Closure* (CC) can be derived from WC and AND.

$$\text{CC}: \frac{O(\alpha_1|\beta), O(\alpha_1 \rightarrow \alpha_2|\beta)}{O(\alpha_2|\beta)}$$

Furthermore, assume as premise sets

$$S = \{O(a \vee p|\top), O(\neg a|\top)\} \quad \textit{and} \quad S' = \{O(a \vee p|\top), O(\neg a|\top), O(\neg p|a)\}$$

where $a$ can be read as 'buying apples' and $p$ as 'buying pears.' A derivation of the obligation $O(p|a)$ from $S$ is represented In Figure 4.4. The intuitive obligation $O(\neg a \wedge p|\top)$ can be derived by AND. From this obligation, the obligation $O(p|\top)$ is derived by WC (hence, from the premise set by CC). From this derived obligation, the obligation $O(p\,|\,a)$ can be derived by $\text{RSA}_O$. The obligation is not derivable from $S'$ by $\text{RSA}_O$, because the CTD obligation $O(\neg p\,|\,a)$ overrides the obligation $O(p|\top)$ for $a$. However, this blocking for $S'$ does not suffice for $S$, just like the blocking in Example 4.2 does not suffice for Example 4.3. □

An alphabetic variant of Example 4.4 is the following version of the Chisholm paradox, in which the conditional obligation is represented as an absolute obligation. However, it is usually argued that the premise $O(a \rightarrow t|\top)$ does not represent the conditional obligation correctly.

**Example 4.5 (Chisholm paradox, continued)** Consider $O(a \mid \top)$ and $O(a \rightarrow t \mid \top)$. The intuitive obligation $O(t|\top)$ is derived from the two premises by CC. However, from this derived obligation the counterintuitive $O(t|\neg a)$ can be derived by SA or $\text{RSA}_O$. □

$$\dfrac{\dfrac{\dfrac{O(a \vee p|\top) \quad O(\neg a|\top)}{O(\neg a \wedge p|\top)}\ \text{AND}}{O(p|\top)}\ \text{WC}}{O(p|a)}\ \text{RSA}_O$$

Figure 4.4: Apples-and-pears example with overridden defeasibility

The examples show that CTD reasoning (i.e., reasoning about sub-ideal behavior) cannot be formalized satisfactorily in a defeasible deontic logic with only overridden defeasibility.

### 4.2.3 Factual defeasibility

As an illustrative example of a formalization of factual defeasibility, we formalize the Chisholm paradox in contextual deontic logic, see Section 3.3. The contextual obligation $O^{cc}(\alpha \mid \beta \setminus \gamma)$ are read as '$\alpha$ should be the case if $\beta$ is the case unless $\gamma$ is the case' and can be compared with the Reiter default rule $\frac{\beta:\neg\gamma}{\alpha}$, where $\neg\gamma$ is the justification of the default rule [Rei80]. The unless clause formalizes a kind of factual defeasibility, because it blocks strengthening of the antecedent (thus it is defeasibility) and it does not refer to any other obligation for this blocking (thus it is factual). The crucial observation of the Chisholm paradox below is that if the premises are valid in all cases (have a context 'unless $\perp$', where $\perp$ is a contradiction), then the derived obligations may still be only valid in a restricted context. The context encodes in such a case the assumptions on which an obligation is derived (i.e. when the obligation is factually defeated). We subscript the inference patterns with a $V$, to emphasize the factual defeasibility (in contrast to the overridden defeasibility used in the previous section). Factual defeasibility is represented in the inference patterns by a consistency check. For example, the derivation of $O^{cc}(\alpha|\beta_1 \wedge \beta_2 \setminus \gamma)$ from $O^{cc}(\alpha \mid \beta_1 \setminus \gamma)$ has a consistency check $\overset{\leftrightarrow}{\Diamond}(\alpha \wedge \beta_1 \wedge \beta_2 \wedge \neg\gamma)$. We call these consistency checks the restriction $C_V$, where $V$ stands for violability. They emphasize the distinction with $C_O$, where $O$ stands for overridden.

$$\text{RSA}_V : \frac{O(\alpha|\beta_1 \setminus \gamma), C_V}{O(\alpha|\beta_1 \wedge \beta_2 \setminus \gamma)}, \quad C_V : \begin{array}{l} \alpha \wedge \beta_1 \wedge \beta_2 \wedge \neg\gamma \text{ is consistent}, \textit{and} \\ \neg\alpha_1 \wedge \beta_1 \wedge \beta_2 \text{ is consistent} \end{array}$$

The following example illustrates that now the Chisholm paradox can be analyzed in contextual deontic logic. The example shows that factual defeasibility of the Chisholm paradox is caused by contextual reasoning, because the *premises* do not have exceptions, only derived obligations have exceptions. Thus, this aspect of factual defeasibility is quite different from defeasibility related to exceptional circumstances or abnormality formalized in logics of defeasible reasoning, because in that case the premises are subject to exceptions. For example, factual defeasibility can be used in a logic for defeasible reasoning to formalize 'birds normally fly, unless they are penguins' by $N(f|b \setminus p)$ *as a premise*.

**Example 4.6 (Chisholm paradox, continued)** Consider the set of obligations

$$S = \{O^{cc}(a|\top \setminus \perp), O^{cc}(t|a \setminus \perp)\}$$

The solution of the counterintuitive derivation of the Chisholm paradox in Example 4.3 is represented in Figure 4.5. The contextual obligation $O^{cc}(t\,|\,\top\setminus\neg a)$ represents that the man should tell his neighbors, unless he does not go to their assistance. □

$$\dfrac{\dfrac{O^{cc}(t|a\setminus\bot) \quad O^{cc}(a|\top\setminus\bot)}{O^{cc}(a\wedge t|\top\setminus\bot)}\ \mathrm{DD}'_V}{O^{cc}(t|\top\setminus\neg a)}\ \mathrm{WC}_V$$
$$\overline{\qquad\qquad}\ (\mathrm{RSA}_V)$$
$$O^{cc}(t|\neg a\setminus\neg a)$$

Figure 4.5: Chisholm paradox solved by factual defeasibility

It can easily be checked that the counterintuitive derivation of $O(p\,|\,a)$ by $\mathrm{RSA}_O$ in Example 4.4 is blocked by $\mathrm{RSA}_V$ too. The examples show that CTD structures sometimes look like overridden defeasible reasoning structures, but a careful analysis shows that they are actually cases of factual defeasibility.

The reader might wonder why we consider condition $C_V$ to be a type of factual defeasibility. In this chapter we only discuss conditional obligations, and how these can be derived from each other. Facts do not seem to come into the picture here. However, a closer analysis reveals that factual defeasibility is indeed the underlying mechanism. The antecedent of a dyadic obligation restricts the focus to possibilities in which the antecedent is *assumed* to be factually true, and the consequent represent what is obligatory, given that these facts are assumed. Hence, the consequent refers to 'the best of the bad lot.' As we discussed in the introduction, these facts can be made explicit with a kind of factual detachment, for example with EFD. From the Chisholm set $\{O(a|\top), O(t|a), O(\neg t|\neg a)\}$ and $\mathcal{A}\top$, we can derive $Ot$ by EFD, and from $\mathcal{A}\neg a$ we can derive $O\neg t$, but not $Ot$. Hence, by adding a fact ($\neg a$) we loose a deontic conclusion ($Ot$).

Moreover, a comparison with, for example, Brewka's prioritized default logic [Bre94] illustrates that $C_V$ is a kind of factual defeasibility. Consider the classical example of non-transitivity of default rules, which consists of the default rules that 'usually, students are adults' ($\frac{s:a}{a}$) and that 'usually, adults are employed' ($\frac{a:e}{e}$). Given that we know that somebody is a student, we can defeat the default conclusion that this person is employed in two ways. Either, it is defeated by the more specific default rule that students are usually unemployed ($\frac{s:\neg e}{\neg e}$), which is a case of overridden defeasibility, or it is defeated by the defeating fact ($\neg a$) that the particular student is known to be no adult. This latter case of defeasibility is the type of factual defeasibility that is analogous to the defeasibility in the Chisholm paradox.

This analogy with default logic also illustrates what we mean by deontic detachment as a defeasible rule. The transitivity of the two default rules above can be blocked either by overridden or factual defeasibility. If neither of the two are the case, then the transitivity holds. In this sense one could say that in default logic transitivity holds as a defeasible rule. Analogously, we say that deontic detachment holds as a defeasible rule. If we only know $O^{cc}(t\,|\,a\setminus\bot)$ and $O^{cc}(a|\top\setminus\bot)$, then we can apply deontic detachment, which results in $O^{cc}(t\,|\top\setminus\neg a)$. But this detachment is defeated if we assume in the antecedent of this conclusion that $\neg a$ is true.

### 4.2.4 Preference semantics

The formalization of the Chisholm paradox in contextual deontic logic CDL gives an intuitive semantic interpretation of factual defeasibility. The following definition repeats the semantic definition from Section 3.3. The preference semantics represent the notion of deontic choice: a preference of $\alpha_1$ over $\alpha_2$ means that if an agent can choose between $\alpha_1$ and $\alpha_2$, she should choose $\alpha_1$. An obligation for $\alpha$ is formalized by a preference of $\alpha$ over $\neg\alpha$. Thus, if the agent can choose between $\alpha$ and $\neg\alpha$, then she should choose $\alpha$. Similarly, a conditional obligation for $\alpha$ if $\beta$ is formalized by a preference of $\alpha \wedge \beta$ over $\neg\alpha \wedge \beta$. This preference is formalized by condition (3) of Definition 4.7 below. The other conditions (1) and (2) of Definition 4.7 formalize the condition that in order to choose between $\alpha$ and $\neg\alpha$, these opportunities must be logically possible.

**Definition 4.7 (Contextual obligation)** Let $M = \langle W, \leq, V \rangle$ be a Kripke model that consists of $W$, a set of worlds, $\leq$, a binary reflexive and transitive relation on $W$, and $V$, a valuation of the propositions in the worlds. Moreover, let $\alpha$, $\beta$ and $\gamma$ be propositional sentences. The model $M$ satisfies the obligation '$\alpha$ should be the case if $\beta$ is the case unless $\gamma$ is the case,' written as $M \models O^{cc}(\alpha|\beta\backslash\gamma)$, iff

1. $W_1 = \{w \in W \mid M, w \models \alpha \wedge \beta \wedge \neg\gamma\}$ is nonempty, and
2. $W_2 = \{w \in W \mid M, w \models \neg\alpha \wedge \beta\}$ is nonempty, and
3. for all $w_1 \in W_1$ and $w_2 \in W_2$, we have $w_2 \not\leq w_1$.

□

The following example explains the factual defeasibility of the Chisholm paradox by preference semantics.

**Example 4.8 (Chisholm paradox, continued)** Consider the set of obligations

$$S = \{O^{cc}(a|\top\backslash\bot), O^{cc}(t|a\backslash\bot), O^{cc}(\neg t|\neg a\backslash\bot)\}$$

A typical[3] model $M$ of $S$ is given in Figure 4.6. This figure should be read as follows. The circles represent non-empty sets of worlds, that satisfy the propositions written contained in them. Each circle represents an equivalence class of the partial pre-ordering $\leq$ of the model (the ordering partitions the worlds of the model into a set of equivalence classes). The arrows represent strict preferences for all worlds in the equivalence classes. We have $M \models O^{cc}(\neg t|\neg a\backslash\bot)$, for example, because for all $w_1 \in| \neg t \wedge \neg a \mid$ and $w_2 \in| t \wedge \neg a \mid$ we have $w_2 \not\leq w_1$. The condition $\neg a$ corresponds to the semantic concept of zooming in on the ordering. In the figure, this zooming in on the ordering is represented by a dashed box. For the evaluation of $M \models O^{cc}(\neg t|\neg a\backslash\bot)$, only the ordering within the dashed box is considered. As we observed in earlier analyses of the Chisholm paradox, the most important thing is that $O(t \mid \neg a \backslash \gamma)$ does not follow from the premises for any $\gamma$. This is true for contextual deontic logic CDL. The crucial observation is that we have $M \not\models O^{cc}(t|\neg a\backslash\gamma)$ for any $\gamma$, because for all $w_1 \in|t \wedge \neg a \wedge \neg\gamma|$ for any $\gamma$, and for all $w_2 \in|\neg t \wedge \neg a|$, we have $w_2 \leq w_1$ (and even $w_2 < w_1$). □

[3]The model $M$ in Figure 4.6 is the unique most connected model of $S$, see Section 2.3.

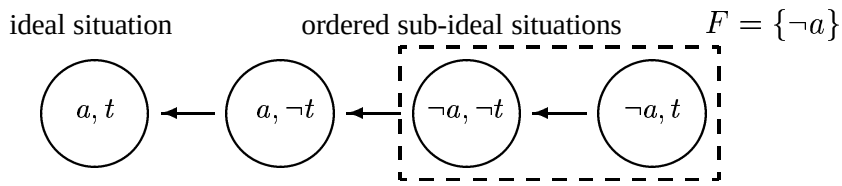



Figure 4.6: Preference relation of the Chisholm paradox

Our discussion of the Chisholm paradox showed the fundamental distinction between overridden and factual defeasibility. Contrary-to-duty reasoning can be formalized as a kind of overridden defeasibility as well as a kind of factual defeasibility, and we showed that it is best formalized by the latter. The preference-based semantics illustrates where this type of factual defeasibility comes from. Semantically, the antecedent zooms in on the context of the preference ordering. The inference pattern WC corresponds semantically to introducing exceptions of this context. In the Chisholm paradox, the derivation of $O^{cc}(t \mid \top \setminus \neg a)$ from $O^{cc}(a \wedge t \mid \top \setminus \bot)$ says that the preference for $t$ is not valid in the context $\neg a$. As shown in Figure 4.6, in this violation context the preferences can be the other way around.

## 4.3 Overridden and factual defeasibility

In this section, we focus on the cancelling aspect of overridden defeasibility and the overshadowing aspect of factual defeasibility. Overridden defeasibility becomes relevant when there is a (potential) conflict between two obligations, i.e. when there are two contradictory obligations. For example, there is a conflict between $O(\alpha_1 \mid \beta_1)$ and $O(\alpha_2 \mid \beta_2)$ when $\alpha_1$ and $\alpha_2$ are contradictory, and $\beta_1$ and $\beta_2$ are factually true. In a defeasible deontic logic, such a conflict is resolved when one of the obligations overrides the other one. In the language of dyadic deontic logic, the overriding of $O(\alpha_1 \mid \beta_1)$ by $O(\alpha_2 \mid \beta_2)$ is formalized by the non-derivability of $O(\alpha_1 \mid \beta_1 \wedge \beta_2)$. An unresolvable conflict is usually called a 'deontic dilemma,' in this case represented by the formula $O(\alpha_1 \mid \beta_1 \wedge \beta_2) \wedge O(\alpha_2 \mid \beta_1 \wedge \beta_2)$.

In particular, we analyze *violated obligations* in a deontic logic that formalizes reasoning about obligations which can be overridden by other obligations. In the language of dyadic deontic logic, an obligation with a contradictory antecedent and consequent like $O(\neg\alpha \mid \alpha)$ represents 'if $\alpha$ is the case, then it is a violation of the obligation that $\neg\alpha$ should be the case.'[4] This representation of violations is related to the more standard representation $\alpha \wedge O\neg\alpha$ as follows. The standard representation of violations is a combination of monadic obligations and factual detachment, see Section 2.5. With the inference pattern EFD discussed in the introduction we can derive the obligation $O\neg\alpha$ from $\mathcal{A}\alpha$ and $O(\neg\alpha \mid \alpha)$. Hence, $O(\neg\alpha \mid \alpha)$ can be read as 'if only $\alpha$ is known, then $O\neg\alpha$ can be derived' and $\alpha \wedge O\neg\alpha$ represents a violation. The contextual obligations we defined in Section 4.2.4 do not represent violated obligations, but in Section 4.3.4 we show how the definition of $O^{cc}(\alpha \mid \beta \setminus \gamma)$ can be adapted to $O^{r}(\alpha \mid \beta \setminus \gamma)$ to derive violated (i.e. overshadowed) contextual obligations. To keep our analysis as general as possible, in this section we only accept the inference pattern $\text{RSA}_O$. Because $\text{RSA}_O$ is the only inference pattern we

[4] Alternatively, such an obligation could represent the obligation to update the present state of affairs. For example, the obligation 'if you smoke in a no-smoking car, then you should not smoke in a no-smoking car' [Han71] can be read as the obligation to quit smoking, see Section 5.1.

assume, we do not have to formalize contrary-to-duty reasoning and its related problems which we discussed in the previous section. Thus, the analyses in this section are independent from our analysis and our solution of the Chisholm paradox.

### 4.3.1 The Fence example

The following so-called Fence example was introduced in [PS96] to illustrate the distinction between contrary-to-duty reasoning and defeasible reasoning (based on exceptional circumstances). It is an extended version of the Forrester (or gentle murderer) paradox: you should not kill, but if you kill, then you should do it gently [For84]. The following example is an alphabetic variant of the original Fence example, see Example 1.3, because we replaced $s$, to be read as 'the cottage is by the sea,' by $d$, to be read as 'there is a dog.' The distinction between 'the cottage is by the sea' and 'there is a dog' is that the latter proposition is controllable, whereas the former is not. This important distinction between controllable and uncontrollable propositions has to be formalized in a deontic (or action) logic, if only because for any uncontrollable $\alpha$ the obligation $O(\alpha \mid \top)$ does not make sense, see Section 5.2 and [Bou94b] for a discussion. For example, it does not make sense to oblige someone to make the sun rise. In this chapter we abstract from this problem and we assume that all propositions are controllable.

**Example 4.9 (Fence)** Assume a dyadic deontic logic that has at least substitution of logical equivalents and the inference pattern $\text{RSA}_O$. Furthermore, assume the premise set of obligations

$$S = \{O(\neg f \mid \top), O(w \wedge f \mid f), O(w \wedge f \mid d)\}$$

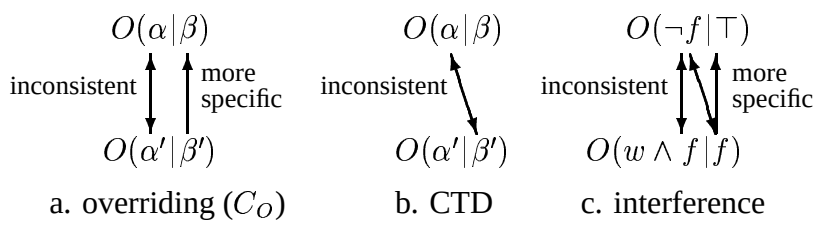
where $f$ can be read as 'there is a fence around your house,' $w \wedge f$ as 'there is a white fence around your house' and $d$ as 'you have a dog.' Notice that $O(w \wedge f \mid f)$ is a CTD obligation of $O(\neg f \mid \top)$ and $O(w \wedge f \mid d)$ is not. If there is a fence and a dog ($\mathcal{A}(f \wedge d)$), then the first premise of $S$ is intuitively overridden, and therefore it cannot be violated. Hence, the obligation $O(\neg f \mid f \wedge d)$ should *not* be derivable. However, if there is a fence without a dog ($\mathcal{A}f$), then the first premise is intuitively not overridden, and therefore it is violated. Hence, the obligation $O(\neg f \mid f)$ should be derivable. Moreover, this is exactly the difference between cancellation and overshadowing that we discussed in the introduction of this chapter. Overriding of $O(\neg f \mid \top)$ by $f \wedge d$ and $O(w \wedge f \mid d)$ means that the obligation to have no fence is cancelled and has no force anymore, hence $O(\neg f \mid f \wedge d)$ should not be derivable. Violation of $O(\neg f \mid \top)$ by $f$ means that the obligation to have no fence has still its force, it is only overshadowed and not cancelled, hence $O(\neg f \mid f)$ should be derivable. The possible derivations of $O(\neg f \mid f \wedge d)$ and $O(\neg f \mid f)$ are represented in Figure 4.7. In the first derivation, the counterintuitive obligation $O(\neg f \mid f \wedge d)$ is not derived from $O(\neg f \mid \top)$ by $\text{RSA}_O$, because the latter obligation is overridden by $O(w \wedge f \mid d)$ for $f \wedge d$. However, in the second derivation the intuitive obligation $O(\neg f \mid f)$ is not derived either from $O(\neg f \mid \top)$ by $\text{RSA}_O$, because it is overridden by $O(w \wedge f \mid f)$ for $f$, according to condition $C_O$. □

The problem in this example is that both $O(w \wedge f \mid f)$ and $O(w \wedge f \mid d)$ are treated as more specific obligations that override the obligation $O(\neg f \mid \top)$, i.e. both are treated as cases of overridden defeasibility. However, this is not correct for $O(w \wedge f \mid f)$. This last obligation should

$$\begin{array}{c} O(w \wedge f \,|\, d) \\ \downarrow \\ \dfrac{O(\neg f \,|\, \top)}{O(\neg f \,|\, f \wedge d)} \; (\mathrm{RSA}_O) \end{array} \qquad \begin{array}{c} O(w \wedge f \,|\, f) \\ \downarrow \\ \dfrac{O(\neg f \,|\, \top)}{O(\neg f \,|\, f)} \; (\mathrm{RSA}_O) \end{array}$$

Figure 4.7: Fence example with $C_O$

be treated as a CTD obligation, i.e. as a case of factual defeasibility. This interference of specificity and CTD is represented in Figure 4.8. This figure should be read as follows. Each arrow is a condition: a two-headed arrow is a consistency check, and a single-headed arrow is a logical implication. For example, the condition $C_O$ formalizes that an obligation $O(\alpha|\beta)$ is overridden by $O(\alpha'|\beta')$ if the conclusions are contradictory (a consistency check, the double-headed arrow) and the condition of the overriding obligation is more specific ($\beta'$ logically implies $\beta$). Case (a) represents criteria for overridden defeasibility, and case (b) represents criteria for CTD. Case (c) shows that the pair of obligations $O(\neg f \mid \top)$ and $O(w \wedge f \mid f)$ can be viewed as overridden defeasibility as well as CTD.




a. overriding ($C_O$) b. CTD c. interference

Figure 4.8: Specificity and CTD

What is most striking about the Fence example is the observation that when the premise $O(\neg f \mid \top)$ is violated by $f$, then the obligation for $\neg f$ should be derivable, but not when the premise $O(\neg f \mid \top)$ is overridden by $f \wedge d$. This means that the CTD or overriding interpretations of $O(\neg f \mid \top)$ are quite different in the sense that they have different consequences. This overriding can be viewed as a type of overridden defeasibility and the violation in the CTD as a type of factual defeasibility. Hence, also the Fence example shows that factual and overridden defeasibility lead to different conclusions. This is a kind of factual defeasibility which differs from its counterpart in default logic in the sense that it is overshadowing factual defeasibility rather than cancelling factual defeasibility.

### 4.3.2 Overridden defeasibility

One obvious analysis of the problem mentioned in Example 4.9 is to observe that condition $C_O$ is too strong. Consider the following ad hoc solution of the problem by weakening the definition of specificity in $C_O$ to $C_O^*$ with an additional condition which represents that a CTD obligation cannot override its primary obligations. The specificity condition $C_O^*$ has three conditions: the two conditions of $C_O$ and the additional condition that the overriding obligation $O(\alpha'|\beta')$ is not a CTD of $O(\alpha \mid \beta)$, i.e. $\beta' \wedge \alpha$ must be consistent. Due to this extra condition the overriding

interpretation in case (c) in Figure 4.8 is no longer valid. The following example shows that the definition of specificity $C_O^*$ gives the intuitive conclusions and avoids the counterintuitive ones.

**Example 4.10 (Fence, continued)** Assume that $\text{RSA}_O$ is replaced by the following $\text{RSA}_O^*$.

$$\text{RSA}_O^* : \frac{O(\alpha|\beta_1), C_O^*}{O(\alpha|\beta_1 \wedge \beta_2)}$$

$C_O^*$: there is no premise $O(\alpha'|\beta')$ such that $\beta_1 \wedge \beta_2$ logically implies $\beta'$, $\beta'$ logically implies $\beta_1$ and not vice versa, $\alpha$ and $\alpha'$ are contradictory and $\alpha \wedge \beta'$ is consistent.

The derivations from $S$ with $\text{RSA}_O^*$ are represented in Figure 4.9. $\text{RSA}_O^*$ does not derive the counterintuitive $O(\neg f \mid f \wedge d)$, just like $\text{RSA}_O$ in Figure 4.7. However, $\text{RSA}_O^*$ does derive the intuitive $O(\neg f \mid f)$ from $O(\neg f \mid \top)$, in contrast to $\text{RSA}_O$. $\text{RSA}_O^*$ solves the problem of Example 4.9, because it does not derive the counterintuitive obligation, but it does derive the intuitive obligation. □

$$\begin{array}{c} O(w \wedge f|d) \\ \downarrow \\ \dfrac{O(\neg f|\top)}{O(\neg f|f \wedge d)} \; (\text{RSA}_O^*) \end{array} \qquad \frac{O(\neg f|\top)}{O(\neg f|f)} \; \text{RSA}_O^*$$

Figure 4.9: Fence example with $C_O^*$

This solution of the Fence example is ad hoc, because there is no *a priori* reason to prefer $C_O^*$ and $\text{RSA}_O^*$ (the violability interpretation) to $C_O$ and $\text{RSA}_O$ (the overridden interpretation). Informally, the reason to prefer the former inference pattern is that with $\text{RSA}_O$, the obligation $O(\neg f|\top)$ can never be violated, which is a highly counterintuitive property of an obligation. In the following subsection, we give a formal analysis of the Fence example, based on the essential property of obligations that they can be violated.

### 4.3.3 Factual defeasibility

Instead of analyzing the problem of Example 4.9 by examining specificity condition $C_O$ (overridden defeasibility), we can also look at properties of violability (factual defeasibility). The following inference patterns *Contrary-to-Duty* (CD) and *According-to-Duty* (AD) formalize the intuitions that an obligation cannot be defeated by only violating or fulfilling it. The CD rule models the intuition that after violation the obligation to do $\alpha$ is still in force (i.e. overshadowing). Even if you drive too fast, you are still obliged to obey the speed limit.[5]

$$\text{CD} : \frac{O(\alpha|\beta)}{O(\alpha|\beta \wedge \neg\alpha)} \qquad \text{AD} : \frac{O(\alpha|\beta)}{O(\alpha|\beta \wedge \alpha)}$$

[5]The inference patterns CD and AD should not be confused with the following inverses of CD and AD, which seem to say that violations or fulfilled obligations do not come out of the blue.

$$\text{CD}^- : \frac{O(\alpha|\beta \wedge \neg\alpha)}{O(\alpha|\beta)} \qquad \text{AD}^- : \frac{O(\alpha|\beta \wedge \alpha)}{O(\alpha|\beta)}$$

We reconsider the Fence example and we show that CD with $\text{RSA}_O$ derives exactly the intuitive conclusions, just like $\text{RSA}_O^*$.

**Example 4.11 (Fence, continued)** Assume the inference patterns $\text{RSA}_O$ and CD. Figure 4.10 represents the same two situations as Figure 4.7. First consider the situation when there is a fence and a dog ($f \wedge d$). The counterintuitive $O(\neg f \,|\, f \wedge d)$ cannot be derived, because the derivation via $O(\neg f \,|\, d)$ from $O(\neg f \,|\, \top)$ is blocked by $C_O$. Now consider the situation when there is a fence but not a dog ($f$). The intuitive obligation $O(\neg f \,|\, f)$ can be derived from $O(\neg f \,|\, \top)$ by CD. □

$$\begin{array}{c} O(w \wedge f \,|\, d) \\ O(\neg f \,|\, \top) \qquad \downarrow \\ \text{- - - - - - - - - } (\text{RSA}_O) \\ \dfrac{O(\neg f \,|\, d)}{O(\neg f \,|\, f \wedge d)}\ \text{CD} \end{array} \qquad\qquad \dfrac{O(\neg f \,|\, \top)}{O(\neg f \,|\, f)}\ \text{CD}$$

Figure 4.10: Fence example with CD

Example 4.10 and 4.11 illustrate that the problem of $\text{RSA}_O$ is that it does not imply CD (because its specificity condition $C_O$ is too strong). In other words, the problem of $\text{RSA}_O$ is that there can be obligations, like $O(\neg f \mid \top)$, that can never be violated. In Example 4.11, CD and $\text{RSA}_O$ yield exactly the same intuitive conclusions as $\text{RSA}_O^*$ in Example 4.10. An advantage of CD is that the inference pattern is very intuitive and not an ad hoc like solution of the problem like the adaptation of $C_O$. Moreover, AD also formalizes an intuitive notion of fulfilled obligations, because it deals with fulfilled obligations in exactly the same way as CD with violated obligations. We illustrate the applicability of our approach by the analysis of the following Reykjavic Scenario, introduced by Belzer [Bel86].

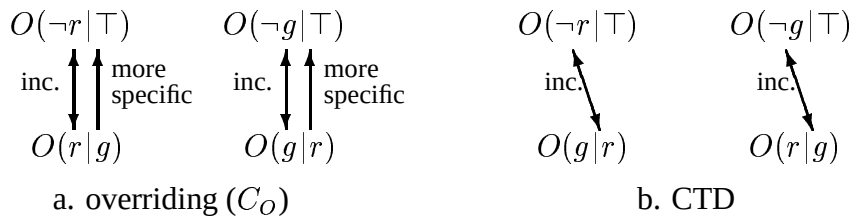
**Example 4.12 (Reykjavic Scenario)** Consider the premise set of dyadic obligations

$$S = \{O(\neg r \,|\, \top), O(\neg g \,|\, \top), O(r \,|\, g), O(g \,|\, r)\}$$

where $r$ can be read as 'the agent tells the secret to Reagan' and $g$ as 'the agent tells the secret to Gorbatsjov.' Figure 4.11 illustrates that the Reykjavic Scenario is a more complex instance of the Fence example, illustrated in Figure 4.8. In the Fence example, $O(w \wedge f \,|\, f)$ can be interpreted as a more specific overriding obligation, and it can be interpreted as a CTD obligation. In the Reykjavic Scenario, the latter two obligations of $S$ can be considered as more specific obligations overriding the former two, and they can also be considered as CTD obligations. □

---

Although these inference patterns seem intuitive at first sight, they are highly counterintuitive on further inspection. Reconsider the Fence example. There should be a white fence, if there is a fence $O(w \wedge f|f)$. Hence, by AD, there ought to be a white fence, if there is a white fence $O(w \wedge f|w \wedge f)$ (a fulfilled obligation). However, this does not mean that there is a unconditional obligation that there ought to be a white fence $O(w \wedge f|\top)$. Hence, the inference pattern $\text{AD}^-$ is not valid. A similar argument can be given for $\text{CD}^-$.




a. overriding ($C_O$) b. CTD

Figure 4.11: Specificity and CTD in the Reykjavic Scenario

The Reykjavic Scenario is a highly ambiguous paradox, as a result of the fact that the latter two obligations can be considered as overriding as well as CTD obligations. Consider the following two interpretations of this paradox.

1. **Overridden interpretation.** In this interpretation, the third sentence of $S$ is an exception to the first sentence, and the fourth sentence is an exception to the second sentence (see Figure 4.11.a). The agent's primary obligation is not to tell Reagan or Gorbatsjov. When he tells Reagan, he should not tell Reagan but he should tell Gorbatsjov. It is a case of overridden defeasibility, because $O(\neg g \,|\, r)$ cannot be derived from $O(\neg g \,|\, \top)$ due to the premise $Og \,|\, r)$. When he tells both, he does not violate any obligations because $r$ and $g$ are considered as exceptions.[6]

2. **Violability interpretation.** In this interpretation the two obligations $O(\neg r|r \wedge g)$ and $O(\neg g|r \wedge g)$ are both derivable from $S$. Hence, when the agent tells both, he should have told neither of them, $O(\neg r|r \wedge g)$ and $O(\neg g|r \wedge g)$, a case of violability. The third sentence of $S$ is a CTD obligation of the second sentence and the fourth sentence is a CTD obligation of the first sentence (see Figure 4.11.b).

In our view the violability interpretation is to be preferred to the overridden interpretation, The following example illustrates that the overridden interpretation conflicts with CD.

**Example 4.13 (Reykjavic Scenario, continued)** Assume a dyadic deontic logic that has at least substitution of logical equivalents and the inference patterns AND, $\text{RSA}_O$, CD and the following disjunction rule OR.

$$\text{OR:}\frac{O(\alpha_1|\beta), O(\alpha_2|\beta)}{O(\alpha_1 \vee \alpha_2|\beta)}$$

Moreover, assume the set of obligations $S$ of Example 4.12. According to the defeasibility interpretation, there is no violation when the agent tells both Reagan and Gorbatsjov. We cannot use $\text{RSA}_O$ to derive a violation for $r \wedge g$ from $S$, because the premises are overridden as represented in Figure 4.11.b. However, we can use CD to derive the violation $O(\neg r \vee \neg g \,|\, r \wedge g)$, as represented in Figure 4.12. Hence, if we accept CD then we have to reject the defeasibility interpretation. Since we gave a general motivation for CD that is independent from particular examples, we reject the overridden interpretation. □

[6]According to the overridden interpretation, it might be argued that the paradox is not modeled correctly by the set of obligations $S$. When the last two conditional obligations should be interpreted as CTD obligations when the agent tells both, the first two obligations should be represented by one conditional obligation $O(\neg r \wedge \neg g|\top)$. In that case, the last two sentences are interpreted as CTD obligations by $C_O^*$.

$$\dfrac{\dfrac{O(\neg r\,|\,\top) \quad O(\neg g\,|\,\top)}{O(\neg r \vee \neg g\,|\,\top)}\ \text{OR}}{O(\neg r \vee \neg g\,|\,r \wedge g)}\ \text{CD} \qquad \dfrac{\dfrac{O(\neg r\,|\,\top) \quad O(\neg g\,|\,\top)}{O(\neg r \wedge \neg g\,|\,\top)}\ \text{AND}}{O(\neg r \wedge \neg g\,|\,r \vee g)}\ \text{CD}$$

Figure 4.12: Reykjavic Scenario with CD

The examples show that the inference patterns CD and AD are adequate tools to analyze conflicts between overridden and contrary-to-duty interpretations. However, they cannot discriminate between the following two violability interpretations of the Reykjavic Scenario. McCarty [McC94a] argues for the first violability interpretation.

1. **Violability-1 interpretation.** When he tells only Reagan, then one could interpret this as an overridden case, i.e. a case of defeasibility. $O(\neg g\,|\,|\top)$ is in this interpretation overridden by $O(g\,|\,r)$ and the fact $r$. Hence, in this interpretation $O(\neg g\,|\,r)$ is not derivable from the premises. The remarkable thing about this interpretation is that $r \wedge g$ is treated as a violability case, whereas $r$ in isolation is treated as an overridden case.

2. **Violability-2 interpretation.** If we accept the reasonable principle that if an obligation is overridden for some situation, that it is then also overridden for a more specific situation, then the obligation $O(\neg g\,|\top)$ cannot be overridden by r only, because it is in the violability interpretation not overridden by the more specific situation $r \wedge g$.[7] According to this interpretation, when the agent tells only Reagan, then he still has the obligation to tell Gorbatsjov $O(g\,|\,r)$, but also he has the derivable obligation not to tell Gorbatsjov $O(\neg g\,|\,r)$. The remarkable thing about this interpretation is that if we accept a reasonable principle, then the Reykjavic Scenario becomes a deontic dilemma.

This again illustrates the fact that this scenario is highly ambiguous, and additional principles have to be accepted if we want to decide between these two interpretations.

### 4.3.4 Preferential semantics: CD and AD

Before we can examine the conflicts between specificity and contrary-to-duty in the semantics, there are two ways in which we have to adapt the definition of contextual obligations. First, in this section we adapt the definition of $O^{cc}(\alpha\,|\,\beta\setminus\gamma)$ to $O^{r}(\alpha\,|\,\beta\setminus\gamma)$. The logic of $O^{r}(\alpha\,|\,\beta\setminus\gamma)$ represents fulfilled and violated obligations, because it has CD and AD. Second, we have to introduce a semantic notion to model specificity, which is done in Section 4.3.5 when we introduce obligations $O^{re}(\alpha|\beta\setminus\gamma)$.

The contextual obligations $O^{cc}(\alpha\,|\,\beta\setminus\gamma)$ do not represent violated and fulfilled obligations, because the first two conditions of Definition 4.7 say that $O^{cc}(\alpha|\beta\setminus\gamma)$ is false if either $\alpha\wedge\beta\wedge\neg\gamma$ or $\neg\alpha \wedge \beta$ is inconsistent. Obviously, we have to relax these two conditions. We allow the set of worlds $W_1'$ and $W_2'$ of $O^{re}(\alpha\,|\,\beta\setminus\gamma)$ to be supersets of $W_1$ and $W_2$ from $O^{cc}(\alpha\,|\,\beta\setminus\gamma)$. If $W_1$ and $W_2$ of Definition 4.7 are nonempty, then the definition of $O$ is equivalent to the definition

[7] This principle certainly holds for defeasible logics. For example, if the 'birds fly' default is overridden by the more specific 'penguins do not fly default, then this latter default also holds for the subset super-penguins of penguins, unless it is explicitly stated that by default 'super-penguins do fly'.

of $O^r$. However, if the set $W_1$ or $W_2$ is empty, then we have $M \not\models O^{cc}(\alpha \mid \beta \setminus \gamma)$, whereas $M \models O^r(\alpha \mid \beta \setminus \gamma)$ if there is any $M \models O^{cc}(\alpha \mid \beta' \setminus \gamma)$ where $\beta$ logically implies $\beta'$ (see Proposition 4.15 and 4.17).

**Definition 4.14 (Contextual obligation, with violations)** Let $M = \langle W, \leq, V \rangle$ be a Kripke model that consists of $W$, a set of worlds, $\leq$, a binary reflexive and transitive relation on $W$, and $V$, a valuation of the propositions in the worlds. The model $M$ satisfies the obligation '$\alpha$ should be the case if $\beta$ is the case unless $\gamma$ is the case,' written as $M \models O^r(\alpha|\beta\setminus\gamma)$, iff

1. there is a nonempty $W_1 \subset W$ such that
   - for all $w \in W_1$, we have $M, w \models \alpha \wedge \neg\gamma$, and
   - for all $w$ such that $M, w \models \alpha \wedge \beta \wedge \neg\gamma$, we have $w \in W_1$, and
2. there is a nonempty $W_2 \subset W$ such that
   - for all $w \in W_2$, we have $M, w \models \neg\alpha$, and
   - for all $w$ such that $M, w \models \neg\alpha \wedge \beta$, we have $w \in W_2$, and
3. for all $w_1 \in W_1$ and $w_2 \in W_2$, we have $w_2 \not\leq w_1$.

□

The distinction between contextual obligations $O^{cc}(\alpha \mid \beta \setminus \gamma)$ and $O^r(\alpha \mid \beta \setminus \gamma)$ is illustrated by the following metaphor, based on an analogy with belief revision. Let the two sets of worlds $W_1 = \{w \mid M, w \models \alpha \wedge \beta \wedge \neg\gamma\}$ and $W_2 = \{w \mid M, w \models \neg\alpha \wedge \beta\}$ be the choice alternatives of $O(\alpha|\beta\setminus\gamma)$. Definition 4.7 in Section 4.2.4 says that $W_1$ and $W_2$ are non-empty, and $w_2 \not\leq w_1$ for every $w_1 \in W_1$ and $w_2 \in W_2$. Thus, we evaluated $O(\alpha|\beta)$ by a choice between $\alpha\wedge\beta$ and $\neg\alpha\wedge\beta$, which can be considered as the AGM expansions of $\beta$ by $\alpha$ and $\neg\alpha$. Now, we evaluate $O^r(\alpha|\beta)$ by a choice between the AGM-style revisions of $\beta$ by $\alpha$ or $\neg\alpha$, which explains our notation $O^r$. Condition (1) and (2) formalize that revision must be possible. The following proposition shows that contextual obligations have strengthening of the antecedent. Hence, the logic also has CD and AD, because CD and AD follow from SA.

**Proposition 4.15** *The logic has unrestricted strengthening of the antecedent.*

$$\text{SA}: \frac{O^r(\alpha|\beta_1\setminus\gamma)}{O^r(\alpha|\beta_1 \wedge \beta_2\setminus\gamma)}$$

**Proof** *Assume $M \models O^r(\alpha \mid \beta_1 \setminus \gamma)$. There are $W_1$ and $W_2$ such that the conditions of Definition 4.14 are fulfilled. The same $W_1$ and $W_2$ also fulfill the conditions for $M \models O^r(\alpha|\beta_1\wedge\beta_2\setminus\gamma)$.*

The following example illustrates the consequences of unrestricted strengthening of the antecedent of contextual obligations.

**Example 4.16 (Chisholm paradox, continued)** We can derive $O^r(t|\neg a\setminus\neg a)$ from $O^r(t|a\setminus\bot)$ and $O^r(a \mid \top \setminus \bot)$ in the Chisholm paradox (see Figure 4.5). There are two ways to view this

derived obligation. The first is to say it is meaningless, because the antecedent $\neg a$ implies the unless clause $\neg a$. The second way is to say that it is counterintuitive, because it looks like the counterintuitive *dyadic* obligation $O(t \mid \neg a)$. We can add a fourth condition to Definition 4.14 if we consider SA too strong, which states that there are $\beta \wedge \neg\gamma$ worlds. In that case, there is a condition $C_V$ on SA and $O^r(t|\neg a \backslash \neg a)$ is not derivable from the Chisholm paradox. □

The following proposition shows the relation between expansion-based contextual obligations $O^{cc}(\alpha \mid \beta \backslash \gamma)$ (Definition 4.7) and the revision-based contextual obligations $O^r(\alpha \mid \beta \backslash \gamma)$ (Definition 4.14).

**Proposition 4.17** *The logic has the following inference pattern.*

$$\frac{O^{cc}(\alpha|\beta\backslash\gamma)}{O^r(\alpha|\beta\backslash\gamma)}$$

**Proof** *Let $M$ be a model such that $M \models O^{cc}(\alpha \mid \beta \backslash \gamma)$, and $W_1$ and $W_2$ the two sets of worlds such that $W_1 = \{w \mid M, w \models \alpha \wedge \beta \wedge \neg\gamma\}$ and $W_2 = \{w \mid M, w \models \neg\alpha \wedge \beta\}$ of Definition 4.7. Then $M \models O^r(\alpha|\beta\backslash\gamma)$, because $W_1$ and $W_2$ fulfill the conditions of Definition 4.14.*

### 4.3.5 Multi preference semantics

In this section we adapt the definition of contextual obligations to model specificity, i.e. overridden defeasibility. Overridden defeasibility can be formalized by introducing a normality ordering in the semantics. Hence, the logic has a multi preference semantics: an *ideality ordering* ($\leq_I$) to model contrary-to-duty structures (factual defeasibility) and a *normality ordering* ($\leq_N$) to model exceptional circumstances (overridden defeasibility). To facilitate the comparison with the definitions of $O(\alpha \mid \beta \backslash \gamma)$ and $O^r(\alpha \mid \beta \backslash \gamma)$, we assume that the preferential orderings are bounded.[8]

**Definition 4.18 (Contextual obligation, with violations and overriding)** Let the model $M = \langle W, \leq_I, \leq_N, V\rangle$ be a Kripke model that consists of $W$, a set of worlds, $\leq_I$ and $\leq_N$, two binary reflexive and transitive relations on $W$, and $V$, a valuation of the propositions in the worlds, such that there are no infinite descending chains. We write $M, w \models_{\leq_N} \alpha$ when $w$ is a $\leq_N$-minimal $\alpha$ world. The model $M$ satisfies the obligation '$\alpha$ should be the case if $\beta$ is the case unless $\gamma$ is the case,' written as $M \models O^{re}(\alpha|\beta\backslash\gamma)$, iff

1. there is a nonempty $W_1 \subset W$ such that
   - for all $w \in W_1$, we have $M, w \models \alpha \wedge \neg\gamma$, and
   - for all $w$ such that $M, w \models_{\leq_N} \alpha \wedge \beta \wedge \neg\gamma$, we have $w \in W_1$, and
2. there is a nonempty $W_2 \subset W$ such that

[8]The fact that $\leq_N$ is bounded, ensures that the set of $w$ such that $w \in W_1$ and $M, w \models_{\leq_N} \alpha \wedge \beta \wedge \neg\gamma$ is well-defined. The more general definition for unbounded orderings is: for all $w$ such that $M, w \models \alpha \wedge \beta \wedge \neg\gamma$, there is a world $w' \leq_N w$ such that $M, w' \models \alpha \wedge \beta \wedge \neg\gamma$ and for all $w''$ such that $M, w'' \models \alpha \wedge \beta \wedge \neg\gamma$ and $w'' \leq_N w'$, we have $w'' \in W_1$.

- for all $w \in W_2$, we have $M, w \models \neg\alpha$, and
- for all $w$ such that $M, w \models_{\leq_N} \neg\alpha \wedge \beta$, we have $w \in W_2$, and

3. for all $w_1 \in W_1$ and $w_2 \in W_2$, we have $w_2 \not\leq_I w_1$.

□

The following example illustrates the multi preference semantics of the Fence example.

**Example 4.19 (Fence, continued)** Consider the set of contextual obligations

$$S = \{O^{re}(\neg f \,|\, \top \backslash \bot), O^{re}(w \wedge f \,|\, d \backslash \bot)\}$$

The typical[9] multi preference model of $S$ is given in Figure 4.13 and can be read as follows. The circles denote equivalence classes of worlds that satisfy the literals inside the circles and the 'horizontal' arrows denote the deontic preference ordering. The boxes denote equivalence classes in the normality ordering and the 'vertical' arrow the normality preference ordering. The set $S$ constructs two preference orderings on the worlds: one ordering for ideality (like before) and one for normality. The idea of the preference ordering on normality is that the worlds with exceptional circumstances (where you have a dog) are semantically separated from the normal situation (where you do not have a dog). The upper box represents the 'normal' worlds, which is determined by the fact that $d$ is false, i.e. you do not have a dog. Deontically, the $\neg d$ worlds are ordered according to the obligation that, usually, there should be no fence. The lower box contains the worlds where $d$ is true and which are therefore exceptional. These worlds are deontically ordered by the obligation that in this situation, there should be a white fence. Because of the exceptional circumstances, the worlds are not subject to the obligation that usually, there should not be a fence. In the ideality ordering, the normal $\neg d \wedge \neg f$ worlds and the exceptional $d \wedge w \wedge f$ worlds are equivalent.

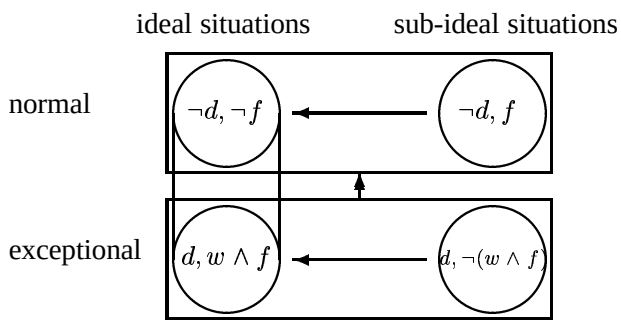



Figure 4.13: Multi-preference relation of the Fence example

For example, we have $M \models O^{re}(\neg f \,|\, \top \backslash \bot)$, because for all $w_1 \in |\neg f \wedge \neg d|$ (the most normal $\neg f$ worlds) and for all $w_2 \in |f \wedge \neg d|$ (the most normal $f$ worlds) we have $w_2 \not\leq_I w_1$. Moreover, we have $M \models O^{re}(w \wedge f \,|\, d \backslash \bot)$, because we zoom in on the $d$ worlds, and $w \wedge f \wedge d$ worlds are preferred over $\neg(w \wedge f) \wedge d$ worlds. □

[9]Computing these typical models in general is difficult, see [TvdT95]. For example, it seems more difficult than defeasible reasoning schemes to complete a single ordering like maximally connected or System Z, see Section 2.3.

Notice that we first minimize in the normality ordering when we evaluate the obligation $O^{re}(\neg f \mid \top \setminus \bot)$ in Example 4.19, because we first determine the sets $W_1 = \mid \neg f \wedge \neg d \mid$ and $W_2 = \mid f \wedge \neg d \mid$, and subsequently we compare the sets $W_1$ and $W_2$ in the ideality ordering. We compare the best most normal worlds and we do not compare the most normal best sets $W_1' = \mid \neg f \wedge \neg d \mid$ and $W_2' = \mid w \wedge f \wedge d \mid$. This is based on the heuristic rule that if an option (like $f$) can be a violation (like $W_2$) or an exception (like $W_2'$), then it is assumed to be a violation. The motivation of this rule is that a criminal should have as little opportunities as possible to excuse herself by claiming that her behavior was exceptional rather than criminal. If an agent has a fence, then it is assumed to be a violation and she cannot excuse herself by claiming that it is an exceptional case (unless, of course, there is a dog). The following proposition shows that the obligations have CD and AD.

**Proposition 4.20** *The logic of the obligations $O^{re}$ does not have* SA*, but it has* CD *and* AD*.*

**Proof** *First, consider the invalidity of* SA*. The contextual obligation $O^{re}(\alpha|\beta_1 \wedge \beta_2 \setminus \bot)$ cannot be derived from $O^{re}(\alpha \mid \beta_1 \setminus \bot)$, because the most normal worlds $\beta_1 \wedge \beta_2$ can contain worlds not among the most normal $\beta_1$ worlds. Thus the logic does not have* SA*. Secondly, consider* CD *and* AD*. Assume $M \models O^{re}(\alpha \mid \beta \setminus \gamma)$. Hence, there are sets $W_1$ and $W_2$ such that the conditions of Definition 4.18 are fulfilled. The same sets $W_1$ and $W_2$ also satisfy the conditions for $M \models O^{re}(\alpha|\beta \wedge \neg\alpha \setminus \gamma)$ and $M \models O^{re}(\alpha|\beta \wedge \alpha \setminus \gamma)$.*

The following example illustrates the conflict between overridden and CTD.

**Example 4.21 (Fence, continued)** Consider the set of contextual obligations

$$S' = \{O^{re}(\neg f|\top \setminus \bot), O^{re}(w \wedge f|d \setminus \bot), O^{re}(w \wedge f|f \setminus \bot)\}$$

The typical multi preference model $M'$ of $S'$ is given in Figure 4.14. The normal worlds have deontically been specified more precisely, compared to the model $M$ in Figure 4.13 of the set of obligations $S$ in Example 4.19. We have $M' \models O^{re}(\neg f|\top \setminus \bot)$, for similar reasons as given in Example 4.19 for $M \models O^{re}(\neg f \mid \top \setminus \bot)$. We also have $M' \models O^{re}(\neg f \mid f \setminus \bot)$, which can be shown as follows. Semantically, the sets $W_1$ and $W_2$ must contain the most normal $\neg f \wedge f$ and $f \wedge f$ worlds, respectively. Hence, $W_1$ can be any subset of $|\neg f|$, and $W_2$ is a subset of $|f|$ that contains at least $|f \wedge \neg d|$. We can choose $W_1$ and $W_2$ as $|\neg f \wedge \neg d|$ and $|f \wedge \neg d|$, and we have $w_2 \not\leq w_1$ for all $w_1 \in W_1$ and $w_2 \in W_2$. However, we do not have $M' \models O^{re}(\neg f \mid f \wedge d \setminus \bot)$, as can be verified as follows. The sets $W_1$ and $W_2$ must contain the most normal $\neg f \wedge f \wedge d$ and $f \wedge f \wedge d$ worlds, respectively. Hence, $W_1$ can be any subset of $|\neg f|$, and $W_2$ is a subset of $|f|$ that contains at least $|f \wedge d|$. Any world $w_2 \in \mid w \wedge f \wedge d \mid$ is deontically preferred, hence there cannot be a world $w_1 \in W_1$ such that $w_2 \not\leq w_1$, thus the first condition cannot be fulfilled. This illustrates that the logic does not have SA, because it does not strengthen $O^{re}(\neg f \mid \top \setminus \bot)$ to $O^{re}(\neg f \mid f \wedge d \setminus \bot)$ (although it does strengthen to $O^{re}(\neg f \mid f \setminus \bot)$). These are precisely the intuitive conclusions that one would draw from $S'$. If one only knows that there is a fence, then one concludes that the first obligation from $S'$ still holds, hence one derives $O^{re}(\neg f|f \setminus \bot)$. However, if one knows that there is a dog as well as a fence, then the first obligation is overridden by the second one, and hence one does not derive $O(\neg f|f \wedge d \setminus \bot)$. □

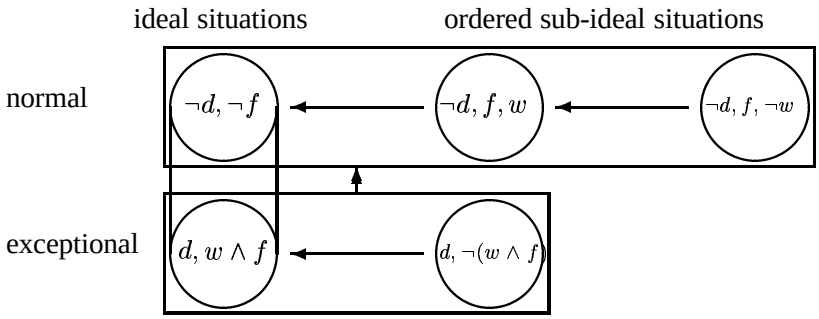



Figure 4.14: Extended multi-preference relation of the Fence example

In this section, we focussed on the cancelling aspect of overridden defeasibility and the overshadowing aspect of factual defeasibility. We argued that the distinction should be reflected by two distinct preference orderings in the semantics: one normality ordering for the cancelling aspect of overridden defeasibility, and one ideality ordering for the overshadowing aspect of factual defeasibility. This is a major distinction between defeasible deontic logics and logics of defeasible reasoning, because in the latter both kinds of defeasibility are cancelling, and they can be modeled by a single preference ordering (see e.g. [Mak93, GP92, Bou94a]).

## 4.4 Strong versus weak overridden defeasibility

In this section we focus on the cancelling aspect and the overriding aspect of overridden defeasibility by formalizing prima facie obligations. First, we show that the overridden defeasibility related to multi preference semantics cannot be used for prima facie obligations. Secondly, we introduce a new kind of preference semantics, based on priorities, to model prima facie obligations.

We call the overridden defeasibility related to multi-preference semantics strong overridden defeasibility, and the overridden defeasibility based on priorities weak overridden defeasibility. The distinction between the different types of overridden defeasibility is shown by three inference patterns which are not valid for the first type, but which are valid for the second type: forbidden conflict and two versions of reinstatement. One of the inferential differences between weak and strong overridden defeasibility is the inference pattern

$$\frac{O(\neg f \mid \top), O(w \wedge f \mid d)}{O(\neg d \mid \top)}$$

which is not valid in strong overridden defeasibility, whereas

$$\frac{O_{pf}(k \mid \top), O_{pf}(p \wedge \neg k \mid d)}{O_{pf}(\neg d \mid \top)}$$

is valid in weak overridden defeasibility. This might look strange, because the premises in both inference schemes have the same syntactic form (obviously the substitution of $\neg k$ for $f$ does not make any difference). However, it simply means that the $O$ that represents obligations like 'there should be no fence' is different from the $O_{pf}$ that represents prima facie obligations.

### 4.4.1 Prima facie obligations

Ross [Ros30] introduced the notion of so-called prima facie obligations. In his own words: 'I suggest '*prima facie* duty' or 'conditional duty' as a brief way of referring to the characteristic (quite distinct from that of being a duty proper) which an act has, in virtue of being of a certain kind (e.g. the keeping of a promise), of being an act which would be a duty proper if it were not at the same time of another kind which is morally significant' [Ros30, p.19]. A prima facie duty is a duty proper when it is not overridden by another prima facie duty. When a prima facie obligation is overridden, it is not a proper duty but it is still in force: 'When we think ourselves justified in breaking, and indeed morally obliged to break, a promise [ . . . ] we do not for the moment cease to recognize a prima facie duty to keep our promise' [Ros30, p.28]. See [Mor96] for a formalization of Ross' theory in a deontic logic. The following example describes the typical kind of defeasibility involved in reasoning about prima facie obligations.

**Example 4.22 (Promises)** Assume the inference pattern $\text{RSA}_O$ and the set of premises

$$S = \{O_{pf}(k|\top), O_{pf}(p \wedge \neg k|d)\}$$

where $k$ can be read as 'keeping a promise,' $p$ as 'preventing a disaster' and $d$ as 'a disaster will occur if nothing is done to prevent it.' There is a potential conflict between the two obligations, because when the facts imply $d$ then the first obligation says that you should keep your promise and the second one implies that you should not. Assuming that the second obligation is stronger than the first one, the first obligation is overridden by the second one. Hence, the inference

$$\frac{O_{pf}(k \mid \top), O_{pf}(p \wedge \neg k \mid d)}{O_{pf}(k \mid d)}$$

is *not* valid. Important here is that this priority does not depend on specificity. In this example the priority is compatible with specificity, but the converse priority could also have been chosen. You do not have an absolute (alias proper) obligation to keep your promise, but you still have the prima facie obligation. The situation is not ideal anymore. All situations where $k$ is false, i.e. where the prima facie obligation for $k$ is violated, are sub-ideal. This can be verified as follows. Consider a person having the obligation to keep a promise to show up at a birthday party, but she does not want to. So, she does something which might result in a disaster later on (leaving the coffee machine on, for instance) and at the moment of the party, she rushes home to turn off the coffee machine. She has the actual obligation to go home and turn off the machine, but leaving the machine on (on purpose) was a violation already. Hence, the inference

$$\frac{O_{pf}(k \mid \top), O_{pf}(p \wedge \neg k \mid d)}{O_{pf}(\neg d \mid \top)}$$

is valid. It says that it is not permitted to do something that might result in a disaster (remember that all propositions are assumed to be controllable). Finally, assume that there may be a disaster but you do not prevent it. Hence, the second obligation has been violated. In this situation, the proper obligation is not fulfilled, but we can still fulfill the prima facie obligation. Violating one obligation is better than violating both. Hence, the inference

$$\frac{O_{pf}(k \mid \top), O_{pf}(p \wedge \neg k \mid d)}{O_{pf}(k \mid d \wedge \neg p)}$$

is valid. □

The following inference pattern is called *Forbidden Conflict* (FC). If the inference pattern is accepted, then it is not allowed to establish a conflict, because a conflict is sub-ideal, even when it can be resolved.

$$\text{FC}: \frac{O(\alpha_1|\beta_1), O(\neg\alpha_1 \wedge \alpha_2|\beta_1 \wedge \beta_2)}{O(\neg\beta_2|\beta_1)}$$

The situation considered in the following inference pattern *Reinstatement* (RI) is whether an obligation can be overridden by an overriding obligation that itself is factually defeated. The obligation $O(\alpha_1|\beta_1)$ is overridden by $O(\neg\alpha_1 \wedge \alpha_2|\beta_1 \wedge \beta_2)$ for $\beta_1 \wedge \beta_2$, but is it also overridden for $\beta_1 \wedge \beta_2 \wedge \neg\alpha_2$? If the inference pattern is accepted, then the first obligation is in force again. Hence, the derivation of the obligation for $\alpha_1$ says that the original obligation is reinstated.

$$\text{RI}: \frac{O(\alpha_1|\beta_1), O(\neg\alpha_1 \wedge \alpha_2|\beta_1 \wedge \beta_2)}{O(\alpha_1|\beta_1 \wedge \beta_2 \wedge \neg\alpha_2)}$$

The following inference pattern RIO is a variant of the previous inference pattern RI, in which the overriding obligation is not factually defeated but overridden. The obligation $O(\alpha_1|\beta_1)$ is overridden by $O(\neg\alpha_1 \wedge \alpha_2|\beta_1 \wedge \beta_2)$ for $\beta_1 \wedge \beta_2$, and the latter is overridden by $O(\neg\alpha_2|\beta_1 \wedge \beta_2 \wedge \beta_3)$ for $\beta_1 \wedge \beta_2 \wedge \beta_3$. The inference pattern RIO says that an obligation cannot be overridden by an obligation that is itself overridden. Hence, an overridden obligation becomes reinstated when its overriding obligation is itself overridden.

$$\text{RIO}: \frac{O(\alpha_1|\beta_1), O(\neg\alpha_1 \wedge \alpha_2|\beta_1 \wedge \beta_2), O(\neg\alpha_2|\beta_1 \wedge \beta_2 \wedge \beta_3)}{O(\alpha_1|\beta_1 \wedge \beta_2 \wedge \beta_3)}$$

Example 4.22 illustrates that the kind of overridden defeasibility related to Ross' notion of 'prima facie' obligations have the inference patterns FC, RI and RIO. In the next section, we show that the type of overridden defeasibility we used to model specificity in the Fence example does not have the inference patterns. Hence, there are two different types of overridden defeasibility; we call the kind related to prima facie obligations *weak overridden defeasibility* in contrast to *strong overridden defeasibility*. In Section 4.4.3, we illustrate this new type of defeasibility by a preference ordering with priorities, instead of the multi preference semantics of strong overridden defeasibility in Section 4.3.5.

### 4.4.2 Strong overridden defeasibility

In the following example, we reconsider the Fence example and we illustrate that it should not have the inference patterns FC, RI and RIO.

**Example 4.23 (Fence, continued)** Reconsider the obligations $O(\neg f \mid \top)$ and $O(w \wedge f \mid d)$ of Example 4.9. There is a potential conflict between the two obligations. When the facts imply $d$, then there is a conflict, because the first obligation says that there should not be a fence, and the second obligation implies that there should be a fence. However, the first obligation is overridden by the second one, because the second one is more specific. Hence, the conflict is resolved and there should be a white fence. The inference

$$\frac{O(\neg f \mid \top), O(w \wedge f \mid d)}{O(\neg f \mid d)}$$

is *not* valid. The first sentence can be read as: 'usually, there should not be a fence around your house.' Hence, in most situations there should not be a fence, but in exceptional circumstances a fence is allowed. Similarly, the second sentence can be read as 'usually there should be a white fence, when you have a dog.' Hence, the situation when you have a dog is one of the exceptional situations in which the first obligation is not in force. The situation is not sub-ideal yet, it is only exceptional. Hence, the inference

$$\frac{O(\neg f \mid \top), O(w \wedge f \mid d)}{O(\neg d \mid \top)}$$

is *not* valid. Finally, assume that there is a dog but there cannot be a white fence (e.g. there might be a black fence or no fence at all). Hence, the second obligation has been violated. In this situation, which is even more specific than the situation where there is a dog ($d$), nothing is said whether no fence is preferred over a non-white fence. Hence, the inference

$$\frac{O(\neg f \mid \top), O(w \wedge f \mid d)}{O(\neg f \mid d \wedge \neg w)}$$

is *not* valid. □

The following example illustrates that the invalidity of the inference patterns FC, RI and RIO can be explained by the multi preference semantics in Section 4.3.5.

**Example 4.24 (Fence, continued)** Reconsider the multi preference model $M$ in Figure 4.13 of the defeasible obligations $O^{re}(\neg f \,|\, \top \setminus \bot)$ and $O^{re}(w \wedge f \,|\, d \setminus \bot)$ in Example 4.19. Figure 4.13 shows why the two inference patterns FC and RI are not valid. First of all, the obligation not to establish a conflict is not valid, we have $M \not\models O^{re}(\neg d | \top \setminus \bot)$, because the $\neg d$ worlds (the most normal $\neg d$ worlds) are no better than the $d \wedge w \wedge f$ worlds (the optimal most normal $d$ worlds). Secondly, reinstatement is not valid, $M \not\models O^{re}(\neg f \,|\, d \wedge \neg w \setminus \bot)$, because all $d \wedge \neg w$ worlds are equivalent. Hence, if we zoom in on these worlds, there is no preference for $f$ or $\neg f$. □

The invalidity of the inference patterns FC, RI and RIO shows that strong overridden defeasibility is not sufficient to model reasoning about prima facie obligations. In other words, the obligations that model the Fence example are a different type of obligations than the obligations that model prima facie obligations. To emphasize this point, we write $O_{pf}$ for prima facie obligations.

### 4.4.3 Weak overridden defeasibility

The notion of weak overridden defeasibility can be formalized in a prioritized system. We do not give the formal definitions of a prioritized system, because they can be found in many papers on defeasible reasoning (see e.g. [Bre94, GP92]), but we illustrate the idea of a prioritized system by our promises example.

**Example 4.25 (Promises, continued)** Reconsider the obligations in Example 4.22. In a prioritized system, a single preference ordering (an ideality ordering) is constructed for the two prima facie obligations $O_{pf}(k|\top)$ and $O_{pf}(p \wedge \neg k|d)$. To construct the ordering, a naming mechanism

is used, similar to the one in conditional entailment [GP92]. When the ordering is constructed, the prioritization of (the violations of) the obligations is taken into account. A typical prioritized preference ordering of Example 4.22 is given in Figure 4.15. The important relations in this preference model are $w_1 < w_2$ for all $w_1 \in |\neg k \wedge p \wedge d| \cup |\neg k \wedge \neg d|$ and $w_2 \in |k \wedge \neg p \wedge d|$, which state that violating the second obligation is worse than violating the first obligation. Without the prioritization, these worlds would be incomparable. Figure 4.15 shows why the inference patterns FC and RI are valid. First of all, forbidden conflict FC is valid, because $M \models O_{pf}(\neg d \mid \top \setminus \bot)$. This follows from the fact that all $d$ worlds are sub-ideal. Secondly, reinstatement is valid because $M \models O_{pf}(k \mid d \wedge \neg p \setminus \bot)$. The $d \wedge \neg p$ worlds are not equivalent. Hence, if we zoom in on these worlds, as represented by a dashed box, there is an obligation for $k$. □

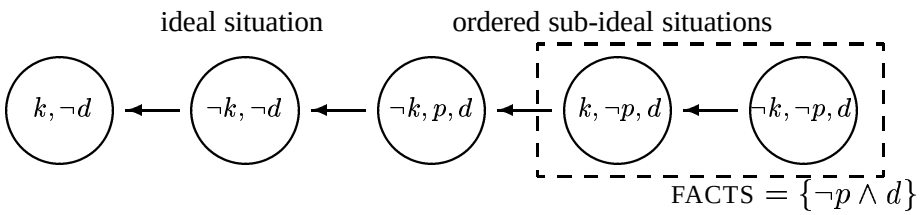



Figure 4.15: Prioritized preference relation

Weak overridden defeasibility is quite close to overshadowing, but these notions are not identical. The typical case of overshadowing is that an obligation $O(p \mid \top)$ is violated by the fact $\neg p$. We can introduce the notion of absolute obligation $Op$ to express that, in spite of the factual violation, the obligation was still in force. In the typical case of weak overridden defeasibility there are two conflicting obligations, say $O_{pf}(p \mid \top)$ and $O_{pf}(\neg p \mid q)$ and the fact $q$, with a priority ordering. To illustrate the difference with overshadowing, let us assume that the second obligation has a higher priority than the first one. We could generalize the logic of absolute obligations to take priority orderings into account, and then these two obligations would imply the actual obligation $O_a \neg p$, but not $O_a p$. This obligation expresses the duty proper, the obligation that should be acted upon. But these obligations would also imply both prima facie obligations $O_{pf} \neg p$ and $O_{pf} p$, which express that both obligations are still in force. These prima facie obligations resemble the absolute obligations of overshadowing. Hence, overshadowing and weak overridden defeasibility are equivalent from the point of view of 'cue for action': once an obligation is violated, it is still fully in force, but no longer a cue for action. Once an obligation is weakly overridden, it is no longer fully in force, but it is still in force as a prima facie obligation.

## 4.5 Related research

The different types of defeasibility have not been studied yet in deontic logic literature. In this section we compare our analysis of the Chisholm paradox with other solutions, and we compare the defeasible deontic logic with dyadic deontic logic and other defeasible deontic logics.

### 4.5.1 Chisholm paradox

Our analysis of the Chisholm paradox is non-standard. Deontic detachment has traditionally been analyzed as transitivity DD. We split DD in two inferences: weak deontic detachment DD′ and weakening of the consequent WC. This is the basis of our solution of the Chisholm paradox. Loewer and Belzer [LB83] argued in a *temporal* framework that deontic detachment should sometimes hold and sometimes not. We argued that deontic detachment should sometimes hold and sometimes not, depending on whether certain assumptions are not violated. Smith [Smi93] also gives an analysis of deontic detachment in the Chisholm paradox in terms of assumptions, and she also notices that in the logical form, such assumptions should not be left implicit. For this reason, she rejects deontic detachment in SDL. In CDL, these assumptions are made explicit and we can accept a restricted form of deontic detachment.

A popular solution of the Chisholm paradox is based on temporal distinctions, see e.g. [vE82, Smi94], analogous to the temporal solution of the Forrester paradox discussed in Section 2.6. This solution demands that the antecedent occurs before the consequent. In Example 4.1 the proposition $t$ is interpreted as telling the neighbors that the man goes to the assistance, and $a$ as going to the assistance. Notice that $t$ occurs before $a$ in the interpretation of the propositional atoms. Hence, the example cannot be represented in a temporal deontic logic that has the antecedent-before-the-consequent assumption.

### 4.5.2 Deontic logic

We compare our contextual deontic logic with dyadic deontic logics. First, Hansson-Lewis minimizing obligations [Han71, Lew74] have too much factual defeasibility, because they do not have any strengthening of the antecedent. This is a result of the fact that every obligation can itself be derived by weakening of the consequent. It is never safe to apply strengthening of the antecedent, because any strengthening can result in an exceptional context. Alchourrón [Alc93] criticizes B. Hansson's logic [Han71] for being a logic of prima facie obligations instead of a logic of CTD obligations. Hansson's logic has FC when the antecedent is a tautology (establishing a conflict is sub-ideal) but not RI (reinstatement). Second, dyadic obligations with a conditional interpretation (like $O(\alpha\,|\,\beta) =_{def} \beta > O\alpha$ [Che80, Alc93]) have too little factual defeasibility, because they have unrestricted strengthening of the antecedent (and factual detachment). Thus they cannot represent contrary-to-duty obligations, because they suffer from the paradoxes.

### 4.5.3 Defeasible deontic logic

Horty [Hor94, Hor93] introduced a deontic logic which is based on non-monotonic logic.[10] In his logic $O\alpha$ and $O\neg\alpha$ are modeled in two different extensions. In Horty's approach, deontic rules can be viewed as (normal) default rules like $\frac{\alpha:\beta}{\beta}$. The default rules, together with a set of facts, yield extensions. These extensions correspond roughly to equivalence classes of

[10] Horty formalizes ideas proposed by Van Fraassen [vF73] in Reiter's default logic. The main difference between the logic of Horty and other deontic logics, is that in Horty's logic deontic dilemmas can be represented in a consistent way. Horty observes that this requires that $O(\alpha \wedge \neg\alpha)$ is not derivable from $O\alpha$ and $O\neg\alpha$ in his logic, see the discussion on consistent aggregation in Section 2.6.3.

preferred models in 2DL and CDL. Horty gives a preferred models semantics which is similar to our semantics only for unconditional obligations. For conditional obligations, Horty defines a notion of deontic consequence, written $\vdash_{CF}$, that derives dyadic obligations from a set of dyadic obligations. Horty's notion of deontic consequence $\vdash_{CF}$ has some serious drawbacks, see [vdT94, Pra96]. Horty also argues for transitivity (i.e. deontic detachment) as a defeasible rule, but he does not implement it in his logic. Finally, Horty extends his logic $\vdash_{cf}$ with a notion of overridden that deals with specificity. It is a weak notion of overridden, because it is based on conflict resolution comparable to priorities. Thus, Horty uses weak overridden defeasibility to formalize specificity. It has reinstatement RI but not overridden reinstatement RIO (although Horty argues for the latter inference too). However, it does not have forbidden conflict FC.

An idea similar to revision-based obligations can be found in a recent proposal of Tan and Pearl [TP94], where a conditional desire $D(l \mid n \wedge \neg l)$ is interpreted as $D(l \mid n)$, representing that 'I desire the light to be on if it is night and the light is off' compares night-worlds in which the light is on with those in which the light is off. However, their formalization is problematic, as is shown in [Bou94b]. Moreover, in our case it is violation detection and revision (it refers to deontic alternatives in the past), in their case it is world improvement and update (it refers to alternatives in the future). For a further discussion, see Section 5.1.

Revision can be considered as a combination of retraction and expansion, known as the Levi identity. In [vdTT95a], we interpreted the essential mechanism to represent violations in terms of a so-called retraction test, see also Section 2.5. Boutilier and Becher [BB95] use a similar kind of retraction to model predictive explanations: 'In order to evaluate the predictive force of factual explanations, we require that the agent (hypothetically) give up its belief in $\beta$ and then find some $\alpha$ that would (in this new belief state) restore $\beta$. In other words, we contract $K$ by $\beta$ and evaluate the conditional $\alpha \Rightarrow \beta$ with respect to this contracted belief state: $\beta \in (K^-_\beta)^*_\alpha$. Thus, when we hypothetically suspend belief in $\beta$, if $\alpha$ is sufficient to restore this belief then $\alpha$ counts as a valid explanation. The contracted belief set $K^-_\beta$ might fruitfully be thought of as the belief set held by the agent before it came to accept the observation $\beta$.'

## 4.6 Conclusions

In this chapter we studied the relation between obligations and defeasibility. We analyzed different types of defeasibility in defeasible deontic logics. We discriminated between two concepts, i.e. overshadowing and cancelling, and three types of defeasibility, i.e. factual defeasibility, strong overridden defeasibility and weak overridden defeasibility. The results we established in this chapter are summarized by Table 4.2.

| | Overshadowing | Cancelling |
|---|---|---|
| Factual defeasibility | X | |
| Strong overridden defeasibility | | X |
| Weak overridden defeasibility | X | X |

Table 4.2: Matrix

1. We observe the distinction between factual and overridden defeasibility in defeasible logics, and we argue that the first should be used to model overshadowing of an obligation by a violating fact, and the second to model cancelling by another obligation.

2. We observe the distinction between weak and strong overridden defeasibility in defeasible logics, and we argue that the first should be used to model overshadowing of a prima facie obligation by a stronger obligation, and the second to model cancelling by a more specific obligation.

There is one relation between obligations and defeasibility which has not been discussed in this chapter. The formalization of the no-dilemma assumption introduces defeasibility, because in Chapter 2 we used non-monotonicity (preferential entailment) to formalize the no-dilemma assumption in 2DL. For example, we can derive $O_D(\alpha_1 | \neg(\alpha_1 \wedge \alpha_2))$ from $O_D(\alpha_1 | \top)$, but not from the two obligations $O_D(\alpha_1 | \top)$ and $O_D(\alpha_2 | \top)$. Moreover, we can derive $O_D(\alpha | \beta_1 \wedge \beta_2))$ from $O_D(\alpha \mid \beta_1)$, but not from the two obligations $O_D(\alpha \mid \beta_1)$ and $O_D(\neg\alpha \mid \beta_2)$. This seems like a kind of overridden defeasibility of the overshadowing type, because it is caused by the introduction of an obligation (thus it is overridden defeasibility) and there is no reason why the obligation should no longer be in force (thus it is not cancelling). However, the intuitions on these two examples seem to be less clear than the examples we discussed in this chapter. We therefore did not include this relation between obligations and defeasibility in Table 4.2.

Moreover, in this chapter we further studied the relation between obligations and preferences. The logic $O^{re}$ with its multi preference semantics illustrates that our bipolar concept of deontic choice is fundamentally different from the classical monopolar interpretation. This distinction is not visible in a single preference ordering. For example, consider the Hansson-Lewis semantics. In the monopolar reading, an obligation $O_\forall(\alpha | \beta)$ is true iff $\alpha$ is true in all preferred $\beta$ worlds. In the bipolar reading, an obligation $O_\forall(\alpha \mid \beta)$ is true iff the preferred $\alpha \wedge \beta$ worlds are preferred to the preferred $\neg\alpha \wedge \beta$ worlds. These two readings are equivalent when we do not consider infinite descending chains. Now consider the multi preference logics. The monopolar reading of a conditional obligation is based on lexicographic minimizing (minimize first $\leq_N$ and then $\leq_I$) like in [Mak93]. In the bipolar reading, an obligation $O_\forall(\alpha | \beta)$ is true iff the $\leq_N$-preferred $\alpha \wedge \beta$ worlds are $\leq_I$-preferred to the $\leq_N$-preferred $\neg\alpha \wedge \beta$ worlds.[11] The distinction can be illustrated by the model in Figure 4.13. For minimizing, the best most normal worlds and the most normal best worlds are both $\neg d \wedge \neg f$. Thus, in the monopolar reading the model satisfies the highly counterintuitive obligation $O(\neg d | \top)$.[12] In the bipolar reading, we have $M \not\models O(\neg d | \top)$, because the $\neg d \wedge \neg f$ worlds are not $\leq_I$-preferred to the $d \wedge w \wedge f$ worlds.

[11] Our approach to multi preference in Definition 4.18 is different, because our second step is not minimizing.

[12] In fact, under certain assumptions lexicographic minimizing is equivalent to minimizing in a single preference ordering (the lexicographic ordering of $\leq_N$ and $\leq_I$).

# Chapter 5

# Applications

In this chapter we consider topics for further research. We discuss two applications that can use deontic logic: qualitative decision theory and a theory of diagnosis. These applications are extensions of deontic logic, because deontic logic only tells us which obligations follow from a set of obligations, but it does not tell us how obligations affect behavior. Qualitative decision theory uses preference-based deontic logic to formalize reasoning about context-sensitive goals. A theory of diagnosis uses deontic logic to represent system rules and violations of these system rules. We discuss reasoning with obligations, but we leave the detailed study of this subject for further research.

The diagnostic framework for deontic reasoning was first presented in [TvdT94c, TvdT94a]. This chapter is a modified and extended version of [vdTRFT97].

## 5.1 Reasoning with obligations

In this chapter we argue that normative reasoning is more than deontic logic. Deontic logic tells us which obligations can be derived from a set of other obligations. In particular, it characterizes the logical relations between obligations. For example, in most logics the conjunction $p \wedge q$ is obliged, if both $p$ and $q$ are obliged. However, it does not explain how obligations affect the behavior of rational agents. From $Op$ you cannot infer whether somebody will actually perform $p$. This is no critique on deontic logic, it is just an observation. Deontic logic was never intended to explain this effect of obligations on behavior. However, if we want to explain all the different aspects of normative reasoning, then we need more formalisms than just deontic logic. In this chapter we discuss two formalisms that can be used to analyze two different types of aspects of how obligations effect behavior, namely the theory of diagnosis and qualitative decision theory.

Two theories that are able to formalize reasoning with obligations are represented in Figure 5.1. A *theory of diagnosis* reasons about violations. In particular, it reasons about the past with incomplete knowledge (if everything is known than a diagnosis is completely known). Diagnostic theories have a modest purpose, because they do not support the decision-making process of the user. They do not derive decisions, they only check systems against given principles. A more expressive framework is *qualitative decision theory*, that describes how obligations influence behavior. It is based on the concept of agent rationality. For example, in a normative system usually sanctions and rewards correspond with obligations, and a rational agent tries to

evade penalties and achieve rewards. In contrast to diagnostic theories, a (qualitative) decision theory reasons about the future. The main characteristic of qualitative decision theory is that it is goal oriented reasoning, usually for planning problems. Moreover, it combines reasoning about goals with uncertainty. This reasoning is based on the application of strategies, which can be considered as qualitative versions of the 'maximum utility' criterion.

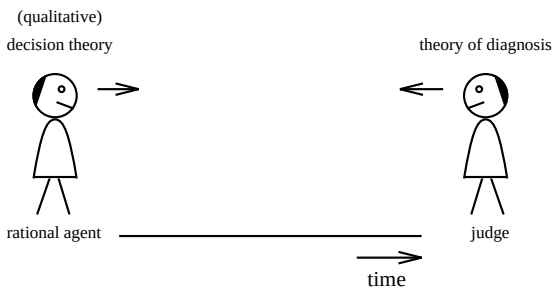



Figure 5.1: Reasoning with obligations

Moreover, *scenario analysis* uses deontic logic to represent the deontic status of agents. Scenario analysis performs simulations of (for example normative) systems, and does not seem to be confined to only the past or the future. Scenario analysis can be based on so-called dead-line obligations. Consider the obligation 'if the obligation to deliver the goods is violated, then a penalty has to be paid.' In Figure 5.2.a, a dead-line obligation can be considered to be a combination of diagnosis (violation of the obligation to deliver the goods $\neg d \wedge Od$), and decision theory (the deontic cue to pay the penalty $Op$). However, not all contrary-to-duty obligations can be interpreted as dead-line obligations. Consider the obligation 'if the man violates the obligation to go to his neighbors assistance, then the neighbors should not be told that he will come' of the Chisholm paradox, see Example 1.16. The diagnosis (violation of the obligation to go to the assistance) is later than the decision moment (telling or not telling), as represented in Figure 5.2.b. Hence, there is no dead-line, and thus no dead-line interpretation of the sentence (it leads to the 'split personality'). The diagnostic and decision-theoretic perspectives can give an interpretation to the sentence. For example, the decision-theoretic perspective make a plan, which contains the intention to go to the assistance. From this plan the obligation is derived to tell the neighbors that he will come.

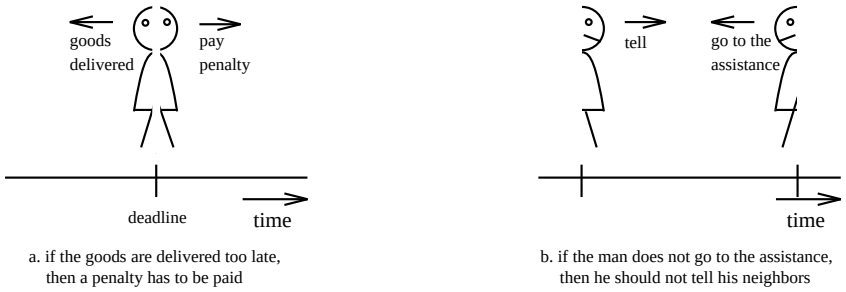



Figure 5.2: Deadline obligations

Logical relations between obligations are an essential component of any formalism that explains the effect of obligations on behavior. Hence, in this chapter we also argue that deontic

logic can be used as a component in the theory of diagnosis as well as qualitative decision theory. Actually, we even argue for the stronger claim that the theory of diagnosis as well as qualitative decision theory can be viewed as extensions of deontic logic. In both cases the formalism contains extra principles that are added to a deontic logic basis. For example, in the case of the theory of diagnosis one of the principles that can be added to deontic logic is the parsimony principle, i.e. the assumption that as few as possible obligations are violated. There is nothing contradictory in the claim that on the one hand these formalisms explain aspects of normative behavior that deontic logic does not, whereas deontic logic is still an essential component of these theories. In the same sense physics can explain phenomena that mathematics cannot, whereas mathematics is still an essential component of physics. There are several structural similarities between preference-based deontic logic and the logics developed for diagnosis and qualitative decision theory, see e.g. [Bou94b, Lan96]. The distinction between the different perspectives and deontic logic raises several important questions.

1. **Obligations and dedicated theories.** The diagnosis of a normative system can use a formalism to represent obligations and additional assumptions or principles to do the diagnosis. For example, Reiter's diagnosis is basically a minimization principle (called the principle of parsimony). Similarly, qualitative decision theory has a formalism for representing obligations (or goals) and additional assumptions or principles to reason with them. Is such a special purpose formalism a deontic logic? How do they stand the test against the Chisholm paradox, the paradox of the gentle murderer, the problem of how to represent permissions, the problem of conflicting obligations? What are the structural similarities and distinctions between the different formalisms?

2. **Obligations and preferences [Lan96].** Qualitative decision theory is based upon the concept of preference. This preference is a kind of desire, i.e. it is an endogenously motivating mechanism (coming from the agent itself). Therefore, it is not a natural candidate for dealing with normative decision-making, since a norm is by definition exogenous, in the sense that it is something the agent would not spontaneously want. How do agents work out norms in terms of gains and losses? What are the gains of observing norms? How do they learn the effects of norms and how do they reason about these effects? Which rules are implied, which ingredients enable agents to make normative decisions? In which way does a normative decider differ from an ordinary decider, if any?

3. **Obligations and norms.** A deontic logic does not derive actual but ideal behaviors. Should we distinguish the obligations derivable from a set of norms and a set of facts, from the norms itself? What is the role of so-called factual detachment in deontic logic?

The distinction between the perspective of a rational agent (qualitative decision theory) and a judge (theory of diagnosis) corresponds to Thomason's distinction between the context of deliberation and the context of justification [Tho81], see Section 1.3.4. Thomason distinguishes between two ways in which the truth values of deontic sentences are time-dependent. First, these values are time-dependent in the same, familiar way that the truth values of all tensed sentences are time-dependent. Second, their truth values are dependent of a set of choices or future options that varies as a function of time. If you think of deontic operators as analogous to quantifiers ranging over options, this dependency on context is a familiar phenomenon. Thus, the context

of deliberation is the set of choices when you are looking for practical advice, whereas the context of justification is the set of choices for someone who is judging you. The following example illustrates that it is important to discriminate between these two contexts, because a sentence can sometimes be interpreted differently in each of them. The original example discussed in [Han71] concerns the obligation 'you should not smoke, if you smoke.'

**Example 5.1** Consider the sentence 'you should not smoke and you smoke.' In the context of justification the obligation is interpreted as the identification of the fact that you are violating a rule, whereas in the context of deliberation, it is interpreted as the obligation to stop smoking. When the context is not known, it is also not known which of these two interpretations (or probably both) is meant. The two perspectives are represented in Figure 5.3. At the present moment in time, $s$ is true. The context of justification considers the moment before the truth value of $s$ was settled, and considers whether at that moment in the past, $\neg s$ was preferred over $s$. The context of deliberation considers the moment the truth value of $s$ can be changed, and considers whether at that moment in the future, $\neg s$ will be preferred over $s$. □

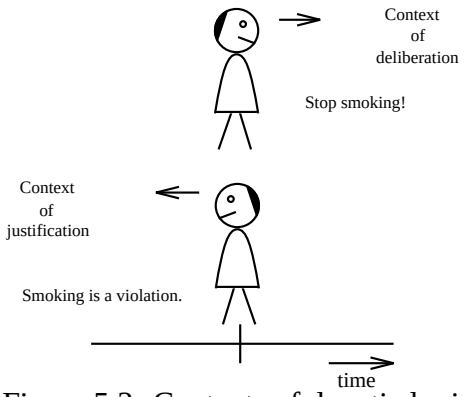



Figure 5.3: Contexts of deontic logic

The distinction between the two interpretations of the obligation is as important as the distinction between Alchourrón-Gärdenfors-Makinson belief revision (or theory revision) [AGM85] and Katsuno-Mendelzon belief update [KM92] in the area of logics of belief. There is a strong analogy, because belief revision is reasoning about a non-changing world and update is reasoning about a changing world. It follows directly from Figure 5.3 that a similar distinction is made between respectively the context of justification and the context of deliberation, because the past is fixed, whereas the future is wide open.

In this chapter, we discuss two applications of (preference-based) deontic logic. The first application of deontic logic is (robot) planning with qualitative decision theory. Boutilier [Bou94b] observes that 'in the usual approaches to planning in AI, a planning agent is provided with a description of some state of affairs, a *goal state*, and charged with the task of discovering (or performing) some sequence of actions to achieve that goal. This notion of goal can be found in the earlier work on planning and persists in more recent work on intention and commitment [CL90]. In most realistic settings, however, an agent will frequently encounter goals that it cannot achieve. As pointed out by Doyle and Wellman [DW91b] an agent possessing only simple goal descriptions has no guidance for choosing an alternative goal state toward which it should

strive. [ . . . ] A recent trend in planning has been the incorporation of decision-theoretic methods for constructing optimal plans [DW91a]. Decision theory provides for most of the basic concepts we need for rational decision making, in particular, the ability to specify arbitrary preferences over circumstances or goals (and hence appropriate behaviors) to vary with context.' Such context-sensitive goals can be represented by dyadic obligations $O(\alpha \mid \beta)$, to be read as '$\alpha$ is a goal if $\beta$.'

The second application of deontic logic we discuss in this chapter is Ramos and Fiadeiro's Deontic framework for Diagnosis of (organizational) process Design DDD. Reiter formalized in [Rei87] the model based reasoning approach to diagnosis. In Reiter's theory of diagnosis, a violation is represented by a predicate expression $Ab(c)$, where $c$ is a component of a system to be diagnosed and $Ab$ an abnormality predicate. For example, this violation can be derived from the system description that $p$ is the correct behavior of a component $\neg Ab(c) \rightarrow p$ and the observation $\neg p$. In a modal deontic logic, a violation can be represented by the sentence $\neg p \wedge Op$. In Section 5.3, we discuss a theory of diagnosis based on deontic logic. An important advantage of the modal deontic language is that the concept of obligation is an intuitive and natural way to represent the kind of principles that arises in process design. The typical diagnostic reasoning with normative systems is performed by a judge, who has to determine whether a suspect is guilty or not. Diagnostic reasoning has to deal with incomplete knowledge, not formalized in a deontic logic. For example, a popular additional assumption of theories of diagnosis is the so-called principle of parsimony: 'you are innocent until proven guilty.' Such a principle about incomplete knowledge is not made in deontic logic; it is an extra-logical assumption about the legal domain.

## 5.2 Qualitative decision theory

A qualitative decision theory formalizes reasoning about goals and can be used for planning problems. It combines reasoning about goals with reasoning about uncertainty. In this section, we discuss two different perspectives on the relation between qualitative decision theory and deontic logic. First we discuss Pearl's logic of pragmatic obligation, that arises out of a criticism of standard deontic logics (like the logics developed in this thesis). Second, we discuss Boutilier's logic of qualitative decision theory, that incorporates a deontic logic.

### 5.2.1 Pearl's logic of pragmatic obligation

Pearl [Pea93] observes that 'obligation statements, also called *deontic* statements, come in two varieties: obligations to act in accordance with peers' expectations or commitments to oneself, and obligations to act in the interest of one's survival, namely, to avoid danger and persue safety.' Moreover, Pearl [Pea93] develops a logic of pragmatic obligation, a decision-theoretic account of obligation statements of the second variety, using qualitative abstractions of probabilities and utilities. 'The idea is simple. A conditional obligation sentence of the form "You ought to do $\alpha$ if $\beta$" is interpreted as shorthand for a more elaborate sentence: "If you observe, believe, or know $\beta$, then the expected utility resulting from doing $\alpha$ is much higher than that resulting from not doing $\alpha$." Pearl observes that 'this decision-theoretic agenda, although conceptually straightforward, encounters some subtle difficulties in practice.

1. First, when we deal with actions and consequences, we must resort to causal knowledge of the domain and we must decide how such knowledge is to be encoded, organized, and utilized.

2. Second, while theories of action are normally formulated as theories of temporal changes [Sho88, DK89], deontic statements invariably suppress explicit references to time, strongly suggesting that temporal information is redundant, namely, it can be reconstructed if required, but glossed over otherwise.

3. Third, decision theoretic methods treat actions as distinct, predefined objects, while deontic statements of the type "You ought to do $\alpha$" are presumed applicable to any proposition $\alpha$.[1]

4. Finally, decision theoretic methods, especially those based on static influence diagrams, treat both the informational relationships between observations and actions and consequences as instantaneous [Sha86, Pea88]. In reality, the effect of our next action might be to invalidate currently observed properties, hence any non-temporal criterion for obligation must carefully distinguish properties that are influenced by the action from those that will persist despite the action.'

Instead of discussing the details of the logic, we discuss two examples from Pearl [Pea93]. The first example considers the assertability of "If it is cloudy you ought to take an umbrella." A $\kappa$ value represents the degree of normality and a $\mu$ value represents a degree of preference.

**Example 5.2 (Umbrella)** [Pea93] We assume three atomic propositions, c - "Cloudy", r - "Rain", and u - "Having an umbrella", which form eight worlds, each corresponding to $c$, $r$ and $u$. To express our belief that the rain does not normally occur in a clear day, we assign a $\kappa$ value of 1 (indicating one unit of surprise) to any world satisfying $r \wedge \neg c$ and a $\kappa$ value of 0 to all other worlds (indicating a serious possibility that any such world may be realized). To express the fear of finding ourselves in the rain without an umbrella, we assign a $\mu$ value of -1 to worlds satisfying $r \wedge \neg u$ and a $\mu$ value of 0 to all other worlds. Thus, $W^+ = false, W^0 = \neg(r \wedge \neg u)$, and $W^- = r \wedge \neg u$. □

The following example is also from Pearl [Pea93]. It demonstrates the interplay between action and observations.

**Example 5.3 (Switch)** We will test the assertability of the following dialogue:

Robot 1: It is too dark here.
Robot 2: Then you ought to push the switch up.
Robot 1: The switch is already up.
Robot 2: Then you ought to push the switch down.

The challenge would be to explain the reversal of the "ought" statement in response to the new observation "The switch is already up." The inferences involved in this example revolve

[1] This has been an overriding assumption in both the deontic logic and the preference logic literatures.

around identifying the type of switch Robot 1 is facing, that is whether it is normal ($n$) or abnormal ($\neg n$) (a normal switch is one that should be pushed up ($u$) to turn the light on ($l$)). □

Tan and Pearl [TP94] observe that the treatment in [Pea93] assumes that a complete specification of a utility ranking is available and that the scale of the abstraction of preferences is the same as the scale of the abstraction of belief. They propose a specification language which accepts conditional preferences of the form "if $\beta$ then $\alpha$ is preferred to $\neg\alpha$", $\alpha > \neg\alpha \mid \beta$. A conditional preference of this form will be referred to as a *conditional desire*, written $D(\alpha|\beta)$, which represents the sentence "if $\beta$ then $\alpha$ is desirable." The output is the evaluation of a preference query of the form $(\phi, \psi_1 > \psi_2)$ where $\phi$ is any general formula while $\psi_1$ and $\psi_2$ may either be formulas or action sequences. The intended meaning of such query is "is $\psi_1$ preferred over $\psi_2$ given $\phi$"? Each conditional desire is given ceteris paribus (CP) semantics; "$\alpha$ is preferred to $\neg\alpha$ other things being equal in any $\beta$ world."' The following example of [TP94] shows that Tan and Pearl formalize the Switch example with the obligation $D(l|n \wedge \neg l)$. This formula is interpreted as an update of the situation (context of deliberation) instead of the identification of a violation (context of justification), see the discussion in Section 5.1.

**Example 5.4 (Switch, continued)** Consider the sentence, "I desire the light to be on if it is night and the light is off," $D(l \mid n \wedge \neg l)$. Clearly such a sentence compares *night*-worlds in which the light is on to those in which the light is off. The former worlds do not satisfy the condition $\beta = \neg l$. Tan and Pearl argue that 'such a reasonable sentence would be deemed meaningless in a restricted interpretation such as [DSW91]' (and Hansson-Lewis minimizing logics). $\beta$ does not act as a filter for selecting worlds to which the desired constraints apply, instead it identifies worlds in which the desires are satisfied. □

Pearl [Pea93] further remarks when he compares his notion of pragmatic obligation with deontic logics that 'exploratory reading of the literature reveals that philosophers hoped to develop deontic logic as a branch of conditional logic, not as a synthetic amalgam of logic of belief, action, and causation.[2] In other words, they have attempted to capture the meaning of "ought" using a single modal operator $O(\alpha|\beta)$, instead of exploring the couplings between "ought" and other modalities, such as belief, action, causation, and desire.' Pearl argues that 'such an isolationistic strategy has little chance of succeeding. Whereas one can perhaps get by without explicit reference to desire, it is absolutely necessary to have both probabilistic knowledge about the effect of observations on the likelihood of events and causal knowledge about actions and their consequences.'

Pearl finally concludes that 'the decision-theoretic account can be used to generate counterexamples to most of the principles suggested in the literature, simply by selecting a combination of $\kappa$ (normality), $\mu$ (preferences) and $\Gamma$ (causal network) that defies the proposed principle. Since any such principle must be valid in all epistemic states and since we have enormous freedom in choosing these three components, it is not surprising that only weak principles such as

[2]The reluctance to connect obligations to causation can perhaps be attributed to a general disappointment with attempts to develop satisfactory accounts for actions and causation. For example, the Stalnaker-Lewis logic of counterfactuals, which promised to capture some aspects of causation (causal relationships invariably invite counterfactuations), ended up as a faint version of the logic of indicative conditionals [Gib80], hiding rather than revealing the rich structure of causation.

$O(\alpha \mid \beta) \rightarrow \neg O(\neg\alpha \mid \beta)$, survive the test. Among the few that survive, we find the sure-thing principle: $O(\alpha|\beta) \wedge O(\alpha|\neg\beta) \rightarrow O(\alpha|\top)$, read as 'if you ought to do $\alpha$ given $\beta$ and you ought to do $\alpha$ given $\neg\beta$, then you ought to do $\alpha$ without examining $\beta$.' But one begins to wonder about the value of assembling a logic from a sparse collection of such impoverished survivors when, in practice, a full specification of $\kappa$, $\mu$ and $\Gamma$ would be required.'

### 5.2.2 Boutilier's logic of qualitative decision theory

Boutilier [Bou94b] develops a logic of qualitative decision theory in which the basic concept of interest is the notion of *conditional preference*. Boutilier writes $I(\alpha \mid \beta)$, read "ideally $\alpha$ given $\beta$," to indicate that the truth of $\alpha$ is preferred, given $\beta$. This holds exactly when $\alpha$ is true at each of the most preferred of those worlds satisfying $\beta$. Boutilier remarks that from a practical point of view, $I(\alpha|\beta)$ means that if the agent (only) knows $\alpha$, and the truth of $\beta$ is fixed (beyond his control), then the agent ought to ensure $\alpha$. Otherwise, should $\neg\alpha$ come to pass, the agent will end up in a less than desirable $\beta$-world. Boutilier mentions that the statement can be *roughly* interpreted as "if $\beta$, do $\alpha$." Moreover, Boutilier observes that the conditional logic of preferences he proposed is similar to the (purely semantic) proposal put forth by B.Hansson [Han71]. He concludes that 'one may simply think of $I(\alpha|\beta)$ as expressing a conditional obligation to see to it that $\alpha$ holds if $\beta$ does.' Thomason and Horty [TH96] also observe the link with deontic logic when they develop the foundations for qualitative decision theory. They consider the problem how to extend the point utilities to utilities on sets, assuming that the utilities of histories are known. They observe that classical decision theory provides a way to do this, but they follow a radically qualitative approach, which assumes that only a linear preference ordering on histories is given, which must be extended to a partial ordering over sets of histories. They observe that this 'utilities lifting' problem is discussed or alluded to, for instance, in the literature on deontic logic [vF72, Jen74, Jen85, Wel88, Hor96].'

Boutilier [Bou94b] introduces a simple model of action and ability. The atomic propositions are partitioned into *controllable* propositions, atoms over which the agent has direct influence, and *uncontrollable* propositions. He ignores the complexities required to deal with effects, preconditions and such, in order to focus attention on the structure and interaction of ability and goal determination. The consequence of this lack of an action model is that 'we should think of a rule as an *evidential rule* rather than a *causal rule*.' Moreover, Boutilier observes that 'the implicit temporal aspect here; propositions should be thought of as *fluents*. We can avoid an explicit temporal representation by assuming that preference is solely a function of the truth values of fluents.' Lang [Lan96] calls controllable and uncontrollable propositions respectively decision variables and parameters. Moreover, he argues that it is necessary to distinguish not only between desires (goals) and knowledge as in [Bou94b] but also between background factual knowledge (which tells which worlds are physically impossible) and contingent knowledge (which tells which of the physically possible worlds can be the actual states of affairs).

The simplest definition of goals is in accord with the general maxim 'do the best thing possible consistent with your knowledge.' Boutilier [Bou94b] dubbed such goals CK goals because they seem correct when an agent has *Complete Knowledge* of the world (or at least of uncontrollable atoms). But Boutilier also shows that CK-goals do not always determine the best course of action if an agent's knowledge is *incomplete*.

**Example 5.5 (Umbrella, continued)**[Bou94b] Consider preferences in the umbrella example,

where 'no umbrella and no rain' $\neg u \wedge \neg r$ is the most ideal, and 'no umbrella and rain' $\neg u \wedge r$ is the worst situation (because the agent gets wet). Moreover, assume that all the agent knows it could rain or not (it has no indication either way). Using CK-goals, the agent ought to do $\neg u$, for the best situation is $\neg r \wedge \neg u$. Leaving its umbrella is the best choice should it turn out not to rain; but should it rain, the agent has ensured the *worst* possible outcome. It is not clear that $\neg u$ should be a goal. Indeed, one might expect $u$ to be a goal, for no matter how $u$ turns out, the agent has avoided the worst outcome. □

The pessimistic perspective of Example 5.5 coincides with Wald's criterion of decision theory, see also [DP95, Lan96].

### 5.2.3 Discussion

Neither Pearl nor Boutilier emphasize the fundamental distinction between deontic logic and decision theory, that decision theory in contrast to deontic logic describes how norms affect behavior. The distinction is observed by McCarty [McC94b], who introduces an assumption to establish the link between deontic logic and planning. In particular, he observes that 'for purposes of planning, it is often useful to assume that actors *do* obey the law.' He calls this the *causal assumption*, since it enables us to 'predict the actions that *will* occur by reasoning about the actions that *ought* to occur.' McCarty concludes that 'if we adopt the causal assumption, we can use the machinery of deontic logic to reason about the physical world.' Lang [Lan96] uses many examples from the deontic logic literature to illustrate his qualitative decision theory. He comments on the distinction between deontic logic and decision theory, and he observes that the two are complementary. The main purpose of a deontic logic is deriving new obligations (and permissions) from an initial specification, while QDT focuses on the search for optimal acts and decisions. Lang concludes that 'deontic logics may be viewed as 'upstream' and QDT 'downstream', since the former provides representations of ideal states, or of a whole preference relation between states, and the latter uses this preference relation ('goalness') to find the best possible actions.' Finally, Lang [Lan96] observes that his methodology contains two phases (generate the preference relation from a set of desires, and then find the optimal feasible worlds, and thus the optimal decision) which, as Lang observes, is in accordance with our argumentation about the two-phase treatment of violated obligations in 2DL.[3]

To consider Pearl's criticism of deontic logics, first observe that his logic of pragmatic obligation allows for exceptions, because it refers to *expected* utilities. Thus, the concept that Pearl intended to reconstruct is completely different from the one which deontic logicians were interested in. In Carnap's terminology, Pearl has different explicata, because he has set out to clarify different explicanda.[4] For example, we can compare Pearl's logic with our logic $O^{re}$ developed in Chapter 4. First, we observe that in our logic $O^{re}$ only weak principles survive (like CD and AD). Thus, Pearl's criticism is nothing but an unsurprising property of *defeasible* deontic logics.

[3]Moreover, Lang uses our distinction between background knowledge and factual contingent knowledge [vdT94], which we introduced to represent specificity.

[4]Alchourrón [Alc93] criticized B.Hansson [Han71] for this reason. Thus, according to Alchourrón, Pearl's logic of pragmatic obligation can be compared to Hansson's deontic logic. However, from the semantics follows that B.Hansson does not consider defeasible deontic logics, because this semantics does not have a normality ordering.

Second, his logic has the sure-thing principle. This principle is not valid in our logics, and its invalidity can be used to analyze dominance arguments, see Example 2.9.

The logics of Boutilier [Bou94b] and Tan and Pearl [TP94] we discussed in this section suffer from CTD paradoxes, as we showed in Section 2.6.4 with the speed limits example. The logic of Lang [Lan96] does not suffer from it. The logic is comparable to a preference logic extended with priorities. As discussed in Section 4.4, such preferences are useful to model prima facie obligations, but it is less obvious that they model an intuitive notion of specificity. Tan and Pearl's logic makes the cigarettes example inconsistent, which they find counterintuitive. They therefore add priorities to their logic in [TP95] in a system analogous to Lang's system (and many other logics, for example the logic of Geffner and Pearl [GP92]).

## 5.3 Diagnosis of organizational process designs

In this section, we discuss Ramos and Fiadeiro's DEontic framework for DIagnosis of process DEsign DDD [RF96b]. This framework is an extension of our DIagnostic framework for DEOntic reasoning DIODE. Moreover, we discuss Ramos and Fiadeiro's diagnosis [RF96a] based on deontic logic and compare their deontic logic LDD with the preference-based deontic logics developed in this thesis.

### 5.3.1 Organizational process design

The work of Ramos and Fiadeiro should be understood as a contribution to the more general purpose to build a formal framework to support organizational process design diagnosis according to predefined process design principles. By principles they mean general rules that characterize the ideal behavior of an organization. They are interested in forms of diagnoses that report violations of such principles. The architecture of their intended framework is represented in Figure 5.4 (taken from [RF96b]).

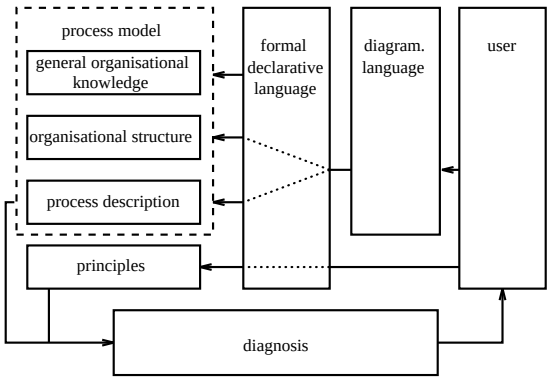



Figure 5.4: Architecture of Ramos and Fiadeiro's framework

The user in Figure 5.4 represents both the designer and the person responsible for defining general principles. As represented in Figure 5.4, the user (supported by a diagrammatic lan-

guage) can describe the structure of the organization and design the process (process description). The diagnosis procedure uses that information, together with general organizational knowledge, to detect violations of the principles indicated by the organization (user). The translation from a diagrammatic language to a declarative formal language is necessary, because Ramos and Fiadeiro want to use logical deduction in the diagnosis procedure. The components of the process model are the following ones:

1. **Organizational structure.** The set of structural concepts that characterize an organization, e.g. *agents, tasks, hierarchies*. These concepts are independent of the processes. They describe the fixed components over which the processes should 'flow'. The structural concepts represent what is fixed in the organization in the sense that it cannot be changed as a consequence of a process (re)design.

2. **Process description.** The description of the process design, made with typical primitives used in organizational process like *assign, output-to-task* etc. Variable concepts are concepts that can be manipulated by the person that designs the process. They can be understood as 'design actions'.

3. **General organizational knowledge.** Definitions (e.g. *available, informed*) and rules common to all organizations (e.g. *if a task is assigned to a collective agent, all the members of the collective agent are assigned to that task*).

The following example of [RF96b] illustrates the design of an order delivering process, and is adapted from [CL92]. In Chen and Lee's framework for the evaluation of internal accounting control procedures, the idea of having general principles guiding organizational diagnosis is already present. However, Ramos and Fiadeiro [RF96b] observe that this framework is not supported by a theory of diagnosis. For instance, it does not deal with either alternative or minimal diagnoses.

**Example 5.6 (Delivering order)** To avoid frauds in organizational accounting procedures, some control rules are often used. In Figure 5.5, the process is designed in order to (partially) fulfill those rules (principles). The process is as follows. The stock manager receives an order (from a salesman, for example), fills up an internal delivery order (IDO) and sends the IDO to agent 1, assigned to the task of verifying the IDO. After receiving the same order the accounting department fills up the invoice and also sends it to agent 1. Agent 1 checks if the values of the IDO and the invoice are the same, stores the invoice in the invoice file and sends the IDO to agent 2, assigned to the task of filling up the outgoing delivery order (ODO). After filling up the ODO agent 2 sends it to the client together with the goods.

Agent 1 is involved in the process in order to avoid a potential fraud between the stock manager and the client, because agent 1 checks if the goods in the IDO matches the values in the invoice. In the process design in Figure 5.5 one general rule, to ensure that the document is not manipulated by other agents is fulfilled: *'all documents must go straight to the control agent after they are created*.' Two other rules that apply to the process are *'an agent should not control a superior in the hierarchy'* and *'socially-close agents should not control each other*.' For example, the stock manager should not be a superior of agent 1 and agent 2 should not be socially-close to the stock manager.

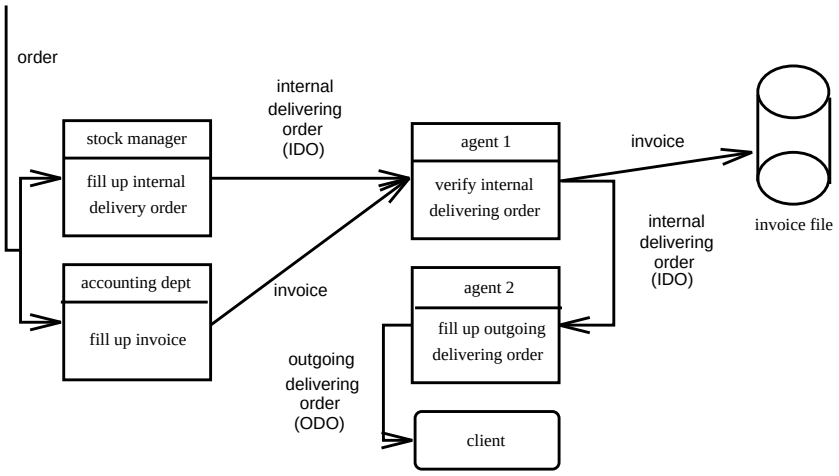



Figure 5.5: Ideal order delivering process

We give a simple formalization of this example in a propositional language, which suffices for our purposes of illustrating DDD. Instead of formalizing the three generic rules as first-order obligations, we formalize several consequences (instances) of these generic rules as propositional obligations. Let us assume the following organization structure: John, Ann and Phil are agents of the organization, Phil is socially-close to John and that the stock manager is hierarchical superior than John. The obligations are (a) the output of the task fill-up-invoice must go to the task verify-IDO. (b) the output of the task verify-IDO must go to the task fill-up-ODO, (c) we must not assign Phil to the task fill-up-ODO, because socially close agents should not be involved in this process, and (d) we must not assign John to the task verify-IDO, because one agent should not control a superior in the hierarchy. We represent the four obligations by $Oa$, $Ob$, $Oc$ and $Od$, respectively. An instance of the general organizational knowledge is that if the output of task verify-IDO goes to fill-up-ODO and Phil is not assigned to fill-up-ODO, then Phil does not receive the ODO, which is represented by $b \wedge c \rightarrow e$. Finally, facts (design) are that Ann is agent 1, Phil is agent 2, John is not assigned to the task verify-IDO and that Phil receives the ODO, i.e., $d \wedge \neg e$. Notice that one of the first or second obligation is violated, the third obligation is fulfilled, and nothing is know about the fourth obligation. □

In the following section we show how this delivering order example is represented in the diagnostic framework for deontic reasoning DIODE, based on Reiter's theory of diagnosis. Ramos and Fiadeiro [RF96a] show that contrary-to-duty scenarios occur in process design. This hints at a possible use of deontic logic, see the discussion in Section 1.2.1.

**Example 5.7** Consider the following rule: *'It should not be the case that the agent that performs an operational task is a direct superior of the agent that controls the operational task'* $O(\neg s|\top)$ and *'if that is nevertheless the case, then instead of storing the control report in the control report file, then the control report should be send to an agent higher (in the hierarchy) than the agent that performs the operational task'* $O(h \mid s)$. The latter is a contrary-to-duty obligation of the former, because the antecedent of $O(h|s)$ is contradictory with the consequent of $O(\neg s|\top)$. □

### 5.3.2 Diagnostic framework for deontic reasoning

The model-based reasoning approach has been studied for several years (for a survey of the topic see [DW88]). Numerous applications have been built, most of all for diagnosis of physical devices. The basic paradigm is the interaction of prediction and observation. Predictions are expected outputs given the assumption that all the components are working properly (i.e. are working according to the model of the structure and behavior of the system). If a discrepancy between the output of the system (given a particular input) and the prediction is found, then the diagnosis procedure will search for defects in the components of the system.

#### Reiter's theory of diagnosis from first principles

The contribution of Reiter to the theory of diagnosis is widely accepted. His *consistency based approach* [Rei87] is the first one to model the model-based reasoning approach to diagnosis. The main goal is to eliminate system inconsistency, identifying the minimal set of abnormal components that is responsible for the inconsistency. That is, reasoning about diagnosis is based on the following assumption.

> **Principle of parsimony** Diagnostic reasoning is based on the conjecture that the set of faulty components is minimal (with respect to set inclusion).

Related to a diagnosis is a set of measurements. Finally, a conflict set is a minimal set of components of which at least one is broken (such sets are used in efficient diagnostic algorithms). In the following definition of diagnosis, a broken component is represented by $Ab(c)$, where $c$ is a component and $Ab$ is short for Abnormal.

**Definition 5.8 (Diagnosis)** A *system* is a pair (COMP, SD) where COMP, the *system components*, is a finite set of constants denoting the components of the system, and SD, the *system description*, is a set of first-order sentences. An *observation* of a system is a finite set of first-order sentences. A system to be diagnosed, written as (COMP, SD, OBS), is a system (COMP, SD) with observation OBS. A *diagnosis* for (COMP, SD, OBS) is a minimal (with respect to set inclusion) set of components $\Delta \subseteq \text{COMP}$ such that

$$\text{CONTEXT}_\Delta = \text{SD} \cup \text{OBS} \cup \{Ab(c) \mid c \in \Delta\} \cup \{\neg Ab(c) \mid c \in \text{COMP} - \Delta\}$$

is consistent. A diagnosis $\Delta$ for (COMP, SD, OBS) *predicts* a *measurement* $\Pi$ iff

$$\text{CONTEXT}_\Delta \models \Pi$$

A *conflict set* for (COMP, SD, OBS) is a minimal (with respect to set inclusion) set $\Delta \subseteq \text{COMP}$ such that $\text{CONTEXT}_\Delta$ is inconsistent. □

Reiter gives the following example of an electronic circuit, which illustrates the diagnosis of a full-adder. In particular, it illustrates that diagnoses are not necessarily unique and that minimality is only with respect to set inclusion.

**Example 5.9 (Full-adder)** Consider the electronic circuit represented in Figure 5.6. The system consists of a set of components $\text{COMP} = \{A_1, A_2, X_1, X_2, O_1\}$ and the system description SD with the following rules:

$$
\begin{array}{l}
ANDG(x) \wedge \neg Ab(x) \rightarrow out(x) = and(in1(x), in2(x)) \\
XORG(x) \wedge \neg Ab(x) \rightarrow out(x) = xor(in1(x), in2(x)) \\
ORG(x) \wedge \neg Ab(x) \rightarrow out(x) = or(in1(x), in2(x)) \\
ANDG(A_1), ANDG(A_2), XORG(X_1), XORG(X_2), ORG(O_1) \\
out(X_1) = in2(A_2), out(X_1) = in1(X_2) \\
out(A_2) = in1(O_1), in1(A_2) = in2(X_1) \\
in1(X_1) = in1(A_1), in2(X_1) = in2(A_1) \\
out(A_1) = in2(O_1)
\end{array}
$$

together with axioms specifying that the circuit inputs and outputs are binary valued (like for example $in1(x) = 0 \vee in1(x) = 1$) and axioms for a Boolean algebra over $\{0, 1\}$:

$$
\begin{array}{lll}
and(in1(x), in2(x)) = 1 & \text{iff} & in1(x) = in2(x) = 1 \\
xor(in1(x), in2(x)) = 1 & \text{iff} & in1(x) \neq in2(x) \\
or(in1(x), in2(x)) = 1 & \text{iff} & in1(x) = 1 \text{ or } in2(x) = 1
\end{array}
$$

Suppose the circuit in Figure 5.6 is given the inputs $\langle 1, 0, 1 \rangle$ and yields the output $\langle 1, 0 \rangle$ in response. This observation can be represented by the following set of first-order sentences.

$$in1(X_1) = 1, in2(X_1) = 0, in1(A_2) = 1, out(X_2) = 1, out(O_1) = 0$$

Notice that this observation indicates that the physical circuit is faulty. Both circuit outputs are wrong for the given inputs. For the electronic circuit, there are three diagnoses: $\{X_1\}$, $\{X_2, O_1\}$, $\{X_2, A_2\}$. Minimality is only with respect to set inclusion, because $\{X_2, O_1\}$ is a diagnosis although the diagnosis $\{X_1\}$ has less elements. For the electronic circuit, the diagnosis $\{X_1\}$ predicts the measurement $out(A_2) = 0$. □

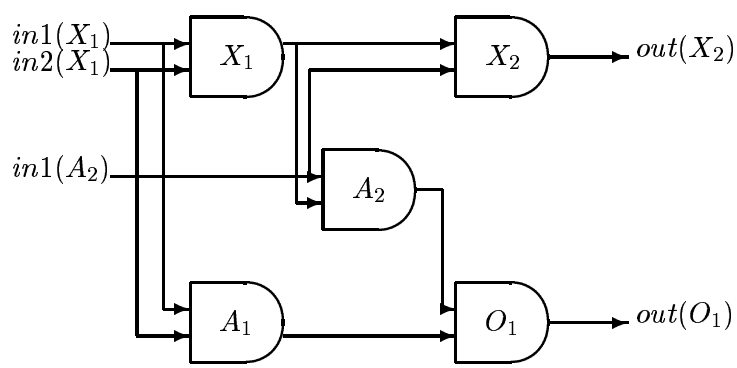



Figure 5.6: Electronic circuit of a full-adder

For the purpose of process design diagnosis, it is not sufficient to capture cases of unfulfilled obligations. This particularity of process design leads Ramos and Fiadeiro in [RF96b] to propose a more general diagnosis, one that distinguishes between potential, benevolent and exigent diagnosis. In order to deal with diagnoses that are not minimal, they extend the representation of obligations by assuming that norms are completely described. With this new approach, more

useful information can be obtained for process design, keeping at the same time all the result of model based reasoning. The following example of [RF96b] criticizes the principle of parsimony for organizational process design.

**Example 5.10 (Delivering order, continued)** Reconsider the first obligation: *'the invoice must go straight from the task fill-up-invoice to the task verify-IDO.'* If it is important that the invoice goes straight to the task verify-IDO, then a design that does not commit itself with the output of the invoice must be avoided. If the principle is not enforced, then it is possible that the invoice goes straight to the invoice file during the implementation of the process in the organization. To avoid this undesired situation, the diagnosis should alert the 'incompleteness' of the design. When it is important to ensure that all obligations are fulfilled, and not only to detect violations of obligations, the principle of parsimony is much too *benevolent*, because it is like the assumption of the fulfillment of obligations in the absence of information. In that case an approach based only on minimal diagnosis is not adequate and an *exigent* diagnosis is more suitable, where unfulfilled obligations are detected. □

To summarize, when Ramos and Fiadeiro interpret the system as a system of norms in DDD, there are two main problems with Reiter's theory of diagnosis.

1. Focus on the *minimal* sets of violations. The underlying assumption of 'innocent until proven guilty' is not always the right one; sometimes 'guilty until proven innocent' is preferred.

2. Lack of fault knowledge. So-called fault knowledge (see e.g. [dKMR90]) describes the consequences of broken components, in general $\beta \wedge Ab(c) \rightarrow \alpha$. Hence, with fault knowledge from the abnormality of a component new information can be derived. If the rules from the system description SD are represented by $\beta \wedge \neg Ab(c) \rightarrow \alpha$, like in Example 5.9, then there is no fault knowledge. In that case, the maximal diagnosis is simply the set of all components. Obviously, for any reasonable definition of a maximal diagnosis, fault knowledge has to be added.[5]

The definitions of diagnosis can easily be adapted such that these two problems are solved, as we show in the next section.

### The diagnostic framework for deontic reasoning DIODE

In this section we discuss the DIagnostic framework for DEontic reasoning DIODE in which deontic reasoning is formalized as a kind of diagnostic reasoning. The framework treats norms as components of a system to be diagnosed; hence the system description becomes a norms description. When a set of norms is translated to DIODE, the following two assumptions (see [RF96b]) are made to incorporate fault knowledge.

- **Assumption 1** As a rule, each (conditional) obligation of a premise set corresponds to a separate norm. Thus, a set of obligations is translated to a set of norms.

[5] As remarked in [dKMR90], with the representation of fault knowledge it is no longer possible to compute all consistent sets of normal and abnormal components based on minimal diagnosis, because not all supersets of minimal sets are consistent. In Reiter's minimal diagnosis that property holds.

- **Assumption 2** Every norm description completely describes an obligation. Thus, a conditional obligation '$\alpha$ should be the case if $\beta$ is the case' is represented in DIODE by the norm description $\neg V(n_i) \leftrightarrow (\beta \rightarrow \alpha)$. The conditional obligation can be read in DIODE as 'if the norm $n_i$ is not violated, then and only then if $\beta$ is the case then $\alpha$ is the case.' The sentence is logically equivalent with $V(n_i) \leftrightarrow (\beta \wedge \neg\alpha)$, which explains why we call $V(n_i)$ a *violation constant* (although, strictly speaking, it is not a constant).

We discriminate between minimal and maximal violated-norm sets. The definition of minimal violated-norm set is analogous to the definition of diagnosis. Just as we can have multiple diagnoses with respect to the same (SD, COMP, OBS), we can have multiple minimal violated-norm sets $\Delta$ with respect to (NORMS, ND, FACTS). The fact that we can have more than one minimal violation state reflects that we can have different situations that are optimal, i.e. as ideal as possible. A contextual obligation of a minimal violated-norm set corresponds to a measurement of a diagnosis and the implicit violated-norm set corresponds to Reiter's conflict set, see [RF96b].

**Definition 5.11 (DIODE)** Let $\mathcal{L}_V$ be the base logic of DIODE and the fragment of $\mathcal{L}_V$ without violation constants $\mathcal{L}$. We write $\models$ for entailment in $\mathcal{L}_V$. A *normative system* is a tuple NS = (NORMS, ND) with:

1. NORMS, a finite set of constants denoting *norms* $\{n_1, \ldots, n_k\}$,
2. ND, the *norms description*, a set of first-order $\mathcal{L}_V$ sentences denoting *conditional obligations* $\neg V(n_i) \leftrightarrow (\beta \rightarrow \alpha)$.

A *normative system to be diagnosed* is a tuple NSD = (NORMS, ND, FACTS) with:

1. NS = (NORMS, ND), a normative system, and
2. FACTS, a set of first-order $\mathcal{L}$ sentences that describe the facts.

Let NSD = (NORMS, ND, FACTS) be a normative system to be diagnosed. A *potential diagnosis* $\Delta$ of NSD is a subset of NORMS such that

$$\text{CONTEXT}_\Delta = \text{ND} \cup \text{FACTS} \cup \{V(n_i) \mid n_i \in \Delta\} \cup \{\neg V(n_i) \mid n_i \in \text{NORMS} - \Delta\}$$

is consistent. A *benevolent (exigent) diagnosis* $\Delta$ is a minimal (maximal) subset (with respect to set inclusion) of NORMS such that $\text{CONTEXT}_\Delta$ is consistent. The set of *contextual obligations* of a benevolent diagnosis $\Delta$ of NSD is

$$\text{CO}_\Delta = \{\alpha \mid \alpha \in \mathcal{L}, \text{CONTEXT}_\Delta \models \alpha\}$$

The *implicit violated-norm set* $\Delta$ of NSD is a minimal subset (with respect to set inclusion) of NORMS such that $\text{CONTEXT}_\Delta$ is inconsistent. □

The set of potential diagnoses can be ordered by set inclusion, of which the benevolent and exigent diagnosis are respectively the lower and upper bounds. Diagnostic reasoning is not restricted to the minimal elements of the graph, but to all elements. Moreover, for the benevolent diagnosis we have the additional information supplied by the implicit violated-norm sets and the contextual obligations. This is illustrated by the example of the delivering order in DIODE, see [RF96b] for a full discussion of this example in DIODE.

**Example 5.12 (Delivering order, continued)** Consider the following normative system:

1. NORMS = $\{n_1, n_2, n_3, n_4\}$, and

2. ND = $\{\neg V(n_1) \leftrightarrow a, \neg V(n_2) \leftrightarrow b, \neg V(n_3) \leftrightarrow c, \neg V(n_4) \leftrightarrow d\}$,

The potential diagnoses of FACTS = $\{b \wedge c \rightarrow e, \neg e\}$ and FACTS′ = $\{b \wedge c \rightarrow e, d \wedge \neg e\}$ are represented in Figure 5.7. We have FACTS $\models \neg b \vee \neg c$ and FACTS $\cup$ ND $\models V(n_2) \vee V(n_3)$. Moreover, we have FACTS′ $\cup$ ND $\models \neg V(n_4)$. In the latter case, there is one exigent diagnosis, $\{V(n_1), V(n_2), V(n_3)\}$, and two benevolent diagnosis, $\{V(n_2)\}$ and $\{V(n_3)\}$. The implicit violation set is $\{V(n_2), V(n_3)\}$, which means that either the second or the third norm has to be violated. □

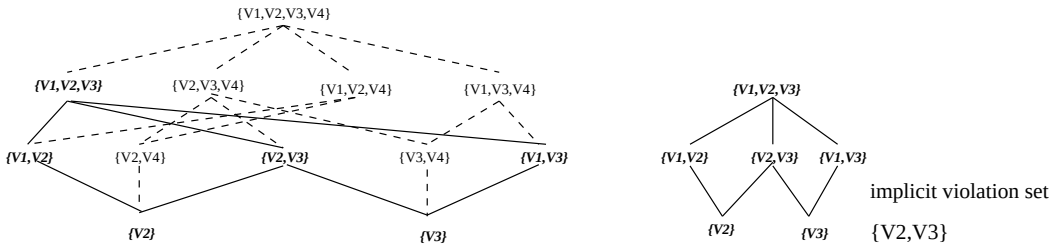



Figure 5.7: Consistent sets of violations

The following example of [RF96a] illustrates that the lack of conditional obligation is a problem of DIODE.

**Example 5.13** Consider that a designer has to follow the rule: *'if an order form is send to a supplier, then a copy of the order form must be send to the department store'* $O(c|o)$. If neither $o$ and $c$ nor their negations can be derived, then an exigent diagnosis contains the violation of the conditional obligation $O(c|o)$. Ramos and Fiadeiro argue that it is not an useful diagnosis, because it is not very intuitive to say that a conditional obligation is violated when the condition is not achieved. □

The diagnosis in Example 5.13 can be avoided if we only consider violations of *actual* obligations in a diagnosis. If 'an order form is send to a supplier' cannot be derived, then the violation will not appear in a diagnosis. In DIODE, there is no distinction between fulfilling a dyadic obligation, and inapplicability of a dyadic obligation. For example, for an obligation $O(\alpha \mid \beta)$ we have $\neg V(n) \leftrightarrow (\beta \rightarrow \alpha)$, i.e. $\neg V(n) \leftrightarrow (\neg\beta \vee (\beta \wedge \neg\alpha))$. Thus, fulfilling an obligation $O(\alpha|\beta)$ by $\beta \wedge \alpha$ and inapplicability $\neg\beta$ are both represented by $\neg V(n)$. We can add applicability information for an obligation $O(\alpha|\beta)$ with $\neg V(n) \leftrightarrow (\beta \rightarrow \alpha) \wedge A(n) \leftrightarrow \beta$. The additional information can be used to determine the applicable obligations by minimizing the $A(n)$. For applicable obligations we can have minimal or maximal violated-norm sets, as before.

**Definition 5.14 (DIODE with applicable norms)** A *normative system* is a tuple NS = (NORMS, ND) with:

1. NORMS, a finite set of constants denoting *norms* $\{n_1, \ldots, n_k\}$,

2. ND, the *norms description*, a set of first-order $\mathcal{L}_V$ sentences denoting *conditional obligations* $\neg V(n_i) \leftrightarrow (\beta \rightarrow \alpha) \wedge A(n_i) \leftrightarrow \beta$.

Let NSD = (NORMS, ND, FACTS) be a normative system to be diagnosed. The *active norms* $\Delta_a$ of NSD is a minimal subset of NORMS such that

$$\text{CONTEXT}_a = \text{ND} \cup \text{FACTS} \cup \{A(n_i) \mid n_i \in \Delta_a\} \cup \{\neg A(n_i) \mid n_i \in \text{NORMS} - \Delta_a\}$$

is consistent. A *potential diagnosis* $\Delta$ of NSD is a subset of some $\Delta_a$ of NSD such that

$$\text{CONTEXT}_\Delta = \text{CONTEXT}_a \cup \{V(n_i) \mid n_i \in \Delta_a\} \cup \{\neg V(n_i) \mid n_i \in \Delta_a - \Delta\}$$

is consistent. □

The following example illustrates the adaptation of DIODE.

**Example 5.15** Consider the normative system of the obligation $O(c|o)$ of Example 5.13.

1. NORMS = $\{n_1\}$,
2. ND = $\{(\neg V(n_1) \leftrightarrow (o \rightarrow c)) \wedge (A(n_1) \leftrightarrow o)\}$.

The set of active norms $\Delta_a$ is empty for FACTS = $\emptyset$, thus there is no potential diagnosis which contains the norm $n_1$. In particular, the only exigent diagnosis is the empty set. Moreover, consider the following normative system of the two obligations $O(p_1|q)$ and $O(p_2|\neg q)$.

1. NORMS = $\{n_1, n_2\}$,
2. ND = $\{(\neg V(n_1) \leftrightarrow (q \rightarrow p_1)) \wedge (A(n_1) \leftrightarrow q), (\neg V(n_2) \leftrightarrow (\neg q \rightarrow p_2)) \wedge (A(n_2) \leftrightarrow \neg q)\}$.

Given the tautology $q \vee \neg q$, we have for FACTS = $\emptyset$ two minimal active sets $\Delta_a = \{n_1\}$ and $\Delta_a = \{n_2\}$. Finally, consider the following normative system of the two obligations $O(p \mid q)$ and $O(q|\top)$.

1. NORMS = $\{n_1, n_2\}$,
2. ND = $\{(\neg V(n_1) \leftrightarrow (q \rightarrow p)) \wedge (A(n_1) \leftrightarrow q), (\neg V(n_2) \leftrightarrow q) \wedge (A(n_2) \leftrightarrow \top)\}$.

The minimal active set for FACTS = $\{\neg p\}$ is $\Delta_a = \{n_2\}$. □

The following definition shows a different solution to formalize conditional norms based on the concept of fulfilled obligations. A theory of diagnosis like DIODE is based on the distinction between violated and non-violated, and a (qualitative) decision theory is based on the distinction between fulfilled and unfulfilled goals. DIO(DE)$^2$ is the DIagnostic and DEcision-theoretic framework for DEontic reasoning. It combines reasoning about violated and fulfilled norms. Hence, it combines reasoning about the past (violated versus non-violated) with reasoning about the future (fulfilled versus not yet fulfilled). As illustrated in Figure 5.1, DIO(DE)$^2$ combines reasoning of a judge with reasoning of a rational agent. It is the extension of DIODE with the elements we discussed in Section 5.2: goal oriented reasoning, the distinction between parameters and decision variables, and uncertainty and strategies. Here we restrict ourselves

to the first extension. Goal-oriented reasoning is introduced by fulfilled-norm constants. For an obligation $O(\alpha|\beta)$ we have $\neg V(n) \leftrightarrow (\beta \rightarrow \alpha) \wedge F(n) \leftrightarrow (\beta \wedge \alpha)$. We minimize the applicable norms by minimizing the relation $(\Delta_f, \Delta_v) \leq (\Delta'_f, \Delta'_v)$.

**Definition 5.16 (**DIO(DE)$^2$**)** A *normative system* is a tuple NS = (NORMS, ND) with:

1. NORMS, a finite set of constants denoting *norms* $\{n_1, \ldots, n_k\}$,

2. ND, the *norms description*, a set of first-order $\mathcal{L}_V$ sentences denoting *conditional obligations* $\neg V(n_i) \leftrightarrow (\beta \rightarrow \alpha) \wedge F(n) \leftrightarrow (\beta \wedge \alpha)$.

Let NSD = (NORMS, ND, FACTS) be a normative system to be diagnosed. A *fulfilled-violated set* $(\Delta_f, \Delta_v)$ of NSD is a pair of subsets of $norms$ such that

$$\begin{aligned} \text{CONTEXT}_\Delta = \text{ND} \cup \text{FACTS} \quad & \cup \{V(n_i) \mid n_i \in \Delta_v\} \cup \{\neg V(n_i) \mid n_i \in \text{NORMS} - \Delta_v\} \\ & \cup \{F(n_i) \mid n_i \in \Delta_f\} \cup \{\neg F(n_i) \mid n_i \in \text{NORMS} - \Delta_f\} \end{aligned}$$

is consistent. Consider the ordering on fulfilled-violated sets $(\Delta_f, \Delta_v) \leq (\Delta'_f, \Delta'_v)$ iff $\Delta_f \subseteq \Delta'_f$ and $\Delta_v \subseteq \Delta'_v$. A *potential diagnosis* $(\Delta_f, \Delta_v)$ of NSD is a pair of subsets of NORMS that is minimal in the ordering. □

The following example illustrates the adaptation of DIO(DE)$^2$ and compares it with DIODE with applicable norms.

**Example 5.17** Consider the following normative system of the obligation $O(c \mid o)$ of Example 5.13.

1. NORMS = $\{n_1\}$,

2. ND = $\{(\neg V(n_1) \leftrightarrow (o \rightarrow c)) \wedge (F(n_1) \leftrightarrow (c \wedge o))\}$

The unique potential diagnosis for FACTS = $\emptyset$ is $(\Delta_f, \Delta_v) = (\emptyset, \emptyset)$. In DIODE, the set of active norms $\Delta_a$ is empty for FACTS = $\emptyset$. Hence, the two systems behave similarly. Moreover, consider the following normative system of the two obligations $O(p_1|q)$ and $O(p_2|\neg q)$.

1. NORMS = $\{n_1, n_2\}$,

2. ND = $\{ \begin{array}{c} (\neg V(n_1) \leftrightarrow (q \rightarrow p_1)) \wedge (F(n_1) \leftrightarrow (p_1 \wedge q)), \\ (\neg V(n_2) \leftrightarrow (\neg q \rightarrow p_2)) \wedge (F(n_2) \leftrightarrow (p_2 \wedge \neg q))\} \end{array}$ .

The potential diagnoses $(\Delta_f, \Delta_v)$ for FACTS = $\emptyset$ are $(\{n_1\}, \emptyset)$, $(\{n_2\}, \emptyset)$, $(\emptyset, \{n_1\})$ and $(\emptyset, \{n_2\})$. In DIODE, we have for FACTS = $\emptyset$ two minimal active sets $\Delta_a = \{n_1\}$ and $\Delta_a = \{n_2\}$. Hence, the two systems behave again similarly. Finally, consider the following normative system of the two obligations $O(p|q)$ and $O(q|\top)$.

1. NORMS = $\{n_1, n_2\}$,

2. ND = $\{(\neg V(n_1) \leftrightarrow (q \rightarrow p)) \wedge (F(n_1) \leftrightarrow (p \wedge q)), (\neg V(n_2) \leftrightarrow q) \wedge (F(n_2) \leftrightarrow \neg q)\}$.

The potential diagnoses for FACTS = $\{\neg p\}$ are $(\Delta_f, \Delta_v) = (\emptyset, \{n_2\})$ and $(\Delta_f, \Delta_v) = (\{n_2\}, \{n_1\})$. In DIODE, the minimal active set for FACTS = $\neg p$ is $\Delta_a = \{n_2\}$. The two systems do not behave similarly, because in DIO(DE)$^2$ it is possible that the first obligation is violated. □

There is an interesting connection between the latter set of obligations of Example 5.17 and deontic detachment. With deontic detachment we can derive the obligation $O(p \mid \top)$ from the two premises $O(p \mid q)$ and $O(q \mid \top)$. Thus, if deontic detachment is valid, then the fact $\neg p$ is a violation. In DIODE, there is only one active set, that contains the second obligation. It is possible that this obligation is fulfilled, and there are therefore no violations. On the other hand, in DIO(DE)$^2$ every potential diagnosis contains violations.

### 5.3.3 Deontic logic as the basis of diagnosis

In [RF96a] Ramos and Fiadeiro show how deontic logic can be used in a theory of diagnosis. Before we discuss their logic, we mention a well-known relation between DIODE and deontic logic. Anderson [And58] showed that Standard Deontic Logic (SDL), see Definition 1.6, can be expressed in alethic modal logic by $O\alpha =_{def} \Box(\neg V \rightarrow \alpha)$, in which $V$ is the so-called violation constant (not a propositional variable!), together with the axiom **D**: $\Diamond\neg V$ (as usual, $\Diamond\alpha =_{def} \neg\Box\neg\alpha$). In SDL, a conditional obligation can be represented by $\beta \rightarrow O\alpha$ or by $O(\beta \rightarrow \alpha)$. The latter is according to the Anderson schema $O(\beta \rightarrow \alpha) =_{def} \Box(\neg V \rightarrow (\beta \rightarrow \alpha))$. The similarity of Anderson's reduction with Reiter's theory of diagnosis is obvious, but there are also two important distinctions. First, in Anderson's reduction every deontic formula is preceded by a box □. Semantically, in the theory of diagnosis distinct *models* represent distinct situations, whereas in a modal system distinct *worlds* within a model represent distinct situations. Second, in Anderson's reduction there is only one violation constant. This means that conflicting obligations are inconsistent with axiom $\Diamond\neg V$. For example, $Op \wedge O\neg p$ is equivalent to $\Box(\neg V \rightarrow p) \wedge \Box(\neg V \rightarrow \neg p)$, from which $\Box V$ can be derived. However, in design there often are conflicting principles. Ramos and Fiadeiro [RF96a] criticize normal modal logics like SDL and Anderson's reduction of SDL with the following example.

**Example 5.18** Consider axiom **K**: $O(\beta \rightarrow \alpha) \rightarrow (O\beta \rightarrow O\alpha)$. Assume the two following rules: *(i) 'the task prepare-budget must be assigned to Ann or to John,'* and *(ii) 'the task prepare-budget must not be assigned to Ann.'* Moreover, assume that the designer assigns the task to Ann (and not to John). A diagnosis should only report the violation of the obligation (ii). However, due to axiom **K** the violation of the obligation *'assign the task to John'* will be inferred.[6] □

Ramos and Fiadeiro follow Chellas [Che74, Che80] to weaken the modal logic $O$ to non-normal models. In normal Kripke models, the accessibility relation relates a world to one set of worlds (the ideal alternatives). In non-normal (or minimal) models, the accessibility relation relates a world to a set of sets of worlds, where each set of worlds represents a distinct norm. For example, consider the two models of $\{Oa, Ob\}$ in Figure 5.8. The normal model only sees $a \wedge b$ worlds, whereas the non-normal model sees a cluster of $a$ worlds, and a cluster of $b$ worlds.

[6] The example is an instance of the apples-and-pears example, see Section 3.1.

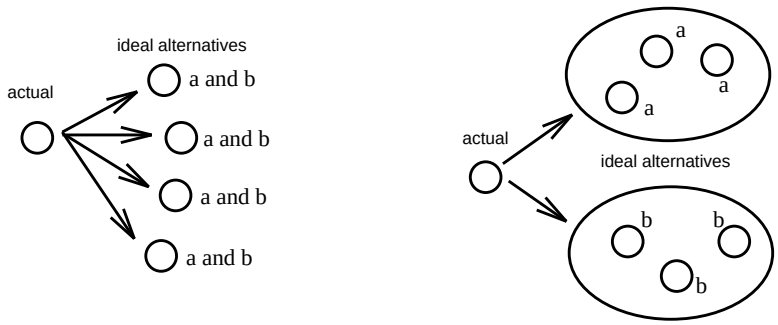



Figure 5.8: Models of $\{Oa, Ob\}$

Chellas' minimal *deontic* logic MDL does not have axiom **K**, but it still has weakening. Hence, we have $Op \to O(p \vee q)$ for any $q$: from an obligation an infinite set of obligations can be derived. Ramos and Fiadeiro argue against weakening with the following example.

**Example 5.19** Consider the rule of inference 'if $\vdash \beta \to \alpha$ then $O\beta \to O\alpha$.' Assume the following rule: *if one agent sends a document to other, then the second receives the document.'* Furthermore, assume the obligation *'Ann is obliged to send a budget to John'* and the fact *'Ann does not send the budget'* (thus John does not receive it). Given the rule, the diagnosis will report two violations. However only one violation really occurs. □

A simple logic without weakening can be defined as follows.

**Definition 5.20 (Minimal logic)** Consider a bimodal logic with $\Box$ and $I$. The logic is the smallest set of formulas that contains the propositional tautologies and the following axioms and is closed under the following rules of inference.

| | | | | |
|---|---|---|---|---|
| **MP**: $\frac{\alpha \to \beta, \alpha}{\beta}$ | **Nes**: $\frac{\vdash \alpha}{\vdash \Box \alpha}$ | **RE**: $\frac{\alpha_1 \leftrightarrow \alpha_2}{I\alpha_1 \leftrightarrow I\alpha_2}$ | $\Box\alpha \to I\alpha$ | |
| **K**: $\Box(\beta \to \alpha) \to (\Box\beta \to \Box\alpha)$ | **T**: $\Box\alpha \to \alpha$ | $\neg I\bot$ | $\neg I\top$ | □ |

**Definition 5.21 (Semantics minimal logic)** Kripke models $M = \langle W, R_1, R_2, V\rangle$ with $W$ a set of worlds, $R_1(w, w')$ a binary reflexive accessibility relation, $R_2(w, W')$ an accessibility relation that gives a non-empty set of set of worlds ($\neq W$) for each world (we write either $R_1(w, w')$ and $R_2(w, W')$, or $w' \in R_1(w)$ and $W' \in R_2(w)$), such that for all $W' \in R_2(w)$ we have $W' \subset R_1(w)$, and $V$ a valuation function for the propositions in the worlds. We have:

$M, w \models \Box p$ iff $\forall w' \in R_1(w)$ we have $M, w' \models p$
$M, w \models Ip$ iff $\exists W' \in R_2(w)$ such that $W' = \{w' \in R_1(w) \mid M, w' \models p\}$. □

In Definition 5.26 we use the minimal logic to define Ramos and Fiadeiro's logic, just like we used the modal preference logic CT4O to define our dyadic deontic logics in Chapter 2. Ramos and Fiadeiro use dyadic deontic logic. They read a dyadic obligation $O(\alpha \mid \beta)$ as '$\alpha$ is obligatory in the context $\beta$.' They consider two components of a conditional obligation. First, the

*design action* that indicates what the designer should do. Second, *the context* (condition) that describes the situation in which the design action should be done. The distinction is analogous to the distinction between decision variables and parameters in a qualitative decision theory, see Section 5.2. The two components must be explicitly represented, because the distinction is crucial in their approach. Moreover, they restrict the scope of the deontic operator to design actions. They assume that it makes no sense to have obligations that oblige a process designer to act in the structure of the organization. It follows from that assumption that, whatever the context, any obligation where the action is represented by a structural concept is not valid. This is illustrated by the following example.

**Example 5.22** The rule *'if the task approve-budget is assigned to John, then John must be the Head of Department (HD)'* is not valid, because John being or not being the HD is not an design action (it is a part of the structure of the organization). The rule should be: *'if John is not the HD, then he cannot be assigned to the task approve-budget.'* □

Ramos and Fiadeiro make a formal distinction between structural and design action variables. The basic idea is the following, inspired by Castaneda's distinction between assertions and actions [Cas81]. The modal language of deontic logic gives the opportunity – not present in Reiter's first order theory of diagnosis – to distinguish between structural variables which are fixed *within a model*, and design action variables which are allowed to vary within the model. For a structural variable $p$, we have $\Box p \vee \Box\neg p$ and therefore $p \rightarrow \Box p$: if the structural variable is true in the actual world, it is true in all worlds see from the actual world (because the accessibility relation related to $\Box$ is reflexive). Notice that $\Box p$ should be read as $p$ is a structural concept, not as $p$ is necessarily true (as in Anderson's proposal). The distinction is illustrated by the following logic for diagnosis LD.

**Definition 5.23 (**LD**)** The logic LD is a minimal logic as defined in Definition 5.20, extended with the following definitions.

$$\begin{array}{lll} struct(\alpha) & =_{def} & \Box\alpha \vee \Box\neg\alpha \\ O(\alpha|\beta) & =_{def} & I(\beta \rightarrow \alpha) \wedge \neg struct(\alpha) \\ O_a\alpha & =_{def} & I(\alpha|\top) \end{array}$$

□

The following proposition shows some typical theorems of the logic LD. The two theorems $O\alpha \rightarrow \Diamond\alpha$ and $O\alpha \rightarrow \Diamond\neg\alpha$ show that the restriction of the consequent to design actions (the clause $\neg struct(\alpha)$) is a formalization of von Wright's contingency clause.

**Proposition 5.24** *The logic* LD *has the following theorems.*

$$\begin{array}{l} O\alpha \rightarrow \Diamond\alpha \\ O\alpha \rightarrow \Diamond\neg\alpha \\ \neg O(\top|\beta) \\ \neg O(\bot|\beta) \\ (\beta \wedge struct(\beta) \wedge O(\alpha|\beta)) \rightarrow O_a\alpha \\ (struct(\alpha_2) \wedge O(\alpha_1 \wedge \alpha_2|\beta)) \rightarrow O(\alpha_1|\beta) \end{array}$$

**Proof** *The theorem* $\beta \wedge struct(\beta) \wedge O(\alpha|\beta) \rightarrow O_a\alpha$ *follows from* $\alpha \wedge struct(\alpha) \rightarrow \Box\alpha$.

Ramos and Fiadeiro [RF96a] argue that the logic LD is too weak, which they illustrate with the following example. According to their desiderata, also the systems DIODE with applicable norms in Definition 5.14 and DIO(DE)$^2$ are too weak.

**Example 5.25** Reconsider the rule *'if the task approve-budget is assigned to the HD, then he cannot be assigned to the task prepare-budget.'* If in the design the tasks approve-budget and prepare-budget are not assigned to anyone yet, then an exigent diagnosis should not report the unfulfillment of that obligation, because it will be an extremely exigent diagnosis. However, if the task prepare-budget is assigned to the HD and the designer has not committed himself about the task approve-budget, then an exigent diagnosis should report that situation. □

A conditional obligation should only be considered in an exigent diagnosis if (i) the condition is true or (ii) the action is not performed and the condition is not fixed. For the second case, we add the axiom $O(\alpha|\beta) \rightarrow O(\neg\beta|\neg\alpha)$. As a consequence only situation (i) has to be considered. In [RF96a] the following logic LDD is proposed.

**Definition 5.26 (LDD)** The logic LDD is a minimal deontic logic as defined in Definition 5.20, extended with the following definitions.

$$\begin{array}{lll} struct(\alpha) & =_{def} & \Box\alpha \vee \Box\neg\alpha \\ O(\alpha|\beta) & =_{def} & I(\alpha \leftrightarrow \beta) \wedge \neg I(\neg\alpha \wedge \beta) \wedge \neg struct(\alpha) \end{array}$$ □

We think the logic LDD is best analyzed in terms of $F$ and $V$ predicates. The definitions of LDD can be decomposed with the following definitions.

$$\begin{array}{lll} F(\alpha|\beta) & =_{def} & I(\alpha \leftrightarrow \beta) \\ V(\alpha|\beta) & =_{def} & \neg I \neg(\beta \rightarrow \alpha) \\ O(\alpha|\beta) & =_{def} & F(\alpha|\beta) \wedge V(\alpha|\beta) \wedge \neg struct(\alpha) \end{array}$$

We can analyze properties of the logic LDD by analyzing the properties of the definitions $F$ and $V$. The definition of fulfilled norm is different in LDD than in DIO(DE)$^2$. In LDD a norm $O(\alpha|\beta)$ is fulfilled iff $(\alpha \leftrightarrow \beta)$, which is equivalent to $(\alpha \wedge \beta) \vee (\neg\alpha \wedge \neg\beta)$. Hence, in LDD not only $\alpha \wedge \beta$ but also $\neg\alpha \wedge \neg\beta$ counts as fulfillment of the norm $O(\alpha \mid \beta)$. The following proposition shows that LDD has the properties Ramos and Fiadeiro desire.

**Proposition 5.27** *The logic* LDD *has the following theorem.*

$$(\neg struct(\beta) \wedge O(\alpha|\beta)) \rightarrow O(\neg\beta|\neg\alpha)$$

*The logic* LDD *does not have the following theorem.*

$$(\neg struct(\beta \rightarrow \alpha) \wedge O(\alpha|\beta)) \rightarrow O(\beta \rightarrow \alpha|\top)$$

**Proof** *The definitions of fulfilled norms* $F(\alpha|\beta) =_{def} I(\alpha \leftrightarrow \beta)$ *and violated norms* $V(\alpha|\beta) =_{def} \neg I \neg(\beta \rightarrow \alpha)$ *have the following properties.*

$F(\alpha|\beta) \leftrightarrow F(\neg\beta|\neg\alpha)$
$F(\alpha|\beta) \leftrightarrow F(\beta|\alpha)$
$F(\alpha|\beta) \leftrightarrow F(\neg\alpha|\neg\beta)$
$F(\alpha|\beta) \leftrightarrow F(\beta \leftrightarrow \alpha|\top)$
$V(\alpha|\beta) \leftrightarrow V(\neg\beta|\neg\alpha)$
$V(\alpha|\beta) \leftrightarrow V(\beta \rightarrow \alpha|\top)$

*The theorem follows from* $F(\alpha|\beta) \leftrightarrow F(\neg\beta|\neg\alpha)$ *and* $V(\alpha|\beta) \leftrightarrow V(\neg\beta|\neg\alpha)$. *The non-theorem follows from lack of* $F(\alpha|\beta) \leftrightarrow F(\beta \rightarrow \alpha|\top)$.

The logic LDD is used for deontics-based diagnosis.

**Definition 5.28 (Deontics-based diagnosis)** An obligation system to be diagnosed is a tuple OSD = (OBL, FACTS, STRUCT) with:

1. OBL, a finite set of modal sentences denoting conditional obligations $O(\alpha|\beta)$,
2. FACTS, a finite set of propositional sentences, and
3. STRUCT, a set of expressions denoting which variables are structural $\Box p \vee \Box\neg p$.

The *actual obligation set* is the set of obligations (without logical equivalents):

$$\text{AO} = \{O_a\alpha \mid \text{OBL} \cup \text{FACTS} \cup \text{STRUCT} \models O(\alpha|\beta) \wedge \beta\}$$

A potential diagnosis $\Delta$ is a subset of AO such that

$$\text{CONTEXT}_\Delta = \text{OBL} \cup \text{FACTS} \cup \text{STRUCT} \cup \{\neg\alpha \mid O_a\alpha \in \Delta\} \cup \{\alpha \mid O_a\alpha \in \text{AO} - \Delta\}$$

is consistent. □

### 5.3.4 Discussion

The system DIO(DE)$^2$ is the extension of DIODE with elements of qualitative decision theory (QDT). We think that the extension of the diagnostic framework DDD proposed by Ramos and Fiadeiro can be considered as a qualitative decision theory too. The first similarity between DDD and logics proposed for QDT is that design actions can be understood as decision variables. Moreover, the exigent diagnosis considers the future, because it reports violations *not yet fulfilled*. Hence, it has a perspective that considers the future. As we argued in Section 5.1, the view on the future is exactly the distinction between a theory of diagnosis and qualitative decision theory. This can also be seen in the architecture of Ramos and Fiadeiro's framework in Figure 5.4. This particular theory of diagnosis can be considered as a theory of decision support, because it is used in a feed back loop.

DIO(DE)$^2$ corresponds to deontics-based diagnosis based on the modal logic 2DL. That is, DIO(DE)$^2$ corresponds to deontics-based diagnosis in Definition 5.28, where the logic LDD is replaced by 2DL. The correspondence follows directly from the semantics. A dyadic obligation $O(\alpha|\beta)$ in DIO(DE)$^2$ is a preference of $\alpha \wedge \beta$ (a fulfilled norm) over $\neg\alpha \wedge \beta$ (a violated norm). In

DIO(DE)$^2$ this preference is defined in two steps: in the base language the fulfilled and violated norm constants are defined, and in the definition of potential diagnosis the set of applicable norms is minimized. In 2DL, the preference is not represented by fulfilled and violated norm constants, but they are defined directly in the semantics. With other words, DIO(DE)$^2$ is the deontic logic 2DL in which certain aspects (fulfillments and violations) are made explicit with the use of a naming convention, i.e. to use names $n_i$ to denote norms.

There are two important similarities between the logic LDD and the ordering logic of 2DL. The most important property is the lack of weakening of the consequent. This property is essential for diagnosis, as illustrated in example 5.19. The desirability of this property also shows that most deontic logics *cannot* be used for diagnosis, because most logics have weakening (see the discussion in Section 1.3.2). The second property of 2DL is $O(\alpha \mid \beta) \rightarrow \Diamond(\alpha \wedge \beta)$ and $O(\alpha \mid \beta) \rightarrow \Diamond(\neg\alpha \wedge \beta)$. Thus LDD as well as 2DL have consistency checks to model the contingency clause. A distinction is that the contingency clause in LDD only checks the consequent, e.g. $\overset{\leftrightarrow}{\Diamond}\alpha$, whereas the contingency clause of 2DL considers antecedent and consequent, e.g. $\overset{\leftrightarrow}{\Diamond}(\alpha \wedge \beta)$. Finally, we observe a similarity between the two-phase deontic logic 2DL and logics proposed for diagnosis, see e.g. [BB95]. The latter logics are also preference-based.

## 5.4 Conclusions

In this chapter we considered topics for further research. We have established the following results.

1. We discussed the relation between qualitative decision theory and deontic logic by discussing two logics of qualitative decision theory. Pearl's logic of pragmatic obligation, a decision-theoretic account of deontic logics, is written as a criticism of deontic logic. However, we showed that this logic of pragmatic obligation is a *defeasible* deontic logic and should not be compared to the traditional deontic logics. Secondly, Boutilier incorporates deontic logic in his approach to qualitative decision theory.

2. We discussed the relation between diagnosis and deontic logic by examing Ramos and Fiadeiro's approach to diagnosis for organizational process design DDD. This framework is built on top of our DIagnostic framework for DEontic reasoning DIODE. We argued that their extension of the diagnostic theory is best understood as a qualitative decision theory.

Moreover, these two applications show why deontic logic is a useful knowledge representation language. In Section 1.2 we argued that knowledge concerning for example legal code, library regulations and trade procedures can be represented in deontic logic. In this chapter we showed that such a formalization can be used for diagnosis and planning. However, not every deontic logic can be used for these applications. For example, Ramos and Fiadeiro argue that the deontic logic used for diagnosis should not have weakening (and they therefore developed the deontic logic LDD). The preference-based deontic logics developed in this thesis can be used, because the preferences of for example 2DL coincide with DIO(DE)$^2$.

The discussion in this chapter illustrates several interesting issues for further research. The relation between deontic logic and qualitative decision theory (QDT) is described by a structural similarity with different perspectives.

1. **Structural similarity** There is a structural similarity between deontic logic and logics for QDT [Bou94b, Lan96, TH96], because both are based on preferences. The use of preferences is introduced in deontic logic by B. Hansson [Han71] and Lewis [Lew74] to deal with contrary-to-duty paradoxes. In QDT, preferences are used for context-sensitive goals [DW91b, Bou94b].

2. **Different perspectives** The main purpose of a deontic logic is deriving new obligations (and permissions) from an initial specification, while QDT focuses on the search for optimal acts and decisions [Lan96]. Deontic logic and a logic for QDT have different perspectives. Obligations are exogenous (they are imposed by a legal or moral code) while desires in logics for QDT are endogenous (coming from the agent) [Lan96]. It is this distinction which we call *the gap between deontic logic and qualitative decision theory*.

The structural similarity suggests that deontic logic can be used in a qualitative decision theory. However, as a consequence of the two different perspectives we first have to bridge the gap between deontic logic and QDT. One of the main features of deontic logic is the fact that actors do not always obey the law. Indeed, it is precisely when a forbidden act occurs, or an obligatory action does not occur, that we need the machinery of deontic logic, to detect a violation and to take appropriate action. For purposes of planning, it is often useful to assume that actors *do* obey the law. McCarty [McC94b] calls this the *causal assumption*, since it enables us to 'predict the actions that *will* occur by reasoning about the actions that *ought* to occur.' If we adopt the causal assumption, we can use the machinery of deontic logic to reason about the physical world [McC94b]. Moreover, preferences can be used to express the gains and losses consequent to normative decisions. Once agents know the losses and gains of their decisions, the normative decision is no longer a problem. Interesting questions are: how do agents work out norms in terms of gains and losses? What are the gains of observing norms? How do they learn the effects of norms and how do they reason about these effects? In which way does a normative decider differ from an ordinary decider, if any?

# Chapter 6

# Conclusions

The main objective of this thesis is the development of deontic logics that formalize contrary-to-duty reasoning and overridden structures. Moreover, the secondary objectives are the explanation of defeasibility in preference-based deontic logics and a classification of the different types of defeasibility in defeasible deontic logic. The classification is necessary to avoid confusion between violations and exceptions. The success test that these objectives are fulfilled is the capacity of the developed logics of dealing with different types of problems, the deontic puzzles usually referred to as deontic paradoxes. In this chapter, we summarize the results established in this thesis. We review the developed logics, we discuss the relation between obligations and defeasibility, and we discuss the different types of problems and their solutions. Finally, we reconsider the applications that motivated our research.

## 6.1 Obligations and preferences

The main objective of this thesis is the development of deontic logics that formalize contrary-to-duty reasoning. In this section we review the logics developed in this thesis. A deontic logic formalizes either obligations that cannot be overridden, or obligations that can be overridden by other obligations. The latter category is sometimes *loosely* called 'defeasible deontic logic' or the 'logic of prima facie obligations.' In this thesis, we developed several preference-based deontic logics of the first and the second category. Of the first category, we developed the ordering logic, two-phase deontic logic 2DL, and contextual deontic logic CDL. Moreover, we developed labeled deontic logic LDL, a simple logic that lacks a semantics. Of the second category we developed contextual deontic logic with overridden defeasibility based on multi preference semantics and we discussed how a prioritized contextual deontic logic can be developed.

We first discuss the two different interpretations of circumstances, based on the distinction between the contextual and the conditional interpretation of the antecedent of dyadic obligations. Then we review the logics developed in this thesis. Finally, we discuss the relation between preferences, actions and time. During our review, we also mention some issues for further research. In particular, we discuss how the preference-based deontic logics can be extended with a temporal notion in update semantics.

### 6.1.1 Circumstances

To place the logics developed in this thesis in deontic logic literature, we started in Chapter 1 with a classification of deontic logics. The most important classification property of *dyadic* deontic logic is the interpretation of circumstances. There are two types of dyadic deontic logics, dependent on how the antecedent of the obligations is interpreted. The contextual interpretation of the antecedent assumes that the antecedent are circumstances which are considered to be fixed, and the consequent refers to the optimal states given these circumstances. Circumstances can be fixed due to external circumstances, or because agents regard them as being settled in determining what to do. The Hansson-Lewis dyadic deontic logics give a contextual interpretation of the antecedent of a dyadic obligation, in contrast to the conditional interpretation of the Chellas-type of dyadic deontic logics. In the semantics this distinction is clear. The contextual interpretation corresponds with the optimality principle, whereas the conditional interpretation corresponds with the ideality principle. The optimality principle corresponds with the semantic concept of zooming in on the ordering. To evaluate a dyadic obligation $O(\alpha|\beta)$, only $\beta$ worlds are considered. Hence, we zoom in on the $\beta$ worlds, and $\neg\beta$ worlds are outside the scope. In the proof theory, the distinction is much less clear. Usually, the dyadic deontic logics with a contextual interpretation have $\neg O(\neg\alpha \mid \alpha)$ as a theorem, whereas dyadic deontic logics with a conditional interpretation do not. This follows from the distinction between the optimality and the ideality principle: the optimal $\alpha$ cannot be $\neg\alpha$, whereas the ideal alternatives of $\alpha$ can be $\neg\alpha$. However, we also showed in Section 4.3.4 that the definitions of a dyadic deontic logic with a contextual interpretation can be adapted such that $O(\neg\alpha|\alpha)$ represents a violation.

The distinction between the two types of dyadic deontic logics also manifests itself in theories that can formalize reasoning with obligations. In Chapter 5, we discuss (qualitative) decision theory and diagnosis. Qualitative decision theory describes how obligations influence behavior and a theory of diagnosis reasons about violations. The two different perspectives of these applications are represented in Figure 6.1, a copy of Figure 5.1. The figure illustrates that a (qualitative) decision theory reasons about the future, whereas a theory of diagnosis reasons about the past with incomplete knowledge (if everything is known than a diagnosis is completely known). A qualitative decision theory is based on the optimality principle, whereas the theory of diagnosis is based on the ideality principle. The typical application of qualitative decision theory is robot planning. Robot planning is based on the optimality principle, because a robot is not bothered by missed opportunities in the past.

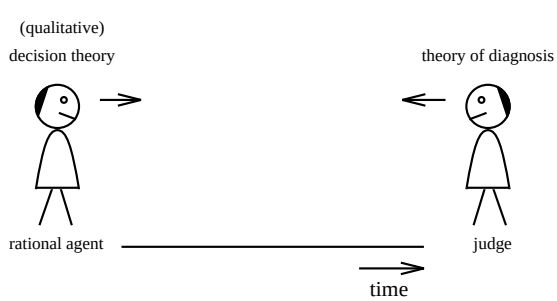



Figure 6.1: Reasoning with norms

Finally, the distinction between contextual and conditional interpretation also appears in temporal deontic logics. The distinction between the perspective of a rational agent (qualitative de-

cision theory) and a judge (theory of diagnosis) corresponds to Thomason's distinction between the context of deliberation and the context of justification [Tho81], see Section 1.3.4. Thomason observes that truth values of deontic sentences are dependent of a set of choices or future options that varies as a function of time. The context of deliberation is the set of choices when you are looking for practical advice, whereas the context of justification is the set of choices for someone who is judging you. It is important to discriminate between these two contexts, because a sentence can sometimes be interpreted differently in each of them. The consequence of the fact that there are two interpretations of the same deontic formula is that there are two distinct types of obligations. Let us call the obligations $O_j$ and $O_d$, respectively. The sentence 'you smoke and you should not smoke' $s \wedge O_j \neg s$ means the identification of the fact that you are violating a rule, whereas $s \wedge O_d \neg s$ means that you should stop smoking. The two perspectives are represented in Figure 6.2. a copy of Figure 5.3. At the present moment in time, $s$ is true. The context of justification considers the moment before the truth value of $s$ was settled, and considers whether at that moment in the past, $\neg s$ was preferred over $s$. The context of deliberation considers the moment the truth value of $s$ can be changed, and considers whether at that moment in the future, $\neg s$ will be preferred over $s$. The two types of obligations are independent: we can have $s \wedge O_j \neg s \wedge \neg O_d \neg s$ as well as $s \wedge \neg O_j \neg s \wedge O_d \neg s$.

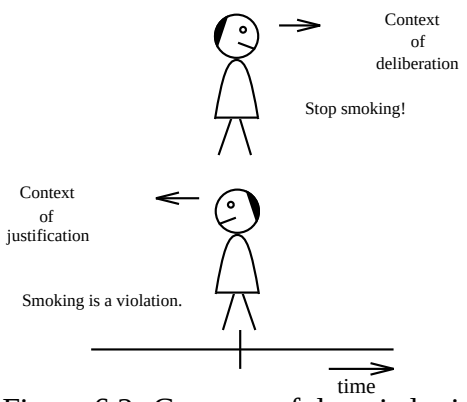



Figure 6.2: Contexts of deontic logic

The distinction between $s \wedge O_j \neg s$ and $s \wedge O_d \neg s$ is as important as the distinction between Alchourrón-Gärdenfors-Makinson belief revision (or theory revision) [AGM85] and Katsuno-Mendelzon belief update [KM92] in the area of logics of belief. There is a strong analogy, because belief revision is reasoning about a non-changing world and update is reasoning about a changing world. It follows directly from Figure 6.2 that a similar distinction is made between respectively the context of justification and the context of deliberation, because the past is fixed, whereas the future is wide open.

### 6.1.2 Preference-based deontic logic

In this section we reconsider the main properties of the preference based deontic logics developed in this thesis. The two most remarkable relations between obligations and preferences are the dynamic interpretation of obligations and the bipolarity related to deontic choice. Moreover, we mention several possible extensions of our logics, like a first-order base language and the representation of '$\alpha$ ought to be (done) if $\beta$ is (done) and it is a violation.'

We discuss the properties of our deontic logic $O_D$, see section 2.3. It is based on a contextual interpretation of the antecedent of the dyadic obligations, like all other logics developed in this thesis, because logics with a conditional interpretation cannot formalize all types of contrary-to-duty reasoning, see Chapter 2. For example, consider our deontic logic $O_D$, see Section 2.3.3. The logic $O_D(\alpha\,|\,\beta) =_{def} O(\alpha\,|\,\beta) \wedge O_\forall(\alpha\,|\,\beta)$ with preferential entailment is quite complicated, because it consists of the following three elements.

1. **Ordering logic.** The ordering logic $O$ is the basic mechanism underlying the logic. An obligation $O(\alpha\,|\,\beta)$ distinguishes between $\alpha \wedge \beta$ and $\neg\alpha \wedge \beta$ worlds. This distinction is represented by the constraint '$w_2 \not\leq w_1$ for all $w_1$ and $w_2$ such that $M, w_1 \models \alpha \wedge \beta$ and $M, w_2 \models \neg\alpha \wedge \beta$.' Proposition 2.22 shows that an ordering obligation can be understood as a set of existential-minimizing obligations $O_\exists$.

2. **Minimizing logic.** The universal-minimizing logic $O_\forall$ is *only* used to make dilemmas (conflicts between obligations) inconsistent. The maximally connected models of an ordering obligation $O(\alpha|\beta)$ are the same as the maximally connected models of $O_D(\alpha|\beta)$, if such orderings exist.

3. **'Maximally connected'.** Preferential entailment based on 'maximally connected' models has several advantages over other popular schemes like 'gravitating towards ideality' or 'towards the center,' as shown by the speed limits example in Section 2.6.4.

We axiomatized the preference-based obligations in a modal preference logic. This technique was introduced by Boutilier [Bou92a] and Lamarre [Lam91] for minimizing conditionals. Advantages of this approach are that it facilitates comparison with other systems, it gives a simple axiomatization, and it enables easy extensions like for example the $c$ and $cc$ conditions. In contrast to Boutilier we defined the operators in two steps. First we defined preference relations between sets of worlds, i.e. between propositions $\alpha_1 \succ \alpha_2$, and secondly we used the preference ordering to define obligations and permissions with $O(\alpha\,|\,\beta) =_{def} (\alpha \wedge \beta) \succ (\neg\alpha \wedge \beta)$. This makes some propositions easier to read, for example Propositions 2.6 and 2.21.

The logic $O_D$ solves the first three problems mentioned in Section 1.4.2, which we studied in Chapter 2. This is discussed in detail in Section 6.3. Von Wright's strong preference problem is solved by the ordering logic, which uses negative preferences ($\neg\alpha$ is not preferred to $\alpha$) instead of positive preferences ($\alpha$ is preferred to $\neg\alpha$), the contrary-to-duty problem is solved by the dyadic representation and the dilemma problem is solved by the combination of the ordering and the minimizing logic. The obligations $O_D$ do not have factual detachment, like the Hansson-Lewis logics that are based on a contextual interpretation of the antecedent. However, it also differs in several respects from these logics. The obligations $O_D$ do not have weakening of the consequent, but they have (restricted) strengthening of the antecedent. In the latter respect, it resembles the Chellas-type logics that give a conditional interpretation of the antecedent. As far as we know, $O_D$ is the first dyadic deontic logic with a contextual interpretation that has strengthening of the antecedent (and the first dyadic deontic logic that lacks weakening of the consequent).

The logic $O_D$ has been used to analyze the relation between obligations and preferences. Below, we discuss two of these properties: the dynamic interpretation of obligations and the bipolarity related to deontic choice.

**Dynamics.** The logic is best understood from the dynamics of the logic, as illustrated by Figure 6.3 (a copy of Figure 2.15). The figure represents the unique preferred models for $S = \emptyset$, $S' = \{O_D(p_1|\top)\}$ and $S'' = \{O_D(p_1|\top), O_D(p_2|\top)\}$. The leftmost figure shows that all worlds are equally ideal when there are no premises. By addition of premise $O_D(p_1|\top)$, the $p_1$ worlds are strictly preferred over $\neg p_1$ worlds. Finally, the rightmost figure shows that by addition of the second premise $O_D(p_2|\top)$, the $p_2$ worlds are strictly preferred over $\neg p_2$ worlds, and the $p_1 \wedge \neg p_2$ and $\neg p_1 \wedge p_2$ worlds become incomparable. An important consequence of the figure is that an obligation $O(\alpha|\top)$ cannot be represented by a preference of all $\alpha$ worlds over $\neg\alpha$ worlds. For example, we have that all $p_1$ worlds are preferred over $\neg p_1$ worlds for $S'$, but this is not the case for $S''$. The dynamic interpretation also relates our logic to update semantics, see Section 6.1.5.

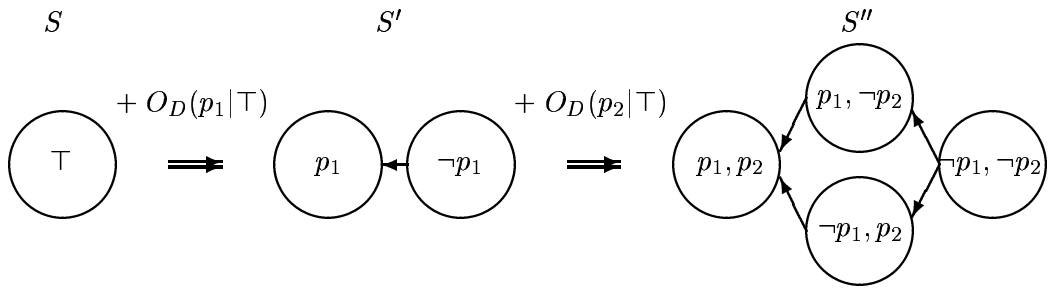



Figure 6.3: Dynamics of the ordering logic

**Bipolarity.** The preference-based deontic logics are based on the concept of deontic choice. An obligation $O(\alpha|\beta)$ is interpreted as follows.

> **Deontic choice** $O(\alpha\,|\,\beta)$: 'If an agent has the choice between $\alpha \wedge \beta$ and $\neg\alpha \wedge \beta$ , then she should choose $\alpha \wedge \beta$.'

The crucial property of the 'deontic choice' interpretation is that the notion of deontic choice is *bipolar* in contrast to the monopolar interpretation of, for example, the Hansson-Lewis and Chellas interpretations of dyadic obligations. The bipolar interpretation of an obligation $O(\alpha|\beta)$ considers its good ideal pole $\alpha \wedge \beta$ as well as its bad violation pole $\neg\alpha \wedge \beta$. The distinction can be illustrated by the following inference pattern *Reasoning-By-Cases* RBC, sometimes called the sure-thing principle.

$$\text{RBC}: \frac{O(\alpha|\beta_1), O(\alpha|\beta_2)}{O(\alpha|\beta_1 \vee \beta_2)}$$

In monopolar logics RBC is accepted, because it implies 'if you ought to do $\alpha$ given $\beta$ and you ought to do $\alpha$ given $\neg\beta$, then you ought to do $\alpha$ without examining $\beta$.' For example, this motivation is given by Pearl [Pea93], see also Section 5.2. Bipolar logics do not necessarily have RBC, as is shown in Example 2.9. It illustrates that the non-validity of RBC can be used to analyze dominance arguments. A common sense dominance argument (1) divides possible outcomes into two or more exhaustive, exclusive cases, (2) points out that in each of these alternatives it is better to perform some action than not to perform it, and (3) concludes that this action is best unconditionally. Thomason and Horty [TH96] observe that, although such arguments are often used, and are convincing when they are used, they are invalid. Example 2.9 is a classic illustration of the invalidity of the dominance argument [Jef83]. From 'we ought to be disarmed if $O(d|w)$ and 'we

ought to be disarmed if there will be no war' $O(d|\neg w)$ we cannot derive $O(d|w \vee \neg w)$, because from $(d \wedge w) \succ_s (\neg d \wedge w)$ and $(d \wedge \neg w) \succ_s (\neg d \wedge \neg w)$ we cannot derive $(d \wedge \top) \succ_s (\neg d \wedge \top)$. In fact, for the model $M$ in Figure 6.4 we have $M \models (\neg d \wedge \neg w) \succ_s (d \wedge w)$, which represents that we ought to be armed if we have peace if and only if we are armed $M \models O(\neg d|w \leftrightarrow d)$.

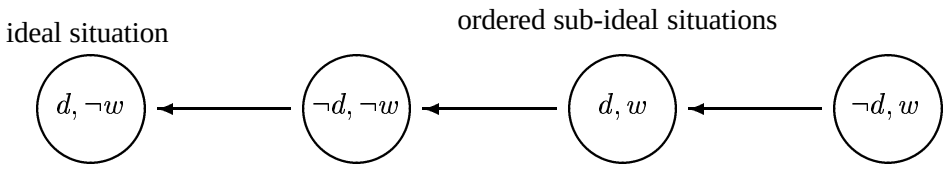



Figure 6.4: Preferential model of cold-war disarmament

There are several interpretations – besides deontic choice – of the preference ordering in the semantics of a preference-based deontic logic like the ordering logic $O_D$ and its extensions 2DL and CDL. In deontic logic literature the distinction between an ideality ordering and a betterness relation has been investigated, see e.g. [Han90a]. A typical property of a betterness relation is that it is not transitive. Another simple interpretation of the preference relation of a deontic logic follows from the diagnostic framework for deontic reasoning DIODE, see Section 5.3. It constructs a preference orderings by ordering sets of violations of norms of a so-called normative system. Thus, a world is more ideal than another when the set of violations of the former is a strict subset of the violation set of the latter. Obviously, such a relation is transitive.[1]

Finally we mention several ways in which the ordering logic has been extended in this thesis, and can be further extended in further research. The logic $O_D$ is the basis of a fully-fledged deontic logic. In Chapter 2 and 3 we have shown how the ordering logic can be extended with weakening of the consequent (in 2DL and CDL), two types of factual detachment EFD and RFD (to deal with the lack of factual detachment FD in deontic logics with a contextual interpretation of the antecedent), and permissions. Moreover, in Chapter 4 we have shown how violations can be represented in the dyadic operator by $O(\neg\alpha\,|\,\alpha)$. Finally, we mention that the extension to a first-order logic is straightforward.[2] A first-order logic can be used to represent act-types in the von Wright tradition. For example, consider the similar extension to first-order logic of Veltman's normally-presumably logic [Vel96].

> "So far, we have been thinking of the language as a propositional language, but we can also give a predicate logical interpretation to it. Think of $p$, $q$ etc. as monadic predicates rather than atomic sentences. Each such predicate specifies a property and each well-formed expression specifies a boolean combination of properties. Think of $W$ as the set of possible objects rather than the set of possible worlds. A possible object $i \in W$ has the property expressed by the atom $p$ if and only if $p \in i$. Note that different possible objects have different properties. Therefore it would be more precise to call the elements of $W$ possible *types* of objects: in reality there can be more than one or no object fitting the description of a given possible object in $W$."

[1] Another interpretation of the preference ordering can be found when we consider the diagnostic and decision-theoretic framework for deontic reasoning DIO(DE)$^2$. In DIO(DE)$^2$, the preference ordering reflects different types of rationality. Compared to DIODE, the ordering is not only determined by violations (the associated rationality is based on penalties), but also on fulfilled norms (where the associated rationality is based on rewards).

[2] The extension is straightforward, but it introduces interesting new problems. For example, such a problem has been discussed by Edelberg [Ede91].

There are several ways in which dyadic deontic logic should be further generalized. The most important extension is a formalization of contrary-to-duty formulas with the following structure in monadic deontic logic.

$$(\beta \wedge O\neg\beta) > O\alpha$$

The formula expresses '$\alpha$ should be (done), if $\beta$ is (done) and it is a violation.' Obviously, this is different from $\beta > O\alpha$, which can be represented by the dyadic formula $O(\alpha|\beta)$. For example, consider the set of two formulas $\{O\neg\beta, (\beta \wedge O\neg\beta) > O\alpha\}$. It is equivalent to $\{O\neg\beta, \beta > O\alpha\}$, which can be represented in dyadic deontic logic by $\{O(\neg\beta|\top), O(\alpha|\beta)\}$. However, suppose we replace $O\neg\beta$ by $\gamma > O\neg\beta$ in the former set. To keep the analogy, we have to replace $\beta > O\alpha$ by $(\gamma \wedge \beta) > O\alpha$ in the latter set. This, however, is not possible if $\gamma$ is itself a modal formula like $O\gamma'$. The point is that this kind of derivations should be done by the logic.

An essential property of our preference-based deontic logics is that the underlying preference ordering on worlds is not (totally) connected. This is illustrated by the models represented in Figure 6.3, because the $p_1 \wedge \neg p_2$ and $\neg p_1 \wedge p_2$ worlds are incomparable. In the logics, we still accept the transitivity axiom **4**: $\Box\alpha \rightarrow \Box\Box\alpha$ (see Section 2.2.1) for the underlying preference logic (the modal logic 2DL). Thus, our modal preference logic is an extension of Boutilier's logic CT4O. However, there are some drawbacks of transitivity. In the ordering logic, c-preferred models are not unique. For example, consider the models of $O_D(p\,|\,q)$. The $\neg q$ worlds can be equivalent to the $p \wedge q$ or to the $\neg p \wedge q$ worlds. That is, for $w_1 \in |\neg q|$, $w_2 \in |p \wedge q|$ and $w_3 \in |\neg p \wedge q|$ we have either $w_1 \leq w_2$ and $w_2 \leq w_1$, or $w_1 \leq w_3$ and $w_3 \leq w_1$. If we drop transitivity axiom **4**, then the $\neg q$ worlds of the c-preferred model of $O_D(p|q)$ are equivalent to the $p \wedge q$ worlds and also to the $\neg p \wedge q$ worlds. However, the $p \wedge q$ worlds and the $\neg p \wedge q$ worlds are not equivalent. The c-preferred model of a set of obligations can be constructed by combining the models of the individual obligations. If we drop axiom **4**, then the preferred models of a set of ordering obligations are unique. For example, the combining of $O(p\,|\,q)$ and $O(q\,|\,\top)$ is illustrated in Figure 6.5. The uniqueness is a similarity with the deontic minimizing logic $O_\forall$ combined with System Z. The first advantage of this uniqueness is that the computational complexity is less, because the unique models can be computed in a similar way as System Z models can be computed explicitly [Pea90]. The second advantage is that an 'axiomatization' based on the use of conditional only knowing becomes possible, in a similar way as System Z is axiomatized in [Bou92a]. Finally, the representation is in accordance with the 'combining preference relations' perspective of [ARS95], see also Section 2.6.4.

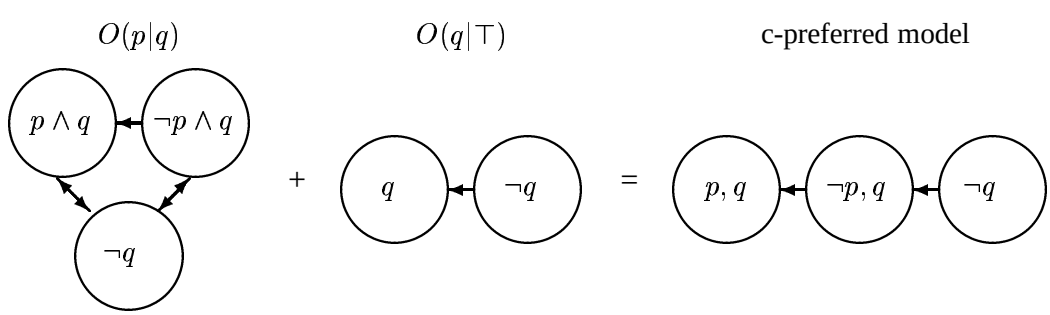



Figure 6.5: Combining preference relations: $O(p|q) + O(q|\top)$

### 6.1.3 Defeasible deontic logic

Defeasible deontic logic formalizes reasoning about obligations that can be overridden by other obligations. The second objective of this thesis is the development of a preference-based defeasible deontic logic that formalizes contrary-to-duty reasoning and overridden defeasibility. We introduced a defeasible deontic logic in Chapter 4. It is an extension of contextual deontic logic CDL that formalizes contextual obligations that can be overridden by other contextual obligations. The deontic logic $O^{re}$ combines the preference-based reasoning of 2DL and CDL with the preference-based reasoning of logics of defeasible reasoning. If the two sets of worlds $W_1 = |\alpha \wedge \beta \wedge \neg\gamma|$ and $W_2 = |\neg\alpha \wedge \beta|$ of a model $M$ are non-empty, then the contextual obligation $O(\alpha|\beta\backslash\gamma)$ is true in $M$ iff the $W_2$ worlds are not as ideal as the $W_1$ worlds. The contextual obligation $O^{re}(\alpha|\beta\backslash\gamma)$ is true in the model iff *the most normal* $W_2$ worlds are not as ideal as *the most normal* $W_1$ worlds. This is our solution of the last two problems in Section 1.4.2, model construction and entailment. The following three properties of the relation between obligations and preferences are discussed below: the heuristic of first minimizing in the normality ordering, and subsequently comparing in the deontic ordering, the bipolarity of deontic choice in a multi preference semantics, and the distinction between multi preference semantics and priorities.

When we evaluate an obligation $O^{re}(\alpha|\beta\backslash\gamma)$, we first minimize in the normality ordering and subsequently we compare the sets $W_1$ and $W_2$ in the ideality ordering. We deontically compare the most normal worlds instead of comparing the best worlds in the normality ordering. This is based on the heuristic rule that if an option can be a violation or an exception, then it is assumed to be a violation. The motivation for this rule is that a criminal should have as little opportunities as possible to excuse herself by claiming that her behavior was exceptional rather than criminal. For example, consider the obligation that normally, there should not be a fence. If an agent has a fence, then it is assumed to be a violation and she cannot excuse herself by claiming that it is an exceptional case.

The logic $O^{re}$ with its multi preference semantics illustrates that our bipolar concept of deontic choice is fundamentally different from the classical monopolar interpretation of, for example, Hansson-Lewis logics. This distinction is not visible in a preference semantics with a single ordering. For example, consider the Hansson-Lewis semantics with a single totally connected ordering $\leq$. In the monopolar reading, an obligation $O_\forall(\alpha \mid \beta)$ is true iff $\alpha$ is true in all preferred $\beta$ worlds. In the bipolar reading, an obligation $O_\forall(\alpha|\beta)$ is true iff the preferred $\alpha \wedge \beta$ worlds are preferred to the preferred $\neg\alpha \wedge \beta$ worlds. These two readings are equivalent (except for infinite descending chains). Now consider the multi preference semantics with two totally connected orderings $\leq_I$ and $\leq_N$. The monopolar reading of a conditional obligation is based on lexicographic minimizing (minimize first $\leq_N$ and then $\leq_I$) like in [Mak93]. Thus, an obligation $O_\forall(\alpha \mid \beta)$ is true iff $\alpha$ is true in all the $\leq_I$-preferred of the $\leq_N$-preferred $\beta$ worlds. In the bipolar reading, an obligation $O_\forall(\alpha \mid \beta)$ is true iff the $\leq_N$-preferred $\alpha \wedge \beta$ worlds are $\leq_I$-preferred to the $\leq_N$-preferred $\neg\alpha \wedge \beta$ worlds. The distinction can be illustrated by the model in Figure 6.6, a copy of Figure 4.13. It is the typical model of the two dyadic obligations $O(\neg f|\top)$ and $O(w \wedge f|d)$ and expresses that normally ($\neg d$) no fence ($\neg f$) is preferred over a fence ($f$), but in exceptional circumstances ($d$) a white fence ($w \wedge f$) is preferred over the absence of a white fence ($\neg(w \wedge f)$). The best most normal worlds and the most normal best worlds are both the $\neg d \wedge \neg f$ worlds. Thus, in the monopolar reading the model satisfies the highly counterintuitive obligation $O(\neg d|\top)$. In the bipolar reading, the sets of most normal worlds $W_1$ and $W_2$

correspond to $|\neg d \wedge \neg f|$ and $|d \wedge w \wedge f|$ respectively. These sets are equivalent in the ideality ordering, thus the model does not satisfy $O(\neg d \,|\, \top)$. The bipolar reading does not validate the counterintuitive obligation, because $d$ is only an exception, not a violation. The monopolar reading cannot discriminate between these two concepts.

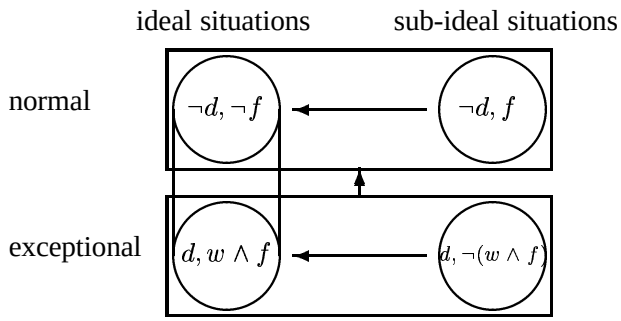



Figure 6.6: Multi-preference relation of the Fence example

The multi-preference deontic logic $O^{re}$ must be distinguished from prioritized deontic logics. The latter can formalize specificity by giving more specific obligations a higher priority than more general ones. However, prioritized deontic logics have weak overridden defeasibility, and weak overridden defeasibility has the following *Forbidden Conflict* FC.

$$\text{FC} = \frac{O(\alpha|\beta_1), O(\neg\alpha|\beta_1 \wedge \beta_2)}{O(\neg\beta_2|\beta_1)}$$

As a consequence, prioritized deontic logics with their weak overridden defeasibility derive the counterintuitive obligation $O(\neg d|\top)$ from the two obligations $O(\neg f\,|\top)$ and $O(w \wedge f\,|d)$ discussed in the previous paragraph. In the multi-preference logic, $\beta_1 \wedge \beta_2$ is not a violation but it is either a violation or an exception. For example, in Figure 6.6 we cannot derive $O(\neg d|\top)$, because having a dog $d$ is an exception and not a violation.[3] Summarizing, we do not want FC. As a consequence, we do not use prioritized deontic logics with their weak overridden defeasibility to model specificity. In Chapter 4 we argue that weak defeasibility can be used for prima facie obligations.

Specificity can be formalized by a priority ordering in the semantics that is not in the language, see e.g. [GP92, Bre94, Lan96, Bou92b, TP95]. That is, the implicit priority ordering in the semantics is determined by constraints derived from specificity. In the artificial intelligence literature, specificity has got a lot of attention, because it is a rule founded in probabilistic inference. We argued that specificity *in a deontic logic* should not be formalized with priorities, because prioritized deontic logics have forbidden conflict (FC) and reinstatement (RI and RIO). In our opinion, we can extend the argument to other logics that formalize specificity structures, like logics of defeasible reasoning and logics for qualitative decision theory. Many of these logics logics are based on priorities. Priorities are introduced to solve the inheritance problem, the derivation of 'penguins have wings' from 'birds have wings' although penguins are exceptional

[3]We can call the model in Figure 6.6 the exception model. Moreover, for any set of dyadic obligations we can say that the exception models are the preferred models, and define a notion of preferential entailment. However, the formalization of this idea is still an open problem. It seems to be much more complex than preferential entailment of one preference ordering like System Z or maximally connected.

birds, see Section 2.6.5. Moreover, 'it is generally accepted that priorities provide a more understandable mechanism for resolving conflicts' [BB95, p.67]. However, it should be observed that the naive use of priorities assumes that one rule is always stronger than another, whereas in general this seems to depend on circumstances (on context). Moreover, we think that priorities are not a good way to model *specificity*, because they also validate inference patterns like forbidden conflict and reinstatement. The only preference-based logic of defeasible reasoning we know of that deals correctly with inheritance and which does not validate these inference patterns (and thus is not based on priorities) is Veltman's normally-presumably logic in update semantics [Vel96].

### 6.1.4 Obligations, actions, time and preferences

In deontic logic literature, the $\alpha$ of $O\alpha$ is interpreted as either a state, an action occurrence (or event), an act-type or a fluent (an action having a duration). In this thesis, we follow this convention. For example, in the Fence example we write $O\neg f$ for the obligation that there should be no fence (a state), in the Forrester paradox we write $O\neg k$ for the obligation of Smith not to kill Jones (an action occurrence) or the obligation not to kill (an act type), and in Example 5.1 we write $O\neg s$ for the obligation not to smoke (a fluent). Obviously, this is a rather crude approach, and a further investigation is needed to study the distinctions between the different logics. An example of a further complexity introduced when the deontic operator has actions or fluents in its scope is that we have to discriminate between the context of justification and the context of deliberation, as illustrated in Example 5.1. Moreover, the *relation* between the different types of obligations can be investigated. For example, does an ought-to-do obligation for action $\alpha$ imply an ought-to-be obligation for the necessary post-conditions of the action $\alpha$?

This unified approach in deontic logic literature can be explained by the observation that the logics of obligations on states, actions and fluents have similarities. This follows from a semantic analysis. Standard deontic logic makes a distinction between the good ideal and bad violation. In this model, there are good and bad states, good and bad actions, and good and bad fluents. This approach can be generalized by taking time into account. At each moment of a temporal model, there are good and bad states, good and bad actions, and good and bad fluents. Alternatively, the semantic distinction between good and bad can be generalized by introducing preferences. The latter generalization is studied in this thesis. We discussed in Section 1.3.7 that we can replace the sentences of the propositional language within the scope of the deontic operators by sentences of an action calculus, as in [Mey88]. There are atomic actions, and connectives like '&' for concurrency, '$\cup$' for choice and ';' for sequencing. In this action language, analogues of the inference patterns studied in this thesis can be identified. For example, weakening can be weakening of a description of the state $O(p \wedge q) \rightarrow Op$ and weakening of the description of a complex action $O(a\&b) \rightarrow Oa$. In a sufficiently expressive language (that contains, for example, dyadic obligations), all problems discussed in this thesis will also appear in this action language, and the solutions do apply.

We illustrate the occurrence of the contrary-to-duty problem in a deontic action logic. Models of action can be found in, for example, qualitative decision theories as discussed in Section 5.2. Such models can be used to analyze the interaction between the different types of obligations, and probably give rise to new puzzles. As an example, we consider a logic recently proposed by Horty [HB95, Hor96]. First, Horty introduces a model of time. The theory is based

on a picture of moments ordered into a tree-like structure, with forward branching representing the openness or indeterminacy of the future, and the absence of backward branching representing the determinacy of the past. An example is represented in Figure 6.7, where the upward direction represents the forward direction of time. This figure depicts a branching temporal frame containing five histories, $h_1$ through $h_5$. Horty writes $H_{(m)}$ for the set of histories through moment $m$. The moments $m_1$ through $m_4$ are highlighted; and we have, for example, $m_2 \in h_3$ and $H_{(m_4)} = \{h_4, h_5\}$.

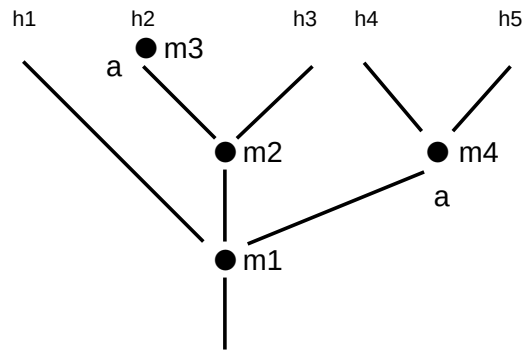



Figure 6.7: Branching time: moments and histories

Second, Horty introduces agents, actions and choices. At any moment in time, an agent can choose between several sets of histories. We write $Choice^m_\alpha$ for the partition of the histories $H_{(m)}$ through $m$. For example, consider the single agent temporal model in Figure 6.8. The choice set of the agent at moment $m_1$ is a partition in the three sets $\{h_1, h_2\}$, $\{h_3\}$, and $\{h_4, h_5, h_6\}$. Two histories are indistinguishable at moment $m$ if they are in the same equivalence class of the partition. There are different ways to define truth of an action at a certain moment of a history. For example, the agent sees to it that (stit) $\alpha$ at moment $m$ on history $h$ iff all histories indistinguishable from $h$ at $m$ make $\alpha$ true, and there is a history at $m$ that does not make $\alpha$ true.

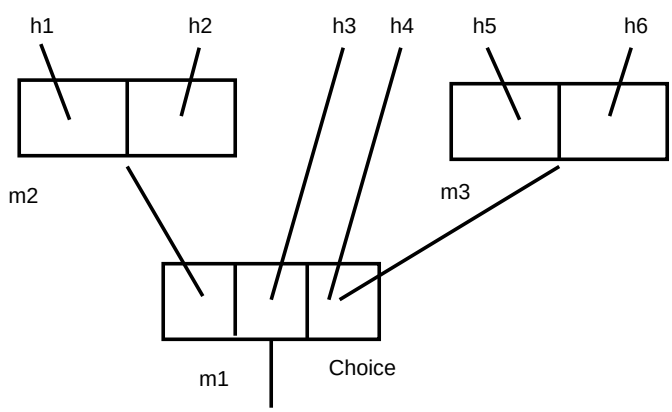



Figure 6.8: An agent's choices

Third, Horty discriminates between good and bad histories, and he generalizes this model by a preference ordering on histories. An example is the Chisholm paradox in Figure 6.9. First,

the agent has to choose between 'telling' $t$ and 'not-telling' $\neg t$, and secondly the agent has to choose between 'going to the assistance' $a$ and 'not-going to the assistance' $\neg a$. The obligation 'go to the assistance' deontically prefers history $h_1$ and $h_3$, and the obligation 'tell that you go if and only if you go' prefers history $h_1$ and $h_4$. Summarizing, history $h_1$ is preferred over all other histories.

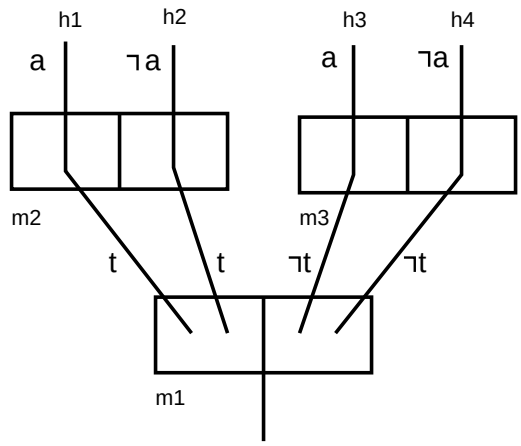



Figure 6.9: The Chisholm paradox

We use the model in Figure 6.9 to illustrate the distinction between the temporal and the preference-based solution of the contrary-to-duty paradoxes. First, the model does not satisfy the sentences of the Chisholm paradox $O(t|a)$, if we take the simple notion of stit defined above. Once the man can see to it that $a$ (moment $m_2$ or $m_3$) he can no longer see to it that $t$, because he can only see to it that $t$ at moment $m_1$. At $m_2$ and $m_3$ the truth value of $t$ is already fixed. This illustrates that we can represent the Chisholm paradox by a temporal model, although the temporal solution – antecedent before the consequent – does not apply. Second, there is no obligation at moment $m_1$ to tell $t$, because history $h_2$ is not at least as good as history $h_3$. Horty observes that complex temporal reasoning can be further analyzed if sequences of actions and strategies are taken into account. If we take two moments together in consideration, then we can say that the agent first ought to see to it that $t$ and thereafter ought to see to it that $a$, because history $h_1$ is preferred to all other histories. This is the preference-based solution of the paradox. The point is that this solution of the paradox does not rely on the temporal representation, but it relies on the fact that the possible strategies can be compared in a preference ordering.

Finally, the model in Figure 6.9 can be used to illustrate the representation of the Chisholm paradox in a dynamic logic. Assume that $a$ and $t$ in Figure 6.9 are atomic actions of an action calculus, and that $\neg a$ is the complement of $a$. In such a logic, we could derive $O(t; a)$ but not $Ot$. This expresses that it is ideal to tell and thereafter go to the assistance, but it is not necessarily ideal to tell. Thus, the inference $O(\alpha_1; \alpha_2) \rightarrow O\alpha_1$ is not valid.

In this section we showed that the contrary-to-duty problem occurs in temporal and action deontic logics, and that a solution of the Chisholm paradox has to compare different strategies in a preference ordering. In Section 1.4.1 we already mentioned the split of deontic logic into two different approaches. The first approach is based on time and actions [Tho81, vE82, LB83, Mak93, Alc93]. The second approach is based on preferences, either in monadic deontic logic [Jac85, Gob90b, Han90b] or in dyadic deontic logic [Han90b, Lew74]. Reasoning about strate-

gies combines the two approaches. In the following section we discuss an alternative way to combine time and preferences: update semantics.

### 6.1.5 Obligations in update semantics

In this section we illustrate how obligations can be formalized with update semantics [Vel96]. In the standard definition of logical validity, an argument is valid if its premises cannot all be true without its conclusion being true as well. In update semantics, the slogan 'you know the meaning of a sentence if you know the conditions under which it is true' is replaced by 'you know the meaning of a sentence if you know the change it brings about in the information state of anyone who accepts the news conveyed by it.' Thus meaning becomes a dynamic notion: the meaning of a sentence is an operation on information states. To define an update semantics for a language $L$, one has to specify a set $\Sigma$ of relevant information states, and a function $[]$ that assigns to each sentence $\phi$ an operation $[\phi]$ on $\Sigma$. The resulting triple $\langle L, \Sigma, []\rangle$ is called an update system. If $\sigma$ is a state and $\phi$ a sentence, then we write '$\sigma[\phi]$' to denote the result of updating $\sigma$ with $\phi$. We can write '$\sigma[\psi_1] \ldots [\psi_n]$' for the result of updating $\sigma$ with the sequence of sentences $\psi_1, \ldots, \psi_n$. Finally, $\phi$ is accepted in $\sigma$, written as $\sigma \vdash \phi$, if and only if $\sigma[\phi] = \sigma$. This notion of acceptance plays the same role as the notion of truth in standard semantics.

We do not discuss the formal definitions of update semantics, but we give an example that illustrates the idea. The formalization of obligations in update semantics is a natural extension of the deontic logics proposed in this thesis. The idea is based on the dynamic interpretation of the preference logics discussed in Section 6.1.2. In our preference-based logics, every model represents a single preference ordering. The information states of update semantics are the preference orderings, but not necessarily transitive. The update $[\text{oblige } \alpha \text{ if } \beta]$ is the deletion of the relations $w_2 \leq w_1$ such that $M, w_1 \models \alpha \wedge \beta$ and $M, w_2 \models \neg\alpha \wedge \beta$. This dynamic interpretation of obligations seems related to Alchourrón's box metaphor we discussed in Section 1.3.1. Alchourrón compares a (monadic) obligation with the action of putting something in a box, and in our case an obligation is compared with an action of creating a partial ordering. It remains to be shown in future research whether the dynamic approach has something to offer over the static approach.

**Example 6.1 (Updates)** Consider the six information states in Figure 6.10. This figure should be read as follows. The six information states are preference orderings like the preference orderings of 2DL. For each equivalence class of worlds, we only write the positive atoms below. The figure also represents five updates: three updates by obligations $[\text{oblige } p]$, $[\text{oblige } q]$ and $[\text{oblige } r]$, and two updates by facts $p$ and $\neg q$. □

A disadvantage of the representation of the three obligations in Example 6.1 is that we did not represent any temporal information in the three obligations. For example, we did not specify which obligation has to be fulfilled first. The following example illustrates that this temporal information can be represented by conditional obligations.

**Example 6.2 (Updates, continued)** Reconsider the three obligations $[\text{oblige } p]$, $[\text{oblige } q]$ and $[\text{oblige } r]$ of Example 6.1. Moreover, assume that the first obligation has to be fulfilled first, then the second one and finally the third one. We can express the temporal sequence with conditional

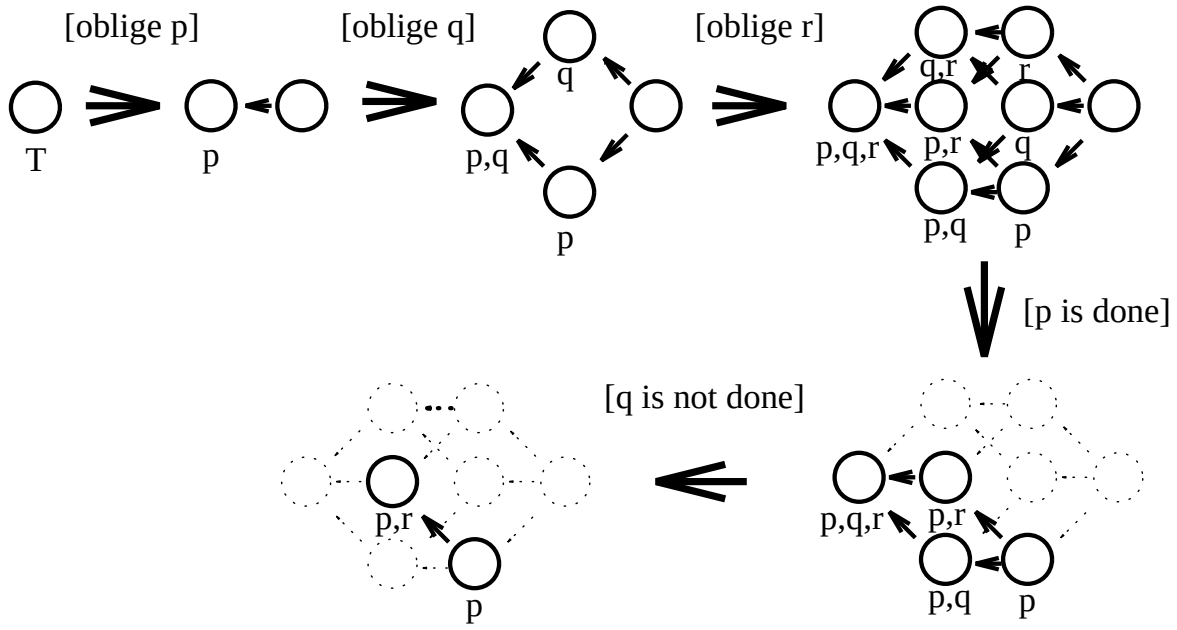



Figure 6.10: Obligations as updates

obligations. That is, we have the three obligations $[\text{oblige } p]$, $[\text{oblige } q \text{ if } p]$ and $[\text{oblige } r \text{ if } p \text{ and } q]$. Figure 6.11 represents the updates of Example 6.1. □

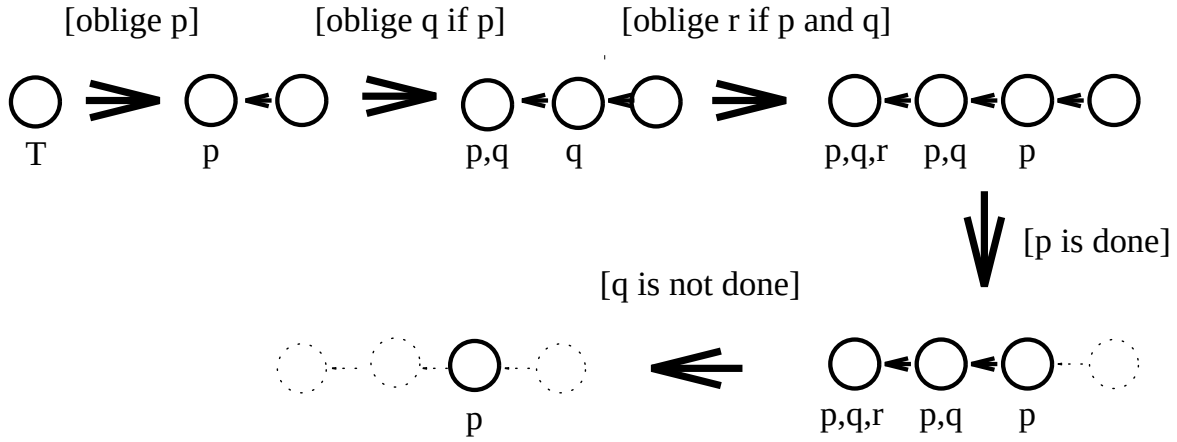



Figure 6.11: Conditional obligations as updates

There are two phases in the example above: the construction of the preference ordering and the zooming in on the ordering. This is in accordance with the occurrence of two phases in several applications. For example, in a theory of diagnosis there is the distinction between phase-1 reasoning about sets of broken components and phase-2 reasoning about *minimal* sets of broken components. In decision theory, there is a distinction between phase-1 reasoning about goals and phase-2 reasoning about reaching goals. Moreover, in the lifetime of a trade procedure discussed in Section 1.2.2, we can distinguish the following two phases. In the first phase, the agents negotiate the terms of the contract. At the end of this negotiation phase, there is a set of agreements

which can be understood as a set of norms or (conditional) obligations. In the second phase, the contract is executed. That is, absolute obligations are detached from the conditional ones, and these absolute obligations are either fulfilled or violated. At the end of the execution, if all ends well, then the set of obligations is empty. These two phases are represented in Figure 6.12 below. As represented in this figure, it is assumed that there are no more obligations at the end of the contractual period. In the first phase of update semantics, the set of equivalence classes increases together with the increased set of obligations. In the second phase of update semantics, the set of equivalence classes decreases with the zooming in on the ordering.

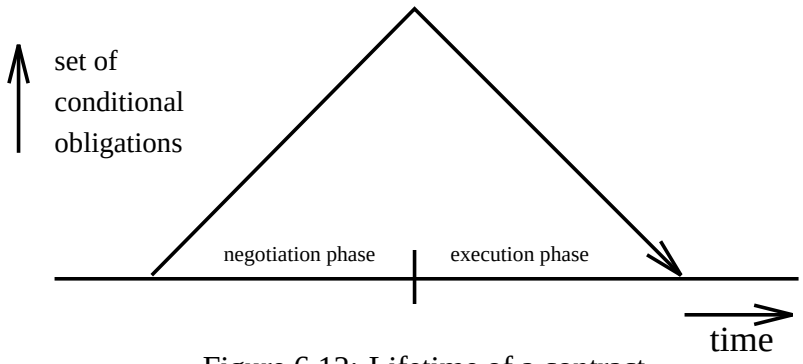



Figure 6.12: Lifetime of a contract

There are several issues for further research. The logic can be further extended by introducing agents, and by introducing different kinds of updates. In that case, we can formalize who (which authority) obliges us to do something. The update $[\text{oblige } \alpha]$ can be interpreted as a so-called speech act. Another kind of update can be used to formalize the phenomenon of 'laying guilt' on someone. For example, assume your wife says to you 'you did not bring me flowers.' You are guilty of violating an obligation you did not know it existed. The distinction is that the speech act does not create a deontic cue, but a violation. The set of equivalence classes increases and at the same time we zoom in on the ordering.

## 6.2 Obligations and defeasibility

The secondary objectives of this thesis are an explanation of the defeasibility in preference-based deontic logic and a classification of the different types of defeasibility in defeasible deontic logic. In this section we review the relations between obligations and defeasibility. First we reconsider the definition of defeasible deontic logic. Then we discuss the three sources of defeasibility we showed in Chapter 2 and 4: overriding (related to specificity or prima facie obligations), the formalization of the no-dilemma assumption, and violability. The latter originates from the inference pattern weakening of the consequent, because this inference pattern introduces exceptions of the context. Moreover, in Chapter 4 we distinguished three types of defeasibility in defeasible deontic logic. First, factual defeasibility that should be used to formalize the violation of an obligation. Second, strong overridden defeasibility that should be used to formalize the cancelling of an obligation by a more specific obligation. Third, weak overridden defeasibility that should be used to formalize the cancelling of a prima facie obligation by a more important obligation. Finally, we discuss the techniques to distinguish different types of defeasibility.

### 6.2.1 What is defeasible deontic logic?

There does not seem to be an agreement in deontic logic literature on the definition of 'defeasible deontic logic.' It seems related to the logic of prima facie obligations and it seems related to default logics studied in artificial intelligence. It is generally accepted that a defeasible deontic logic has to formalize reasoning about conflict resolution. Defeasibility becomes relevant when there is a (potential) conflict between two obligations. For example, there is a conflict between $O(\alpha_1 \mid \beta_1)$ and $O(\alpha_2 \mid \beta_2)$ when $\alpha_1$ and $\alpha_2$ are contradictory, and $\beta_1$ and $\beta_2$ are factually true. There are several different approaches to deal with deontic conflicts, see Section 1.3.6. In von Wright's so-called standard deontic logic SDL [vW51] a deontic conflict is inconsistent. In weaker deontic logics, like Chellas minimal deontic logic MDL [Che74], a conflict is consistent and called a 'deontic dilemma.' In defeasible deontic logic a conflict can be *resolved*, because one of the obligations overrides the other one. However, defeasibility in deontic logic is *not* defined as conflict resolution. This definition is much too restricted, because there is more to defeasibility than conflict resolution.

We argued that violability has to be considered as a type of defeasibility too. The defeasible aspect of contrary-to-duty obligations is different from the defeasible aspect of, for example, Reiter's default rules [Rei80]. Different types of defeasibility in a logic of defeasible reasoning formalize a single notion, whereas defeasible deontic logics formalize two different notions. Consider again the logics of defeasible reasoning and the famous Tweety example. In the case of factual defeasibility, we say that the 'birds fly' default is *cancelled* by the fact $\neg f$, and in the case of overridden defeasibility by the 'penguins do not fly' default. By cancellation we mean, for example, that if $\neg f$ is true, then the default assumption that $f$ is true is null and void. The truth of $\neg f$ implies that the default assumption about $f$ is completely falsified. The fundamental difference between deontic logic and logics for defeasible reasoning is that $O(\neg r \mid \top) \wedge r$ is *not* inconsistent. That is the reason why the deontic operator $O$ had to be represented as a modal operator with a possible worlds semantics, to make sure that *both* the obligation and its violation could be true at the same time. Although the obligation $O(\neg r \mid \top)$ is violated by the fact $r$, the obligation still has its force, so to say. This still being in force of an obligation is reflected, for example, by the fact that someone has to pay a fine even if she does $r$. Even if you are robbed, you should not have been robbed. But if penguins cannot fly, it makes no sense to state that normally they can fly. We refer to this relation between the obligation and its violation as *overshadowing* to distinguish it from *cancellation* in the case of defeasible logics. By the overshadowing of an obligation we mean that it is still in force, but it is no longer to be acted upon. The conceptual difference between cancelling and overshadowing is analogous to the distinction between 'defeasibility' and 'violability' made by Smith in [Smi93] and by Prakken and Sergot in [PS96]. Our definition of 'defeasibility' is much more liberal. It does not mean that a defeated obligation is out of force in all respects (it is not necessarily cancelled), but it only means that it is out of force in some respects.

The main advantage of the violability-as-defeasibility perspective is that it explains the distinctions *and the similarities* between cancelling and overshadowing. Moreover, it can be used to analyze complicated phenomena like prima facie obligations, which have cancelling as well as overriding aspects. This perspective has been very useful in the analyses of the puzzles, see Section 6.3 below. In the contrary-to-duty paradoxes, we showed that weakening of the consequent corresponds to introducing exceptions of the context, i.e. to defeasibility. In the cigarettes

example, we showed that we cannot accept unrestricted strengthening of the antecedent (hence we had to use defeasibility), because otherwise several intuitively consistent sets would become inconsistent. Finally, in defeasible deontic logic we showed that the conflicts between specificity and contrary-to-duty can be analyzed if both violability and cancelling are interpreted as different faces of defeasibility.

### 6.2.2 Why is deontic logic defeasible?

There are several sources of defeasibility in deontic logic. In this thesis, we have discussed the following three.

1. **Overriding.** Overriding defeasibility occurs in logics of prima facie obligations and logics based on specificity. For example, $O(\neg f \mid d)$ can be derived from $O(\neg f \mid \top)$, but not from $O(\neg f \mid \top)$ and $O(f \mid d)$.

2. **Conflicts.** The formalization of the no-dilemma assumption introduces defeasibility. For example, we can derive $O_D(\alpha_1 \mid \neg(\alpha_1 \wedge \alpha_2))$ from $O_D(\alpha_1 \mid \top)$, but not from the two obligations $O_D(\alpha_1 \mid \top)$ and $O_D(\alpha_2 \mid \top)$. Moreover, we can derive $O_D(\alpha \mid \beta_1 \wedge \beta_2))$ from $O_D(\alpha \mid \beta_1)$, but not from the two obligations $O_D(\alpha \mid \beta_1)$ and $O_D(\neg\alpha \mid \beta_2)$.

3. **Violability.** The formalization of contrary-to-duty reasoning introduces a kind of defeasibility. For example, Hansson-Lewis minimizing deontic logics do not have strengthening of the antecedent.

The latter source is the most surprising. Even deontic logics of obligations that cannot be overridden and that do not have the no-dilemma assumption are defeasible. This can be explained from a analogy between deontic logic and default logic. There is a similarity between these logics, because both can be preference-based logics. In particular, there is an analogy between the treatment of violations in preference-based deontic logics and the treatment of exceptions in preference-based default logics. This analogy is not a very satisfactory explanation of the defeasibility. A violation makes the ideal unreachable, but a violated obligation is still in force. The obligation is not cancelled. It is only no longer a cue for action.

The source of the defeasibility can be found when the exceptions are represented explicitly. In our contextual deontic logic we showed that weakening of the consequent corresponds to introducing exceptions. A contextual obligation is written as $O(\alpha \mid \beta \backslash \gamma)$ and read as '$\alpha$ ought to be the case if $\beta$ is the case unless $\gamma$ is the case.' Thus, the unless clause $\gamma$ formalizes exceptions of the obligation for $\alpha$. In contextual deontic logic, the inference pattern weakening of the consequent only derives $O(\alpha_1 \mid \top \backslash \neg\alpha_2)$ from $O(\alpha_1 \wedge \alpha_2 \mid \top \backslash \bot)$. For example, in the apples-and-pears example in Figure 6.16 we only derive $O(p \mid \top \backslash a)$ from $O(p \wedge \neg a \mid \top \backslash \bot)$. This formalizes that from 'you should buy pears and not apples' we can derive 'you should buy pears unless you buy apples.' The obligation 'you should buy pears' cannot be derived. Summarizing, defeasibility in deontic logic is caused by weakening of the consequent.

### 6.2.3 Types of defeasibility

In defeasible reasoning one can distinguish at least three types of defeasibility, based on different semantic intuitions. Consider the famous Tweety example. The 'birds fly' default can be

defeated by the fact $\neg f$, or it can be overridden by the more specific 'penguins do not fly' default. We call the first case *factual defeasibility* and the last case *overridden defeasibility*. The distinction between factual and overridden defeasibility is only the start of a classification of different types of defeasibility. To illustrate the distinction between different types of overridden defeasibility, we consider the adapted 'penguins do not fly and live on the Southern Hemisphere' default. Assume that there is a penguin that does not live on the Southern Hemisphere. In the first logics the 'birds fly' default is not reinstated, whereas in the second logics it is, because it was only suspended. In other words, in the latter case the penguin default overrides the bird default only when it is applicable itself. We call the first case *strong overridden defeasibility* and the second case *weak overridden defeasibility*. The different types of overridden defeasibility are based on different semantic intuitions. Strong overridden defeasibility is usually based on a probabilistic interpretation of defaults (most birds fly, but penguins are exceptional), like in Pearl's $\epsilon$-semantics [Pea88]. Weak overridden defeasibility is usually based on an argument-based conflict resolution interpretation (there is a conflict between the two rules, and the second one has highest priority), for example in conditional entailment [GP92] and Brewka's prioritized default logic [Bre94]. In Table 6.1 below the three different faces of defeasibility in defeasible deontic logic are represented with their corresponding character.

| | overshadowing | cancelling |
|---|---|---|
| Factual defeasibility | X | |
| Strong overridden defeasibility | | X |
| Weak overridden defeasibility | X | X |

Table 6.1: Matrix

The distinction between different types of defeasibility is crucial in logics that formalize reasoning about obligations which can be overridden by other obligations. In a defeasible deontic logic a conflict can be *resolved*, because one of the obligations overrides the other one. For example, overridden structures can be based on a notion of specificity, like in Horty's well-known example that 'you should not eat with your fingers,' but 'if you are served asparagus, then you should eat with your fingers' [Hor93]. In such cases, we say that an obligation is *cancelled* when it is overridden, because it is analogous to cancelling in logics of defeasible reasoning. The obligation not to eat with your fingers is cancelled by the exceptional circumstances that you are served asparagus. A different kind of overridden structures have been proposed by Ross [Ros30] and formalized, for example, by Morreau in [Mor96]. In Ross' ethical theory, an obligation which is overridden has not become a 'proper' or actual duty, but it remains in force as a prima facie obligation. For example, the obligation not to break a promise may be overridden to prevent a disaster, but even when it is overridden it remains in force as a prima facie obligation. As actual obligation the overridden obligation is cancelled, but as prima facie obligation it is only overshadowed. Because of this difference between cancellation and overshadowing, it becomes essential not to confuse the types of defeasibility in analyzing the deontic paradoxes.

The no-dilemma assumption is also a source of defeasibility, as discussed in Section 6.2.2. It seems like a kind of overridden defeasibility of the overshadowing type, because it is caused by the introduction of an obligation (thus it is overridden defeasibility) and there is no reason why the obligation should no longer be in force (thus it is not cancelling). However, the intuitions

on the two examples mentioned in Section 6.2.2 seem to vary. We therefore did not include this relation between obligations and defeasibility in Table 6.1.

### 6.2.4 Distinguishing different types of defeasibility

We used two techniques to analyze the defeasibility in defeasible deontic logic. First, we used inference patterns and derivations, see for example Figure 6.14. The distinction between inference patterns and logical inferences is that the inference patterns are much more general. For example, the conditions $C_O$ and $C_V$ we used in the analyses of the defeasibility paradoxes contain conditions like 'there is no overriding obligation' and 'the union of the label and the antecedent is consistent.' Moreover, the inference patterns only consider simple formula, and thus are similar to the Kraus, Lehmann and Magidor type of analysis of logics of defeasible reasoning [KLM90]. The derivations emphasize sequences of derivation steps instead of single derivation steps. For example, we analyze the *combination* of SA and WC in our analysis of the contrary-to-duty paradoxes in Figure 6.14. Second, we used (multi-)preference semantics. The semantic analysis is based on the distinction between violations and exceptions. The result is the general analysis of different types of defeasibility in defeasible deontic logics, where the intuitions behind the various distinctions are illustrated with preference-based semantics. The general analysis can be applied to any defeasible deontic logic, because we use inference patterns to analyze the different types of defeasibility.

We interpreted CTD structures as a relation between dyadic obligations. Traditionally, CTD obligations are considered as obligations that refer to a subideal situation, i.e. a conditional obligation $O(\alpha|\beta)$ is a contrary-to-duty obligation if there is an absolute obligation $O\neg\beta$. However, this runs into the problematic issue of absolute obligations and factual detachment. Moreover, we also interpreted specificity as a relation between dyadic obligations. This is more standard, because in logics of defeasible reasoning it is common practice to analyze conditionals (either at the level of inference relations [KLM90] or in the object language [Bou94a]). This structural analysis reveals that solutions of the contrary-to-duty problem based on specificity reasoning can be rejected. They are two different things.

An example of an inference pattern to analyze defeasible deontic logics is the following inference pattern FC. The inference pattern is used to discriminate between different types of overridden defeasibility, because it is valid in weak overridden defeasibility but not in strong overridden defeasibility.

$$\text{FC} = \frac{O(\alpha|\beta_1), O(\neg\alpha|\beta_1 \wedge \beta_2)}{O(\neg\beta_2|\beta_1)}$$

There are two different interpretations of $\beta_1 \wedge \beta_2$ in $O(\alpha|\beta_1)$ and $O(\neg\alpha|\beta_1 \wedge \beta_2)$: it can be an exception or a violation. The second interpretation is given by a prioritized system, whereas a multi-preference semantics does not commit to one of these interpretations, see Section 6.1.3. The acceptance of the inference pattern FC corresponds to the second interpretation. In general we do not want the inference pattern, thus we do not want weak overridden defeasibility.

A second example of an inference pattern to analyze logics is the inference pattern CD, for which we have to assume that the dyadic deontic logic can represent dyadic obligations with contradictory antecedent and consequent like $O(\alpha|\neg\alpha)$.

$$\text{CD} = \frac{O(\alpha|\beta)}{O(\alpha|\beta \wedge \neg\alpha)}$$

Logics that do not derive the obligation $O(\alpha \mid \beta \wedge \neg\alpha)$ from $O(\alpha \mid \beta)$ are counterintuitive, see Section 6.3.3. The choice between the two interpretations in the defeasibility puzzles is made in favor of the interpretation that validates CD.

## 6.3 The puzzles

In this thesis, we analyzed three different types of deontic puzzles, and the preference-based deontic logics we introduced were tested by their capacity for dealing with these puzzles. The most important deontic puzzles are the contrary-to-duty paradoxes. Besides the notorious Forrester and Chisholm paradoxes, we introduced the new Apples-and-Pears example. The second type of puzzle we discussed is the Cigarettes example, a puzzle related to the representation of dilemmas. Finally, we discussed the Fence example and the Reykjavic Scenario, two deontic puzzles typical for defeasible deontic reasoning. They are related to the interaction between specificity and contrary-to-duty structures.

### 6.3.1 Contrary-to-duty puzzles

We analyzed the contrary-to-duty (CTD) paradoxes as a problem of combining strengthening of the antecedent with weakening of the consequent. An obligation $O(\alpha \mid \beta)$ is a CTD obligation of the primary obligation $O(\alpha_1 | \beta_1)$ iff $\alpha_1 \wedge \beta$ is inconsistent.

$$\begin{array}{c} O(\alpha_1|\beta_1) \\ \text{inconsistent} \updownarrow \\ O(\alpha|\beta) \end{array}$$

Figure 6.13: $O(\alpha|\beta)$ is a contrary-to-duty obligation of $O(\alpha_1|\beta_1)$

The following example illustrates that the derivation of the obligation $O(\alpha_1 | \neg\alpha_2)$ from the obligation $O(\alpha_1 \wedge \alpha_2 \mid \top)$ is a fundamental problem underlying several contrary-to-duty paradoxes. Hence, the underlying problem of the contrary-to-duty paradoxes is that a contrary-to-duty obligation can be derived from its primary obligation.

**Example 6.3 (Contrary-to-duty paradoxes)** Assume a dyadic deontic logic that validates at least substitution of logical equivalents and the (intuitively valid) inference patterns *Weakening of the Consequent* (WC), *Restricted Strengthening of the Antecedent* (RSA), *Conjunction* (AND) and the following version of *Deontic Detachment* DD′, in which $\overset{\leftrightarrow}{\Diamond}$ is a modal operator and $\overset{\leftrightarrow}{\Diamond}\alpha$ is true for all consistent propositional formulas $\alpha$.

$$\text{WC}: \frac{O(\alpha_1|\beta)}{O(\alpha_1 \vee \alpha_2|\beta)} \qquad \text{RSA}: \frac{O(\alpha|\beta_1), \overset{\leftrightarrow}{\Diamond}(\alpha \wedge \beta_1 \wedge \beta_2)}{O(\alpha|\beta_1 \wedge \beta_2)}$$

$$\text{AND}: \frac{O(\alpha_1|\beta), O(\alpha_2|\beta)}{O(\alpha_1 \wedge \alpha_2|\beta)} \qquad \text{DD}': \frac{O(\alpha|\beta), O(\beta|\gamma)}{O(\alpha \wedge \beta|\gamma)}$$

Furthermore, consider the premise sets of obligations

$$S = \{O(\neg k|\top), O(g \wedge k|k)\}$$

$$S' = \{O(a|\top), O(t|a), O(\neg t|\neg a)\}$$
$$S'' = \{O(\neg a|\top), O(a \vee p|\top), O(\neg p|a)\}$$
$$S''' = \{O(\neg r \wedge \neg g|\top), O(r \wedge g|r)\}$$

$S$ formalizes the Forrester paradox [For84] when $k$ is read as 'killing someone' and $g \wedge k$ as 'killing someone gently.' $S'$ formalizes the Chisholm paradox [Chi63] when $a$ is read as 'a certain man going to the assistance of his neighbors' and $t$ as 'the man telling his neighbors that he will come.' $S''$ formalizes an extension of the apples-and-pears example introduced in Section 3.1, when $a$ is read as 'buying apples' and $p$ as 'buying pears.' Finally, $S'''$ formalizes a part of the Reykjavic Scenario, when $r$ is read as 'telling the secret to Reagan' and $g$ as 'telling the secret to Gorbatsjov.' The last obligation of each premise set is a contrary-to-duty obligation of the first obligation of the set, because its antecedent is contradictory with the consequent of the latter. The paradoxical consequences of the sets of obligations are represented in Figure 6.14. The underlying problem of the counterintuitive derivations is the derivation of the obligation $O(\alpha_1|\neg\alpha_2)$ from $O(\alpha_1 \wedge \alpha_2|\top)$ by WC and RSA: respectively the derivation of $O(\neg(g \wedge k)|k)$ from $O(\neg k\,|\top)$, $O(t\,|\neg a)$ from $O(a \wedge t\,|\top)$, $O(p\,|a)$ from $O(\neg a \wedge p\,|\top)$, and $O(\neg(r \wedge g)\,|r)$ from $O(\neg r \wedge \neg g|\top)$. □

$$\dfrac{\dfrac{\dfrac{O(\neg k|\top)}{O(\neg(g \wedge k)|\top)}\,\text{WC}}{O(\neg(g \wedge k)|k)}\,\text{RSA} \qquad O(g \wedge k|k)}{O(\neg(g \wedge k) \wedge (g \wedge k)|k)}\,\text{AND}$$

$$\dfrac{\dfrac{\dfrac{\dfrac{O(t|a) \quad O(a|\top)}{O(a \wedge t|\top)}\,\text{DD}'}{O(t|\top)}\,\text{WC}}{O(t|\neg a)}\,\text{RSA} \qquad O(\neg t|\neg a)}{O(t \wedge \neg t|\neg a)}\,\text{AND} \qquad \dfrac{\dfrac{\dfrac{\dfrac{O(\neg a|\top) \quad O(a \vee p|\top)}{O(\neg a \wedge p|\top)}\,\text{AND}}{O(p|\top)}\,\text{WC}}{O(p|a)}\,\text{RSA} \qquad O(\neg p|a)}{O(p \wedge \neg p|a)}\,\text{AND}$$

$$\dfrac{\dfrac{\dfrac{O(\neg r \wedge \neg g|\top)}{O(\neg(r \wedge g)|\top)}\,\text{WC}}{O(\neg(r \wedge g)|r)}\,\text{RSA} \qquad O(r \wedge g|r)}{O(\neg(r \wedge g) \wedge (r \wedge g)|r)}\,\text{AND}$$

Figure 6.14: Four contrary-to-duty paradoxes

The following example illustrates the formalization of the contrary-to-duty paradoxes in the two-phase deontic logic 2DL, contextual deontic logic CDL and labeled deontic logic LDL. We only discuss the apples-and-pears example, the formalization of the other contrary-to-duty paradoxes is analogous.

**Example 6.4 (Apples-and-Pears, continued)** Consider the sets of obligations

$$S = \{O^c(a \vee p|\top), O^c(\neg a|\top)\}$$

$$S' = \{O^c(a \vee p|\top\setminus\bot), O^c(\neg a|\top\setminus\bot)\}$$

$$S'' = \{O(a \vee p|\top)_{\{a\vee p\}}, O(\neg a|\top)_{\{\neg a\}}\}$$

where $\neg a$ does not entail the negation of $p$ and where $a$ can be read as 'buying apples' and $p$ as 'buying pears.' We have $S \models \overset{\leftrightarrow}{\Diamond}(\neg a \wedge p)$, $S \models O^c(\neg a \wedge p \mid \top)$ and $S \models O^c_\exists(\neg a \wedge p \mid \top)$, $S \not\models O^c(p \mid \top)$ and $S \models O^c_\exists(p \mid \top)$. The crucial observation is that $O^c_\exists(p \mid a)$ is not entailed by $S$. Consider a typical 2DL countermodel $M$ of $O^c_\exists(p \mid \top)$ in Figure 6.15 below. We have $M \models O^c(a \vee p|\top)$ and $M \models O^c(\neg a|\top)$, because $|\neg a \wedge \neg p| \not\leq |a \vee p|$ and $|a| \not\leq |\neg a|$. Moreover, we have $M \not\models O^c_\exists(p \mid a)$, because all $a$ worlds are equivalent. The semantic representation of CDL is similar to the representation in 2DL. The crucial observation is that we do not have $O^{cc}(p|a\setminus\gamma)$ for any $\gamma$, and a typical countermodel is again the model in Figure 6.15.

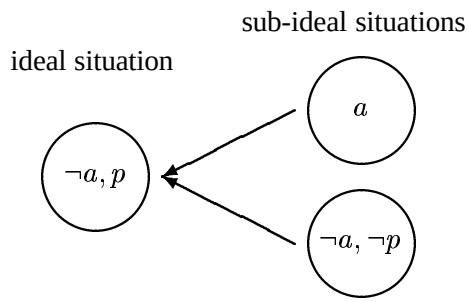



Figure 6.15: The apples-and-pears example in 2DL (semantics)

For the proof-theoretic analysis of the underivability of $O^c_\exists(p|a)$ from $S$, see the derivations in Figure 6.16. First of all, $O^c_\exists(p|a)$ is not second-phase entailed by $S$ via a first-phase derivation of $O^c_\exists(\neg a \wedge p|a)$, because $O^c(\neg a \wedge p|a)$ is not entailed by $O^c(\neg a \wedge p|\top)$ due to the restriction in **RSA**. Secondly, $O^c_\exists(p|a)$ is not first-phase entailed by $S$ via $O^c_\exists(p|\top)$, because $O^c_\exists(p|\top)$ is not first-phase entailed by $S$. Finally, $O^c_\exists(p|a)$ is not second-phase entailed by $S$ via $O^c_\exists(p|\top)$ either, because in second-phase entailment $O^c_\exists$ does not have strengthening of the antecedent at all. The proof-theoretic representation of contextual deontic logic CDL does not depend on two distinct phases like 2DL. We have $S' \models O^c(p|\top\setminus a)$, as is shown in Figure 6.16, which expresses that our little sister should buy pears, unless she buys apples. Finally, in Figure 6.16 it is also shown how the derivation in labeled deontic logic LDL is blocked. The representation in LDL shows that the context is restricted to non-violations of premises. If the antecedent is a violation, i.e. if the derived obligation would be a CTD obligation, then the derivation is blocked. □

### 6.3.2 Dilemma puzzles

Logics with the no-dilemma assumption make deontic dilemmas inconsistent. However, it has turned out that it is difficult to make exactly the right set of formulas inconsistent. We analyzed the dilemma puzzles as a problem of strengthening of the antecedent. Hence, we analyzed the problems as defeasibility problems. The following cigarettes problem, adapted from Prakken and Sergot [PS96], illustrates the defeasibility in the no-dilemma assumption.

$$\dfrac{\dfrac{\dfrac{O^c(a \vee p|\top) \quad O^c(\neg a|\top)}{O^c(\neg a \wedge p|\top)}\,\text{RAND}}{\dfrac{O^c(\neg a \wedge p|a)}{O^c_{\exists}(\neg a \wedge p|a)}\,\text{REL}}\,(\text{RSA})}{O^c_{\exists}(p|a)}\,\text{WC} \qquad \dfrac{\dfrac{\dfrac{\dfrac{O^c(a \vee p|\top) \quad O^c(\neg a|\top)}{O^c(\neg a \wedge p|\top)}\,\text{RAND}}{O^c_{\exists}(\neg a \wedge p|\top)}\,\text{REL}}{O^c_{\exists}(p|\top)}\,\text{WC}}{O^c_{\exists}(p|a)}\,(\text{RSA})$$

$$\dfrac{\dfrac{\dfrac{O^c(a \vee p|\top\backslash\bot) \quad O^c(\neg a|\top\backslash\bot)}{O^c(\neg a \wedge p|\top\backslash\bot)}\,\text{AND}}{O^c(p|\top\backslash a)}\,\text{WC}}{O^c(p|a\backslash a)}\,(\text{RSA})$$

$$\dfrac{\dfrac{\dfrac{O(a \vee p|\top)_{\{a\vee p\}} \quad O(\neg a|\top)_{\{\neg a\}}}{O(\neg a \wedge p|\top)_{\{a\vee p,\neg a\}}}\,\text{AND}}{O(\neg a \wedge p|a)_{\{a\vee p,\neg a\}}}\,(\text{SA/RSA})}{O(p|a)_{\{a\vee p,\neg a\}}}\,\text{WC} \qquad \dfrac{\dfrac{\dfrac{O(a \vee p|\top)_{\{a\vee p\}} \quad O(\neg a|\top)_{\{\neg a\}}}{O(\neg a \wedge p|\top)_{\{a\vee p,\neg a\}}}\,\text{AND}}{O(p|\top)_{\{a\vee p,\neg a\}}}\,\text{WC}}{O(p|a)_{\{a\vee p,\neg a\}}}\,(\text{SA/RSA})$$

Figure 6.16: The apples-and-pears example (proof theory)

**Example 6.5 (Cigarettes problem)** Assume a dyadic deontic logic that validates at least substitution of logical equivalents, no inference pattern for strengthening of the antecedent and one of the following axioms which make dilemmas inconsistent.[4]

**D***: $\neg \overset{\leftrightarrow}{\Diamond}(\alpha_1 \wedge \alpha_2 \wedge \beta) \rightarrow \neg(O(\alpha_1|\beta) \wedge O(\alpha_2|\beta))$

**D*′**: $\neg \overset{\leftrightarrow}{\Diamond}(\alpha_1 \wedge \alpha_2) \rightarrow \neg(O(\alpha_1|\beta) \wedge O(\alpha_2|\beta))$

Consider the obligations $S = \{O(p_1 \mid \top), O(p_2 \mid \top)\}$, $S' = \{O(p \mid q_1), O(\neg p \mid q_2)\}$ and $S'' = \{O(\neg c|\top), O(c|k)\}$. The latter set formalizes Prakken and Sergot's [PS96] cigarettes example, when $k$ is read as 'killing someone (the witness)' and $c$ as 'offering someone a cigarette.' The cigarettes example is a dilemma, whereas the other two sets are not dilemmas. Thus, in a logic with the no-dilemma assumption, $S$ and $S'$ should be consistent and $S''$ should be inconsistent. Total lack of strengthening of the antecedent is too weak, because then $S''$ is consistent. The inference pattern RSA (see 6.3) is too strong, because it makes $S$ (for **D***) and $S'$ (for **D*** and **D*′**) inconsistent. □

A solution of the cigarettes problem in Example 6.5 above is to weaken RSA such that the obligation $O(p_1|\neg(p_1 \wedge p_2))$ cannot be derived from the obligation $O(p_1|\top)$ when $O(p_2|\top)$ is another premise (set $S$), and such that $O(p|q_1 \wedge q_2)$ cannot be derived from the obligation $O(p|q_1)$ when $O(\neg p|q_2)$ is another premise (set $S'$). However, RSA may not be weakened too far, because the set $S''$ has to remain inconsistent. Hence, $O(\neg c|k)$ has to be derivable from $O(\neg c|\top)$, even

[4] In a dyadic deontic logic with a contextual interpretation of the antecedent, the two axioms are equivalent.

$$\dfrac{\dfrac{O(p_1|\top)}{O(p_1|\neg(p_1 \wedge p_2))}\,\text{RSA} \quad \dfrac{O(p_2|\top)}{O(p_2|\neg(p_1 \wedge p_2))}\,\text{RSA}}{\bot}\,\text{D*}$$

$$\dfrac{\dfrac{O(p|q_1)}{O(p|q_1 \wedge q_2)}\,\text{RSA} \quad \dfrac{O(\neg p|q_2)}{O(\neg p|q_1 \wedge q_2)}\,\text{RSA}}{\bot}\,\text{D* / D*'}$$

$$\dfrac{\dfrac{O(\neg c|\top)}{O(\neg c|k)}\,\text{RSA} \quad O(c|k)}{\bot}\,\text{D* / D*'}$$

Figure 6.17: Cigarettes problem

when $O(c \mid k)$ is another premise. This solution is incorporated in the ordering logic $O_D$ with preferential entailment. The solution in our logic uses preferential entailment for the defeasible reasoning scheme maximally connected. Preferential entailment is a typical mechanism from non-monotonic reasoning. The following example illustrates why the solution of the cigarettes problem leads to non-monotonicity.

**Example 6.6 (Cigarettes problem, continued)** Consider the three sets of obligations $S = \emptyset$, $S' = \{O^c_D(p_1 \mid \top)\}$ and $S'' = \{O^c_D(p_1 \mid \top), O^c_D(p_2 \mid \top)\}$. The three unique c-preferred models of the sets $S$, $S'$ and $S''$ are represented in Figure 6.3. We have $S' \models_{\sqsubset} O^c_D(p_1|\neg(p_1 \wedge p_2))$ and $S'' \not\models_{\sqsubset} O^c_D(p_1|\neg(p_1 \wedge p_2))$. Hence, by addition of a formula we loose conclusions. Moreover, it shows that the cigarettes problem is solved by weakening RSA, because with $S''$ the obligation $O^c_D(p_1|\top)$ is not strengthened to the obligation $O^c_D(p_1|\neg(p_1 \wedge p_2))$. □

### 6.3.3 Defeasibility puzzles

We analyzed two defeasibility problems, the Fence Example and the Reykjavic Scenario, as specificity versus CTD problems. We analyzed specificity as a kind of restricted strengthening of the antecedent.

$$\text{RSA}_O : \frac{O(\alpha|\beta_1), C_O}{O(\alpha|\beta_1 \wedge \beta_2)}$$

where condition $C_O$ is defined as follows:

> $C_O$: there is no premise $O(\alpha'|\beta')$ such that $\beta_1 \wedge \beta_2$ logically implies $\beta'$, $\beta'$ logically implies $\beta_1$ and not vice versa and $\alpha$ and $\alpha'$ are contradictory.

**Example 6.7 (Defeasibility paradoxes)** Consider the sets of dyadic obligations

$$S = \{O(\neg f|\top), O(w \wedge f|f), O(w \wedge f|d)\}$$

$$S' = \{O(\neg r|\top), O(\neg g|\top), O(r|g), O(g|r)\}$$

In a defeasible deontic logic, there are (at least) two interpretations of $S$ and $S'$. In the overridden interpretation of $S$, both $O(w \wedge f \mid f)$ and $O(w \wedge f \mid d)$ are treated as more specific obligations

that override the obligation $O(\neg f \,|\, \top)$, i.e. both are treated as cases of overridden defeasibility. In the violability interpretation, $O(w \wedge f \mid f)$ is treated as a CTD obligation, i.e. as a case of factual defeasibility. This interference of specificity and CTD is represented in Figure 6.18, a copy of Figure 4.8. This figure should be read as follows. Each arrow is a condition: a two-headed arrow is a consistency check, and a single-headed arrow is a logical implication. For example, the condition $C_O$ formalizes that an obligation $O(\alpha|\beta)$ is overridden by $O(\alpha'|\beta')$ if the conclusions are contradictory (a consistency check, the double-headed arrow) and the condition of the overriding obligation is more specific ($\beta'$ logically implies $\beta$). Case (a) represents criteria for overridden defeasibility, and case (b) represents criteria for CTD. Case (c) shows that the pair of obligations $O(\neg f \,|\, \top)$ and $O(w \wedge f \,|\, f)$ can be viewed as overridden defeasibility as well as CTD.

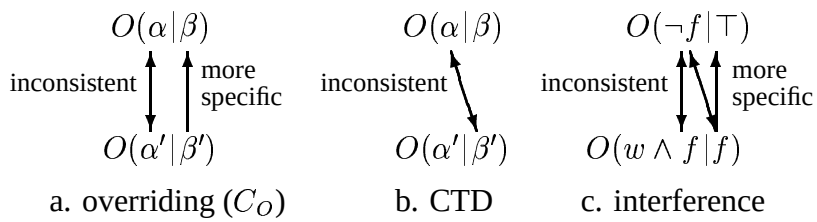



Figure 6.18: Specificity and CTD

Figure 6.19 illustrates that the Reykjavic Scenario is a more complex instance of the Fence example. In the Reykjavic Scenario, the latter two obligations of $S'$, $O(r|g)$ and $O(g|r)$, can be considered as more specific obligations overriding the former two $O(\neg r|g)$ and $O(\neg g|\top)$, and they can be considered as CTD obligations. □

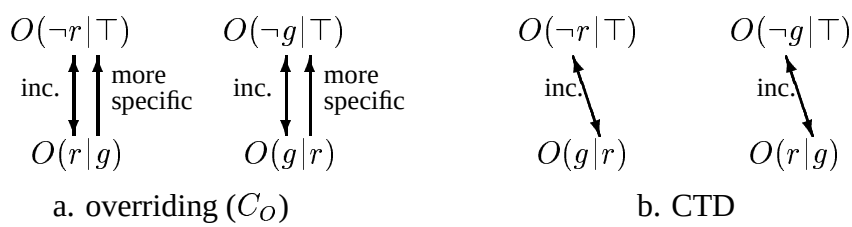



Figure 6.19: Specificity and CTD in the Reykjavic Scenario

Both defeasibility puzzles have at least two interpretations, and the problem of the puzzles is to decide between the different interpretations. We proposed that any solution has to validate the inference pattern CD. The inference pattern assumes that it is possible to represent a violation by $O(\neg\alpha|\alpha)$.

$$\text{CD} = \frac{O(\alpha|\beta)}{O(\alpha|\beta \wedge \neg\alpha)}$$

The rationality of this inference pattern is that obligations can be violated. They cannot be overridden by only violating it. For example, according to the overridden interpretation in the Fence Example, the obligation $O(\neg f \mid \top)$ is never violated. According to our argument, we have to

reject the overridden defeasibility and choose the violability interpretation. If we know the interpretation, then we can formalize the solution of the defeasibility paradoxes by adapting the definition of specificity $C_O$.

## 6.4 Applications

The formalization of the trade scenario we discussed in Section 1.2.2 involves the formalization of normative reasoning with contractual obligations. In Chapter 5 we discussed two theories that can formalize reasoning with obligations, a theory of diagnosis and qualitative decision theory. Scenario analysis extends the applications of diagnosis and decision theory. This become obvious if one tries to use these logics to model the legal risks of the numerous agents (buyer, seller, transporter, forwarder, bank, custom offices etc.) in an international trade scenario. For an adequate logical framework for modeling legal risk analysis one needs the following elements:

- model obligations, and in particular violations of obligations. This can be done by *deontic logics*. In particular, we argue that the deontic logics introduced in this thesis are useful, because they can deal with contrary-to-duty paradoxes and discriminate between exceptions and violations.

- model how obligations affect actions and vice versa. In particular, analyze how behavior of an agent is influenced by her obligations. This can be done with qualitative *decision theory*, as discussed in Section 5.2.

- model how an agent argues that she has sufficient evidence, obtained via documents, to be convinced that the other agent has fulfilled her obligations. This reasoning is usually a precondition for the fulfillment of her own obligation. For example, a buyer in New York will only fulfill her obligation to pay, if she has sufficient evidence that the seller in Rotterdam has indeed shipped the goods to her. This reasoning about the other agent's strategy will involve *game theory*, but also deontic, non-monotonic and epistemic logic to model the inference based on incomplete but sufficient evidence.

- model reasoning about the other agent's behavior, in particular with respect to her obligations towards you. This can be modeled with multi-agent logics. In particular, the formal framework that was developed by the group of Rosenschein [RZ94] is suitable for this purpose.

The last three points can be summarized as the dynamic aspects of obligations in a multi-agent environment. Basically, the problem is to analyze how the fulfillment of an obligation of agent $A$ by doing a certain action leads to another agent $B$ fulfilling her obligation by doing another action. And what kind of documents are required to give the agents enough confidence that they are not deceived. What are the reasoning processes that drive these dynamics? Additional expressiveness can be found in decision theory, because risk is the core of it. Decision theory can be used to determine the *optimal* strategy of agents. For example, a decision-theoretic analysis of an EDI-based trade procedure checks that an agent has *enough* evidence of her legal status, such that she will be able to prove her rights in court. In this analysis, the critical factor is that additional controls are more expensive than detecting fraud. For example, any bill-of-lading can

be a forgery; however, in everyday practice it is assumed that a piece of paper that looks like a bill-of-lading is a bill-of-lading, unless there is evidence of the contrary.

The dynamics of obligations are a novel extension of deontic logic. We can distinguish two phases in the lifespan of a contract: the creation of obligations and the execution of the contract (in which obligations are either fulfilled or violated). The creation of contractual obligations is subject to rationality, because an agent will only accept an obligation if she gets something in return (a right, money etc). Contractual obligations differ in this respect from other speech acts like promises, because in general a promise does not have to be profitable. This example illustrates that the introduction of decision-theoretic notions is also important for the dynamic modeling of the EDI-procedures. Another important ingredient to describe the dynamics of obligations is defeasibility. In practice, obligations are valid 'given normal circumstances' and can be overridden in exceptional circumstances.

The dynamic analysis of procedures is related to protocols being investigated by, for example, Rosenschein [RZ94] and in agent-oriented frameworks [vL96, FHMV95, ch.5]. Rosenschein's approach shows how analytical techniques from the world of game theory and decision analysis can be applied to the dynamic organization of autonomous intelligent agents. The formal analysis of rules governing the high-level behavior of interacting computer systems can also be used to analyze the procedures based on EDI messages. The result of this analysis is that cryptography is not enough, we need an additional agent, the notary (a so-called trusted third party). The Rosenschein-style analysis of protocols is very useful for a classification of protocols, and a normative analysis of EDI procedures.

# Bibliography


[AB81] C.E. Alchourrón and Buligyn. The expressive conception of norms. In R. Hilpinen, editor, *New studies in Deontic Logic: Norms, Actions and the Foundations of Ethics*, pages 95–124. D. Reidel, 1981.

[AB96] N. Asher and D. Bonevac. Prima facie obligation. *Studia Logica*, 57:19–45, 1996.

[AGM85] C.E. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: partial meet contraction and revision functions. *Journal of Symbolic Logic*, pages 510–530, 1985.

[Alc93] C.E. Alchourrón. Philosophical foundations of deontic logic and the logic of defeasible conditionals. In J.-J. Meyer and R. Wieringa, editors, *Deontic Logic in Computer Science: Normative System Specification*, pages 43–84. John Wiley & Sons, 1993.

[Alc96] C.E. Alchourrón. Detachment and defeasibility in deontic logic. *Studia Logica*, 57:5–18, 1996.

[And58] A.R. Anderson. A reduction of deontic logic to alethic modal logic. *Mind*, 67:100–103, 1958.

[Åqv67] L. Åqvist. Good Samaritans, contrary-to-duty imperatives, and epistemic obligations. *Noûs*, 1:361–379, 1967.

[Arr50] K.J. Arrow. A difficulty in the concept of social welfare. *Journal of Political Economy*, 58:328–346, 1950.

[Arr63] K.J. Arrow. *Social Choice and Individual Values*. Wiley, New York, 2nd edition, 1963.

[ARS95] H. Andreka, M. Ryan, and P.-Y. Schobbens. Operators and laws for combining preference relations. In *Information Systems: Correctness and Reusability (Selected Papers)*. World Publishing Co, 1995.

[BB95] C. Boutilier and V. Becher. Abduction as belief revision. *Artificial Intelligence*, 77:43–94, 1995.

[BC89] T.J.M. Bench-Capon. Deep models, normative reasoning and legal expert systems. In *Proceedings of the Second International Conference on Artificial Intelligence and Law (ICAIL'89)*, pages 37–45, 1989.

[BDP92] S. Benferhat, D. Dubois, and H. Prade. Representing default rules in possibilistic logic. In *Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning (KR'92)*, pages 673–684, Cambridge, MA, 1992.

[Bea73] H. Beatty. On evaluating deontic logics. In M. Bunge, editor, *Exact philosophy*, pages 173–178. D.Reidel Publishing Company, 1973.

[Bel86] M. Belzer. A logic of deliberation. In *Proceedings of the Fifth National Conference on Artificial Intelligence (AAAI'86)*, pages 38–43, 1986.

[BLWW95] R.W.H. Bons, R.M. Lee, R.W. Wagenaar, and C.D. Wrigley. Modeling interorganizational trade procedures using documentary Petri nets. In *Proceedings of the Hawaii International Conference on System Sciences (HICSS'95)*, 1995.

[BM91] A.L. Brown and S. Mantha. Preferences as normative knowledge: Towards declarative obligations. In *Proceedings of the First Workshop on Deontic Logic in Computer Science ($\Delta$eon'91)*, pages 142–163, Amsterdam, 1991.

[Bou92a] C. Boutilier. Conditional logics for default reasoning and belief revision. Technical Report 92-1, Department of Computer Science, University of British Columbia, 1992.

[Bou92b] C. Boutilier. What is a default priority? In *Proceedings of the Ninth Canadian Conference on Artificial Intelligence (CAI'92)*, pages 140–147, Vancouver, 1992.

[Bou94a] C. Boutilier. Conditional logics of normality: a modal approach. *Artificial Intelligence*, 68:87–154, 1994.

[Bou94b] C. Boutilier. Toward a logic for qualitative decision theory. In *Proceedings of the Fourth International Conference on Principles of Knowledge Representation and Reasoning (KR'94)*, pages 75–86, 1994.

[Bre94] G. Brewka. Adding specificity and priorities to default logic. In *Proceedings of European Workshop on Logics in Artificial Intelligence (JELIA'94)*. Springer Verlag, 1994.

[Cas81] H. Castañeda. The paradoxes of deontic logic: the simplest solution to all of them in one fell swoop. In R. Hilpinen, editor, *New Studies in Deontic Logic: Norms, Actions and the Foundations of Ethics*, pages 37–85. D.Reidel Publishing company, 1981.

[Che74] B.F. Chellas. Conditional obligation. In S. Stunland, editor, *Logical Theory and Semantical Analysis*, pages 23–33. D. Reidel Publishing Company, Dordrecht, Holland, 1974.

[Che80] B.F. Chellas. *Modal Logic: An Introduction*. Cambridge University Press, 1980.

[Chi63] R.M. Chisholm. Contrary-to-duty imperatives and deontic logic. *Analysis*, 24:33–36, 1963.

[CL90] P.R. Cohen and H.J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42:213–261, 1990.

[CL92] K.-T. Chen and R.M. Lee. Schematic evaluation of internal accounting control systems. Technical Report RM-1992-0801, EURIDIS, Erasmus University Rotterdam, 1992.

[Con82] E. Conee. Against moral dilemmas. *The Philosophical Review*, 91:87–97, 1982.

[Cre67] M.J. Creswell. A Henkin completeness theorem for $T$. *Notre Dame Journal of Formal Logic*, 8:187–190, 1967.

[Dav67] D. Davidson. The logical form of action sentences. In N. Rescher, editor, *The Logic of Decision and Action*. University of Pittsburg Press, 1967.

[Del88] J.P. Delgrande. An approach to default reasoning based on a first-order conditional logic: revised report. *Artificial Intelligence*, 36:63–90, 1988.

[DK89] T. Dean and K. Kanazawa. A model for reasoning about persistence and causation. *Computational Intelligence*, 5:142–150, 1989.

[dKMR90] J. de Kleer, A.K. Mackwort, and R. Reiter. Characterizing diagnosis. In *Proceedings AAAI'90*, pages 324–330, Boston, MA, 1990.

[DP95] D. Dubois and H. Prade. Possibility theory as a basis for qualitative decision theory. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI'95)*, pages 1924–1930. Morgan Kaufmann, 1995.

[DSW91] J. Doyle, Y. Shoham, and M.P. Wellman. The logic of relative desires. In *Sixth International Symposium on Methodologies for Intelligent Systems*, Charlotte, North Carolina, 1991.

[DW88] R Davis and H. Walter. Model based reasoning: troubleshouting. In *Exploring Artificial Intelligence: Survey talks from the National Conferences on Artificial Intelligence*, pages 297–346, San Mateo, California, 1988. Morgan Kaufmann.

[DW91a] T. Dean and M. Wellman. *Planning and Control*. Morgan Kaufmann, San Mateo, 1991.

[DW91b] J. Doyle and M.P. Wellman. Preferential semantics for goals. In *Proceedings of AAAI-91*, pages 698–703, Anaheim, 1991.

[Ede91] W. Edelberg. A case for a heretical deontic semantics. *Journal of Philosophical Logic*, 20:1–35, 1991.

[FH71] D. Føllesdal and R. Hilpinen. Deontic logic: an introduction. In R. Hilpinen, editor, *Deontic Logic: Introductory and Systematic Readings*. D. Reidel Publishing company, Dordrecht, Holland, 1971.

[FHMV95] R. Fagin, J.Y. Halpern, Y. Moses, and M.Y. Vardi. *Reasoning About Knowledge*. MIT press, 1995.

[For84] J.W. Forrester. Gentle murder, or the adverbial Samaritan. *Journal of Philosophy*, 81:193–197, 1984.

[Gab91] D. Gabbay. Labelled deductive systems. Technical report, Centrum fur Informations und Sprachverarbeitung, Universitat Munchen, 1991.

[Gär88] P. Gärdenfors. *Knowledge in Flux*. MIT Press, Cambridge, 1988.

[Gib80] A. Gibbard. Two recent theories of conditionals. In *Ifs*, Dordrecht, Holland, 1980. D. Reidel.

[Gob89a] L. Goble. A logic of better. *Logique et analyse*, 32:297–318, 1989.

[Gob89b] L. Goble. The logic of obligations, 'better' and 'worse'. *Philosophical Studies*, 70:133–163, 1989.

[Gob90a] L. Goble. A logic of good, would and should, part 1. *Journal of Philosophical Logic*, 19:169–199, 1990.

[Gob90b] L. Goble. A logic of good, would and should, part 2. *Journal of Philosophical Logic*, 19:253–276, 1990.

[Gob91] L. Goble. Murder most gentle: the paradox deepens. *Philosophical Studies*, 64:217–227, 1991.

[GP92] H. Geffner and J. Pearl. Conditional entailment: bridging two approaches to default reasoning. *Artificial Intelligence*, 53:209–244, 1992.

[Gre75] P.S. Greenspan. Conditional oughts and hypothetical imperatives. *Journal of Philosophy*, 72:259–276, 1975.

[Han71] B. Hansson. An analysis of some deontic logics. In R. Hilpinen, editor, *Deontic Logic: Introductory and Systematic Readings*, pages 121–147. D. Reidel Publishing Company, Dordrecht, Holland, 1971.

[Han89] S.O. Hansson. A new semantical approach to the logic of preference. *Erkenntnis*, 31:1–42, 1989.

[Han90a] S.O. Hansson. Defining "good" and "bad" in terms of "better". *Notre Dame Journal of Formal Logic*, 31:136–149, 1990.

[Han90b] S.O. Hansson. Preference-based deontic logic (PDL). *Journal of Philosophical Logic*, 19:75–93, 1990.

[HB95] J.F. Horty and N. Belnap. The deliberative stit: a study of action, omission, ability, and obligation. *Journal of Philosophical Logic*, pages 583–644, 1995.

[HC84] H.G. Hughes and M.J. Creswell. *A Companion to Modal Logic*. Methuen, London, 1984.

[Hil93] R. Hilpinen. Actions in deontic logic. In J.-J. Meyer and R. Wieringa, editors, *Deontic Logic in Computer Science: Normative System Specification*, pages 85–100. John Wiley & Sons, 1993.

[Hin71] J. Hintikka. Some main problems of deontic logic. In R. Hilpinen, editor, *Deontic Logic: Introductory and Systematic Readings*. D. Reidel Publishing company, Dordrecht, Holland, 1971.

[Hor93] J.F. Horty. Deontic logic as founded in nonmonotonic logic. *Annals of Mathematics and Artificial Intelligence*, 9:69–91, 1993.

[Hor94] J.F. Horty. Moral dilemmas and nonmonotonic logic. *Journal of Philosophical Logic*, 23:35–65, 1994.

[Hor96] J.F. Horty. Agency and obligation. *Synthese*, 108:269–307, 1996.

[Hum71] I.L. Humberstone. Two sorts of 'ought's. *Analysis*, 32:8–11, 1971.

[Hum83] I.L. Humberstone. Inaccessible worlds. *Notre Dame Journal of Formal Logic*, 24:346–352, 1983.

[Jac85] F. Jackson. On the semantics and logic of obligation. *Mind*, 94:177–196, 1985.

[Jef83] R. Jeffrey. *The Logic of Decision*. University of Chicago Press, 2nd edition, 1983.

[Jen74] R.E. Jennings. A utilitarian semantics for deontic logic. *Journal of Philosophical Logic*, 3:445–465, 1974.

[Jen85] R.E. Jennings. Can there be a natural deontic logic? *Synthese*, 65:257–274, 1985.

[Jon90] A.J.I. Jones. Deontic logic and legal knowledge representation. *Ratio Iuris*, 2:237–244, 1990.

[Jon93] A.J.I. Jones. Towards a logic of defeasible deontic conditionals. *Annals of Mathematics and Artificial Intelligence*, 9:151–166, 1993.

[JS92] A.J.I. Jones and M. Sergot. Deontic logic in the representation of law: Towards a methodology. *Artificial Intelligence and Law*, 1:45–64, 1992.

[JS93] A.J.I. Jones and M. Sergot. On the characterisation of law and computer systems: The normative systems perspective. In J.-J. Meyer and R. Wieringa, editors, *Deontic Logic in Computer Science*. John Wiley & Sons, 1993.

[Kam74] H. Kamp. Free choice permission. In *Proc. Auristotelian Society*, volume 74, pages 57–74, 1973/74.

[KH96] Krogh and Herrestad. Getting personal: Some notes on the relation between personal and impersonal obligation. In *Deontic Logic, Agency and Normative systems. Proceedings of the Third Workshop on Deontic Logic in Computer Science ($\Delta$eon'96)*, pages 134–153. Springer Verlag, 1996.

[KLM90] S. Kraus, D. Lehmann, and M. Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44:167–207, 1990.

[KM92] H. Katsuno and A.O. Mendelzon. On the difference between updating a belief base and revising it. In P. Gärdenfors, editor, *Belief Revision*, pages 183–203. Cambridge University Press, 1992.

[Knu81] S. Knuuttila. Deontic logic in the fourteenth century. In R. Hilpinen, editor, *New studies in Deontic Logic: Norms, Actions and the Foundations of Ethics*, pages 225–248. D.Reidel Publishing company, 1981.

[Lam91] P. Lamarre. S4 as the conditional logic of nonmonotonicity. In *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning (KR'91)*, pages 357–367, Cambridge, 1991.

[Lan96] J. Lang. Conditional desires and utilities - an alternative approach to qualitative decision theory. In *Proceedings of the European Conference on Artificial Intelligence (ECAI'96)*, pages 318–322, 1996.

[LB83] B. Loewer and M. Belzer. Dyadic deontic detachment. *Synthese*, 54:295–318, 1983.

[LB86] B. Loewer and M. Belzer. Help for the good Samaritan paradox. *Philosophical Studies*, 50:117–127, 1986.

[Lee91] R. M. Lee. CASE/EDI: Edi modeling, user documentation. Technical report, EURIDIS, Erasmus University Rotterdam, 1991.

[Lee92] R. M. Lee. Dynamic modeling of documentary procedures: A case for edi. Technical Report 92.06.01, EURIDIS, Erasmus University Rotterdam, 1992.

[Lev90] H. Levesque. All I know: a study in autoepistemic logic. *Artificial Intelligence*, 42:263–309, 1990.

[Lew73] D. Lewis. *Counterfactuals*. Blackwell, Oxford, 1973.

[Lew74] D. Lewis. Semantic analysis for dyadic deontic logic. In S. Stunland, editor, *Logical Theory and Semantical Analysis*, pages 1–14. D. Reidel Publishing Company, Dordrecht, Holland, 1974.

[LM92] D. Lehmann and M. Magidor. What does a conditional knowledge base entail? *Artificial Intelligence*, 55:1–60, 1992.

[Mak93] D. Makinson. Five faces of minimality. *Studia Logica*, 52:339–379, 1993.

[Mal26] E. Mally. *Grundgesetze des Sollens. Elemente der Logik des Willens*. Leuschner & Lubensky, Graz, 1926.

[McC80] J. McCarthy. Circumscription: A form of non-monotonic reasoning. *Artificial Intelligence*, 13:27–39, 1980.

[McC94a] L.T. McCarty. Defeasible deontic reasoning. *Fundamenta Informaticae*, 21:125–148, 1994.

[McC94b] L.T. McCarty. Modalities over actions: 1. model theory. In *Proceedings of the Fourth International Conference on Principles of Knowledge Representation and Reasoning (KR'94)*, pages 437–448, San Francisco, CA, 1994. Morgan Kaufmann.

[Mey88] J.-J.Ch. Meyer. A different approach to deontic logic: deontic logic viewed as a variant of dynamic logic. *Notre Dame Journal of Formal Logic*, 29:109–136, 1988.

[Mor95] M. Morreau. Allowed argument. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI'95)*, pages 1466–1472, 1995.

[Mor96] M. Morreau. *Prima Facie* and seeming duties. *Studia Logica*, 57:47–71, 1996.

[Mot73] P.L. Mott. On Chisholm's paradox. *Journal of Philosophical Logic*, 2:197–211, 1973.

[MW93] J.-J.Ch. Meyer and R.J. Wieringa. Deontic logic: a concise overview. In J.-J. Meyer and R. Wieringa, editors, *Deontic Logic in Computer Science*, pages 1–16. John Wiley & Sons, Chichester, England, 1993.

[NY97] D. Nute and X. Yu. Introduction. In D. Nute, editor, *Defeasible Deontic Logic*. Kluwer, 1997.

[Pea88] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, 1988.

[Pea90] J. Pearl. System Z: A natural ordering of defaults with tractable applications to default reasoning. In M. Vardi, editor, *Proceedings of Theoretical Aspects of Reasoning about Knowledge (TARK'90)*, pages 121–135, San Mateo, 1990. Morgan Kaufmann.

[Pea93] J. Pearl. A calculus of pragmatic obligation. In *Proceedings of Uncertainty in Artificial Intelligence (UAI'93)*, pages 12–20, 1993.

[Pet81] J. L. Peterson. *Petri Net Theory and the Modeling of Systems*. Prentice-Hall, 1981.

[Pow67] L. Powers. Some deontic logicians. *Noûs*, 1:380–400, 1967.

[Pra96] H. Prakken. Two approaches to the formalisation of defeasible deontic reasoning. *Studia Logica*, 57:73–90, 1996.

[Pri62] A.N. Prior. *Formal Logic*. Oxford University Press, 2nd edition, 1962.

[PS96] H. Prakken and M.J. Sergot. Contrary-to-duty obligations. *Studia Logica*, 57:91–115, 1996.

[PS97] H. Prakken and M.J. Sergot. Dyadic deontic logic and contrary-to-duty obligations. In D. Nute, editor, *Defeasible Deontic Logic*. Kluwer, 1997. To appear.

[Rei80] R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13:81–132, 1980.

[Rei87] R. Reiter. A theory of diagnosis from first principles. *Artificial Intelligence*, 32:57–95, 1987.

[Res67] N. Rescher. The logic of preference. In *Topics in Philosophical Logic*. D. Reidel Publishing Company, Dordrecht, Holland, 1967.

[RF96a] P. Ramos and J.L. Fiadeiro. A deontic logic for diagnosis of organisational process design. Technical report, Department of Informatics, Faculty of Sciences – University of Lisbon, 1996.

[RF96b] P. Ramos and J.L. Fiadeiro. Diagnosis in organisational process design. Technical report, Department of Informatics, Faculty of Sciences – University of Lisbon, 1996.

[RL93] Y.U. Ryu and R.M. Lee. Defeasible deontic reasoning: A logic programming model. In J.-J. Meyer and R. Wieringa, editors, *Deontic Logic in Computer Science: Normative System Specification*, pages 225–241. John Wiley & Sons, 1993.

[Ros30] D. Ross. *The Right and the Good*. Oxford University Press, 1930.

[Ros41] A. Ross. Imperatives and logic. *Theoria*, 7:53–71, 1941.

[Roy96] L. Royakkers. *Representing legal rules in deontic logic*. PhD thesis, University of Brabant, 1996.

[RTvdT96] J.-F. Raskin, Y.-H. Tan, and L.W.N. van der Torre. How to model normative behavior in Petri nets. In *Proceedings of the Second Modelage Workshop on Formal Models of Agents (Modelage'96)*, 1996.

[RZ94] J.S. Rosenschein and G. Zlotkin. *Rules of Encounter*. MIT press, 1994.

[SA85] W. Sinnot-Armstrong. A solution to Forrester's paradox of gentle murder. *Journal of Philosophy*, 82:162–168, 1985.

[Seg70] K. Segerberg. Modal logics with linear alternative relations. *Theoria*, 36:301–322, 1970.

[Sha86] R.D. Shachter. Evaluating influence diagrams. *Operations Research*, 34:871–882, 1986.

[Sho88] Y. Shoham. *Reasoning About Change*. MIT Press, 1988.

[Smi93] T. Smith. Violation of norms. In *Proceedings of the Fourth International Conference on AI and Law (ICAIL'93)*, pages 60–65, New York, 1993. ACM.

[Smi94] T. Smith. *Legal expert systems: discussion of theoretical assumptions*. PhD thesis, University of Utrecht, 1994.

[Spo75] W. Spohn. An analysis of Hansson's dyadic deontic logic. *Journal of Philosophical Logic*, 4:237–252, 1975.

[Sta81] R.C. Stalnaker. A theory of conditionals. In W. Harper, R. Stalnaker, and G. Pearce, editors, *Ifs*, pages 41–55. D. Reidel, Dordrecht, 1981.

[Ste92] W. Stelzner. Relevant deontic logic. *Journal of Philosophical Logic*, 21:193–216, 1992.

[Sus87] R.E. Susskind. *Expert Systems in Law: A Jurisprudential Inquiry*. Oxford University Press, 1987.

[TH96] R. Thomason and R. Horty. Nondeterministic action and dominance: foundations for planning and qualitative decision. In *Proceedings of the Sixth Conference on Theoretical Aspects of Rationality and Knowledge (TARK'96)*, pages 229–250. Morgan Kaufmann, 1996.

[Tho81] R. Thomason. Deontic logic as founded on tense logic. In R. Hilpinen, editor, *New studies in Deontic Logic: Norms, Actions and the Foundations of Ethics*, pages 165–176. D. Reidel, 1981.

[Tom81] J.E. Tomberlin. Contrary-to-duty imperatives and conditional obligation. *Noûs*, 16:357–375, 1981.

[TP94] S.-W. Tan and J. Pearl. Specification and evaluation of preferences under uncertainty. In *Proceedings of the Fourth International Conference on Principles of Knowledge Representation and Reasoning (KR'94)*, pages 530–539, 1994.

[TP95] S.-W. Tan and J. Pearl. Specificity and inheritance in default reasoning. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI'95)*, pages 1480–1485, 1995.

[TvdT94a] Y.-H. Tan and L.W.N. van der Torre. DIODE: Deontic logic based on diagnosis from first principles. In *Proceedings of the Workshop 'Artificial normative reasoning' of the Eleventh European Conference on Artificial Intelligence (ECAI'94)*, Amsterdam, 1994.

[TvdT94b] Y.-H. Tan and L.W.N. van der Torre. Multi preference semantics for a defeasible deontic logic. In *Legal Knowledge-based Systems. The relation with Legal Theory. Proceedings of the JURIX'94*, pages 115–126, Lelystad, 1994. Koninklijke Vermande.

[TvdT94c] Y.-H. Tan and L.W.N. van der Torre. Representing deontic reasoning in a diagnostic framework. In *Proceedings of the Workshop on Legal Applications of Logic Programming of the Eleventh International Conference on Logic Programming (ICLP'94)*, Genoa, Italy, 1994.

[TvdT95] Y.-H. Tan and L.W.N. van der Torre. Why defeasible deontic logic needs a multi preference semantics. In *Symbolic and Quantitative Approaches to Reasoning and Uncertainty. Proceedings of the ECSQARU'95. LNAI 946*, pages 412–419. Springer Verlag, 1995.

[TvdT96] Y.-H. Tan and L.W.N. van der Torre. How to combine ordering and minimizing in a deontic logic based on preferences. In *Deontic Logic, Agency and Normative Systems. Proceedings of the $\Delta$eon'96. Workshops in Computing*, pages 216–232. Springer Verlag, 1996.

[vdA92] W.N.P. van der Aalst. *Timed Colored Petri Nets and their Application to Logistics*. PhD thesis, Technical University Eindhoven, 1992.

[vdT94] L.W.N. van der Torre. Violated obligations in a defeasible deontic logic. In *Proceedings of the Eleventh European Conference on Artificial Intelligence (ECAI'94)*, pages 371–375. John Wiley & Sons, 1994.

[vdTRFT97] L.W.N. van der Torre, P. Ramos, J.L. Fiadeiro, and Y.-H. Tan. The role of diagnosis and decision theory in normative reasoning. In *Proceedings of the Second Modelage Workshop on Formal Models of Agents (Modelage'97)*, 1997. To appear.

[vdTT95a] L.W.N. van der Torre and Y.H. Tan. Cancelling and overshadowing: two types of defeasibility in defeasible deontic logic. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI'95)*, pages 1525–1532. Morgan Kaufman, 1995.

[vdTT95b] L.W.N. van der Torre and Y.H. Tan. Preference orderings for defeasible conditional logics. Technical Report WP 95.12.02, EURIDIS, Erasmus University Rotterdam, 1995.

[vdTT97a] L.W.N. van der Torre and Y.H. Tan. Contextual deontic logic. In *Proceedings of International and Interdisciplinary Conference on Modeling and Using Context (Context'97)*, Rio de Janeiro, 1997. To appear.

[vdTT97b] L.W.N. van der Torre and Y.H. Tan. The many faces of defeasibility in defeasible deontic logic. In D. Nute, editor, *Defeasible Deontic Logic*. Kluwer, 1997. To appear.

[vdTT97c] L.W.N. van der Torre and Y.H. Tan. Prohairetic deontic logic (PDL). In *Proceedings of the AAAI Spring Symposium on Qualitative Approaches to Deliberation and Reasoning*. AAAI Press, 1997. To appear.

[vE82] J. van Eck. A system of temporally relative modal and deontic predicate logic and its philosophical application. *Logique et Analyse*, 100:249–381, 1982.

[Vel96] F. Veltman. Defaults in update semantics. *Journal of Philosophical Logic*, 25:221–261, 1996.

[vF72] B.C. van Fraassen. The logic of conditional obligation. *Journal of Philosophical Logic*, 1:417–438, 1972.

[vF73] B.C. van Fraassen. Values and the heart command. *Journal of Philosophy*, 70:5–19, 1973.

[vL96] B. van Linder. *Modal Logics for Rational Agents*. PhD thesis, University of Utrecht, 1996.

[vW51] G.H. von Wright. Deontic logic. *Mind*, 60:1–15, 1951.

[vW63] G.H. von Wright. *The logic of preference*. Edinburgh University Press, 1963.

[vW68] G.H. von Wright. *An Essay on Deontic Logic and the General Theory of Action*. North-Holland Publishing Company, Amsterdam, 1968.

[vW71a] G.H. von Wright. Deontic logic and the theory of conditions. In R. Hilpinen, editor, *Deontic Logic: Introductory and Systematic Readings*, pages 159–177. D. Reidel Publishing Company, Dordrecht, Holland, 1971. An earlier version of this paper appeared in *Critica* 2:3–25, 1968.

[vW71b] G.H. von Wright. A new system of deontic logic. In R. Hilpinen, editor, *Deontic Logic: Introductory and Systematic Readings*, pages 105–120. D. Reidel Publishing company, Dordrecht, Holland, 1971. A reprint of 'A New System of Deontic Logic,' *Danish Yearbook of Philosophy* 1:173–182, 1964, and 'A Correction to a New System of Deontic Logic', *Danish Yearbook of Philosophy* 2:103–107, 1965.

[vW81] G.H. von Wright. On the logic of norms and actions. In R. Hilpinen, editor, *New studies in Deontic Logic: Norms, Actions and the Foundations of Ethics*, pages 3–35. D.Reidel Publishing company, 1981.

[Wel88] M. Wellman. Formulation of tradeoffs in planning under uncertainty. Technical Report TR-427, Massachusetts Institute of Technology Laboratory for Computer Science, Cambridge Massachusetts, 1988.

[WM93] R.J. Wieringa and J.-J.Ch. Meyer. Applications of deontic logic in computer science: A concise overview. In J.-J. Meyer and R. Wieringa, editors, *Deontic Logic in Computer Science*, pages 17–40. John Wiley & Sons, Chichester, England, 1993.

# Index