CHISHOLM'S PARADOX IN ACTION DEONTIC LOGICS

MSc Thesis *(Afstudeerscriptie)*

written by

Pietro Pasotti
(born 20/03/1991 in Monza, Italy)

under the supervision of **dr. ir. Jan Broersen and dr. Sonja Smets**, and submitted to the Board of Examiners in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam.*

Date of the public defense:	**Members of the Thesis Committee:**
23/06/2015 – 16:00	Maria Aloni (Chair)
	Frank Veltman
	Fenrong Liu
	Roberto Ciuni
	Sonja Smets
	Jan Broersen

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

Abstract

The thesis focuses on Chisholm's paradox and offers a philosophical comparative analysis using three logical frameworks in which the paradox can be formalised. Chapter 1 considers *Standard Deontic Logic* and tracks down the causes of its well-known inadequacies in coping with contrary-to-duty obligations to the lack of expressive power concerning time, preference orders and action. Consequently in chapter 2 we follow J-J.Ch. Meyer in giving an action-based *Propositional Dynamic Deontic Logic* (PD_eL) analysis of Chisholm's paradox. Finally, acknowledging that the PD_eL analysis has many open issues, we move to the *seeing-to-it-that* (*stit*) framework exemplified by one of the main existing proposals from P. Bartha. Finally, we show that Bartha's solution is problematic and try to fix it by adopting a variant of J. Broersen's temporal *next-state stit*. In this logic, under some assumptions, many deontic paradoxes including Chisholm's paradox are avoided.

to P.K. Chomsky

for his priceless insights throughout the work on this thesis,
and on my life in general.

Table of Contents

Introduction.

«most of us need a way of deciding, not only what we ought to do, but also what we ought to do after we fail to do some of the things we ought to do» [Chi63, p. 36]

In this thesis, I focus on deontic logics and the ways they deal with the so-called *Chisholm paradox*. The paradox, ever since its 'discovery', has been shown to affect many if not most deontic systems. Our aim is to give a philosophical comparative analysis using three logical frameworks in which the paradox can be formalised. Two of them are action-based. Given their very different tackle on the notions of time, action and agency, we expect a comparative analysis of their behaviour regarding Chisholm's scenario to be informative. Chisholm wrote in [Chi63, p. 34] that:

> «[...] any four statements of the following form are mutually inconsistent: (1) it ought to be that a; (2) it ought to be that if a then b; (3) if not-a, then it ought to be that not-b; and (4) not-a.»

A contradiction could in fact be derived from (1-4) in Standard Deontic Logic (SDL), which is one of the first and most studied deontic logics. Still discussed nowadays, the reasons behind the robustness of the paradox are unclear. In the following §0.1 we describe the paradox in a pre-formal fashion (i.e. without relying on any formalization of it) so as to guide our further analysis in the following chapters. Then in §0.2, we describe the *desiderata*: Chisholm's scenario is just one of many paradoxes that plague deontic logics; we will explain *why* paradoxes are bad for deontic logics, and what we want deontic logics to be eventually able to do. Finally in §0.3, we will describe the structure of the three chapters that make up the bulk of this thesis.

0.1 The problem: Chisholm's paradox.

Chisholm's [Chi63] meant to underline the inability of contemporary deontic logics to cope with so-called *contrary-to-duty obligations*.[1] *Contrary-to-duty* (CTD) are those conditional obligations whose antecedent is forbidden by some other norm. E.g. suppose (F1) it is forbidden to kill. Furthermore, assume that (F2) if someone

[1] For other famous paradoxes, see [Åq67], [For84].

kills somebody, he ought to do it gently. This is the so-called *gentle murderer*- or *Forrester's* paradox (cfr. [For84]). F2 expresses a CTD: it prescribes what to do when F1 is not complied with.

Perhaps surprisingly, most formal systems cannot handle this sort of situations. It is difficult to pinpoint what goes wrong in these logics without understanding their behaviour; and to do this we need to grasp their mechanics. Different logics deal with CTDs in their own way, and once we have them all at hand we will in a better position to see why those who fail, do, and those who don't fail, don't.

The Chisholm set. Now let's turn to the Chisholm set proper:

«suppose: (1) *it ought to be that a certain man go to the assistance of his neighbours;* (2) *it ought to be that if he does go he tell them he is coming; but* (3) *if he does not go then he ought not to tell them he is coming; and* (4) *he does not go.»* [Chi63, p. 34-35]

We will refer to (1-4) uniformly throughout the thesis with the codes C1, C2, C3, C4 respectively. The sentences are clearly *consistent*: we can easily imagine a situation in which all four assumptions are actually the case. The four premises are, moreover, usually argued to be *independent* from one another: none is redundant and could be removed without loss of information. C1 and C3, together, form a pair of sentences very similar to the *gentle murderer* scenario. In the Forrester paradox, however, what ought to be done in case of violation entails what is forbidden. This is not the case in Chisholm's scenario.

It is customary (e.g. [MDW94, p. 12-15]) to classify various versions of the paradox depending on whether the CTD action that needs to take place according to C2 has to be executed before, after or concurrently with the action that should (not) have been performed. The original paradox makes much more sense in the *backwards* case: you should go to assist your neighbours, and *before* going you should tell them you will go.

An informal way to contradiction. since you ought to go to assist your neighbours (C1), and it ought to be that (if you go, then you tell you are going to assist them) (C2), it is also true that you ought to tell them you are going (i).[2] Now since in fact we are not going (C4) and since if we are not going, we ought not to tell we are (C3), we also have that it is true that we ought not to tell we are going (ii). At this point, we conclude there are two conflicting obligations: you should tell them you are going (i), and you should *not* tell them that you are going (ii).

The crucial bit is when we deduce (i). I myself would probably not endorse that step. The point is, under some circumstances that inference seems quite

[2]This conclusion is dubious, for we might deny that C2 and C1 imply that you have to tell them you are coming. However, for the sake of the argument, let us go on.

infelicitous: once we add C4 to the picture, we *display the possibility of compliance failure*. Once we know that obligations may fail to be fulfilled, the conclusion that all their consequences be obligatory breaks down (and this may suggest that there is some non-monotonic reasoning going on). In a very simplified scenario, where all duties are assumed abided, we could certainly conclude from C2 and C1 that you ought to tell the neighbours you are coming. In 'real life' this may just not be done.

Forrester's *gentle murderer*, informally. suppose it is forbidden to kill (F1), but still if someone kills somebody, he ought to do it gently (F2). Furthermore, necessarily, *if you kill gently then you kill* (F3): a gentle murder is still a murder! And if it is obligatory to perform a gentle murder (since a gentle murder is *a murder*) it is in fact obligatory to murder. Hence, it is obligatory to murder.

This *paradox of the gentle murderer* is similar to Chisholm's in many respects, e.g. F1 and F2 are almost identical to C1 and C3. The scenario is simpler than Chisholm's, but it seems to stumble on very similar problems. Why is reasoning about compliance failure so troublesome? Compliance failure being so common in 'real life', we need deontic logics to be able to account for CTD obligations (this was precisely Chisholm's point).

0.2 Deontic logics.

The first modern (axiomatised) deontic logic is Mally's [Mal26]. His formal system was explicitly intended to capture the notion of 'ought to', which according to him was "the basic concept of the whole of ethics". He tried to transform into an operator on sentences the construct 'it ought to be the case that'. Mally's system was soon discovered to be deeply flawed[3]. Namely 'it ought to be that p' and 'p' were provably equivalent in it. Many, aware of Mally's and others'[4] difficulties would argue that deontic logic as a whole is not possible, and to attach truth-values to deontic sentences is an enterprise doomed to fail.[5]

However, not everybody agreed to give up on deontic logic altogether. Consequently, and especially after von Wright's seminal [vW51] the field evolved quickly. On the one side researchers started to work on dyadic deontic logics[6], where obligations are always relative (conditionalized) to some circumstance. On the other side, following [vW51], which was deeply influenced by developments in modal

[3] See [Lok04, Lok13, Men39] for an overview.

[4] [Gre39, HM39, Ran39] in fact had similar problems.

[5] For ex. see [Ros41, Jø 8] or the general 'expressivist' trend concerning the metaphysics of norms cf. [vR14].

[6] First proposed, I believe, still by von Wright in [vW56].

logic, the so-called *Standard Deontic Logic* (SDL) system emerged.[7] That is nothing but the normal modal logic **KD**. Von Wright's influence was enormous and research on modal logics in general contributed to making SDL well understood too[8]. Certainly, SDL has been for a long time the most influential deontic logic paradigm around.

Dyadic deontic logics were devised to avoid some of the paradoxes of the monadic ones, but have some problems of their own.[9] The deontic operator of the logics we will consider in this thesis are all, for contingent reasons, monadic. Consequently, this is the end of our discussion of dyadic deontic logic. "deontic logic" will be a shortcut for "monadic deontic logic" henceforth.

0.2.1 What is a deontic logic?

Assuming we all know what a 'logic' is, we explain here what 'deontic' means. The *possibility* of a deontic logic has been a hotly debated issue; its aims no less. We believe *deontic logic should (minimally) describe/model the common usage of the concepts of* obligation, permission, forbiddance, optionality. I write 'minimally' because we would appreciate a deontic logic to also capture other important moral/legal notions such as *supererogation* and maybe even *liberties* or *powers* and, in general, any agent-based notion.[10] Some modern deontic logics in fact have been used to formalise (parts of) legal systems[11]. I will, in what follows, regard deontic logic as if this were its main goal. An ideal deontic logic would thus be able to capture all relevant agent-based notions that occur in a legal corpus.

Norms versus imperatives: two sides of the same coin. *Norms* are statements that describe how a set of persons (agents) should behave. Hence a norm, whether alone or embedded in a broader normative system (such as a set of norms) can be said to *describe an ideal behaviour*.[12] E.g., by asserting what is true in an ideal world relative to the actual one.[13]

[7] Of course there have been many more approaches to solve Mally's difficulties. E.g. Menger suggested to tackle them by adding the *doubtful* truth-value to the picture, beside truth and falsity. Many-valued deontic logics have been moderately researched ever since.

[8] Remark: research has focused on SDL especially as **KD**: that is, not from a deontic perspective, but from a modal logic one. So its proof theory is very well studied, just like its model theory, but its semantic specificities (deontically speaking) are not as much.

[9] See 2.5^{159} for an overview. Also see [PS97, Gob03].

[10] Cfr. [McN96, Tho81a].

[11] Theory: [NR14, Ser90, WM91], practice: [AHBG+13, CJ96].

[12] Even more specifically, we can conceive deontic logic as the *logic of the actual versus ideal behaviour*, cfr. [DMW96, WM91]. This perspective will prove itself useful in §2.

[13] What counts as *ideal* is an entirely different question, that needs not be addressed here. We will assume in what follows that the normative systems we deal with are contextually given, for

Imperatives are natural language statements usually expressed in the imperative mood, such as 'open the door!' or 'tell her immediately!'; but also 'you ought to apologize to him'. Some authors see an essential difference between norms and imperatives, deep enough to require different logics to account for the two.[14]

It is customary[15] to distinguish between a *descriptive* and *prescriptive* function of obligations. Descriptive are those sentences that assert that there is some obligation x in force and request, invite or command abidance thereof. Prescriptive, symmetrically, are sentences that create or stipulate some new norm, that is to be followed thereafter. Natural language is often ambiguous between the two, and in fact what is descriptive and what is prescriptive is clear only once the context is sufficiently specified. The sentence "do not park your bike on this office desk!" is descriptive when the normative context (e.g. the Dutch Civil law, or the Office Regulations of the building we are in) already specifies that bikes cannot be parked on tabletops. It is prescriptive when such rule does not exist, and the speaker possesses the authority to establish it afresh.[16] *Descriptive imperatives stem from the normative context whereas prescriptive imperatives, crystallizing into norms, are able to influence it.*

Norm and imperative are thus just the two facets of one and the same linguistic/cultural coin. Hence, the *logic of norms* and *logic of imperatives* need to describe the same class of phenomena, although, perhaps from different perspectives. The main concrete difference I can tell between the two is that imperatives are linguistic phenomena[17] akin to gestures which prompt the recipient to *do* something. On the other hand, norms are abstract entities (unable to prompt, by themselves, anything). The two are intertwined and very closely related, but also different. E.g. revising a norm (adjusting the law) is very different a procedure from retracting a previously issued order (imperative).[18] Despite the differences, I am convinced norms and imperatives can be treated as two interacting layers that, together, make up a deontic structure.

An unified framework being still missing, in this thesis we will be concerned exclusively with logics devised (albeit implicitly) to reason only about *descriptive* obligations and their interactions.

---

example by an institution or by some ethical theory or cultural tradition.

[14] See [BvdT12, vdTH08] for an overview.

[15] E.g. see [AB81, ABvdT10].

[16] This difference parallels, in a way, the sharper difference between saying that there's an anvil on that mozzarella and actually putting an anvil on the mozzarella.

[17] Technically, a form of *speech act*: [Gre14].

[18] For the many facets of retracting a law versus retracting an imperative, see: [BRea12, p.22 §C.].

0.2.2 Making a logic deontic.

Here we explain briefly the most common way to *make* a deontic logic. Anderson and Kanger are credited for coming up with a reduction of deontic modalities to alethic modalities[19] now known as 'Andersonian reduction' or 'Anderson-Kanger reduction'.

The core observation is that the normative system in force determines what is obligatory. We can therefore say that 'something is obligatory iff it follows from the norms in force'. If we conceive norms as giving a description of a 'perfect world', then it is obligatory to do p iff p is true at the perfect world of reference, or, alternatively, iff p is entailed by the set of norms.

Equivalently, we could define p to be obligatory iff $\neg p$ implies that the norms are not abided (i.e. the world is *imperfect*). Arguably, this is true whenever *there is wrongdoing* and so we introduce V, a propositional constant that loosely means that something is (deontically) wrong. Commonly, scholars read V as 'there is wrongdoing', 'there is liability to punishment' or 'a violation of the norms has occurred'. The only common constraint on V's meaning is that it should be possible, at any given moment, to act in a way that does not imply a violation.

Implementations of Anderson-Kanger's reduction vary from logic to logic, but commonly involve a definition similar to the following one:

$$O\varphi := \Box(\neg\varphi \to V) \qquad \text{(A-K reduction)}$$

where $\Box$ is an alethic (historical) necessity operator. The intuition is that φ is obligatory iff (in all deontically perfect worlds, or *inevitably*) $\neg\varphi$ entails a violation. Introducing deontic operators in a logic in this way is very common, and in fact different approaches are seldom found.[20]

No-pardon. It is sometimes suggested (e.g. [Mey87]) in temporal deontic logics to make sure that whenever a violation occurs it is 'carried along' all future states. This is called *"no-pardon"*, and is meant to ensure that violations are never 'forgotten'. Semantically, the principle can be enforced by:

$$w \in \pi(V) \Rightarrow \forall v\,(vRw \Rightarrow v \in \pi(V)) \qquad \text{(V-inheritance)}$$

where R is some 'later than' accessibility relation. The intuition is: once something wrong is done, the violation atom V is true henceforth. The principle has its advantages and its drawbacks; consequently we will consider each time whether to accept it or reject it.

[19] Tracing back their contributions to: [And58, And67a, Kan71].

[20] With notable exceptions including [Hor01].

What makes a deontic logic a good one.

Broadly, a deontic logic is *intuitive* whenever it captures our intuitions about normative reasoning. As far as a logic sets out to capture some 'real-world' (linguistic or cognitive) phenomenon by giving an accurate description of the reasoning patterns that back it up, a logic is the more intuitive the more these reasoning patterns are faithfully reproduced:

1. does this system capture the (*objective*) patterns our common-sense deontic reasoning follows?[21]
2. are the theorems of the logic (*subjectively*) intuitive?
3. conversely, are all of our pertinent (again, *subjective*) intuitions, once formalized, theorems of the logic?

Obviously, the perfect deontic logic would answer 'yes' to all three questions. Addressing them concretely requires some empirical investigation and this is not the work we set out to do here. However, it will be valuable to keep in mind these tests, lest we ignore what we are looking for. In what follows we will consider 2. as a baseline requirement for the logics we investigate to be a viable deontic logic. This means that we will check (both syntactically and semantically) that the syntactic/semantic treatment of the deontic part of the logic makes sense. Finally, a look at the theorems of the logic will possibly reveal implausible consequences of the definitions. *Remark:* our main concern will be the features which are specifically *deontic*. Thus, we will not discuss features that are inherited from propositional calculus, such as the classical tautologies and *modus ponens*.[22]

Paradox hunting.

However implausible the axioms of a logic may seem under some analysis, the most common way to show that a deontic logic is not a good deontic logic is to show that it gives rise to paradoxes. That by itself is not much of a proof of anything. Furthermore: what classifies as a paradox?

The SDL-validity $Op \supset O(p \vee q)$ (*Ross' Paradox*) is sometimes listed as a paradox (e.g. [MDW94]), sometimes not (e.g. [Cas81, HF70]). For an overview, see [McN14]. We say it *is* because the natural language disjunction 'or' gives rise to a strong *choice implicature* that our common sense reasoning inevitably

[21] Unless the goal of the logic is different. But here we are concerned with logics that are targeted at either capturing our concrete earthly ethical reasoning patterns or, more broadly, our normative ones.

[22] Of course, the question whether classical logic is or not a desirable part of a deontic logic (paradoxes of material implication, double negation elimination, non-contradiction) is nontrivial. However, we will not address it here. E.g. cfr. [McN14].

picks up: if I tell to Jane that she is "obliged to a or b", she will understand that she can choose freely between the two. Clearly it is possible to treat this paradox as a formalization issue that should be dealt with by pragmatic means. A more promising strategy would involve encoding some notion of 'choice' into the semantics of disjunction.[23] This highlights how we tend to consider paradoxes as displaying a flaw in the logic; clues that the phenomenon we want to analyse is not being modelled properly.

Some paradoxes are much more prolific than others in displaying flaws in logics; an example of a particularly nasty brand of troubles is certainly Chisholm's paradox. Widely debated in the literature as it is, Chisholm's paradox can be used to focus the discussion about the strengths and limits of the logical systems in which it has been, or can be, expressed.

0.3 The structure of the thesis.

Following the thin red line of Chisholm's paradox, the thesis will investigate SDL and two of the major modern, action-based frameworks used to investigate agency: Propositional Dynamic d$_e$ontic Logic (PD_eL) and "Seeing To It That" logic (*stit*). Furthermore, being Chisholm's a CTD puzzle, some attention will be devoted to a famous paradigmatic CTD paradox: Forrester's *gentle murderer* (cf. [For84]). Each chapter discusses a framework and is (broadly) structured as follows:

- A quick, informal syntactic and semantic introduction to the logic, pointing to a technical appendix for the formalities.
- Presentation/formalization of Chisholm's paradox in the framework.
- Discussion:
  - Analysis of the (deontically) relevant features.
  - Analysis of how the Chisholm setting is treated.
  - Possibly, a detour into the *gentle murderer* paradox, finally drawing back the results to Chisholm's.
- Wrapping-up and conclusions about the logical system. Comparison with other logics and final remarks.

Given this structure, the chapters are designed to be overall self-contained. Only the final section of every chapter, where comparisons are drawn between the various logics, will explicitly refer back to earlier chapters.

Roadmap. This thesis will start in **§1** with a discussion of SDL and its features relative to the paradox. The problems we will encounter with SDL will lead us to

[23] E.g. [Mey88] or §2.

examine the PD_eL framework in chapter **§2**. We will discover there that the PD_eL formalization of Chisholm's paradox present in the literature is not as satisfactory as it should be, and cannot be fixed in a straightforward way. Consequently we will pinpoint the origins of the troubles and, in **§3**, attempt an analysis in a different framework: *stit*. Analysis of *stit* will take more than half of the whole thesis. We will examine one of the main existing proposals, due to Paul Bartha, which will turn out to be not as unproblematic as he had argued. His *stit* being insufficient for our purposes, in §3.4 we will extend it to a temporal ('next') deontic *xstit*. We will see how there, under some assumptions, Chisholm's paradox disappears along with many deontic paradoxes (§3.5). The final **§4** is devoted to wrapping up the results of the thesis, pointing directions for future work and highlighting this research's limitations.

Chapter 1

SDL.

1.1 Introduction.

SDL is simply the normal modal logic **KD** (i.e. the logic of the class of *serial* Kripke frames). The classical modal box and diamond $\Box, \Diamond$ are replaced by O, P to mirror their intended deontic readings: "it is obligatory that" (O) and "it is permissible that" (P). The language is backed up by a so-called perfect worlds semantics (Kripke-style), where accessibility encodes an 'is deontically better than' relation.

A similar approach would be to take some suitable[1] alethic modal logic and introduce O, P as defined operators.[2] In that case, the fragment of the logic without alethic operators would be equivalent to SDL. Here we choose to focus on SDL alone (instead of giving an Andersonian reduction from some alethic modal logic) because we are only interested in the deontic part of the logic.

In §1.2 we sketch informally syntax and semantics of SDL. For the interested reader, Appendix A will contain some more formal material. In §1.3 we show how Chisholm's setting is problematic in SDL, and in §1.4 we search SDL for the origin of the troubles. A conclusion follows in §1.5.

1.2 Informal syntax and semantics of SDL.

For the reader unfamiliar with the SDL language and semantics, Appendix A will contain some detailed material. An informal overview follows.

The language of SDL is just a propositional language plus a monadic O operator, whose intended reading is "it ought to be that". Finally, we have a spe-

[1] See [And58].

[2] The definition would be in this case $O\varphi := \Box(\neg\varphi \supset V)$. Also see §0.2.

cial propositional atom V, which loosely stands for 'a violation occurs', 'there is wrongdoing' or something similar. If V is true at a world, some obligation has been infringed. A sound and complete axiom system extends any one for propositional logic with the two following axioms:

$$O(\varphi \supset \psi) \supset (O\varphi \supset O\psi) \tag{O-K}$$
$$O\varphi \supset P\varphi \tag{O-D}$$

and a 'necessitation' inference rule $\varphi / O\varphi$ which we call O-NEC.

The semantics is based on relational structures $\langle W, D\rangle$ where W is a set of worlds and D an accessibility relation whose intended reading is 'deontic ideality': if Dww', we say that "w' is deontically ideal with respect to w". Enriching the relational structure with an interpretation function I yields an SDL model. The truth conditions of deontic formulae ($O\varphi$) will depend on what is true at the deontically ideal worlds accessible from the world of evaluation. Embedding such '*world-switching*' operators into each other would of course extend the 'search space' further (D is *serial*). On the other hand, boolean operators are entirely static.

The truth definition of p being true at w (write $\mathcal{M}, w \models p$) is classical, and the other boolean cases are equally traditional. The only interesting case is $O\varphi$, which is true at world w iff φ is true *at all deontically ideal worlds relative to* w, that is, iff $w' \models \varphi$ for all $w' : Dww'$. The intuition is that $O\varphi$ is true at some world w iff at all ideal worlds relative to w (at all worlds which are better than w), φ is.

Its dual P, which reads 'it is permissible that', is introduced as $P\varphi := \neg O\neg\varphi$. This should respect the intuition that something is permitted iff its negation is not obligatory.

1.3 Chisholm's paradox in SDL.

We now turn to the first formalization of the Chisholm set [3]. As we explained more extensively in the introduction, Chisholm's paradox stems from the formalization of the four sentences C1, C2, C3 and C4. These four assumptions, or *premises*, appear to be *independent* from one another and *consistent*. In the remaining of this section we will show how a paradox arises and pinpoint the features of SDL that are responsible it.

Chisholm's paradox: the standard formalization. We use g to denote the proposition 'one **g**oes to assist his neighbours', and t to denote 'one **t**ells them

[3] As laid down by he himself in [Chi63].

he is coming'. Given this, Chisholm's set of four sentences we presented above is usually formalised as follows in SDL:

$$\text{C1} \mapsto \quad Og \tag{1.1}$$
$$\text{C2} \mapsto \quad O(g \supset t) \tag{1.2}$$
$$\text{C3} \mapsto \quad \neg g \supset O\neg t \tag{1.3}$$
$$\text{C4} \mapsto \quad \neg g \tag{1.4}$$

Now from (1.3) and (1.4), by MP, we can deduce that one ought not to t:

$$O\neg t \tag{1.5}$$

And from (1.1) and (1.2) we derive that Ot in the following way:

$$Og \supset Ot \qquad \text{from (1.2), by K-O} \tag{1.6}$$
$$Ot \qquad \text{from (1.6) and (1.1), by MP} \tag{1.7}$$

Now we have both that Ot (1.7) and that $O\neg t$ (1.5). Suppose a world w exists in some SDL model such that C1-C4 hold at w. Given that SDL frames are serial, Dww' for some w'. that $w \models Ot$ entails that $w' \models t$. That $w \models O\neg t$, on the other hand, entails that $w' \not\models t$. This is quite a contradiction.

Missing uniformity. Here (1.2) and (1.3) are treated differently despite their very similar surface form (i.e. natural language formulation)[4]. Clearly the logical form needs not always fit the surface form in an intuitive way. However formalization issues should be settled not by silently adopting a different (and seemingly less natural) logical form but, instead, by closing in on the correspondence between logical form and surface form.

The classical argument[5] is usually that the premises are *prima facie* independent from one another, and rendering (1.3) like (1.2), that is as

$$O(\neg g \supset \neg t) \tag{9'}$$

would make (9') derivable from (1.1) in SDL.[6].

[4]It is in fact tempting to avoid the problem by just stating the paradox in different terms. In their original formulation, C2 uses an 'ought to be' construct, whereas C3 has an 'ought to' in the consequent position of a conditional statement. Their difference is unclear: to my mind, 'it ought to be that if p, then q' and 'if p, it ought to be that q' are pretty much equivalent. It is tempting to adequate the natural language formulation to the formalization. However this would seem quite suspicious a move: thus I chose to stick to Chisholm's original formulation (which sounds natural enough).

[5]E.g. [McN14].

[6]In fact: $(g \supset (\neg g \supset \neg t)) \supset (O(g \supset (\neg g \supset \neg t)) \supset (Og \supset O(\neg g \supset \neg t))$ Then by MP from (1.1) we obtain $O(\neg g \supset \neg t)$.

Similarly we could replace (1.2) by

$$g \supset Ot \tag{8'}$$

but (8') itself would be derivable from (1.4): $\neg\varphi$ implies $\varphi \supset \psi$ in SDL.

However, supposing for a moment that $p \supset (\neg p \supset q)$ were an intuitive principle, the derivability of (8') from C4 is not much of a problem. C2 tells us what to do if g obtains. But since we know that $\neg g$ obtains (C4), all conditional obligations that have g as antecedent, including C2, suddenly become irrelevant (i.e. trivially/vacuously true). This counterintuitive fact (one of the *Paradoxes of Derived Obligation*) is well known[7]. In SDL, anything false commits you to anything whatsoever: $\neg\varphi \supset (\varphi \supset O\psi)$ is in fact a theorem. Chisholm's C2 is slightly different from C3, the difference being literally captured in their formalizations (1.2), (1.3). However in natural language the two formulations seem interchangeable and not different in any relevant way. If that is so, and if C2's logical form is (8'), then the fact that C2 is derivable from C4 is as true and intuitive as classical logic's *explosion*[8] is. However, perhaps we better realize the premises don't really feel as independent from each other as they are usually argued to be.

Variations on C2 and C3. It has been argued that conditional obligations are just not adequately expressible in SDL: neither $a \supset Ob$ nor $O(a \supset b)$ represent well-behaved formalizations.[9] So, maybe the logic just cannot adequately model them. This is a view many authors have endorsed over time[10] and that we shall accept as well.

However suppose we prefer uniformity over independence and choose to use (9') instead of (1.3). The set of premises is thus $\{Og, O(g \supset t), O(\neg g \supset \neg t), \neg g\}$. Now the system allows us no more to deduce that $O\neg t$. We only have that $O\neg g \supset O\neg t$, but $O\neg g$ is false. So, formalizing conditional obligations in this way prevents us from reasoning about such obligations at all! ($\neg\varphi$ and $O(\neg\varphi \supset \psi)$ do not entail ψ in SDL)

Finally, what if we choose to use (8') instead of (1.2)? Now the set of premises is: $\{Og, g \supset Ot, \neg g \supset O\neg t, \neg g\}$. But then, as we have seen, everything of the form $g \supset \psi$ such as (8') follows from (1.4). So, even if independence is lost (but not necessarily we care), we are still left with the problem that since $\neg g$, the sentence "if you go to your neighbour's assistance, then you ought to shave a baboon" is now predicted valid. However, since we know that you are not assisting your neighbours

[7] Cfr. [Pri54, McN14].

[8] I.e. the 'ex falso sequitur quodlibet' theorem $\bot \supset \varphi$.

[9] For a review, and an explanation of the paradoxes that arise under both formalizations, see [HF70].

[10] Cfr. [McN14, §4.5].

(by (1.4)), we also know these 'oughts' will never be detached: that is, "you ought to shave a baboon" will not be derivable from these premises.

So, using (9') instead of (1.3) seems to disable the paradox and lead to no worse effects than the usual undesirable SDL validities. However, the Chisholm scenario is still not adequately modelled: we cannot obtain the desired conclusion that we ought not to tell we are going. Unsurprisingly, most authors think that the Chisholm set is just not adequately formalizable in SDL. Since we accept the original formulation and we are aware that reformulating (1.2) or (1.3) will not solve much of the pile of troubles SDL has, we will now try to focus closer on the semantics of SDL in light of the intended semantics for a logic of norms, hoping this will shed some light on the reasons behind this trouble.

1.4 Discussion.

In this section we try to dig a bit deeper into the (semantic) features of SDL that make it prone to paradox. We will start by an analysis of the central syntactic elements of SDL. Then, moving towards a semantic perspective, we will discuss some crucial design choices such as making propositions the objects of obligation instead of, how von Wright had suggested, act types. There, we will try to trace back the origin of troubles to SDL's semantic treatment of obligations. Next we will discuss these same issues relative to CTDs and Forrester's paradox of the gentle murderer. Finally, we will link back these considerations to Chisholm's scenario.

1.4.1 Philosophical analysis of SDL's syntax's core elements.

Here we will discuss whether SDL's core elements make sense from a deontic perspective. As we argued in §0.2, this preliminary analysis can already reveal some implausibilities in how the deontic modalities are (syntactically) treated in SDL. The axiom **O-D** 'obligation to x entails permission to x' is clearly desirable and unproblematic. The others require some thought.

O-K is a deontically weird axiom. We have already seen in the introduction that an argument can be given in favour of O-K. We now propose an example with no agency involved: suppose that we are in a situation where it is obligatory that, if there is a dog on the bed (a), then there is a rag below it (b). That is, $O(a \supset b)$. Assuming Oa, would we endorse Ob?

Clearly, since if a it ought to be the case that b, if it ought to be the case that a, then b is something that will end up being the case anyway (because of $O(a \supset b)$). But does this make b obligatory *on its own*?

One very important feature of deontic modals is that they are, if not defeasible, *breakable*. What if in fact $\neg a$ holds? Then nothing follows from $O(a \supset b)$, since the obligation's content is vacuously fulfilled by the given state of affairs. Oa being true but a being false, we are in a sub-ideal world which however tells us nothing about an alleged obligation to b. It seems in this case that although a and b are related by an obligation, the obligations to a and to b are not.

However, if the semantics of O is given as plain 'truth in all ideal worlds relative to the world of evaluation', then it is intuitively true that an obligation to b obtains. This interpretation makes sense up to some extent, and is clearly desirable if, for the sake of simplicity, we want to stick to O as a KD modality. Furthermore, if we take up Meyer's suggestion to consider deontic modals as describing an actual versus ideal class of behaviours, the following argument seems sound:

$$\frac{\textit{Ideally, if } a \textit{ then } b. \qquad \textit{Ideally, } a.}{\textit{Ideally, } b} \tag{1.8}$$

For these two reasons, we shall accept O-K as a desirable axiom of SDL.

O-NEC. If something is true at all worlds, is it obligatory? Is it obligatory that $p \vee \neg p$ or that the planet Earth currently exists? (assuming no consistent world can exist where these sentences are false) Unless we are very Hegel[11] in thinking that what is ought to be (*pace* Hume[12]) this is not something we would like to see happening.

A philosophical motivation for O-NEC is in fact hardly found in the literature, the only one usually given being the pragmatic need for logical simplicity. However, O-NEC is just as intuitive as the modelling choice to make D a 'perfect world' accessibility relation. If φ is true at all worlds, this implies it is true at all *perfect* worlds. Consequently, there is no world which has a better alternative at which $\neg\varphi$ is the case. This means that also $O\varphi$ is true everywhere. This axiom is quite problematic. However, it doesn't play a role in Chisholm's paradox (it does in Forrester's).

Undesirable features of SDL. SDL has many undesirable features. For example many (e.g. [PS96]) think the SDL-valid $\neg(O\varphi \wedge O\neg\varphi)$ (*inconsistency of incompatible obligations*) is undesirable. However, in 'real life' it is all too common to find oneself in situations where different obligations conflict with each other, e.g. the famous *trolley problem*.[13] There we have a situation where two duties clash. The dilemma is not only apparent: *people really do have trouble*[14] in

[11] Cit. "we must first of all know what the ultimate design of the world really is, and secondly, we must see that this design has been realized" [Heg75].

[12] The famous rule 'no ought from is' is known as *Hume's Law*. Ref. [Hum10].

[13] Cfr. [Tho85, Wik15c].

[14] See for example the experiment of [LCR08].

deciding what they would do in such situations. A deontic logic should be able to cope with this, e.g. by ordering obligations in hierarchies or layers[15], or to live with this e.g. by having $Pq \wedge P\neg q$ as a theorem. Striving to track down the source of the shortcomings of SDL, we shall start from the bottom and see what sort of tackle SDL has on the notions of *act* and *fact*.

1.4.2 What is it that ought to be the case?

Some tried to dispel the many troubles of SDL (chiefly Chisholm's paradox) by disambiguating *the paradoxical statements* it in various ways. Still, the premises do not seem at all to be ambiguous to begin with. But one could argue there is ambiguity in the *object* of the obligation. In a common-sense reading of "A must go to assist his neighbours", for example, it is clear who is obliged to do what, and what this obligation consists of. Namely we have to choose between the following interpretations of the sentence:

1. *obligatory state reading:* an ideal (*normatively perfect*) world would be such that A goes to assist his neighbours.
2. *obligatory action reading:* in an ideal world, A would carry out the action "going to assist his neighbours".

In case 1. the focus is on the status of a state of affairs 'A-goes-to-assist', whereas in 2. what matters most is the agency of 'A-going-to-assist': there is an action that A ought to do, namely going to assist.

In a setting where acts and facts are kept apart, we could tell 1. from 2. by applying the O operator to propositions describing states (i.e. that prescribe 'the next world must be such and such') versus applying the O operator to actions, e.g. as transitions from state to state. In this latter case, what is obligatory is the execution of an action (regardless perhaps of the consequences it has) and not the subsistence of a state of affairs. This sort of disambiguation needs heavy philosophical gunnery and may or may not lead to interesting considerations about Chisholm's or other deontic paradoxes. Maybe there is some plausible interpretation of natural language obligations, or a smart and plausible formalization, under which all paradoxes of this kind disappear. However, certainly it has not been found yet.

These confusions may be due to a poor treatment of the notion of obligation: what is it that ought to be the case? An action, or perhaps a state? Should we follow the early suggestions of von Wright and read p as an act type, or follow the modern usage and read p as an atomic proposition?

[15]For a somewhat related approach, see [BvdT03]. Also, analysing *prima facie* norms such as the ones just presented as having a logical form different from one another and from 'absolute' norms that are exceptionless seems a viable perspective. Also see: [PS96].

Act types versus propositions. In von Wright's 1951 system[16] the atomic particles of the language were ranging over *action types*, not propositions. Action types are, for example, (the acts of) *killing*, *eating*, *pulling a thread*.

> «First a preliminary question must be settled. What are the "things" which are pronounced obligatory, permitted, forbidden, etc.?
>
> We shall call these "things" *acts*.
>
> The word "act", however, is used ambiguously in ordinary language. It is sometimes used for what might be called act- qualifying properties, e.g. theft. But it is also used for the individual cases which fall under these properties, e.g. the individual thefts.
>
> The use of the word for individual cases is perhaps more appropriate than its use for properties. For the sake of verbal convenience, however, we shall in this paper use "act" for properties and not for individuals. We shall say that theft, murder, smoking, etc. are acts. The individual cases that fall under theft, murder, smoking, etc. we shall call *act-individuals*. It is of acts and not of act-individuals that deontic words are predicated.» [vW51, p. 2]

Action types, and not propositions, were in [vW51] the focus of obligations and prohibitions: the O operator would take (the name of) an act type and return a sentence. On the other hand, SDL's O is defined on propositions and yields sentences. Now: from a deontic perspective (or from a technical one), what difference does it make to express obligations and prohibitions on actions rather than on states of affairs / propositions? Suppose we want to formalise

"You ought to go to bed early and turn off the lights!" (1.9)

Which one of the following formulae is most suitable to formalize (1.9)?

$$O(p \wedge q) \tag{1.10}$$
$$Op \wedge Oq \tag{1.11}$$

On the one hand, (1.10) can be argued to contain only one 'ought' just like (1.9) does. On the other hand, one could say that (1.9) is specifying two distinct duties and not a single though compound one. Natural language is ambiguous here. Consider the related example:

"You ought to go to bed early and Uma ought to kill Bill!" (1.12)

Probably because of the repetition of 'ought', which is required by the English language because of the change of subject, (1.11) seems more plausible a formalisation of (1.12). A good question to ask, in this case, is if there is a reading of (1.12) which can be formalised as (1.10). The answer is: yes, if we stretch a bit the ought-to-do flavour of (1.12). The key lies in reading (1.12) as 'it ought to be the case that/ideally, you go to bed early and Uma kills bill': now the state of affairs

[16] [vW51].

that should obtain is a compound state where both things hold, as you have gone to bed early and Uma has finally killed Bill.

What if we tried to stick to an ought-to-do analysis of imperatives? (1.12) simply asserts that there are two actions that ought to be carried out (by the respectively relevant agents). The point is, (1.9) can be analysed in very much the same way! Where is the difference then?

These questions arise also with the other boolean connectives that build up complex actions or states of affairs from basic building blocks. We will thus now climb up one level of generality and inquire whether SDL's treatment of meanings does justice to the way we commonly reason about norms.

1.4.3 SDL's treatment of meanings.

Whereas Mally had thought of p standing for a state of affair, von Wright had conceived them as denoting *act types*.[17] One crucial consequence is that if p is an act type, then iteration of deontic operators is not allowed: O is defined on acts only, and 'Op' is not an act. So, the choice of taking p to be an act or a state of affair is not devoid of consequences; in von Wright's system p being an act, e.g., of *killing*, Op is the proposition that it is obligatory to perform act p. Then OOp is not well-formed because O is defined on act types but returns propositions, so Op is a proposition and cannot fall in the scope of another O. Also, not well-formed are formulae that combine deontic and non-deontic parts such as $p \supset Oq$.

If we say p denotes a state of affairs (or a sentence) and say that O returns entities of the same kind it receives, then the iteration of Os becomes plausible. Op would then say that a state where p holds is deontically desirable (or that it is obligatory to see to it that p); similarly, OOp would say that a state such that p is desirable is in turn desirable. This seems perfectly understandable, and is thus an argument in favour of an interpretation of p that makes such 'OOp' constructions meaningful.

Boolean operations on act types. If p refers to an act type, the boolean operations become suddenly difficult to interpret. E.g. if p, q refer to some specific act types $\boldsymbol{p}$,$\boldsymbol{q}$, then $p \wedge q$ would stand for *a compound act type*, i.e. the class of acts of both type $\boldsymbol{p}$ and type $\boldsymbol{q}$ ($\boldsymbol{p} \cap \boldsymbol{q}$). If for example $\boldsymbol{p}$ is the act-type *killing somebody* (i.e. the class of acts that involve killing somebody) and $\boldsymbol{q}$ is the class of acts one does with a teaspoon, $\boldsymbol{p} \cap \boldsymbol{q}$ is a complex act predicate whose act-individuals lie in the however implausible intersection of the two.

[17] The emphasis on *types* is important: other deontic logic systems such as *stit* also focus on what they sometimes call 'event types', drawing from the dynamic logic tradition, cf. [BHar].

But then, what sort of act type is denoted by $p \supset q$? There is no such thing as a 'conditional act', unless that is how we want to call acts that are performed under certain preconditions[18], not unlike to *lift a vase* there needs be a vase in your hands. In conclusion, *it is desirable to let the variables such as p, q denote states of affairs / propositions*, which is in fact what most people do nowadays.

1.4.4 A semantic glance over SDL's treatment of contrary-to-duty obligations: to Forrester's paradox and back.

Assuming the p, q variables of SDL denote states of affairs, are SDL meanings adequate for our deontic purposes? This is what we set out to investigate now. A CTD statement such as, in a context where killing is forbidden, *if you kill, you ought to kill gently* (F2) cannot be easily captured by a semantics like the one SDL relies on.

Forrester's *gentle murderer* paradox.

Let k denote 'you kill' and g 'you kill gently'. A common (cfr. [For84, PS96, vdTT99]) SDL formalization of Forrester's paradox is:

$$\text{F1} \mapsto \quad O\neg k \tag{1.13}$$
$$\text{F2} \mapsto \quad k \supset Og \tag{1.14}$$
$$\text{F3} \mapsto \quad g \supset k \tag{1.15}$$
$$\text{F4} \mapsto \quad k \tag{1.16}$$

Given this formalization, a paradox is reached in a little number of steps.

Og	MP (1.16,1.14)	(1.17)
$O(g \supset k)$	NEC(1.18)	(1.18)
$Og \supset Ok$	K-O axiom + MP on 1.18	(1.19)
Ok	MP(1.17,1.19)	(1.20)

And here we are, at the undesirable conclusion that it is (unconditionally, or *categorically*) obligatory to kill someone.

SDL's expressive inadequacies, semantically. Nobody would hold that F2 means that killing gently *in general* becomes a moral duty once you killed someone. What it says is, *some particular killing act* should be a gentle killing instance, and not a gruesome murder. We let k, g range over specific act-instances, and so g

[18] Meyer, in [Mey88], uses 'conditional acts' in this sense. See §2.

denotes this particular killing specimen, and k denotes (another) specific murder. The two variables should share part of their reference: the person who dies and the person who kills, for instance, are probably expected to be the same. E.g. let a be *lady A.*, the murderer, and b *Bill the lawyer*, now corpse. k will now stand for "a kills b", and g will stand for "a murders b gently".

That $\mathcal{M}, w \models p$ just means that p is true at w: SDL-models holds no information on the temporal reference of propositions. Many have tried to focus in the dynamics of deontic agency from a temporal perspective, including [Tho81b] and more recently [vdTT98]. Here we will stick to bare SDL and ignore all temporal aspects; however, we should keep in mind that these could play a crucial role.

We have argued above that from a syntactic perspective neither $k \supset Og$ nor $O(k \supset g)$ are plausible candidates for formalizing F2 in SDL. This is also true from a semantic perspective.

CTDs and preference hierarchies. In Forrester's scenario, the obligation is about *a way of* performing something; a prescription about actions, and not about states. It seems that best would have been not to kill b in the first place, but since this is unavoidable for some reason we better ask a about, still it is more desirable for actions of a certain kind to lead to that ending rather than others (cfr. [SA85]). F2 seems to be laying out a hierarchy of preferences:

Best: do not kill anyone: $\neg k$ (and, obviously, $\neg g$)

Bad: kill someone gently: g (and, clearly, k)

Worst: kill someone (not gently): k (but $\neg g$)

Can this hierarchy of preferences be expressed in SDL? Not in any straightforward way. Let's turn once more to the two formalizations of F2 (and C2).

- $O(k \supset g)$ seems to be saying that in all deontically ideal worlds, all *a kills Bill* instances are in fact *a kills Bill gently* instances. Now, since killing Bill is not morally advised ($O\neg k$), all ideal worlds relative to it are $\neg k$-worlds. And there, g cannot be true either since $g \supset k$. Hence, in these worlds, $\neg g$ and $\neg k$ are the case, and these are the deontically 'good' worlds.

However since $O\neg k$, all deontic alternatives force $\neg k$. So, there can be no deontic alternative where k, and so $O(k \supset g)$ is always vacuously satisfied, without there being a single g world. So, in no deontic alternative g is true and the secondary obligation a has of killing gently cannot be captured by the model in any way.

- $k \supset Og$ Suppose lady A. at w has indeeed killed Bill the lawyer, b. Then, at all deontically ideal worlds relative to w $(d_0, d_1...)$, a kills b gently. However, at the same time, it is still true at w that $O\neg k$; so at all d_i it also holds that $\neg k$. Still, b implies k. So, since $d_i \models k \wedge \neg k$, $d_i \models \bot$. This is the model-theoretic explanation of one of the paradoxes of derived obligation: here doing something forbidden

produces an inconsistency in all the deontically perfect worlds, trivializing them all in one fell swoop. We could truthfully say that yes, lady a is 'obliged to perform a murder', but this is misleading; she is obliged to perform *a specific* murder, namely that of poor Bill the lawyer. But she wouldn't be obliged to kill b gently, weren't she killing him already. So, hers is a secondary duty, that kicks off only as she fails her primary duty that is not to kill anyone.

Chisholm in the light of Forrester.

After this analysis of the *gentle murderer*, a paradigmatic CTD puzzle, it is time to go back to Chisholm's paradox and draw some conclusions.

Chisholm's paradox as a contrary-to-duty puzzle. That C1 and C2 (and C4) coincide with the formulae of Forrester's paradox (in virtually all SDL formalizations I have seen) shows there is a great deal of overlap, to say the least. This notwithstanding, the question is whether having a flawless treatment of CTDs would make Chisholm's scenario *totally* unproblematic in SDL. Given the absence of flawless treatments of CTDs in SDL, this is a question that remains for the moment unanswerable. Nonetheless, I believe there are good hopes that the answer would be positive.

Further remarks. Since Forrester's paradox is to some extent a fraction of Chisholm's paradox, all the considerations that we have sketched in this section also apply to that fraction of the latter. A semantic analysis of (equivalent versions of) C1, C2 and C3 and their rendering in SDL was able to shed some light on why such rendering is in many ways faulty.

After all this, it seems quite obvious that SDL is unable to deal with the Chisholm scenario in a natural way. To see it, we had to dive into the mechanics of SDL and analyse bit by bit its treatment of the O operator.

1.5 Conclusion.

Both syntactically and semantically, SDL seems in principle unable provide an intuitive account for CTDs, and *a fortiori* for Chisholm's and many similar deontic paradoxes.

From a semantic perspective the problem can perhaps be traced back to SDL frames, that do not contain enough structure to model faithfully the situations we are trying to cope with here. We have seen that all obvious ways to formalise CTD sentences fall to paradoxes and implausibilities.

SDL lacks fault tolerance, norm fine-grainedness, and time. To name a pertinent triviality, SDL is certainly not fault tolerant and not norm-fine-grained enough. Concerning norm-fine-grainedness, what I mean is that there is no difference between breaking a norm or breaking two, and there is no difference of severity between breaking a norm and breaking another. Our intuitions tell us that there is a difference between an obligation not to kill and an obligation not to insult people at random. Similarly, we expect a difference between "do not kill" and "do not kill brutally and mercilessly"[19], not to mention intentionality differences. SDL is not at all able to capture these though macroscopic distinctions.

Concerning fault tolerance, this is something that has been pointed at to explain SDL's failure with Chisholm and with CTDs in general ([Hor93]). CTDs can in fact be seen as failure recovery instructions; and fault tolerance is a critical feature in deontic logics meant to express, e.g., database integrity constraints (cfr. [CJ96]). Fault (in)tolerance may come in many forms. SDL is a system that would work well (i.e. would point the 'right thing to do') under the assumption that the agents do what they ought to. If this assumption is dropped we allow for 'mistakes' to occur in the system. CTDs are rules on how to take the best out of the worst, and the fact that they are basically useless in SDL shows the point.

Finally, we note that no notion of time is present in the logic, and this may be a fatal flaw. SDL might be good still as a logic of synchronic obligations[20], but Chisholm's paradox involves some dynamic normative reasoning: once you fail to go to assist your neighbours, the normative landscape changes and *new obligations take place of the old ones*, your initial failure notwithstanding. This just cannot be captured in SDL.

Consequently, we hope a dynamic logic (PDL-style) can be a powerful tool for addressing the issues we met in this chapter. PD_eL, the system the next chapter is about, is still not 'fault-tolerant' in any obvious way. However, Meyer in [Mey88] claimed that the Chisholm scenario were unproblematic in it; consequently, taking a look at the framework will be worthwhile. Not only PD_eL is based on a (quite complicated) action logic, but also it (implicitly) encodes some notion of *time*. Every action takes place in time, so the time-unit of the logic is *the time it takes to perform an action*. These two features, which are central to Chisholm's paradox, may be crucial in solving it.

[19] Capturing these can be attempted once one has a deontic logic able to express preferences, such as the one presented in [vdTT99].

[20] As opposed to *diachronic* (or 'dynamic', I guess) obligations.

Chapter 2

PD_eL.

In this chapter we will deal with Propositional Dynamic Deontic Logic (PD_eL henceforth), which is a deontic Propositional Dynamic Logic (PDL) variant. PD_eL is an action-based logic developed concurrently (and often in competition) with *stit*, first proposed in 1988 by J-J. Meyer([Mey88]). Very broadly the idea is to use the transition systems which are specific to dynamic logics (and PDL in particular) to model agents' actions. Graph-theoretically speaking, nodes are states of affairs and edges correspond to actions, conceived as 'programs' that transform a state into another.

2.1 Introduction.

In §2.2 we sketch informally syntax and semantics of PD_eL, pointing as before to Appendix B for the formalities. In §2.3 we present a dynamic Chisholm variant and in §2.4 we discuss its features. We will see how not only formalization is problematic, but also the logic itself has some implausible theorems. To answer some of this problems, in §2.4.4 we follow Meyer in extending PD_eL to $PD_eL(n)$, where multiple violation atoms are added to distinguish different 'sorts' of obligations. A conclusion follows in §2.5, where we wrap-up on the features of PD_eL and $PD_eL(n)$, and consider whether a similar extension of SDL to an hypothetical $SDL(n)$ would be beneficial.

> «The system for (Propositional) Deontic Logic presented in this paper [...] does not contain the very nasty paradoxes that often appear in other systems in the literature, especially where the connection between actions and assertions is concerned. Although based upon Anderson's idea for reduction (cf. [And67b], [McA81]), it lacks the undesirable consequences of Anderson's original reductions.» [Mey88, p.110]

unfortunately, PD_eL is *not* devoid of paradoxes. In §2.4.1 we examine one of them, discovered by Anglberger, and discuss its relevance for Chisholm's paradox. The latter, however, has apparently been solved: by looking into why Chisholm's scenario is not a problem in PD_eL (if it really is not), we can hopefully learn more about why it *is* a problem in other systems. To address this question, in §2.4.2 we inquire whether the formalization Meyer suggests does justice to Chisholm's set (and we will argue that it does not).

2.2 Informal syntax and semantics of PD_eL.

Following the intuitions made famous by Castañeda in [Cas81], PD_eL distinguishes syntactically between assertions and practitions. Roughly, the main difference is that *assertions* can be asserted, but not performed whereas *practitions* can be performed, but not asserted. Given two sets of atomic propositions and atomic actions, some operators combine them in two sets of *actions* (analogous to *practitions*) and *assertions*.

Actions. The operators $\underline{\cup}, \&, ;, ^-$ are used to build up inductively out of the simple ones $(\alpha, \beta...)$ the set of complex actions *Act*. $\&$ is the *simultaneous execution* operator, and $\alpha \& \beta$ reads 'α together with β'. $\underline{\cup}$ is the classical dynamic *choice* operator, and $\alpha \underline{\cup} \beta$ reads 'α or β'. ; is the also classical *sequential composition* operator, so $\alpha; \beta$ reads 'α followed by β'. $^-$ is the *negation* operator, thus $\bar{\alpha}$ reads 'not-α'.[1] Conditional actions $(\varphi \rightarrow \alpha/\beta)$, which in this thesis will play a minor role, will be introduced only when they will become relevant. For the moment, it is best to focus on the rest.

Assertions. The set *Ass* of assertions is obtained similarly. The language of assertions is a propositional language plus a modal box operator $[\alpha]\varphi$ (with α an action, φ an assertion). It is similar to the standard 'necessity' box $\Box$, but is labelled with the name of an action. E.g., $[a]\varphi$ denotes in PD_eL that after performing action a, φ holds. We will usually read $[a]\varphi$ as "if action a is done, φ will hold (afterwards)" ([Mey88, p. 110]). Differently put, sufficient condition for φ to hold is executing α. Meyer explains that "$[\alpha]\varphi$ is a more refined version of $\alpha \supset \varphi$ in traditional deontic logic with the difference that now actions and assertions are separated, and a notion of time-lag is built in" ([Mey88, p. 110]). $[\ ]$ has a dual: $\langle\alpha\rangle\varphi := \neg[\alpha]\neg\varphi$. Just like the K diamond, $\langle\alpha\rangle$ existentially

[1] Following [Mey88], we interpret $\bar{\alpha}$ as 'any (set of) action(s) which does not include α'. The semantics of negation is controversial (e.g., [Bro04a, Bro03b, Wan04]). However being this discussion so little specific to what we are interested in, we shall not go into it.

quantifies over the outcomes of executing α. $\langle\alpha\rangle\varphi$ thus means that it is possible that executing α will result in a state where φ holds.

Next, we add to *Ass* the 'violation' atom V (see §0.2.2). Then the deontic operators are defined via an Andersonian reduction. First the 'it is forbidden to' F is introduced as $F\alpha := [\alpha]V$. The intuition is that performing action α is forbidden iff any execution of α results into a state of violation. Its dual, the 'it is obligatory to' O operator is defined as $O\alpha := F\bar{\alpha}$.

Semantics.

[Mey88] gives first an informal, then a formal semantics for PD_eL. Being the formal part admittedly long and complicated, here we will give a mere impression of it. Broadly, PD_eL's semantics is based on a labelled transition system, where labels are not act types as in for example PDL, but *sets* of them. These are called *sincronicity sets* (s-sets for short). Bearing resemblances with LTL-like linear models, *actions* are identified with bundles of histories (or *traces*); an action bears information of not only its outcomes, but also all further actions that will be possible after executing it. We will identify here 'action' with the notion of s-set; *carrying out an action* is equivalent with concurrently executing a finite and nonempty multitude of atomic actions. Thus $s \models [\alpha]\varphi$ means roughly the following 'all s-sets containing α, when simultaneously executed in s, transform it into some s' such that $s' \models \varphi$'. However, for the purposes of this thesis, it will be sufficient to read $s \models [\alpha]\varphi$ simply as 'in s, after α, φ'.

Remark on simultaneity. Simultaneous execution, usually simulated by interleaving, in PD_eL is built-in: at a given state one can in principle perform any (finite) set of actions. By contrast, in PDL executing $\alpha \cup \beta$ is equivalent to executing any of the following: α, β, $\alpha;\beta$, $\beta;\alpha$. That is, the closest we can come to simultaneous execution is in fact sequential composition.In a way, we could say that while PDL is the logic of a single-core computing machine, Meyer's PD_eL is the logic of a many-cores one. More precisely, since only finitely many actions can be executed simultaneously, it is the logic of a finite-cores machine.

2.3 Dynamic Chisholm's paradox.

In [Mey88] a PD_eL formalization is put forth of Chisholm's paradox and then is argued that the paradox "just vanishes". This is the formalization he proposes:

$$\text{C1} \mapsto \quad O\alpha_1 \tag{d1}$$
$$\text{C2} \mapsto \quad [\,\alpha_1\,]O\alpha_2 \tag{d2}$$
$$\text{C3} \mapsto \quad [\,\bar{\alpha}_1\,]O\bar{\alpha}_2 \tag{d3}$$

The assumption ($\simeq$ C4) that in fact $\bar{\alpha}_1$ is executed cannot be expressed in basic PD_eL: we would need a $done(\alpha)$ operator. To see how the system behaves in this context we have to play along the dynamics of PD_eL and suppose we are in a world where (d1-d3) hold, in an otherwise arbitrary PD_eL-model $\mathcal{M}$. Then we would see that, as Meyer himself remarks, that there is no paradox: assume in a model $\mathcal{M}$ (d1) through (d3) are true in some world w. $O\alpha_1$ holds iff $[\,\bar{\alpha}_1\,]V$; since $[\,\bar{\alpha}_1\,]O\alpha_2$ (d3), we conclude that $[\,\bar{\alpha}_1\,](V \wedge O\bar{\alpha}_2)$. So we obtain the derived obligation that you should not tell that you are coming, despite being in a state of violation, and this is good. Furthermore we have a few other desirable entailments, i.e. that at w it holds that: $O(\alpha_1;\alpha_2)$, and $[\,\bar{\alpha}_1\,]F\alpha_2$ (proof omitted).

2.4 Discussion.

Where all doubts about PD_eL and its treatment of Chisholm are either dispelled or deepened (mostly the latter).

2.4.1 "No possible action is forbidden".

In [Ang08], Anglberger notes that the following is a PD_eL validity[2]:

$$F\alpha \to [\,\alpha\,]F\beta \tag{2.6}$$

[2]The deduction is as follows:

$\vdash F\alpha \to F(\alpha \,\&\, \beta)$	proof in [Mey88, p. 116]	(2.1)
$\vdash F\alpha \to F(\alpha;\beta)$	"Theorem of excluded Robin Hood" from previous line, proof in [Ang08, p. 434]	(2.2)
$\vdash F\alpha \to [\,\alpha;\beta\,]V$	Meaning of F	(2.3)
$\vdash F\alpha \to [\,\alpha\,]([\,\beta\,]V)$	From previous line and axiom $(;)$	(2.4)
$\vdash F\alpha \to [\,\alpha\,]F\beta$	Meaning of F	(2.5)

Just like in SDL we had $(Fp \wedge p) \supset Fq$, here we face similar problems. But Anglberger's critique goes further: from (2.6) and the *no-conflicting-obligations* (NCO)[3] assumption

$$\neg(O\alpha \wedge O\bar{\alpha}) \tag{NCO}$$

it is in fact possible to derive the rather undesirable

$$F\alpha \rightarrow [\,\alpha\,]\bot \tag{CON}$$

that in turn results into even more paradoxical consequences such as $\neg(F\alpha \wedge \langle\alpha\rangle \supset p)$, that says that no possible action is forbidden.

Anglberger argues that the source of the troubles in PD_eL lies in its unclear stance between a *goal-oriented* and a *process-oriented* conception of norms.[4] Whereas the general definition of obligation (as V takes place in the target state only) seems to go by the first conception, the action algebra seems process-oriented (insofar as, for example, not executing $\alpha;\beta$ is equivalent to executing either $\alpha;\bar{\beta}$ or $\bar{\alpha}$ straight away).

This notwithstanding, what is the import of what we are discussing concerning Chisholm's paradox? Problem-specifically, very little. Of course we can now show that assuming $O\alpha_1$ (d1), which means exactly $F\bar{\alpha}_1$, together with the assumption that α_1 is done immediately after, entails that $\bot$ holds in the outcome state. So, clearly the "Chisholm situation" is not handled properly, but this is not a problem specific to it or to CTDs. In other words, the paradox that Anglberger outlines in the paper just discussed is way more general than Chisholm's paradox, and so discussing the latter in terms of the former would shed very little light on either issues. However, what Anglberger sorts out is a quite disruptive critique to the whole PD_eL.

[3]NCO is, by all authors I am aware of, in most contexts, deemed a desirable property of a deontic logics. In most cases, however, NCO is not derivable in the logic and has to be added as an additional axiom. Adding it often results into paradoxes, and PD_eL is no exception. In PD_eL, furthermore, NCO is equivalent to the PD_eL form of SDL's O-D axiom: $O\alpha \supset P\alpha$. So, since O-D is clearly desirable (even in PD_eL), NCO should be as well. Besides this, NCO is desirable for we would like a moral system to state clearly what we should or should not do, without any overlapping between the two. Our moral reasoning seems to break down in case there are *dilemmas*, which are precisely those situations where $Op \wedge O\bar{p}$. In a way, the perfect ethics would be, according to the standard view, like a program in this respect: it should not lead those who comply with it to deadlocks. The point is, like many point out e.g. [AB81], the best-so-far (human) normative systems are full of "contradictions" of this sort. A system of deontic logic displaying misbehaviour as a consequence of accepting NCO might thus simply be not fault-tolerant enough to model real-world normative systems.

[4]For more on the distinction see e.g. [Bro03a, p. 177].

2.4.2 Chisholm's paradox formalization issues.

Meyer claims rightfully that the formalization he gives of Chisholm's set is unproblematic. Here, we discuss whether *the formalization* itself is rightful.

Independence. Anglberger remarks in [Ang08] that independence of the sentences of the Chisholm set is not preserved in Meyer's formalization. The point is, (d1) entails (d3).[5] Now, as we already argued above, independence of the sentences of Chisholm set is not *per se* an absolute value to us. Furthermore, PD_eL has much worse problems than this.

Uniformity. A *positive* feature of the formalization of Chisholm's paradox presented above is that uniformity is achieved between (d2-d3). This is possible because PD_eL models more finely than PDL can the logic of actions. However, we shall ask ourselves whether this is really the uniformity we want. Iis the Chisholm's set accurately captured by this proposed formalization?

Problem I: isn't there a conditional in C2?

To see where the first problem lies, we will take a closer look at C2. The assertion C2, at a first glance, appears to contain a conditional. Where is the conditional in (d2)? Maybe we do not need one, for the modal box already contains some implicit notion of conditionality, which reveals itself when you read $[a]x$ as "once done a, x holds".[6] Then the issue becomes whether this implicit notion does justice to the apparent logical form of C2. Meyer does not discuss the formalization he suggests at all; we will have to do the job.

Modal box as a temporalized material implication. First of all, we see that the intuitive understanding we are offered of the modal box is very much close to a temporalized material implication; that is, just a material implication with a built-in time lag between antecedent and consequent. $[\alpha_1]O\alpha_2$ means that executing any set of actions that includes or entails α_1 in the present state results into a ('later') state where $O\alpha_2$ holds. On the other hand C2 says that if we go (or anyway do anything that implies going) to assist our neighbours then we are obliged to warn them of our arrival. Stretching a bit the terminology, we can easily

[5] See previous section: $F\alpha_1 \to [\alpha_1]F\beta$ for arbitrary β. So assuming that $O\alpha_1$ (d1), which is equivalent to $F\bar{\alpha}_1$, and instantiating (2.6) as $F\bar{\alpha}_1 \to [\bar{\alpha}_1]F\bar{\alpha}_2$ we obtain by MP that $[\bar{\alpha}_1]F\bar{\alpha}_2$, that is equivalent to $[\alpha_1]O\bar{\bar{\alpha}}_2$, which in turn is nothing but (d3).

[6] On a related note, we might remark that usually an Anderson-Kanger reduction involves a conditional made 'strict' by an historical necessity operator. In PD_eL, instead, $O\alpha := [\alpha]\neg V$. This also suggests reading $[\ ]$ as a disguised implication.

read off C2 precisely (d2). So the intended reading of (d2) bears no significant difference with the one of C2 and thus qualifies as a plausible formalization for it.

Problem II: translation.

The second question that pops to mind is: what exactly should α_1 be denoting? If α_1 is the (possibly compound) action *going-to-assist*, (d2) would be asserting that some obligation (presumably that to tell the hosts about an incoming guest) kicks off only once the action *going-to-assist* is carried out. I.e. (d1-d3) is a formalization of the *forward* reading of Chisholm's paradox, the one where you first go and then tell. We shall inquire, here, how we could possibly formalize the (much more intuitive) *backwards* Chisholm set, where you are going *after* telling. Since it is hardly possible to assign to α_2 a backwards temporal reference such as 'having warned them before coming' we have to look for other ways of achieving the same.

A 'commitment' reading of α_1. We could try to read α_1 as 'being set out to (committed to) going to assist', by which I mean that you have undertaken a course of actions that will lead you to assisting them (such as getting on the bike, turning on the lights...). We can picture this ourselves as an internal 'final decision/resolution', or as a 'acting in such a way that the outcome is inevitable'. α_1 *should* entail that you will be going to assist in the future, but *not* that you are already there or done with it.

We could read (d2) as "once you have committed yourself to going to assist your neighbours, it becomes your duty to tell them you are coming". However, not in all situations (or meanings of 'commitment') this would make sure that you will go. We can presume in any case that Chisholm's set is not a future contingents problem: nobody expects you to predict the future and tell your prospective hosts whether you will or will not actually go. So, we can also presume that the notion of commitment we are concerned with here is a purely internal one that is mainly or exclusively a matter of (your) intentions. In this light, the 'commitment' reading of (d2) seems the most plausible one. Now, this was a disambiguation of C2 rather than (d2). Having clear in mind what we just found out, is (d2) a worthy formalization of C2?

Another reading. Under these assumptions on α_1's reference, had not we better render C2, C3, jointly, as (d2-3'):

$$O((\alpha_1;\alpha_2)\sqcup(\bar{\alpha}_1;\bar{\alpha}_2)) \qquad \text{(d2-3')}$$

What (d2-3') appears to be saying is: "it is obligatory that one of the following is executed: either you first commit yourself to going, and (immediately after-

wards)[7] you tell them you are going; or you first commit yourself not to go, and (immediately after) you do anything but telling them you are going".

The semantics of $\alpha_1, \bar{\alpha}_1$ prevents them from being carried out simultaneously, so we have a faithful rendering of a binary choice situation: either α_1 then α_2, or $\bar{\alpha}_1$ and then $\bar{\alpha}_2$, but not both. However, (d2-3') is implied by $O(\alpha_1; \alpha_2)$ (this is a variant of Ross' Paradox, cfr. [Mey88, p. 116]). So this proposal is particularly uninteresting.

Formalizing the *backwards* Chisholm set in PD_eL.

The main problem we have with the formalization of the 'backwards' C2 as

$$[t]Og \tag{$m2$}$$

is that this looks like a translation of "after you told them you are going, you ought to go" rather than of C2. Is there a way to obtain a more intuitive translation?

A solution with conditional actions. The full PD_eL logic allows also for *conditional actions*, of the form $\varphi \to \alpha/\beta$. Roughly, a conditional action $\varphi \to \alpha/\beta$ can denote either one of two actions α, β; it denotes α in case φ evaluates to true in the current world, β otherwise. So what if we rendered C2 (and similarly C3) as

$$O(\varphi \to t/p) \tag{$m2'$}$$

What ($m2'$) says is, roughly, that t ought to be performed conditional on φ being true. The only problem is now to fill in the φ. What we would like to plug into φ is, plainly, "you will go to assist your neighbours". I can see no use of this formalization without tenses.[8] The conclusion is, this solution is viable only if we are willing to extend PD_eL with temporal operators.

A solution with sequential action composition. It is tempting to adopt the following translation for C2:

$$O(t; g) \tag{$m2''$}$$

[7] What we really would like to say is that you have to tell them you are going any time soon, or simply any time before actually going. But we will ignore these subtleties here.

[8] Here we are rendering separately C2 and C3. Conditional actions could however be exploited to render the two in a single expression such as $O(\varphi \to t/\bar{t})$, to the effect that, provided φ is some expression that goes true iff you are settled on going to assist and goes false iff you are settled on *not* going, you will tell them you are going only in the first case, and not in the second. This could work, but finding such a φ looks, again, rather hard.

What we want to express, with the two of C2 and C3 together is that there are two possible 'good' courses of actions that one should follow: either you t *followed by* g, or you $\bar{t}$ *followed by* $\bar{g}$. It comes quite natural to try to use a sequential chaining operator. We could say that it is obligatory that either "t and then g", or "$\bar{t}$ and then $\bar{g}$", deferring the choice between the two options to the agent (even though in fact the choice is constrained by C1).

C2 is stating that a suboptimal course of action would be to first not tell you are going but then (when the time comes) actually go to assist.[9] We have also called this a 'commitment' reading of C2: telling you are going *commits you to* actually going (when the time comes). In this respect, $(m2'')$ seems to provide quite an intuitive translation. However, we must spend a minute wondering whether the same also works flawlessly for C3. The point is C3 ideally means that that not telling you are going *commits you to* not actually going. Going is a duty of yours in the first place, so only a 'commitment reading' can avoid the implication that not going is a duty as a consequence of $\bar{t};\bar{g}$ being a duty. However, in PD_eL, $O(\bar{t};\bar{g})$ implies $O\bar{t} \wedge [\bar{t}]O\bar{g}$[10] so an $(m2'')$-style translation of C3 would be inadequate, for it implies that $\bar{t}$ is a duty, which is clearly not what we want.

2.4.3 Time flow and dynamics of PD_eL.

Consider the sentence $[\alpha]\varphi$. It means that, whenever α is executed, a state is reached where φ holds. As we have remarked already in the previous section, there is a time-lag built into this formalism. Executing α means to execute simultaneously a set of atomic actions.[11] Only *after* that sequence will have been fully carried out, φ will hold.

PD_eL's treatment of time. Many authors (e.g.[Mot73, Tho81b, Bro04c]) have claimed that time is an essential component of deontic reasoning. As far as I can tell tense modalities cannot be *expressed* in the object language of PD_eL. In other words, PD_eL is a dynamic logic and not a temporal logic, and this is an important distinction. Still, what sort of dynamism is PD_eL concerned with?

PD_eL's dynamics. Actions take place in time. At the core of a logic of action is thus reasonable to expect a proper treatment of time. Time is so relevant because actions change how the world looks like, and changes (before–after) involve a

[9] C3 would state exactly the converse: that you are forbidden to tell that you are going, and then not go.

[10] Cfr. [Mey88, p. 116].

[11] Actually, α is a bundle of sequences of *syncronicity sets*, that are nonempty sets of atomic actions. For simplicity, we will speak here of 'atomic actions' only, but a precise reader might want to read 'bundles of sincronicity traces' instead.

time gap. So arguably not time in itself but the logics of *changing*, or what is usually called a logic of *dynamics*, lies at the foundations of a logic of actions and obligations.

PD_eL is a logic of how actions change a state into another. Recalling the distinction between prescriptive and descriptive norms, in PD_eL $F\alpha$ is *purely descriptive*. In fact, it means that the (current) world is such that all possible changes which involve executing α are bad. Actions can change the world, but not the model. That is, no new connections can be created, no new worlds can become accessible, no forbidden actions can become permissible.

PD_eL as a logic of instantaneous prescriptive norms. The time-lag built into the modal box is too implicit to be useful in practice, and PD_eL's dynamics is concerned with how a single world (not a model) changes due to action. So what kind(s) of norms can be expressed? Clearly some form of hypothetical reasoning is possible: e.g. we can express that after doing α, doing β is always possible or forbidden. Still, CTD scenarios are out of the picture. Doing something forbidden prevents any further deontic reasoning, and this is something Meyer got soon aware of (see next section).

Given these constraints, PD_eL is a logic of instantaneous prescriptive norms. By 'instantaneous' I mean that all we can express in it is the point of view of a single agent, at a single moment, reasoning about what doing some action would entail in terms of changes to the present world.

2.4.4 An extension of PD_eL to cope with CTDs: $PD_eL(n)$.

In [Mey87], Meyer acknowledges that the 'original' PD_eL framework, which is the one we have been concerned with here, is unable to cope with CTD scenarios. While discussing the Forrester Paradox we already touched upon in the previous chapter (the *gentle murderer*), he writes something similar to what we have been remarking about SDL:

> «This paradox is caused by the fact that in PD_eL by the definition of Fa as $[a]V$ only signals the event that doing a leads to some liability to sanction. So it only signals violation of some prohibition. It cannot tell whether by doing a perhaps more than one violation of the laws is committed. »[Mey87, p.86]

Acknowledging the troubles. What he means is that the logic is unable to capture the intuition that if killing someone is wrong, killing someone cruelly is even more wrong (or, better, 'of a different brand of wrong'); violation of different norms calls for different liabilities. The intuition is, then, that this can be obtained

by having many *obligation* and *forbiddance* operators.[12] By introducing (finitely) more violation atoms[13] we would be able to dodge some of PD_eL's troubles with CTDs. In fact, this is precisely what Meyer hints at in [MDW94], and spells out formally in [Mey87].[14]

If, as [vBGL10] titles, "Deontics = Betterness + Priority", we can rightfully hope an extension of PD_eL which will make it able to capture more than just one degree of betterness will give good results.

Solving some troubles. Quoting [Mey87], the extended PD_eL system which they call $PD_eL(n)$ is simply defined:

> «The system $PD_eL(n)$ is the system PD_eL, but instead of one propositional variable V, indicating some (state of) liability to sanction, we have n distinct variables $V_1, V_2 ... V_n$ indicating a specific liability to the first to n-th sanction. Furthermore, the abbreviations $F_k\alpha \equiv [\alpha]V_k, O_k\alpha \equiv F_k\bar{\alpha}$ and $P_k\alpha \equiv \neg F_k\alpha$ are introduced for $k = 1, 2, ...n$. Moreover the axiom (NP) is replaced by
>
> $$V_k \to [\alpha]V_k \qquad (\mathrm{NP}_k)$$
>
> for $k = 1, 2, ...n$. PD_eL is now the special case that $n = 1$, so $PD_eL = PD_eL(1)$.» [Mey87, p. 87]

Meyer suggests the following formalization for C1 and C3 in $PD_eL(2)$, respectively:

$$O_1 g \tag{2.7}$$
$$F_2(t; \bar{g}) \tag{2.8}$$

Here t stands for 'telling you are going' and g stands for 'going'. So (2.7) says that you are obliged$_1$ to go to assist, and (2.8) that you are forbidden$_2$ to tell you're going and then fail to go. Clearly, since you are obliged$_1$ to go, you also are forbidden$_1$ to execute $t; \bar{g}$, since doing it entails doing $\bar{g}$, which is precisely what is forbidden$_1$ by (2.7).

[12] Kuijer, in [Kui12], argues that no logic based on an Anderson-like reduction can deal satisfactorily with CTD obligations. The reason he puts forth is similar to what we are complaining about here: only two kinds of situations can be told apart as a matter of deontic properties in the basic Andersonian setting. Namely, those in which a violation has occurred, and those in which it didn't. We can dub 'bad' the former states and 'good' the latters, but this is all we can say. The problem is that, as we have argued, CTDs can describe hierarchies with *more than two layers*, which thus cannot be captured by plain PD_eL formulae. By increasing the number of violation atoms, we overcome (or dodge) this issue: we can now discriminate between having zero violations, having one, having two... and induce the desired hierarchy.

[13] And, why not, an ordering of the Vs reflecting a hierarchy of moral code violation expressing 'V_1 is worse than V_2'.

[14] Another author that chooses the same strategy is [CM09].

So now, what if after all we execute $\bar{g}$? Then we surely incur into the sanction V_1 because we have failed the obligation$_1$ stated in (2.7). However, if before doing $\bar{g}$ we have done $\bar{t}$, we do not incur into sanction V_2 because we have complied with (2.8).

More formally,

$[\bar{g}]V_1$	translation of (3.1)	(2.9)
$[t][\bar{g}]V_1$	by axiom NP and (2.9)	(2.10)
$[t][\bar{g}]V_2$	translation of (3.3)	(2.11)
$[\alpha]\varphi \wedge [\alpha]\psi \to [\alpha](\varphi \wedge \psi)$	theorem, cfr. [Mey88]	($\Box\wedge$)
$[t][\bar{g}](V_1 \wedge V_2)$	from (2.10,2.11) by ($\Box\wedge$)	(2.12)

So, the conclusion (2.12) says that if not only we do not go to assist but also we told them that we would have gone, then we are liable to two sanctions. This is clearly a desirable consequence.

presumably, he would then translate C2 analogously:

$$F_3(\bar{t}; g) \tag{2.13}$$

so that in the end we end up with a hierarchy of obligations. If we tell we are going and then fail to go, we are liable to V_1 and V_2. If we do not tell we are coming and still we go, we are only liable to V_2. If we do not tell we are coming and we do not go, we are only liable to V_1. If we do all we should and both warn that we are coming and actually go, we get no sanction.

Under the weak assumption that two sanctions are less desirable than just one, we already at this point have a partial order of executable actions depending on their consequences' desirability (i.e. the amount of sanctions they result into).[15]

[15] In case a stronger order is desirable for some applications, then a way has to be found to induce it; a simple way to let this happen is to stipulate that a primary violation is worse than a secondary one. So for example, not going to assist is *worse* than (going and) not telling that you are coming.

However, supposing that 'you should not kill' is an universal duty, it is also true in the Chisholm setting that the sentence "still, if you don't go to assist, you should not take a kalashnikov and shoot everyone there, especially the kids" is true. Violating a norm does not exempt you from other duties you might have. Still, clearly, shooting down kids sounds worse than just failing to help. So in this case a 'secondary obligation' (which is *also* a primary one though) is more important than the primary obligation.

Thus, we need to be careful with the criteria we lay down. In any case, taking the necessary cautions, I believe a linear ordering of duties or prohibitions is achievable even in this simple PD_eL setting.

2.5 Conclusive remarks.

As we have seen in the previous section, by adding multiple violation atoms and thus layering the obligations, Meyer was able to give a promising formalization of Chisholm's set.

The questions we would like to ask here are the following:

1. Did the introduction of multiple violation atoms solve all of the troubles of PD_eL?

2. Could we use the same 'trick' and add multiple violation atoms to SDL? If yes, does this also solve the paradox?

2.5.1 So is $PD_eL(n)$ devoid of problems?

No. This is easy to see; each 'layer' of $PD_eL(n)$ is at all effects equivalent to standard PD_eL. So, for every i, every PD_eL validity is a validity of $PD_eL(i)$. Thus not only all the troubles we had with plain PD_eL pop up again in $PD_eL(n)$, but they also do exactly n times. For example, $O_1 g \supset [\bar{g}]O_1\varphi$ for all φ in $PD_eL(1)$, and so $F_1\bar{g} \supset F_1(\bar{g};\bar{t})$, the infamous *theorem of excluded Robin Hood*. Similarly Anglberger's result applies to each layer: $F_1\bar{g} \supset [\,g\,]\bot$ (if we accept NCO: $\neg(O\alpha \wedge O\bar{\alpha})$). Furthermore, the limits of the logic when it comes to time and dynamics, as we noted in §2.4.3 are all still there. This means that the class of situations we can hope to capture in $PD_eL(n)$ is just a little bigger than in PD_eL. Adding multiple violation atoms has proven itself useful to avoid implausible entailments between the obligations in the Chisholm set. Nonetheless *we don't think this is sufficient*: $PD_eL(n)$ has still too many implausible theorems.

2.5.2 Layering violations in SDL: a preliminary investigation.

Despite the large differences between SDL and PD_eL, maybe it is possible to trade between them some useful techniques. One of them is certainly the one we have seen exploited in the last section, where PD_eL was extended to $PD_eL(n)$, the only difference being that in $PD_eL(n)$ norms can be layered up indefinitely by using different violation atoms.

To my knowledge, $SDL(n)$ does not yet exist. Is it even possible? Clearly, we would have to use here a SDL variant where the deontic operators have been introduced via an Anderson-Kanger reduction, in order to be able to add multiple violation atoms. Nevertheless, we can do something very similar in SDL by having many accessibility relations. Each accessibility relation will model *ideality*

relative to some norm. By contrast, the only accessibility relation in SDL modelled 'perfection': what ought or ought not be the case is modelled by what is or is not the case in the D-accessible worlds, that are supposed to be the 'deontically ideal' ones. Could we, for example, have a family of $\{D_i\}_{i\in\mathbb{N}}$ relations that encode hierarchies of perfect worlds?

We found out (too late, unfortunately) an interesting paper that seems to be proposing a logic similar to the *SDL(n)* we are hypothesising. That is Goble's [Gob00]. There, the idea is in fact to extend SDL with a family of accessibility relations and interpret each one as 'relative (to a standard) ideality'. Then he would define two operators, 'obligatory in general' (O_g) and 'obligatory relative to some standard' (O_r). Then roughly:

- $w \models O_g\varphi$ iff for *all* accessibility relations D, wDw' implies $w' \models \varphi$
- $w \models O_r\varphi$ iff for *some* accessibility relations D, wDw' implies $w' \models \varphi$

O_r is a 'weak' obligation, where for example $O_r\varphi \not\equiv \neg O_r\neg\varphi$. On the other hand, O_g is stronger and $O_g\varphi \equiv \neg O_g\neg\varphi$.

This is different from what we (and Meyer) had in mind. Still it would be interesting to see if there is any overlap and, more in general, in what ways does this approach contribute. We procrastinate its detailed discussion to future work.

SDL to *SDL(n)*: what advantages? We could define the O operator:

$$O_i\varphi := \Box_i(\neg\varphi \supset V)$$

Where $\Box_i$ is one of the n accessibility relation we have in *SDL(n)*. Now, if we define an 'absolute ideality' modal $\Box_{abs}$:

$$\Box_{abs}\varphi := \bigwedge_{x\leq n} \Box_x\varphi$$

it is clear that $\vdash \varphi$ entails $\vdash \Box_{abs}\varphi$ which entails $\vdash \Box_x\varphi$ for all x.
For example, let us consider briefly Forrester's paradox:

F1 $O_1\neg k$ F2 $k \supset O_2 g$ F3 k F4 $g \supset k$

Now, modus ponens on F2, F4 outputs $O_2 g$. F3 $\Rightarrow \Box_{abs}(g \supset k) \Rightarrow O_2(g \supset k) \Rightarrow (O_2 g \supset O_2 k)$; thus $O_2 k$ (because of F2). Similarly we obtain $O_1 g \supset O_1 k$. By modus tollens with F1 (since $O_i\varphi \supset \neg O_i\neg\varphi$) this entails that $O_1\neg g$.

Summing up, the norm/ideal n.1 forbids us to kill and to kill gently; the norm/ideal n.2 obliges us to kill and to kill gently. This doesn't seem much of an improvement.

Conclusions. We have tried to imagine whether *SDL(n)* would solve SDL's issues like *PD_eL(n)* solved (some of) *PD_eL*'s issues, but the first results were not

encouraging. However, we think more research is needed especially to see how [Gob00] could contribute to the discussion.

As we have seen, formalizing Chisholm's set in PD_eL can be as tricky as it would have been in SDL. However, the feeling we get after all this discussion is that we may need some temporal expressivity other than the 'once a, p' that comes for free with the action modal box of PD_eL. We also met several formalization issues.

For these reasons in the next chapter we will move to the *stit* paradigm, which was born as a logic of action *and* time and tense operators are readily available and meant to be used. Hopefully these factors will enable a more fine-grained translation of the Chisholm sentences. Furthermore formalization is a much more straightforward task in *stit*, for reasons we will shortly see.

Chapter 3

stit.

The first 'seeing to it that' (*stit*) logic was put forth in [BP90], following a series of papers by Chellas (e.g. [Che69]) who had laid out informally the setting.[1] *stit* more than a logic is a paradigm, with instances ranging from the relatively simple Chellas *stit* or *cstit* ([Che69, Che92]) to much more complicated *strategic, coalition stit*s as [BHar]. In the first section of this chapter we will discuss, at a general level, what the *stit* paradigm is grounded on, and what is the inspiration that keeps driving the research in this direction. Then we will turn to the formal syntax and semantics and finally, as usual, we will examine CTD-related problematics and a *stit* version of the Chisholm paradox.

The analyses of this section will partly build on [SMK13], but a specific attention (§3.2) will be devoted to Paul Bartha's work in [Bar93], which contains one of the very few off-the-shelf (and fleshed-out) deontic variants of *stit* which we will call $d_e stit$. What makes Bartha's work even more apt to open our discussion in this chapter, is that he claimed that Chisholm's scenario was unproblematic in his $d_e stit$. In §3.4.1 we will show that he was wrong. Consequently, as we try to solve the problems $d_e stit$ had, in §3.5 we will move towards a more complex, temporal *stit* logic called *xstit*. The second part of this chapter will be devoted to producing a deontic variant of *xstit* (tailored to the purpose of solving the problems we will have found in Bartha's $d_e stit$). In §3.5.2-§3.5.3, we will assess its worth against Chisholm and Forrester's puzzles.

Bartha's $d_e stit$ and Broersen's *xstit* are both *stit* logics, as the names suggest. Consequently, we choose to treat them in a single chapter and the following introduction applies to them both.

[1] For an elaborate and in-depth analysis of the paradigm, see [BPX01].

3.1 Introduction: philosophical background.

Unlike many logical systems, *stit* takes off from explicitly philosophical grounds, and is very careful in respecting some basic intuitions such as *indeterminism* and *free will*. Though in fundamental[2] disagreement[3] with these philosophical premises, I cannot deny the attractiveness of the logic and, most importantly, its wide influence and usefulness.

In the remaining paragraphs of this section we will sketch the philosophical basis of the logic, and see how the concept of *action* is consequently carved out, before moving to the formalities and the paradoxes.

3.1.1 Indeterminism and free agency.

> « Our project assumes the indeterminism of the causal order in which agency is embedded, it assumes that actions are based on real choices, and it assumes that choices are therefore not predetermined. [BPX01, p. VII]

The fundamental assumption of the whole *stit* research is, without much doubt, that of *indeterminism*. Described in a causal fashion, the *indeterminism assumption* roughly says that the future is not causally determined (as a matter of natural law) by the present and past.

Indeterminism in *stit*.

In science, indeterminism is approximately "the belief that no event is certain and the entire outcome of anything is a probability" [Wik15b]. Putting together this notion with the aforementioned causality-flavoured one, we can see how determinism is sometimes thought to undermine *free will* theses: freedom of choice is strictly related with the possibility of randomness, or chance. If, like Laplace argued a long time ago, every mental process were deterministic, then true freedom would be impossible: everything we think or do would be part of the effects of *the* universe-long causal flow of events. The possibility to freely choose (where we say that 'freedom' lies precisely in being devoid of any causal influence towards either alternative the choice consists of) is in a way equivalent to the possibility of having an internal generator of 'true randomness' that makes the choice in our name. However, this is just one of the possible views. We refer to [Esh14] for other opinions and a broader discussion.

[2] Never heard of Bohmian quantum mechanics? [Wik15a].

[3] *The cat* ([Sch35]) *is either dead or alive*. We just don't yet know which.

***stit* logic is an *endogenous* logic of agency and actions;** its purpose is to express what agents can do in what situations, and what is the case after the agents have carried out their choices in the form of actions. Actions are abstract (i.e. not named, i.e. not expressible in the object language) and the focus is instead on their outcomes. In *stit* we can talk about what agents see to (not only about what they *can* see to)[4], but not explicitly about actions (i.e. *whatever by mean of which agents can see to things*). In *stit*, you can say that *a* closes the door only by saying that *a* sees to it that the door is closed (a *fact*); you *cannot*, instead, express 'closing the door' (an *act*).[5]

Being the world indeterministic, and given that no agent is the only agent capable of free will and action, the outcomes of action (the 'next state') is subjectively undetermined. No agent, in other words, necessarily has certainty that his actions will lead to such and such consequences.[6] What is the case in the next state is, on the other hand, jointly determined by the choices of all the involved agents *plus* these of nature, which can be pictured as a special (deterministic) agent.

3.1.2 Choices and freedom.

Concretely, these considerations result in the following two fundamental features that are shared by all *stit* logics:

1. agents can, by acting, rule out some of the histories that were causally possible at the moment of the choice.[7]
2. agents' choices are *real* in the sense that they are independent of other agents' ones. This also means that each choice possible to an agent is consistent (i.e. compatible) with all the choices that are possible to some other agent.[8]

Some more words need be spent to clarify point 2.

Independency of choices. Certainly marking one of the peculiarities of *stit*, the choices of agents are independent from one another. This means that there

[4]That is what happens in ATL: there, formulae are evaluated against moment alone, without taking the history into account.

[5]In Broersen's radical terms: "actions, as [we are describing them in this paragraph], simply do not exist: any action is an agental effort to see to an effect." (private conversation).

[6]I write 'necessarily' here because there can be situations in which an agent is (and *knows*, since the basic *stit* does not restrict or even model agent knowledge) the only active agent in the current situation.

[7]«When Jones butters the toast, the nature of his act, on [our] view, is to constrain the history to be realized so that it must lie among those in which he butters the toast. Of course, such an act [...] cannot determine a unique history; but it does rule out all those histories in which he does not butter the toast.» [BPX01, p. 33]

[8](and with the laws of nature, obviously)

is no way by which an agent a's choices can influence what choices are available to another agent at the same moment. In a metaphor, each agent chooses his strategy independently from the other agents, and then the choices are 'merged' and resolved simultaneously.[9]

This entails that no incompatible choices can be available at the same moment to different agents: if a has the power to ensure φ, b cannot have the power to ensure $\neg\varphi$. *stit* is a logic of (α-)effectivity: a can see to it that φ only in case a has a winning strategy (that cannot be disrupted by anything, not even another player's choices) to achieve φ.

Freedom of choice. Partly encoded by indeterminism, partly by independency of choices, such is the *stit* treatment of freedom. It is worth remarking that independency is in some sense equivalent to the impossibility of coercion: no choice of a may influence the range of choices available to b nor the way b will choose. Instead, every agent's choice can be pictured as a refinement of every other agents' choices. Suppose a sees to it that φ, and b sees to it that ψ. The future will follow a history where $\varphi \wedge \psi$ is true (validity, as we will see, is relativised to histories and moments). In other words, every choice available to a overlaps with any choice available to any other agent.

3.1.3 Picking one out of many *stit*s.

Regardless of whether this exhausts our though idealized conception of *free will*, *stit* is a logic of agents' powers more than anything else. However, unlike other agents' powers logics such as ATL (cfr. [AHK98]), in *stit* not only we can quantify over agents' strategies, but we can also express which one is 'selected' and being followed. Next we will describe informally the meaning of the main operators of some *stit* logics, from the most basic and famous ones (*cstit* and *dstit*) to the next-time *xstit* and the backwards-looking *astit*. The reader is advised to pay specific attention to *dstit* and *xstit*, for these are the operators we will deal with in the remainder of this chapter.

cstit stands for 'Chellas *stit*', honouring the name of the founder of the *stit* tradition. The intended meaning of $[a\ cstit]\varphi$ is that "the [future] fact that φ is guaranteed by a *present* choice of the agent a" (cit. [Bar93]). In other words, a will act in such a way that in the next moment φ will hold. Game-theoretically: $[a\ cstit]\varphi$ iff a presently is following a winning strategy to obtain φ.

[9] In game-theoretic terms: there can be *no action profile gap*.

dstit stands for 'deliberative *stit*', and is definable in terms of *cstit* and alethic possibility as follows: $[a\ dstit]\varphi := [a\ cstit]\varphi \wedge \Diamond\neg\varphi$. In words, *dstit* adds to the *cstit* requirement that φ is now not settled true. The intuition is, in order to be capable of genuine *deliberation* there must be an actual possibility that what you are trying to ensure does not happen. So a can (be said to) see to it (deliberatively) that φ only if, if a wouldn't do anything to ensure φ, there would be a possibility that $\neg\varphi$ happens instead. Such possibility is usually called the *counter*. In game-theoretic terms, we can say that $[a\ dstit]\varphi$ iff $[a\ cstit]\varphi$ (i.e. 'a has a winning strategy') but not all strategies available to a are winning strategies.[10]

xstit in *xstit* time is explicitly assumed to be discrete, and a relation R_X assigns to any world the next one. The main operator is similar to the *cstit* one: $[a\ xstit]\varphi$ means that 'a sees to it that φ in the next state'.

astit a quite different *stit* operator, *astit* (for "achievement *stit*"), often appears in the literature. $[a\ astit]\varphi$ roughly means that "the *present* momentary fact that φ is guaranteed by a *prior* choice of the agent a."[11] In other words 1) now φ holds, and 2) it holds because a acted in a way that ensured φ would result. The focus here is on φ as something which has been achieved (i.e. is now true) as a result of a by-now-over course of actions taking off from a. What is seen to by a, in *astit*, depends on what a *previously* had chosen to do.[12]

Choosing a *stit*. Firstly, we will not consider *astit* for a philosophical and a technical reason. The former is that *astit* has been applied mainly to philosophical research, and never (as far as I know) to deontic logics. The technical reason is that *astit* is much more complex than its sisters[13]. In fact, *astit* has been virtually abandoned and recently little or no research resources have been invested in that direction.[14]

Secondly, we prefer *dstit* over *cstit* since the notion of *deliberation* that *dstit* sets out to express seems appropriate for dealing with deontic obligations. Requiring the existence of the counter makes sure that a's actions were (in some sense) truly *responsible* for the outcome φ. It is in fact a necessary truth of *cstit*, but not of *dstit*, that I am now seeing to it that tomorrow the sun rises. This is, in our opinion, not quite desirable.

However, as we will find out, *dstit* is not quite enough to our purposes. Consequently, we will turn our attention to the next-state-*stit* *xstit* and see that it gives

[10] Cfr. [vK86, Hor89, BPX01].

[11] Quoted from [HB95, pp. 587-588], who apparently had taken it from [Bel91].

[12] This brand of *stit* is mainly due to Belnap and Perloff: [BP90].

[13] Cfr. [BPX01]. Note: workability is near the top of logicians' concerns.

[14] Broersen pointed out (private conversation) that *astit* could naturally fit in formalizing the *backwards* version of Chisholm's paradox.

much better results. Furthermore, a *dstit*-like operator (with counter) can easily be devised in *xstit* too.

To get an Andersonian reduction up and running on a *stit* basis, the only thing we really need is an (alethic) necessity operator: then we define 'a ought to do p' iff 'necessarily, if a does not do p, a bad thing (V) occurs' just as we did in PD_eL. Consequently, any *stit* logic with a modal necessity box would suffice for introducing O in terms of necessity. Still when there is some ready-made logic around, it is wise to examine it. To my knowledge, very few deontic variants of *stit* have been investigated other than Broersen's and Bartha's.[15] The most relevant is due to Horty.

Horty's deontic *stit* variant, and why we will not discuss it. Horty has proposed in [Hor01] a deontic logic framework based on *stit*. Unlike the authors we will consider, Horty evaluates the truth of $O[\,a\ stit\,]p$ (at w) relative to the *optimal choices* available to a at w.[16] What choice is optimal depends solely on the model. A model comes with a function that distributes payoffs among the choices available to the agents. Consequently, what is optimal in some state is 'objectively so' and dependent uniquely on the model - not on the agent. The world being indeterministic, the payoff agents aim for is to be determined relative to the expected outcome of their actions. In fact, Horty uses the words 'expected value utilitarianism' to describe his theory. His aim was in fact precisely to show the power of *stit* by giving a straightforward formalization of the utilitarian theory. However, we wish not to commit ourselves to something like utilitarianism, although adopting Horty's framework would solve Chisholm's paradox rather easily.[17]

However, as we will see near the end of this chapter, perhaps inducing an ideal-to-sub-ideal hierarchy of histories is necessary for addressing CTD reasoning in *stit*. But, taking a from-simple-to-complex approach, we prefer to start with Bartha's *dstit* variant and then, if necessary, complexify at will.

3.2 d_e*stit*.

We will dub [Bar93]'s system (which he called SA) $d_e stit$ in an attempt to call back to the PD_eL name. Now we will present its syntax and semantics. Then, we will see how far we can go with Chisholm's paradox in $d_e stit$, and eventually move to a different, more complex *stit* basis.

[15] [Xu15], the only review I could find, only reports a handful.

[16] Also see the related [CJ05, vdTT98].

[17] CTDs try to lay out a hierarchy of outcome states (good-average-bad). In an utilitarian *stit* model, what is good, bad or so-and-so is *given by the model*, so in a way *trivial* from an external point of view. The same goes for [vdTT98].

3.2.1 Informal syntax and semantics.

As usual, we only explain what is strictly required to understand what will follow. Find more detailed material in Appendix C.

Syntax. The $d_e stit$ language is a propositional language plus the unary operators $\Box, F, P, [a\ dstit]$ and S. Also, we have a special propositional atom V that unambitiously stands for 'there is wrongdoing' or 'liability to punishment obtains' as in PD_eL. The $[a\ dstit]$ operator was already discussed in the previous section. About the intended readings of the other operators:

- $\Box$ is a modal *historical necessity*, 'it is necessary that' operator.
- S reads 'it is settled that'.
- F reads 'it will be the case that'.
- P reads 'it used to be the case that'.

Then, a 'possibility' $\Diamond$ operator is defined as dual of $\Box$ in the usual way. Finally, we introduce an 'it is obligatory that' operator O *à la* Anderson:

$$O\varphi \quad := \quad S(\neg\varphi \supset V)$$

But since φ is expected to always be a *dstit* statement[18], this in fact amounts to: $O[a\ dstit]\psi := S(\neg[a\ dstit]\psi \supset V)$.[19] As a matter of fact, O never occurs 'alone': any sentence of the form $O\varphi$ is such that φ is of the form $[a\ dstit]\psi$ for some agent a and some sentence ψ. The intuition is clear: a is obliged to see to it that φ iff it is settled that, if a does not see to it that φ, then there is wrongdoing.

Semantics. *stit* conceives time as a branching tree of *moments* ordered by an 'earlier than' strict partial order $(\leq)$[20]. Let a *history* be a chain of moments ordered by $\leq$. The tree is assumed not to branch backwards and to be historically connected: any two moments have a common past. In other words, any two divided moments' histories join in the past (but not in the future). $\langle moment, history\rangle$ tuples are called *dynamic states*, or *worlds*.

$d_e stit$ formulae are evaluated against a *model, moment, history* triple. A model is a structure made up by the following elements: a set of moments, their $\leq$ ordering, a set Ag of agents and a function **Choice**. **Choice** roughly returns, for every moment m and agent a, a *partition* of the histories passing through m.

[18] Because of the *restricted complement thesis* discussed later.

[19] The most common way to introduce O is to use a $\Box$ in place of the S we use. The reason for this change is that in this setting $\Box$ is a 'global' operator: quantifies over all moments and histories, whereas we only need here to quantify over histories. If we were to quantify also over moments, we would have the implausible result that all obligations are the same throughout the model. That is, they never change from a moment to another. For more, see [Bar93, p. 6].

[20] Where $\forall mm'm''(m' < m \wedge m'' < m \supset m' \leq m'' \vee m' \geq m'')$.

Only histories *not* in the same cell of the partition can be told apart by a choice of a at m.[21] Formal notation for this is $h \equiv_m^a h'$. Finally, a model comes with an interpretation function that maps atomic propositions to *dynamic states*.

A semantic entailment relation $\models$ is defined in an obvious way for the boolean cases. Less obvious are the other operators:

S $\mathcal{M}, m, h \models S\varphi$ iff $\mathcal{M}, m, h' \models \varphi$ for all h' s.t. $m \in h'$.
$\Box$ $\mathcal{M}, m, h \models \Box\varphi$ iff $\mathcal{M}, m', h' \models \varphi$ for all m' and all h' s.t. $m' \in h'$.
P $\mathcal{M}, m, h \models P\varphi$ iff there is $m' < m$ with $\mathcal{M}, m', h \models \varphi$.
F $\mathcal{M}, m, h \models F\varphi$ iff there is $m' > m$ with $\mathcal{M}, m', h \models \varphi$.
$[a\ dstit]$ $\mathcal{M}, m, h \models [a\ dstit]\varphi$ iff **1.** *and* **2.**:

1. $\mathcal{M}, m, h' \models \varphi$ for all $h' : h \equiv_m^a h'$. In words, we require that the choice cell to which h belongs (at m) is such that all next states validate φ. To put it differently, if agent a constrains the next state to be on a history h_n which is in the cell to which h belongs, then the next moment on h_n is such that φ holds there.
2. $\mathcal{M}, m, h \not\models S\varphi$. In words, we require φ not to be settled true at m, h.

Remarks on obligations. The so-called *restricted complement thesis*[22] says that "a variety of constructions concerned with agents and agency – including deontic statements [...] – must take agentives as their complements."[23] This, concretely, means that if you have a sentence like "it ought to be that p", or "a ought to do q", they both should have a logical form like $O[x\ dstit]\psi$, where $x \in Ag$ and ψ is any formula. E.g. we formalize "a ought to do q" as $O[a\ dstit]q$.

The point is that deontic constructions, insofar they are 'concerned with agents and agency', must take as complement an *agentive*. Here, *agentive* is a predicate that is true of any sentence which can be paraphrased with a *stit* statement. However, this is circular: one of the main tenets of the *stit* theory is precisely that a statement of the form $[a\ stit]\varphi$ is always agentive for a (cfr. [BPX01, p. 6]). Schematically, that a sentence is agentive for someone (an agent) simply means that the sentence depicts the agent as someone who is carrying out an action. That is, someone being *agentive* for something else; that is, having had some (causal) role in the chain of events that brought it about.

[21] This is a simplification. For the full story, look at the Appendix or the literature.
[22] Or "Thesis 5", in [BPX01, p. 13].
[23] Cfr. [BPX01, p. 13].

3.3 Chisholm in *stit* sauce.

Applying the restricted complement thesis to C1, C2 and C3, we can motivate the following translation of the Chisholm set in $d_e stit$, which is the one proposed in [Bar93]. We use g to abbreviate 'go to the neighbour's assistance' and t for 'tell the neighbours that you are going (to their assistance)'. In the sake of readability, we stipulate the shortcut $[a]\varphi := [a\ dstit]\varphi$.

$$\begin{aligned}
\text{C1} \mapsto &\quad O[a]g && \text{(s1)}\\
\text{C2} \mapsto &\quad O[a]([a]g \supset [a]t) && \text{(s2)}\\
\text{C3} \mapsto &\quad O[a](\neg[a]go \supset [a](\neg[a]tell)) && \text{(s3)}\\
\text{C4} \mapsto &\quad \neg[a]go && \text{(s4)}
\end{aligned}$$

(s1) comes quite close to C1: "it is obligatory that a sees to it that he goes to the assistance of his neighbours".
(s2) reads, quite straightforwardly: 'it ought to be that a sees to it that if a (sees to it that he) goes, then he (sees to it that he) tells them he is coming.'
(s3) roughly means that it is obligatory for a to see to it that if a does not go, then a sees to it that a does not tell he is going.
(s4) simply reads 'a does not see to it that he goes.'

A condition for factual detachment. There is a specific condition under which, in $d_e stit$, $\varphi \supset O[a]\psi$ is entailed by $O(\varphi \supset [a]\psi)$. And this is precisely when φ is a *circumstance*: that is, something that a cannot prevent from being true. Formally, φ is a circumstance iff $S\neg[a]\neg\varphi$: it is settled that a cannot see to it that $\neg\varphi$.[24] Formally, the following rule is admissible[25]:

$$\frac{O[a](\varphi \supset [a]\psi) \qquad \varphi \qquad S\neg[a]\neg\varphi}{O[a]\psi} \qquad \text{(DD)}$$

The Chisholm set is unproblematic in $d_e stit$ (?) Bartha argues that in Chisholm's scenario it is plausible to assume that $\neg[a]g$ is *not* a circumstance, i.e. that agent a *can* prevent $\neg[a]g$ from being true. In yet simpler words, we assume that there is some history at which $[a]g$ is true. Consequently, we cannot use (DD) to detach $O[a]\neg[a]t$ from (s3).

Bartha's discussion of the paradox does not go any further: he claims that because of this fact (DD) the Chisholm set is unproblematic in $d_e stit$. However,

[24] Proof in [Bar93, pp. 12-13].

[25] Here 'admissible' is a technical term. It means that adding the rule to the formal system, i.e. taking the set of formulae generated by the axioms and closure under the inference rules we already have and closing under this new rule, would not result in a different set.

let us work out a bit more the details. Assume we have a model of C1-C4 plus the assumption that a can see to it that he goes. Now since (DD) is blocked, (OR) is *not* generally true:

$$O[a]\neg[a]t \qquad \text{(OR)}$$

If $\neg[a]go$ were a circumstance, from (s3) we could deduce (s3') (via (DD)) and then, via (s4) and MP, obtain (OR). $\neg[a]g$ is a circumstance exactly in case $S\neg[a]\neg\neg[a]g$, which is equivalent to $S\neg[a]g$. So, under the assumption that it is possible for a to choose to go, we can make sure that $\neg[a]g$ is not a circumstance and so avoid detaching (OR) from C3. If it were impossible for a to go, we could argue for a genuine conflict of obligations.[26]

But, unfortunately, (OR) would have been a much desirable conclusion! Assuming C4 and C3 *should* result into an obligation to refrain from telling that you are going, since you in fact are *not* going.

A more desirable result would have been that you are obliged to refrain from telling that you are going (since you are not), and you are not obliged to tell that you are going. That is, we want to prevent the detachment of $O[a]tell$ from C2, but we *do* want to detach (OR) from C3.

Bartha only notes that being (OR) false, the paradox does not arise because we cannot detach an obligation to refrain from telling.[27] Unfortunately in this way we obtain the result that (OR) is not valid in a model of Chisholm's set. In conclusion *Bartha avoids a contradiction, but not the deontic paradox, because* d_e*stit blocks the wrong detachment.*

On a related note, Bartha 'solves' Forrester's gentle murderer in exactly the same unsatisfactory way: by blocking the deontic detachment of 'you ought to kill gently'. However, we will focus on his treatment of Chisholm.

3.4 Discussion.

As usual, we consider in this section the most prominent good and bad features of the d_e*stit* treatment of deontic modalities and in particular, of Chisholm's paradox.

3.4.1 d_e*stit*'s treatment of Chisholm's puzzle is problematic.

Bartha's explanation of how Chisholm's paradox is solved in d_e*stit* is quite hasty and in my opinion misses a few points. Here we will address them.

[26] However in *dstit*, $O[a]\varphi \rightarrow \Diamond[a]\varphi$.

[27] However, I can't see how even if (OR) were true $\bot$ could be derived: $O[a]t$ and $O[a]\neg[a]t$ *are not contradictory.* Since Bartha refuses the axiom $\neg SV$, there is a model where SV is true and everything is obligatory.

One thing Bartha got right is that we cannot detach from C2 the obligation to tell we are going. However, we cannot detach from C3 the obligation to refrain from telling either. Still, that obligation holds (for entirely different reasons though). The problem is, *there is a model in which also the obligation to tell holds at the same time.*

A (counter)model of the Chisholm set in d_e*stit*. Consider the model $\mathcal{M}$ drawn in Figure 3.1, consisting of a single moment with two histories h_0, h_1 determining two choice cells.

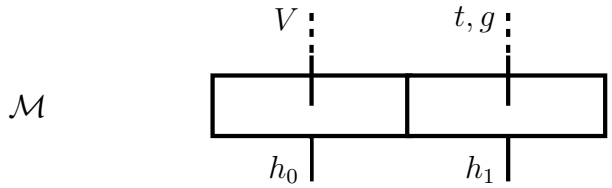

Figure 3.1: The model $\mathcal{M}$.

Firstly, we notice how at $\mathcal{M}, h_0$ both of the following hold:

$$\neg[a]g \qquad \text{since } \mathcal{M}, h_0 \not\models g \tag{3.1}$$

$$O[a](\neg[a]g \supset [a]\neg[a]t) \tag{3.2}$$

(3.2) is $S(\neg(\neg[a]g \supset [a]\neg[a]t) \supset V)$. V is true at h_0 and $\neg(\neg[a]g \supset [a]\neg[a]t)$ is false at h_1.[28] So (3.2) holds at $\mathcal{M}, h_0$.

Now, since $\neg[a]g$ is not a circumstance (as a can see to it that g by selecting the h_1 choice cell), we *cannot* apply (DD) and detach

$$O[a]\neg[a]t \tag{3.3}$$

In fact, (3.3) is *false* at h_0. (3.3) holds iff $\mathcal{M}, h_0 \models S(\neg[a]\neg[a]t \supset V)$. But $\mathcal{M}, h_1 \models \neg[a]\neg[a]t$[29] and $\mathcal{M}, h_1 \not\models V$; so $\mathcal{M}, h_0 \not\models O[a]\neg[a]t$. However, it is *true* at h_1 that

$$O[a]t \tag{3.4}$$

In conclusion, even though $\neg g$ is true and C3 holds, it is true that a ought to t. Notably, this model works even if we include the axiom $\neg SV$. That is, we are not exploiting a weakness of Anderson's reduction in general, but of d_e*stit* in particular. The model shows that Bartha's strategy to avoid the paradox by blocking the detachment of (3.3) from (3.2) indeed dodges a contradiction but fails to give the 'right' result. Thus d_e*stit* fails to meet the 'intuitiveness' requirements we declared in the introduction.

[28] $\neg(\neg[a]g \supset [a]\neg[a]t) \equiv \neg[a]g \wedge \neg[a]\neg[a]t)$, and $\neg[a]g$ is (3.1).

[29] Intuitively: $\neg[a]\neg[a]t$ means that 'a does not refrain from t'. And in fact at h_1 t is true.

Telling apart categorical and conditional obligations in d_e*stit*. Categorical obligations should be independent from the agent's choices, and in fact are in d_e*stit*. What a is categorically obliged to do should not depend on what a in fact does.

On the other hand, conditional obligations depend on what the agent chooses to do; so perhaps they should be allowed to hold at different choice cells. And, as we have already argued[30], C2 and C3 (and CTDs in general) are conditional obligations.

Graphically, we could be tempted to tell categorical obligations from conditional ones by looking at the outermost operator: if it is an O, then we have a categorical obligation; if it is a $\supset$, we do not. This however would be misleading: in some sense, $O(\varphi \supset \psi)$ is a conditional obligation.[31]

At a different level, I would suggest to tell conditional obligations from categorical ones in the following way: in the version of Chisholm's scenario we are examining here, the obligations laid out by C2 and C3 are meant to depend on a choice of the agent. On the other hand, categorical obligations (such as C1) hold regardless of what the agent does. In *dstit*, what an agent does is modelled by looking at the outcomes of an instantaneous choice. We could thus say that conditional obligations are history-dependent, whereas categorical ones are not.

Bartha's proposed solution does not lend itself to this kind of treatment: in his formalization, C2 and C3's outermost operator is an O; which is just a shortcut for a statement whose outermost operator is an S. Consequently, the conditional obligations expressed by C2 and C3 are not, in Bartha's solution, history dependent.

In conclusion, we can find no way in the present setting to distinguish meaningfully (at a formal level) conditional and categorical obligations.

Bartha, whitouth explicit motivation, did not use temporal operators in his formalization of Chisholm's sentences. Probably this has something to do with the fact that the standard F operator is too loose to be useful with obligations.[32] However, if choices were not instantaneous and their effects were visible at later moments, we could say that conditional obligations kick out only at later moments, whereas unconditional ones are true *right here*.

3.4.2 d_e*stit*'s troubles: origins and (tentative) solutions.

The fact that Bartha did not manage to give a good account for Chisholm's scenario does not mean that we have to give up *stit* altogether. We feel, instead, that the

[30] Backed up by most of the literature, e.g. [Bar93, Chi63].

[31] This would be true *a fortiori* if $O(\varphi \supset \psi) \supset (O\varphi \supset O\psi)$.

[32] The procrastination problem, and time-of-validity issues.

power of *stit* has been underused by Bartha. We will try, complexifying a bit the logic, to overcome the troubles he highlighted. First, still, we will have to pin them down.

In this (sub)subsection, we (1) consider alternative definitions of O, (2) consider alternative formalizations for CTDs (3) examine $d_e stit$'s troubles and (4) conclude we need a different *stit* to tackle Chisholm's puzzle and related CTD puzzles. This will lead, in the next section, to present the *xstit* paradigm and undertake again (this time, more successfully) the enterprise of formalizing Chisholm's puzzle.

Weakening the definition of O. We could[33] weaken the definition (O):

$$O[\,a\ dstit\,]\varphi \quad := \quad S([\,a\ dstit\,]\neg\varphi \supset V) \qquad (O'')$$

in case we want to say that a violation occurs only in case a sees to the 'bad outcome' instead of simply allowing it. This choice seems to me to depend on the specific obligation one may want to model. In some contexts it seems reasonable to expect that a will be punished iff the 'bad thing' happens (regardless of what a did); in other contexts, a will be punished iff he didn't do what he was expected to, namely to see to it that something was the case. The former case corresponds to (O''), the latter to (O). Regardless of one's philosophy of ethics, both operators can be used for slightly different purposes; (O'') seems to capture a notion of obligation focused on the outcome, whereas (O) highlights the agent's actual behaviour. So we may expect both 'oughts' to play a role in different contexts! E.g., (O'') seems perfect for formalising sentences laying out 'agentless' responsibilities such as "the room must be tidy by 9pm". On the other hand, (O) seems more suitable for more ordinary commands, such as the ones that make up the Chisholm paradox. Consequently, we see no reason to change it.

Alternative definitions of O. If we want to save $d_e stit$, the main *desideratum* is to avoid situations where $O[\,a\,]\varphi$ holds just because φ holds at some history where no violation occurs, and violations occur everywhere else. Such situation is clearly one where φ is desirable, because the only 'perfect' world is one where φ holds. Still, the obligation to φ might be a conditional one. Conditional, perhaps, on something that does not occur where we are.

1. We could hard-wire some norms in the model and having the semantics do only the 'deductive' part of the normative reasoning. This would take us one step closer to Horty's *stit*, where the oughts are given via an expected utility function. This is, in my opinion, the last stand. It makes some sense but may be overkill. We will come back to this towards the conclusion of the chapter.

[33] As suggested both in Broersen's and Bartha's works, respectively [Bro09b, Bar93].

2. We could treat conditional norms as defaults in the following sense: an obligation to $p \to q$ would detach the obligation to p only if p is true. More precisely, introduce an axiom that says that if we have a conditional norm $O[a]([a]\varphi \to [a]\psi)$, and $\neg[a]\varphi$ holds, then $\neg O[a]\psi$ holds. However, this would make it impossible to have multiple conditional norms with different antecedents and the same consequent. Finally, it would make the logic inconsistent (unless we adjust elsewhere).

3. I would be tempted by the option of redefining O as a biconditional:

   $$O[a]\varphi := S(\neg[a]\varphi \equiv V)$$

   This would certainly make O more picky. However, this would quickly lead to a complete breakdown of the deontic part of the logic in situations where we have more than one obligation. If they are compatible and there is a choice cell which satisfies both, it's ok. If not, no obligation will be true in the model.

Instead of modifying the definition of O, we might try to give an alternative formalization of CTD statements, to make sure that their consequent can not be true (be detached) at histories unless the antecedent is also true.

Alternative translation of CTDs. We have already discussed the differences between C2 and C3. In SDL the most straightforward possible formalizations of C2 were $O(g \to t)$ and $g \to Ot$. As we have seen, they both fall short of fully capturing the notion of conditional obligation. However, maybe in $d_e stit$ the picture is different. In $d_e stit$ we can choose between $O[a]([a]g \supset [a]t)$ (s2) and

$$[a]g \supset O[a]t \tag{s2'}$$

and similarly an alternative form for (s3) is:

$$\neg[a]g \supset O[a]\neg[a]t \tag{s3'}$$

this would make the detachment history-dependent. In two different histories (through the same state) we could have different obligations. This is not by itself a problem. A problem is, instead, that even (s2') fails to ensure that the conditional obligation is detached only when the antecedent is true (and not by virtue of the meaning of the obligation alone). Suppose for example we are in a history where we don't go. Still, say, all of the worlds in which we don't tell are sub-optimal. So we ought to tell.

In conclusion, this wouldn't solve $d_e stit$'s problems.

Conclusions about d_e*stit*.

As we have seen, Bartha's proposed treatment of Chisholm's scenario is not very satisfactory. The problems were mainly due to conditional obligations, that is, statements which describe what will become obligatory once some conditions are met. In Bartha's *dstit we couldn't prevent the conditional obligations from being* categorically *obligatory at the wrong moment*, and regardless whether their triggering condition were true. In the countermodel we gave, this is precisely what happened. Discussing a few key features of Bartha's d_e*stit* can help us understanding what goes wrong.

Problem I: Bartha's models are temporally flat. The critical point of conditional obligations is that they describe some obligation that should *become enforced* at some point (possibly in the future) when the triggering condition is satisfied. Even though *dstit* comes with temporal operators, Bartha used a tenseless sub-language to formalize Chisholm's paradox. However, even if he had not, *dstit*'s $F\varphi$ operator can only express φ being true some time in the future. Especially when dealing with obligations, the 'sometime in the future' tense operators are problematic. The main problem is that it is impossible to know when exactly φ will be true, since F existentially quantifies over an infinite (R is serial) set of *future* states. Furthermore, rendering future obligations like "you will have to clean your room, at some point" with F is hard if not impossible, because the action of cleaning your room (and the violation which stems from failing to do so) can procrastinated indefinitely (which is kind of realistic anyway).[34]

From another perspective, Bartha's models can be said to be *temporally flat*. Since tense operators were not being used, and the other operators only allowed[35] to express shifts over histories along the same moment, Bartha's models are indistinguishable from models having only one moment. Consequently, all obligations (including conditional ones) cannot but be in force at that same moment. This prevents us from modelling conditional obligations, a fortiori if their trigger is *a choice which is yet to become effective*.

dstit is an 'instantaneous' *stit*. What agents can see to is relative to the present moment only. Had we a built-in time gap to work with, we could model what an agent can bring about (but has not yet brought about). This way, a more precise notion of *choice* would become available.

[34] [Bro04b] makes this same point.

[35] With the exception of the $\Box$ operator, that universally quantifies over *all histories and moments*. However, $\Box$ is still useless in expressing temporal relations because of its 'universal' nature.

Problem II: Bartha's V lacks structure. Assuming we had a logic (a *stit* variant) in which temporal relations could be rendered, in a way that also avoids the 'procrastination' troubles we hinted at in the above paragraph; e.g., by employing 'next state' operators, deadlines or intervals. Even then, maybe not all of our troubles would be solved. Suppose, we are in Chisholm's scenario and we are in a two-cells moment (one cell being made of $\neg g$-worlds, the other of g-worlds). We have just chosen to $\neg g$, that is to *not* go to our neighbours' assistance, which means we have just ensured V is true at the present world. There are two main problems with d_e*stit*'s treatment of the violation V.

Philosophical point having only one kind of 'violation' is not enough. If murder is not as bad as theft, then we want different brands of wrongdoing to express the respective obligations.

Technical point having all obligations based on the same violation atom means that conditional obligations will be more likely to be satisfied all at the same worlds (because their consequents, V, will all be true at the same worlds). We would like, instead, to have conditional obligations to be fairly independent from one another, unless clearly they trigger or are triggered by the same kinds of actions or violations. Furthermore if we accept *no-pardon* (cfr. §0.2.2), unless we have multiple violation atoms, we cannot properly model CTD scenarios: once a violation has occurred, all subsequent states are V-states. Consequently, we cannot discriminate between 'less worse' worlds and 'more worse' ones, as we only have one violation atom which is true at both.

Conclusion: going to *xstit*. As a consequence of all we have argued so far, it seems natural to try to resort to a *stit* logic in which time is introduced, but in a more *discrete* fashion than the usual Prior-style F and P operators as unbounded existentials. The perfect candidate seems to be Broersen's XSTIT, in which the main operator $[a\ xstit]\varphi$ expresses that a sees to it that *in the next state* φ holds.

Furthermore it seems a good idea to introduce multiple violation atoms in the language to model CTD obligations. Being just a technicality with no import on the properties of the logic[36] we can do it without cost. Consequently in the following section we turn our attention to *xstit* (a variant of Broersen's XSTIT), which as already anticipated will prove itself useful in tackling the problems left unsolved by d_e*stit*.

[36] They are just propositional constants: adding them will not result in a different logic.

3.5 *xstit*.

Jan Broersen proposed a few next-state *stit* logic variants, some of which can express obligations. For example, in [Bro09b] a logic dubbed XSTIT with nice properties is presented which however was devised for the specific purpose of 'distinguishing modes of *mens rea*.' To that purpose it can also express knowledge and intentionality modalities we don't need. Also, no XSTIT paper we are aware of offers a philosophical analysis of its *deontic* properties.[37] Consequently, we will only use a single-agent fragment of XSTIT, distilling it from [Bro09b, Bro11]. The only modal operator we will need is $[a\ xstit]$, so we will also disregard any epistemic or intentionality modalities. Finally, we add to the language many violation atoms following $PD_eL(n)$'s example and (tend to) accept *no-pardon*. We name the resulting logic *xstit*, and we describe it formally in the appropriate appendix (D).

3.5.1 The language of *xstit*.

xstit is similar to *dstit*. The main difference is that frames come with an ordering R_X of ⟨state, history⟩ pairs, and not just a generic 'later than' relation on states.[38] Also, instead of the function **Choice**, the frames have a family of R_a relations for each agent $a \in Ags$. R_a is an *effectivity function* that gives, under some classical constraints, what agents are seeing to at any *state, history* pair. The 'classical constraints' on the family of relations $\{R_a\}_{a\in Ags}$ include that there can be no choice between undivided histories and that no agent can deprive another agent of the possibility of exercising a choice (cfr. §3.1). In *xstit* agents' choices depend on the next state only: the main operator is in fact defined as follows:

$$\mathcal{M}, \langle s, h\rangle \models [a\ xstit]\varphi \quad \Longleftrightarrow_{df} \langle s, h\rangle R_a \langle s', h'\rangle \text{ implies } \mathcal{M}, \langle s', h'\rangle \models \varphi$$

However, $wR_aw' \Rightarrow wR_Xw'$; in words, states related by R_a are always the latter 'next of' the former. So $[a\ xstit]\varphi$ is valid in a state iff a can ensure that the next state is a φ-state. For additional details, the reader is advised to read Appendix D. Now we shall discuss a few formal details of *xstit*. 1) will discuss the

[37] Broersen has wrote quite a bit on Chisholm's paradox. Namely he has put forth some promising results in CTL/ATL settings in [Bro10, Bro06b]. However, the logics he employs are in my opinion quite overkill. We hope to give a satisfactory treatment of the paradox in the much simpler *xstit*. Furthermore, they tend to overlook the dimension of *agency*, which is central to *stit* logics. Deeming agency a crucial element of deontic modalities, we prefer *xstit* over a bulky CTL/ATL treatment.

[38] In Broersen's XSTIT R_X is quite unconstrained. E.g. its transitive closure is allowed not to be irreflexive. This means that it is possible that time is cyclic. Anyway, we are only interested in acyclic models. *Thus, we will pretend R_X's transitive closure is always a tree-like ordering of dynamic states.*

multiple violation atoms, then 2) how to formalize *obligations* and finally 3) how to formalize *CTDs* (C2 in particular).

Multiple violation atoms. Drawing from $PD_eL(n)$, we introduce multiple violation atoms (MVA) in the language. The idea is to have a V_φ atom for each sentence φ in the language (including sentences which involve other violation atoms). Still, every V_φ is evaluated like an atomic proposition; i.e. its truth-value depends on the model (is given by the evaluation function π) and is totally independent from φ.[39] V_φ can be read as "there is wrongdoing *because an obligation to* φ *is/was violated*", "an obligation to φ is/was violated", or "someone is to be punished for letting $\neg\varphi$ be the case". So failing to fulfil an obligation to p will result in a state where V_p is true. This way we can go on violating obligations (stacking V_ψ atoms) without losing track of what has been violated (because of *no-pardon*: $V_\varphi \supset XV_\varphi$).

We add no constraint to $\pi(V_\varphi)$ beside *no-pardon*. However it is possible to enforce the requirement, commonly found in Anderson-Kanger-reduction-based logics, that acting rightfully be possible ($\Diamond\neg V$ in SDL). We might require that for all φ, $\Diamond\neg V_\varphi$. But this is too strong: if we have violated an obligation to φ in the past and *no-pardon* is valid, then all future states necessarily are V_φ-states. For the same reason we cannot require that there is some accessible state w where no violation holds $\forall\varphi : w \not\models V_\varphi$. Something we can require, on the other hand, is ($\forall\varphi$) that if V_φ doesn't hold now, then there be some future state in which it doesn't hold either:

$$\neg V_\varphi \supset \Diamond\neg XV_\varphi \qquad (xstit\text{-pr})$$

In this way we can make sure not to break *no-pardon*. The intuitive meaning of this requirement is that for any norm, if we did not break it yet then it is possible not to break it next.[40] [41]

Defining O_a. There are *many* ways to formalize the 'agent a ought to do' operator O_a. Here we examine the two most obvious ones.

1. $O_a\psi := S[a\ xstit](\neg\psi \supset V_\psi)$ $\qquad (O_x)$

[39] To generalize this to a multi-agent or coalition setting, we might index V also with names of coalitions, so that different coalitions will be allowed to have different obligations in a more natural way. Also, we might want to index V with the name of the moment in which it first occurs (the bottom element of the 'violation' branch). This would result in a more accurate modelling of situations where you can go on violating the same obligation over and over; this in turn would help us render obligations which can be violated multiple times and remain in force even when violated.

[40] The formula ensures that if $\neg V_\varphi$ is true at $\langle s, h\rangle$, there exists some simultaneous dynamic state $\langle s, h'\rangle$ such that at its 'next' state $\langle s', h'\rangle$, $\neg V_\varphi$ is also true.

[41] An equivalent requirement would be that the sub-model based on only those dynamic states at which V_φ is *false* be *serial*.

the intuition is, in line with the other Andersonian reductions we have seen, that ψ is obligatory (to an agent a) iff it is settled that a's present choice will ensure that next, if ψ is false, a violation will occur.[42]

2. $O_a\psi := S(\neg[\,a\ xstit\,]\psi \supset XV_\psi)$ (O_x')

Here, the intuition is that ψ is obligatory for a iff it is settled that if a cannot ensure ψ then at the next state there will be a violation. In other words, something is obligatory iff at all future moments at which a did not ensure that ψ will hold, V_ψ will be true. In even simpler words, $O_a\psi$ iff it is settled that if a doesn't have the power to ensure that (next) ψ, then the next state will be sub-optimal. This is probably the most intuitive definition; it is also very close to the one suggested in [Bro09b] (but he models V as something for which a is agentive). *Hence, we will use* (O_x') *as our definition of* O_a.

Formalizing CTDs. There are also many ways to formalize CTDs in *xstit*. Here we will enumerate the most straightforward ones. Suppose we want to formalize C2 in its 'forward' interpretation.

- $Xg \supset XO_at$ (3.5)

which reads: if you are going to the party, you (will then) ought to tell that you are going. This is not so intuitive because it does not contain no $[\,a\ xstit\,]$ operator, and so there is no explicit involvement of agency. However, we could obtain a very similar, more promising (slightly weaker, but with similar properties) definition:

- $[\,a\ xstit\,]g \supset XO_at$ (3.6)

if we feel like CTDs should be history-independent, we could turn this into

- $S([\,a\ xstit\,]g \supset XO_at)$ (3.7)

to yield a very promising formalization for C2. A related option is:

- $[\,a\ xstit\,](g \supset O_at)$ (3.8)

[42] We might as well have used the following definition: $O_a\psi := S[\,a\ xstit\,](\neg\psi \equiv V_\psi)$. I like the idea of making sure that there cannot be a violation of an obligation to ψ unless ψ is *actually* false. Having now infinitely many violation atoms, this is less problematic than it used to be: multiple obligations won't influence each other as each one has its own violation atom. However, for the moment let's stick to the more standard definition. Furthermore, we follow the general trend of requiring obligations to be *moment-determinate* (i.e. *history-independent*), as [Bro09b, Hor01]. However, see [Wan01] for a contrary opinion.

This is stronger than (3.5): it implies it.[43] However it is less intuitive because the obligation is in the scope of $[a\ xstit]$, and the obligation's enforcement is not supposed to be responsibility of the agent a. We want to say that if a does x, then there is an obligation. Not that a brings about that (if a does x, then there is an obligation). To obtain a history-independent variant, we can modify it to

- $S[a\ xstit](g \supset O_a t)$ (3.9)

which is a stronger version of (3.7).

Another option (a weakening of (3.9)) is:

- $SX(g \supset O_a t)$ (3.10)

this is also quite straightforward and enforces the following: 'it is settled that if the next state you go, then the next state you will be obliged to tell.' The detached obligation will be true not here, but at the next state, so that it will constrain the then-next step and not the presently-next one.

A final option worth mentioning is:

- $O_a(g \supset O_a t)$ (3.11)

though nobody to my knowledge has tried, we could argue that CTDs are second-order obligations. After all, they *prescribe*, not describe, what ought to be the case under certain circumstances. That 'if you go you ought to tell' is perhaps an obligation in its own right, and could be re-stated as: 'the following is obligatory: that if you go, then you ought to tell.' The CTD can then rightfully be said to 'be violated' in two cases: when you go but there is no obligation to tell, or when you go and are consequently obliged to tell, but you fail to. This may seem weird, and it is: compliance with the outermost obligation does not depend on any agent's choices: it only depends on the distribution of violation atoms. Namely, it depends on $\pi(V_{g \supset O_a t})$. Consequently, it seems that it does not describe an obligation of a, but some sort of 'global' obligation whatever may this mean. For this reason, we conclude that (3.11) is hard to justify from a philosophical perspective, and prefer the other two formalizations instead.

Now, we would like to follow the Anderson-Kanger tradition and require that obligations are history-independent; this rules out (3.5). Adding an S in front of (3.5), since $X(\varphi \supset \psi) \equiv (X\varphi \supset X\psi)$[44] , simply results in (3.10). However (3.10)

[43] In fact $[a\ xstit]\varphi \supset X\varphi$, and X distributes over $\supset$: so, $([a\ xstit](\varphi \supset \psi)) \supset (X(\varphi \supset \psi)) \supset (X\varphi \supset X\psi)$.

[44] Semantically (where $X(s) := \{s' \mid sXs'\}$): $\mathcal{M}, s \models X(\varphi \supset \psi) \Longleftrightarrow \forall s' \in X(s).\mathcal{M}, s' \models \varphi \supset \psi \Longleftrightarrow \forall s' \in X(s).\mathcal{M}, s' \models \varphi \Rightarrow \forall s' \in X(s).\mathcal{M}, s' \models \psi \Longleftrightarrow \mathcal{M}, s \models X\varphi \Rightarrow \mathcal{M}, s \models X\psi \Longleftrightarrow \mathcal{M}, s \models X\varphi \supset X\psi$.

is stronger than (3.7): it implies it.[45] Since they results they give are very similar, the weaker alternative shall be preferred. So, we shall use (3.7) as a formalization of CTD conditionals such as C2, C3.

3.5.2 An *xstit* model for Chisholm's scenario.

Given the previous paragraphs, we propose the following *xstit* formalization of Chisholm's set.[46] [47] [48]

$$\text{C1} \mapsto \quad O_a g \tag{3.12}$$

$$\text{C2} \mapsto \quad S([\,a\ xstit\,]g \supset XO_a t) \tag{3.13}$$

$$\text{C3} \mapsto \quad S([\,a\ xstit\,]\neg g \supset XO_a \neg t) \tag{3.14}$$

$$\text{C4} \mapsto \quad [\,a\ xstit\,]\neg g \tag{3.15}$$

These formulae have an intuitive model in *xstit*. That is represented in Figure 3.2.[49] As we can see, the model develops through three moments m_0, m_1, m_2 (made of 8 static states) and we have four (bundles of) histories h_1, h_2, h_3, h_4. There are a few important assumptions backing up this model of the Chisholm set; we will now make them explicit.

Assumption 0 is that the validity time of the obligations to go and to tell are different. *Now* you are obliged to go, and depending on whether you actually do, at a later moment you will be obliged to tell or not. If this seems counterintuitive, it is because this is the 'forward' Chisholm's paradox:

Assumption 1 is that first an agent has to choose whether to go or not; and only afterwards he has to choose whether to tell or not. This is objectionable on its

[45] In fact $SX(g \supset O_a t) \supset S(Xg \supset XO_a t) \supset S([\,a\ xstit\,]g \supset XO_a t)$. This is true because X distributes over $\supset$ and $[\,a\ xstit\,]\varphi \supset X\varphi$.

[46] Remark: we cannot render C1 as $O_a[\,a\ xstit\,]g$ because this would result into the obligation being enforced one moment too late. We can defend our choice (and point out that the restricted complement thesis is not in fact violated) by noting that O_a is defined as a *xstit* statement. Consequently the complement of O_a in any statement of the form $O_a\varphi$ is *in any case* agentive. This is made apparent by the fact that O comes indexed with the name of an agent.

[47] Remark: we need the X in (3.13) to have $O_a t$ be true at the right state: that is, the next one (if g holds *there*). Just like before, we cannot render it as $SX(g \to O_a t)$ because this would result the detached obligation to be enforced one moment too late.

[48] Remark on C4's formalization: we could have used the weaker $\neg[\,a\ xstit\,]g$. We don't, because we think it's more intuitive to interpret C4 in a 'strong' way ('a sees to it that he doesn't go') and because we want to match C3's antecedent. However, reverting this choice would have no adverse effects.

[49] We adopt the following graphic conventions. What holds at states (that little has to do with the choices available to the agent) is written in blue outside of the outer rectangle. In red, violation atoms V_φ. If a rectangle is divided in smaller rectangles, the innermost ones represent the choices available to the agent. By selecting a cell whose next states are p-states, the agent ensures that at the next state p holds.

own right, but even worse it entails that the former takes place before the latter (as far as the choices are about the next state only). as we have argued before, this does not seem to be the most obvious analysis (and most true to Chisholm's intentions).[50] However, being the present 'forward' Chisholm scenario simpler to model, we shall start with it.

Assumption 2 is that the agent can freely chose to see to it that g, that $\neg g$, and then that t or that $\neg t$. That is, the agent has a great deal of control over the situation. We think this is the most interesting situation. However, this assumption can be dropped free of charge.

Assumption 3 is the *crucial* assumption that *no violation atoms are true other than these we need to make true the Chisholm sentences*. This is a way to say that no other infringement obtains beside those we stipulate. We will see later why we need this (remarkably strong) assumption.

Now, let us go through the formalities to see how the obligations we have distribute the violation atoms around the model (see Figure 3.2).

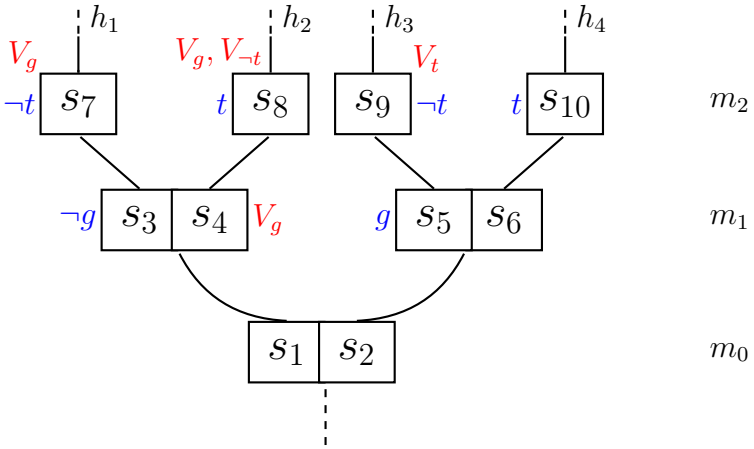

Figure 3.2: A model of Chisholm.

Formalities. (3.15) tells us that the kind of situation we are interested in is where we are in s_1 or s_2; that is, where $\neg[\,a\ xstit\,]g$ holds.

Categorical obligations such as (3.12) are history-independent (the outermost operator being an S) and in fact at s_1, s_2 (any history) it holds that $O_a g$. In fact,

[50] Else we might treat *xstit* not as a model for action, but of deliberation. In this case we would argue that since the 'telling' choice depends on the 'going' one, in the deliberation cycle it is naturally expected to come after.

$O_a g \iff S(\neg[\, a \ xstit\,]g \supset XV_g)$ and all dynamic states based on s_1, s_2 satisfy $\neg[\, a \ xstit\,]g \supset XV_g$.[51]

Conditional obligations, instead, become effective only in two steps from now even though their validity is history-independent. E.g.: (3.13) is true at s_0, s_1 (regardless of the history), but the obligation it prescribes kicks off only in those next static states in which g was not seen to; namely s_3, s_4.

Consider (3.14); formally,

$$\langle s_1, h_1\rangle \models S([\, a \ xstit\,]g \supset XO_a t) \iff \text{for all } h_i \in \{h_1, h_2, h_3, h_4\}: \\ \langle s_1, h_i\rangle \models [\, a \ xstit\,]g \supset XO_a t$$

For example in case $h_i = h_1$, we have that $\langle s_1, h_1\rangle \models [\, a \ xstit\,]g \supset XO_a t$ iff $\langle s_1, h_1\rangle \models [\, a \ xstit\,]g$ implies $\langle s_1, h_1\rangle \models XO_a t$. However $\langle s_1, h_1\rangle \not\models [\, a \ xstit\,]g$, so no conclusion follows. In contrast, if $h_i = h_3$, since $\langle s_2, h_3\rangle \models [\, a \ xstit\,]g$ we obtain that $\langle s_2, h_3\rangle \models XO_a t$. Therefore we conclude $\langle s_5, h_3\rangle \models O_a t$, i.e. $\langle s_5, h_3\rangle \models S(\neg[\, a \ xstit\,]t \supset XV_t)$. This ultimately enforces that $\langle s_9, h_3\rangle \models V_t$.

This is why at $\langle s_7, h_1\rangle$ we have only V_g (due to no-pardon) but not $V_{\neg t}$ and at $\langle s_8, h_2\rangle$ we have both V_g (still due to no-pardon) and $V_{\neg t}$ (because $\langle s_8, h_2\rangle \models t$).

Properties of the model. In line with what we suggested when discussing the introduction of multiple violation atoms, both at the beginning of this section and at the end of the section on PD_eL, we now have a situation in which a hierarchy of best-to-worse states (and *histories*!) can be extracted from the model. The 'best' states are those where no violation obtains; so h_4 is a history made up by optimal states. Next is h_3, where from m_2 on we have $V_{\neg t}$. Similarly, in h_1 we violate only one obligation and so this history is as bad as h_3. h_2 is clearly the worst, because not only we fail to go, but we also tell we are going. In fact, at h_2 two violations $V_{\neg t}$ and V_g are true.

No undesired obligation obtains at any state in the model. What matters is not that violations take place *after* the choice has been made, but that choices are in *xstit* not instantaneous; that is, they become effective later. In contrast, in $d_e stit$ violation atoms were true at the same worlds where the obligations (and what made them violated) were. *xstit* obligations distribute violation atoms depending not on the choices which are available (as it used to be in *dstit*) but on the choices as

[51] We show how the proof goes for $\langle s_1, h_1\rangle$. $\langle s_1, h_1\rangle \models S(\neg[\, a \ xstit\,]g \supset XV_g)$ iff $\langle s_1, h_i\rangle \models \neg[\, a \ xstit\,]g \supset XV_g$ for all h_i such that $h_1 R_S h_i$; e.g. if $h_i = h_1$:

$$\langle s_1, h_1\rangle \models \neg[\, a \ xstit\,]g \supset XV_g \iff \langle s_1, h_1\rangle \not\models [\, a \ xstit\,]g \Rightarrow \langle s_1, h_1\rangle \models XV_g$$

In fact, $\langle s_1, h_1\rangle \models XV_g$ iff $\langle s_3, h_1\rangle \models V_g$; which is true since $\langle s_3, h_1\rangle \in \pi(V_g)$ as depicted in Figure 3.2. The same goes for $h_i = h_2, h_3, h_4$.

they become effective. Also, having a V_φ for every obligation $O_a\varphi$ dodges vacuous satisfaction (due to no-pardon) of obligations if a violation has been committed in the past.

Another way to clarify the point is by noticing that in *dstit*, $O[a\ dstit]p$ depends on what *a is seeing to* at the world of evaluation. This ensures that in the present world, if a cannot (deliberatively) see to it that p, V is true. In *xstit*, an equivalent statement would simply make sure that (regardless of what *a can* see to in the world of evaluation) if a in fact does not make sure that p in the next state, then a violation occurs *there*.

Discussing Assumption 1. In the PD_eL chapter we stressed quite a bit that the most straightforward reading of the original statement of the paradox cries for a 'backwards' interpretation of the time-flow in the Chisholm scenario. This is something we have ignored here. The formalization we gave was the 'forward' version of Chisholm, which is known to be 'easier' at least in action and time logics. What if we wanted to analyse the *backwards* version? I think that would be impossible in the present setting. We would need time operators to formalize 'a ought to sometime'. Of course a tailored solution is possible:

$$\text{C1} \mapsto \quad O_aXXg \tag{3.16}$$
$$\text{C2} \mapsto \quad S([a\ xstit]XXg \supset XO_at) \tag{3.17}$$
$$\text{C3} \mapsto \quad S([a\ xstit]XX\neg g \supset XO_a\neg t) \tag{3.18}$$
$$\text{C4} \mapsto \quad [a\ xstit]\neg XXg \tag{3.19}$$

These sentences also have a nice *xstit* model that yields the desired conclusions. Being it so similar to the previous one, we leave the details to the reader. Still, one would need to justify this formalization. The sentences are obtained by prefixing $g/\neg g$ with XX in (3.12-3.14). This has the effect of making g obligatory in two time-steps from now so that we can ensure that t is true before g takes place (albeit without 'before' operators), and we can also make sure that V_t and $V_{\neg t}$ hold exactly when we want (by 'looking ahead' for g being true).

This is quite a simplification, and we can only skim briefly through some of the underlying complexities. Firstly, we may remark that $[a\ xstit]XXg$ merely says that a is following a strategy for obtaining g in two steps.[52] On the other hand, C2 seems to mean that if you will actually go/ intend to go/ are committed to going, then you ought to tell. The first reading, as we have already argued in §2.4.2, we will set aside not to entangle ourselves in sea battles we might never come out from (semicit. [Gas95]). The second one is more promising, because there are

[52]Equivalently, we could have rendered C2 as $S([a\ xstit][a\ xstit][a\ xstit]g \supset XO_at)$.

xstit variants that can capture some forms of 'intentionality'.[53] The commitment reading is more complicated, and we will not discuss it here. Similarly would go the argument for C3.

So, one possible way to look for a generalization of the present setting that would solve simultaneously the backwards and forward version of Chisholm's paradox is to extend the present deontic *xstit* to include intentionality. We will not do that, because this is already page 62 and space is running out.

3.5.3 What about the Gentle Murderer?

The first observation that comes to mind concerning the differences between the gentle murderer and Chisholm's paradox is that the former, unlike the latter, unambiguously takes place in a single moment. There is only one choice involved, this time between three alternatives: don't murder (gently or not), murder brutally or murder gently. An *xstit* formalization of Forrester's paradox however shows how some more thought concerning *xstit*'s treatment of CTDs is needed. We now propose a formalization of Forrester's paradox:
The first premise, F1, that murder is forbidden, is straightforward:

$$O_a\neg g \tag{3.20}$$

The second premise, F2, that if we kill (k) we ought to do it gently (g), is more problematic. We might try to formalize it as we did with C2 or C3 in *xstit* (see (3.13)). That would yield:

$$S([\,a\ xstit\,]k \supset XO_a g) \tag{3.21}$$

The problem is that, being the choice simultaneous, this would result in the obligation to g being enforced one moment too late.[54] We would need, to obtain the desired model (as drawn in Figure 3.3), something like:

$$O_a(k \supset g) \tag{3.22}$$

but this would now open a philosophical problem: how to justify the different treatment for Forrester's second premise and Chisholm's C2? We might want to argue that their logical forms are in fact different. Chisholm's C2 is about a new obligation popping out in the future in case something will then be true. Forrester's F2 is more about a (present) obligation that issues no 'new' obligation sometime

[53] Again, we are referring to [Bro09b].

[54] A quick fix would be: F2 be $S([\,a\ xstit\,]k \supset O_a g)$. However that not only would entail an unconditional obligation to g, but would also make all possible choices sub-optimal (also the ones in which we don't kill!).

in the future.[55] Still it's a long way to go between $SX(\varphi \supset O_a\psi)$ and $O_a(\varphi \supset \psi)$. We might want to revise our formalization of C2 to $O_a(g \supset O_a t)$ (that is, (3.11)), but we will then have to explain why our formalization of C2 contains two nested obligations. For the moment, let us accept (3.22) and go on.
Finally, the assumption F3 that murder is committed is rendered as

$$[\,a\ xstit\,]k \tag{3.23}$$

and says, in terms of Figure 3.3, that we are in $\langle s_2, h_2\rangle$ or $\langle s_3, h_3\rangle$.

A look at the model. Now, again, let us look at the structure of a model of Forrester's scenario; and in particular, at what the obligations we assume enforce in terms of violation atoms (Figure 3.3).

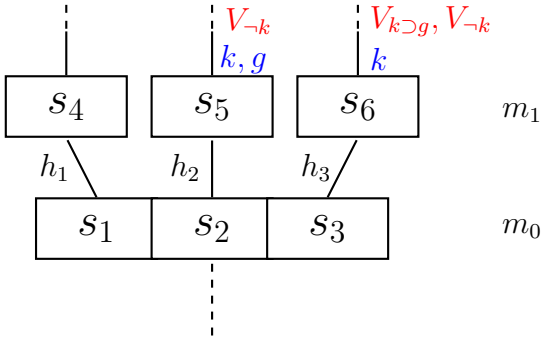

Figure 3.3: A model for Forrester's scenario.

The treatment of (3.20) and (3.23) being entirely alike to how C1 and C4 were treated in the Chisholm *xstit* model we have just discussed, we will turn immediately to (3.22). Expanding the formula results into:

$$S(\neg[\,a\ xstit\,](k \supset g) \supset V_{k\supset g})$$

which is true for $t \in \{\langle s_1, h_1\rangle, \langle s_2, h_2\rangle, \langle s_3, h_3\rangle\}$ iff $t \models \neg[\,a\ xstit\,](k \supset g) \supset V_{k\supset g}$. This enforces the following requirement: if a doesn't see to it that next $k \supset g$, then $V_{k\supset g}$ is true at the next state. As we can see, the only world at which $V_{k\supset g}$ comes out true is $\langle h_3, m_1\rangle$. We obtain a clear hierarchy of strategies: best is h_1, with no violation atoms true at m_1, then comes h_2, with only one violation atom true, and finally (worse) comes h_3 which hosts two violations.

[55] So perhaps, despite the surface form very similar to C2, F2 might just *not be a CTD*.

No undesirable consequences in sight. Also in this case, the model does not seem to entail oddities of any sort. $O_a t$ is not generally true, unlike in SDL. So, the 'main puzzle' of the *gentle murderer* scenario is certainly avoided.

3.6 Conclusion.

As we have seen, in *xstit* natural and unproblematic formalizations could be given for Chisholm's and Forrester's scenarios. However, the two formalizations were non-uniform. Furthermore, it is still to be discussed what was the benefit of moving from $PD_eL(n)$ to *xstit*, and in particular of having multiple violation atoms. By comparing the two frameworks we can shed some light on both issues for the price of one. Finally, the crucial point is and remains that a strong *minimality* assumption (Assumption 3) was required to make things work. In this order, these are the issues that we will address in this section.

The quest for a uniform treatment of contrary-to-duty statements in *xstit*. All obligations in *xstit* are, in a way, about the future. They are concerned not with choices, but with their outcomes; and in *xstit* outcomes are visible in the next state only. However, C2 describes what will become obligatory once something will be done, and so may trigger violations only two ticks from the state of evaluation. On the other hand F2 describes just a present obligation that tells you that some act (killing) should always be of a particular kind (gentle).[56]

Besides this, non-uniformity is not the Evil. Different sentences may require different formalizations. However, it would be nice if we had a CTD formalization able to capture both C2, F2 with minimal formal differences e.g. a stack of X operators making sure the obligations and their triggers are properly timed. This is achieved in the forwards/backwards formalizations of Chisholm's paradox, but not in the *gentle murderer*. We leave to further research the question whether such a formalization exists.

A comparative analysis with PD_eL: the importance of adding multiple violation atoms to the logic. As we saw in Chapter 2, PD_eL managed to offer a consistent formalization of Chisholm's scenario when multiple violation atoms (MVA) were added to the language. The scenario was not paradoxical only in $PD_eL(n)$. In that case, that was the winning move.

In *xstit*, on the other hand, what is the import of having MVA? The answer is twofold: in modelling Chisholm's scenario, *xstit* does a good job also without MVA, *provided that no-pardon is dropped*. If, once a violation is committed, the whole

[56] Cf. [SA85] for a discussion of this.

branch is made of V-worlds, this prevents us from carrying out further reasoning about obligations (from the next state on). Namely we have the following theorem:

$$(O_a p \wedge [\, a\ xstit\,]\neg p) \supset XO_a\varphi \qquad \text{(coll)}$$

If on the other hand no-pardon is dropped, then we avoid (coll)ateral damage and Chisholm's paradox can be modelled almost as before. We do not, however, obtain the nice hierarchy of worlds we wanted.

On the other hand, in Forrester's paradox the situation is different. In this case, without MVA also *xstit* does poorly: assuming $O_a\neg k$ results in V being true at all k-worlds, *and so all g-worlds*. Consequently $O_a\neg g$ is also (unconditionally) true.

We could try to reformulate Forrester's scenario in a way that allows for its choices to be distributed along two moments instead of a single one, but that would miss the point. *xstit* without MVA does poorly on reasoning with obligations based on a single-moment structure (i.e. where we have only one choice). *stit* and PD_eL are very different logical frameworks, based on wholly different mechanisms. Still, the *gentle murderer* is solvable by the same approach in them both. This is an indication that when we have only one choice (between any number of alternatives) in order to reason about them we need to be able to talk about *various kinds of wrongdoing*. If we only have one brand of wrong, the Anderson-style O will vacuously satisfy many unneeded obligations.

All the rest aside, the variant of *xstit* we laid down here is still a better deontic logic than PD_eL. Firstly 'no possible action is forbidden' (cf. §2.4.1) is not a theorem of the logic, and this is good. Secondly, formalization is a much more straightforward task in *xstit*. Having at hand clear-cut notions of agency and action, we spare the need to wonder what exactly does $[\,\alpha\,]\psi$ mean. *xstit* is backed by powerful, clear philosophical assumptions, which make disambiguation and formalization easier. We acknowledge in this regard the admirable efforts of all those who have contributed to it.

3.6.1 Not all *xstit* models are good models: a discussion of *Assumption 3.*

Even though the models we gave for Chisholm's and Forrester's scenarios looked good and gave the right predictions, not all *xstit* models are good in this respect. To see where the problem is, we present as an example a simpler scenario depicted in Figure 3.4. A binary choice situation: *to go or not to go*.

Assume that $O_a g$ is true at $\langle s_1, h_1\rangle$. Consequently, V_g will be true at s_3. The problem is that, unless additional constraints are put on the evaluation function π, nothing prevents $V_{\neg}g$ from being true at s_4. This would result into $O_a\neg g$ being true at s_1.

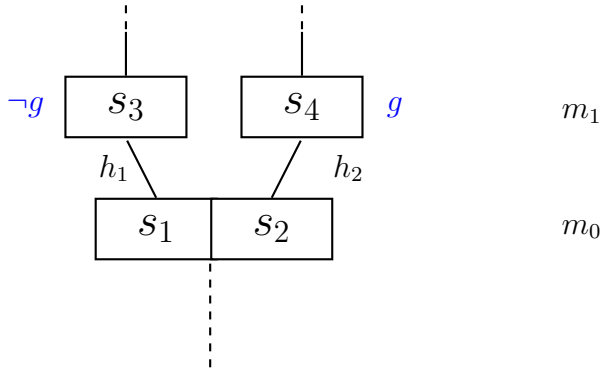

Figure 3.4: A binary choice model: *to go or not to go*.

To require that $V_\varphi \supset \neg\varphi$ is implausible if no-pardon is assumed: it would entail that if you fail to do something obligatory then you will be unable to do it forever since. Without no-pardon that becomes a viable option, but does not solve the issue (and we lose the preference order).

A solution would be to add axiom (nco) to the system[57]:

$$O_a\varphi \supset \neg O_a\neg\varphi \tag{nco}$$

which says that per every couple of $V_\varphi, V_{\neg\varphi}$ atoms, every state and every agent a it is settled true that there is an R_a-next state where φ holds and $V_{\neg\varphi}$ doesn't, or $\neg\varphi$ holds and V_φ doesn't.

On a related note, $SX\neg\varphi \supset O_a\neg\varphi$[58] is a theorem of *xstit* we don't really like. However this can be solved easily by adopting a deliberative variant of the main *xstit* operator or by re-defining O_a in a suitable way. But this wouldn't solve completely the issue of 'spurious obligations': if at s_4 we have V_ψ (for some ψ true at s_3, but false at s_4), then at m_0 it is true that $O_a\psi$.

The cause of the troubles, and two possible solutions. The problem is that have left unconstrained the occurrence of the V_φ atoms. In other words, *the violation atoms don't have a logic*: they are propositional constants and as such are totally *wild*.[59] There are two solutions:

- **The first solution** is to adopt, as we did, the strong 'minimality' *Assumption 3* that says that the evaluation of the V_φ atoms only depends on the assumptions we want the model to capture. This can also be seen as a sort of

[57] Most of the literature accepts it .

[58] In fact, $SX\neg\psi \supset S[a\ xstit]\neg\psi$.

[59] Broersen pointed out that Halpern's A_i constant has similar features, e.g. in [FH88].

default: the V_φs default to false, and are true only if we assume the enforcement of some obligation as we did to model Chisholm and Forrester's scenarios. The main argument for this solution is that which norms are broken is arguably part of what the model tries to capture. A model of Chisholm's set (take that of Figure 3.2) where at s_0 it is obligatory to shave a baboon *is just not a good model of Chisholm's set*. Interpreting V_φ as '*a is punished* for not fulfilling his obligation to φ' makes the point more clear: in Chisholm's scenario is quite unlikely that you are punished for not shaving a baboon, whatever choice you make at s_1.

The risk is that of circular reasoning: since the obligation to *not tell* after *going* is 'not part of Chisholm's scenario' we have similarly assumed that e.g. $V_{\neg t}$ is false at s_1. Had we not assumed this, we would have allowed for models of Chisholm where (at s_1) if you go, you ought to tell and you ought to not tell.

- **A second solution**, more subtle, is to constrain the evaluation of the V_φs via axioms such as (nco). After all V is an operator, perhaps modal in nature, that can depend not only on other violation atoms but also on the proposition they are indexed with. Perhaps a suitable 'logic of violations' can be found which spares the need for *Assumption 3* and enforces plausible, weaker constraints on the occurrence of violations. A list of axioms which need to be considered for this purpose include:

$$V_\varphi \equiv \neg V_{\neg\varphi} \qquad \textit{No conflicting violations} \tag{3.24}$$
$$\neg V_\varphi \supset \Diamond X \neg V_\varphi \qquad \textit{No necessary violation} \tag{3.25}$$
$$\Box X V_\varphi \supset \neg \Box X V_{\neg\varphi} \tag{3.26}$$

Especially if we read V_φ as 'there is liability to punishment because an obligation to φ has been violated', it is apparent how $V_\varphi \equiv \neg V_{\neg\varphi}$ is a plausible axiom.[60] You cannot be simultaneously punished for doing and not doing something. We expect (3.24) to entail (nco) under the additional constraint that $SX\varphi \not\supset O_a\varphi$. Also at the moment $O_a(\varphi \wedge \psi)$ does not entail $O_a\varphi$. In some situations this is desirable. In some others, it isn't. Therefore axioms such as $V_{\varphi\wedge\psi} \supset V_\varphi$ need be considered as well, just like we may want violations to depend on what has (not) just been done.

In conclusion, research is needed to determine whether a logic of the V_φ atoms is possible and beneficial. A suspect that comes to mind is that the logic of violation may contain, in a 'smaller' scale, the same problems that we encountered in SDL. This is a risk worth investigating. On the other hand, the additional structure encoded by the two interacting layers of violations and following obligations may have a positive *antiparadox* effect.

[60] Which still if we don't drop *no-pardon* would give problems.

Conclusion. In this thesis we investigated the first option and leave the second one open for future work. The logic we obtain (*xstit* + *Assumption 3*) can be made fully formal in a straightforward way. *xstit* + Assumption 3 includes no oddities except for $SX\neg\varphi \supset O_a\varphi$ being a theorem. The problem here is that the implication in the O_a is satisfied vacuously. To solve this, either we modify the definition of $O_a\varphi$ to require a *counter* (that φ be *possible*) or we modify it to use, instead of an $[a\ xstit]$, some sort of *deliberative xstit* operator (requiring, also, a counter) e.g. the $[a\ dxstit]$ operator of [Bro09b]. We call this fix *del*. Given this, the most common paradoxes of deontic logic will be unproblematic in the logic. W.r.t. the list of deontic paradoxes laid down in [MDW94], *xstit* +Assumption 3 + *del* only falls to n.8 and n.11 (n.12 being inexpressible in the logic) out of 12!

This is quite a remarkable result.

Chapter 4

Conclusion.

As we have seen in the last chapter, even though Bartha's $d_e stit$ was not able to fully capture the complexity of Chisholm's scenario, $xstit$ does a much better job. In this final section we wrap-up our findings and point out some directions for future research and the limitations of this work.

4.1 Logics of prescriptive *and* descriptive obligations: towards a unified framework.

Many deontic logic theorists have argued that deontic logic needs a foundation in a logic of action (e.g. [vW63, LS07, Hor01]). There is some agreement that, in turn, a logic of action needs to be a dynamic logic (i.e. a logic of *change*), e.g. [Mey00]. Still, the 'PDL' sort of dynamics is not the only one around: modern research on dynamics (on the style of Dynamic Epistemic Logic) is not much about changing worlds and states but about changing *models*.

The models of the logics we have seen so far were static. This means that the norms we could model were only descriptive: true or false depending on some characteristics of the model.

In SDL the truth of Op at w is evaluated against what is true at the deontically perfect alternatives to w. The function D that gives the 'deontical optimality' relation, was fixed by the model.

In PD_eL what is obligatory at some state w also only depends on what is true at the worlds accessible from w through some action. The accessibility function is again part of the model. There is no way to change the model; formulae of PD_eL can only 'move around' agents in the current model.

In *stit* the same applies. *stit* models carry immutable information about the conditions under which a violation occurs. So for example, no choice can ensure that something which is forbidden (now) becomes permitted (now).

In a way, *a model of either these logics encodes a snapshot of a normative system*; all that agents can do is try to act by the rules. We cannot reason about what would be true if the norms were different, because we have no formula able to change the norms. One system in which that is possible is Aucher et al.'s DEDL[1], which stands for *Dynamic Epistemic Deontic Logic*. The epistemic component of the logic was not the main focus of our interest here, but of course it would be an interesting addition to the picture. We deem fruitful to look into a dynamic extension of *xstit*. Define model-changing operations to encode the norm-changing potential of some assertions. This would open a whole new variety of options when it comes to formalizing CTDs e.g. "if the antecedent is true, then the consequent becomes enforced (as if it were a new norm established afresh)".[2] This would also probably solve the problem of vacuously satisfying obligations because the Vs would be true in other models, not in the present one. On a related note, we would consider the possibility to have 'deontically perfect *models*' against which to evaluate obligations in *xstit*.

4.2 What is missing in *xstit*.

Firstly, we were not able to find a uniform translation for CTDs. We argued this is because they refer to different future times at which either the triggering condition or the subsequent obligation are supposed to take place. This makes them different brands of CTDs. However, a uniform analysis (if possible at all) would sure look better.

Secondly, the *xstit* we presented was relatively basic. There are versions of *xstit* which include *intentionality* and *knowledge* (see [Bro11, Bro09a, SW08]) and more explicit notions of *strategy* ([BHar]). That would be an interesting extension, but extending *xstit* models with model-update operators as suggested above would take these efforts to a whole new level.

Finally, once more we have to underline our abidance of *Assumption 3*, that is, the assumption that the evaluation of violation atoms is 'minimal', i.e. tailored so as to make true only the *assumptions* of the situation we want to model. This is partly justifiable because the conditions at which you are liable to punishment are part of the model as much as the choices you have available. Furthermore, it is plausible that violations are not completely unconstrained as to where they are allowed to occur. It is however true that this assumption is *very* strong. Consequently, the fact that the resulting logic contains very little 'paradoxes' should be acknowledged with careful enthusiasm. More research is needed to find out whether it is possible to obtain the same results by means of weaker requirements.

[1] Cfr. [ABvdT10]. Other similar approaches include [Yam08, vBL10, vBLG14].

[2] This suggests a nonmonotonic solution of some deontic puzzles such as Chisholm's.

Examples worth considering (and a more precise explanation of what we mean) can be found in §3.6.1.

4.3 What is missing in this thesis.

To keep this section's length below that of the whole thesis, I will only mention the main points.

We are aware of, but did not consider because of mixed reasons, research in the following directions: dyadic deontic logics of all sorts[3], update semantics-based approaches[4], arguing that the paradoxical statements are in fact inconsistent[5], unnecessarily though beautifully complicated CTL/ATL-based approaches[6], *non-monotonic* deontic logics[7], *relevance* deontic logics[8] and even fancier proposals such as [CC86]. Furthermore V-based, Anderson-style deontic logics are not the only way to go. Some argue it is not the best either ([Kui12] even argues that they are structurally unable to capture CTDs), still the version of *xstit* we laid down does not display striking problems.

We took a from-simple-to-complex approach to avoid overly complicated solutions, and to pinpoint the cause of the troubles as they disappear (or get worse). It is known that complexifying the logic is not the only way to get rid of paradoxes, the symmetrically opposite solution being to *weaken* it. This is something we did not try, because we believe that the closer the model comes to reality, the more it becomes capable of novel and intuitive predictions.

A final remark: in this thesis I favoured a semantic approach to the logics I have examined. The reason is that, as it is apparent in Bartha's paper (cf. §3.4.1), employing a syntactic approach bears the risk of losing sight of the desiderata. When dealing with deontic reasoning, a semantic approach provides much clearer intuitions about what we are doing relative to what we mean to do.

[3] Including: [vdTT99, CJ12].
[4] Namely [Mar13].
[5] E.g. cf. [SA85, Pas92] and somewhat similarly [CP09].
[6] E.g. [Bro06a, Bro10].
[7] Cf. [Hor93].
[8] Cf. [AB75, Lok06, Mar92, Gob99].

Acknowledgements.

My deepest gratitude goes to Jan Broersen and Sonja Smets for patiently supervising the research that led to this thesis. More specific gratitude goes to Sonja for undertaking the disproportionate task of forcing me into a less verbose (and Italian) writing style and trying to make my sentences look shorter; unsure whether she succeeded or not.

One more truck of gratitude is to be delivered to Roberto Ciuni, for helping me out especially with discussions on the *stit* framework and proving an useful ally when it came to break down wrong-looking arguments.

Appendix A: SDL

A.1 Syntax.

Let Φ_{SDL} be a set of atomic propositions. The BNF specification for the language $\mathcal{L}_{\mathbf{SDL}}$ is given as:

$$\mathcal{L}_{\mathbf{SDL}} : \varphi ::= p \mid \varphi \mid \neg\varphi \mid \varphi \vee \varphi \mid O\varphi$$

where p ranges over Φ_{SDL}. We then can define $\supset, \wedge$ from $\neg, \vee$ as usual.[9] Also, we introduce O's dual P:

$$P\varphi := \neg O \neg \varphi$$

An axiom system for SDL is given by:

all propositional tautologies	(TAUT)
$O(\varphi \supset \psi) \supset (O\varphi \supset O\psi)$	(O-K)
$O\varphi \supset P\varphi$	(O-D)

and the two inference rules:

$$\text{MP}: \quad \frac{\varphi \quad (\varphi \supset \psi)}{\psi} \qquad\qquad \text{O-NEC}: \quad \frac{\varphi}{O\varphi}$$

Then we define **SDL** as the smallest subset of $\mathcal{L}_{\mathbf{SDL}}$ containing all instances of the axioms TAUT, O-K, O-D and closed under MP and O-NEC.

Finally, we introduce a syntactic entailment relation $\vdash_{\mathbf{SDL}}$, where A is a set of **SDL**-formulae and ^{c} denotes closure under MP and O-NEC:

$$A \vdash_{\mathbf{SDL}} \varphi \quad := \quad \varphi \in (\mathbf{SDL} \cup A)^c$$

In particular, if $A = \varnothing$, we also write $\vdash_{\mathbf{SDL}} \varphi$ and this would mean that $\varphi \in \mathbf{SDL}$. The intuition is that $A \vdash_{\mathbf{SDL}} \varphi$ iff from the rules and axioms of SDL, plus any additional assumption in A, we can derive φ in, say, a Hilbert-style proof calculus.[10]

[9] Uniformly, we will use "$\supset$" to denote material implication throughout the thesis. The other notation is more standard.

[10] For additional technicalities, proofs and remarks, see [BdRV02].

Remarks. The logic is entirely standard (propositional calculus), the only detail worth noticing being the O and P operators. These are (modal) operators whose intended reading is, respectively, *Obligatory* and *Permissible*. For example, $O\varphi$ will read 'it is obligatory that φ'.

A.2 Semantics.

Let W be a countable set of *worlds* or 'states', I an interpretation function that maps every world w to a given set of atomic propositions in Φ_{SDL}, and D a serial 'deontically ideal alternative' relation with:

$$D := \langle w, v \rangle \mapsto \{0, 1\}$$

where w, v range over W. In other words, D is (the characteristic function of) a subset of $W \times W$. The intention is to have $D(w, v) = 1$ iff $\forall \varphi \in \mathbf{SDL}$: if $w \models_{\mathbf{SDL}} O\varphi$ then $v \models_{\mathbf{SDL}} \varphi$. So, we write that $D(w, v) = 1$, or wDv for short, when all obligations in w are fulfilled in v, or, so to say, v is an ideal world relative to w.

SDL-models are then tuples $\langle W, D, I \rangle$, whose typical element we will denote $\mathcal{M}^{\mathbf{SDL}}$.

Finally, we define a semantic entailment relation $\models_{\mathbf{SDL}}$ in the usual way. The only clause worth mentioning here is the one for O:

$$\mathcal{M}^{\mathbf{SDL}}, w \models_{\mathbf{SDL}} O\varphi \quad \leftrightarrow \quad \forall v \in W \text{ if } wDv \text{ then } v \models_{\mathbf{SDL}} \varphi$$

Appendix B: PD_eL

Remark: the main source for this material is [Mey87]. The parts concerning conditional actions are instead drawn from [Mey88].

B.1 Syntax.

Let A be a finite alphabet denoting atomic (*elementary*) actions, with typical elements $a, b, c....$ Next, let B be a set containing atomic propositions $p, q, r....$

The grammars for the set Act of actions and the set Ass of assertions are simultaneously defined by the BNFs:

$$Act : \underline{a} ::= a \mid \underline{a} \cup \underline{a} \mid \underline{a} \,\&\, \underline{a} \mid \underline{a}; \underline{a} \mid \bar{\underline{a}} \mid \varphi \to \underline{a}/\underline{a}$$

$$Ass : \varphi ::= p \mid \varphi \vee \varphi \mid \neg\varphi \mid [\underline{a}]\varphi$$

where a ranges over A and p over B. Finally we call V a special propositional variable in A. Its intended meaning is 'liability to sanction', or 'there is wrongdoing'. We also define the other boolean connectives $\wedge, \supset, \equiv$ out of $\vee$ and $\neg$, and we define $[\,a\,]$'s dual as usual: $\langle \underline{a} \rangle \varphi := \neg[\,\underline{a}\,]\neg\varphi$

A Hilbert-style proof system The basic system PD_eL is given by:

All tautologies of classical propositional logic		(PC)
$[\alpha](\varphi \supset \psi) \supset ([\alpha]\varphi \supset [\alpha]\psi)$		$(\Box\supset)$
$V \supset [\alpha]V$	**	(NP)
$[\alpha_1;\alpha_2]\varphi \equiv [\alpha_1]([\alpha_2]\varphi)$		$(;)$
$[\alpha_1 \cup \alpha_2]\varphi \equiv [\alpha_1]\varphi \vee [\alpha_2]\varphi$		$(\cup)$
$[\alpha_1 \,\&\, \alpha_2]\varphi \equiv [\alpha_1]\varphi \wedge [\alpha_2]\varphi$	*	$(\&)$
$[\varphi \to \alpha_1/\alpha_2]\psi \equiv (\varphi \supset [\alpha_1]\psi) \wedge (\neg\varphi \supset [\alpha_2]\psi)$		$(\to/)$
$\langle\alpha\rangle\varphi \equiv \neg[\alpha]\neg\varphi$		$(\Diamond)$
$[\overline{\alpha_1;\alpha_2}]\varphi \equiv [\bar{\alpha}_1]\varphi \wedge [\alpha_1][\bar{\alpha}_2]\varphi$		$(\bar{;})$
$([\bar{\alpha}_1]\varphi \vee [\bar{\alpha}_2]\varphi) \supset [\overline{\alpha_1 \cup \alpha_2}]\varphi$	*	$(\bar{\cup})$
$[\overline{\alpha_1 \,\&\, \alpha_2}]\varphi \equiv [\bar{\alpha_1}]\varphi \wedge [\bar{\alpha_2}]\varphi$		$(\bar{\&})$
$[\overline{\varphi \to \alpha_1/\alpha_2}]\psi \equiv (\varphi \supset [\bar{\alpha}_1]\psi) \wedge (\neg\varphi \supset [\bar{\alpha}_2]\psi)$		$(\overline{\to/})$
$[\bar{\bar{\alpha}}]\varphi \equiv [\alpha]\varphi$		$(^{=})$
$[\underline{\varnothing}]\varphi$		$(\underline{\varnothing})$

*:provided that the durations of α_1, α_2 are the same.[11]

**:this axiom is listed as 'optional' in standard PD_eL. If we want to add it, we will have to constrain the interpretation function of PD_eL-models to evaluate V accordingly.

As inference rules we have: MP : $\dfrac{p \quad (p \to q)}{q}$ and NEC : $\dfrac{p}{[\alpha]p}$

Then PD_eL is defined as the smallest set containing all instances of the above axioms, with $\alpha, \alpha_1, \alpha_2$ ranging over Act and φ, ψ ranging over Ass, and closed under N and MP.

Relative to the semantics given in the next section, this system is sound and can be made complete by adding a *done* predicate that keeps track of what action is executed at each step.[12]

[11] *Generic* actions are, semantically, rendered as infinite sequences of *atomic* actions. Such sequences have a 'prefix' of *relevant* atomic actions, that are the ones actually executed to carry out the action, and an infinite suffix of *irrelevant* atomic actions (that are there just to make sure all sequences are infinite). The duration of an action is then defined as the length of the prefix. One can expect to have infinitely long actions, as well. For a more precise definition, refer to [Mey88].

[12] Cfr. [Mey86] for the proof.

B.2 Semantics.

In a nutshell, we have states $s, s', s''...$ and actions $a, a', a''...$ that are grouped up in finite and nonempty *synchronicity sets* (or *s-sets*) that, in turn, label the transitions between states. Intuitively, the transition $s \xrightarrow{\{a,a',a''...a^n\}} s'$, means that executing simultaneously all the a^is in state s we change the state of the system to s'.

Next, we define *synchronicity traces* (or *s-traces*); these are *finite sequences* $S_0, S_1...S_n$ *of s-sets* S_i. Denotations of action expressions will be collections of infinite s-traces[13], which are specified only up to a certain length. To mark whether a s-set is relevant for the trace (i.e. is part of the specification of the trace, or, instead, has just been added there to make the trace infinite) we will use superscripts; a $^{(1)}$ means the s-set is relevant/specified, a $^{(0)}$ that it is not.[14] So the semantic domain $\mathfrak{C}$ is *the collection of s-traces which have a finite prefix of relevant s-sets and an infinite prefix of irrelevant s-sets.* We call these *admissible* traces. if $[a]$, $[b]$, $[c]$... are s-sets, an example of a s-trace of this kind is:

$$\{[a]^{(1)} \circ [b]^{(1)} \circ [c]^{(1)} \circ [d]^{(0)} \circ [e]^{(0)}...\}$$

Then a binary operation $\cap$ is defined on admissible s-traces in the following way: if T is an s-trace, we define $\pi_n(T)$ to be the nth projection of the sequence T; then $\cap$ is a function

$$\cap := \langle T_1, T_2 \rangle \mapsto T_3$$

where T_3 is such that:

if $\pi_n(T_1) = [a]^{(i)}$ and $\pi_n(T_2) = [b]^{(j)}$, then

$$\pi_n(T_3) = \begin{cases} [c]^{(\text{MAX}(i,j))} & \text{if } a = b = c \\ \varnothing & \text{otherwise} \end{cases}$$

In words, $\cap$ takes two traces T_1, T_2 and outputs their 'intersection' T_3 in this way: the nth position in T_3 there will be what T_1 and T_2 have at their nth position if that is the same s-set; and $\varnothing$ if the nth positions of the input traces contain different s-sets. Finally, the nth position of T_3 is relevant only if at least one of T_1, T_2 contains a relevant s-set in its nth position.

[13] They need be infinite because time never stops running; but as we will see only a finite initial part of each s-trace will be semantically relevant and specified.

[14] Formally, we would have to re-define s-sets as pairs $\langle S, i \rangle$ where $i \in \{0, 1\}$ denotes the relevance of S and S itself is the s-set proper.

Next, we need an action complement operation $\sim$ to be the counterpart of $^-$. The complement $\widetilde{[a]^{(i)}}$ of a s-set $[a]^{(i)}$ is defined as the s-set $[a^c]^{(i)}$, where ^{c} denotes, in this case, the complement relative to the powerset of A (excluding $\varnothing$).[15]

Action semantics. First, let $cut(T)$ be a function that returns the relevant part of an s-trace. We define the semantic function $[\![.]\!]$, that takes elements of Act to elements of $\mathfrak{C}$:

$$
\begin{aligned}
&[\![\underline{a}]\!] = \{S \mid a \in S\}^{(1)} \circ (\mathcal{P}^+(A)^{(0)})^\omega && ** \\
&[\![\alpha_1; \alpha_2]\!] = cut([\![\alpha_1]\!]) \circ [\![\alpha_2]\!] \\
&[\![\alpha_1 \underline{\cup} \alpha_2]\!] = [\![\alpha_1]\!] \cup [\![\alpha_2]\!] \\
&[\![\alpha_1 \,\&\, \alpha_2]\!] = [\![\alpha_1]\!] \cap [\![\alpha_2]\!] \\
&[\![\bar{\alpha}_1]\!] = \widetilde{[\![\alpha_1]\!]} \\
&[\![\varnothing]\!] = \varnothing \\
&[\![\underline{U}]\!] = \mathcal{P}^+(A)^1 \circ (\mathcal{P}^+(A)^{(0)})^\omega && **
\end{aligned}
$$

**: $(\mathcal{P}^+(A)^{(0)})^\omega$ stands for 'an infinite suffix of irrelevant nonempty subsets of the set of act types A'.

Conditional actions have been omitted because we had not yet assigned an interpretation to assertions.

Assertion semantics. Let Σ denote the universe of states. Then suppose we are given a function ρ that interprets s-sets in terms of state transitions; $\rho := \mathcal{P}^+(A) \to (\Sigma \to \Sigma)$.[16] So, $\rho(S)(s)$, where S is a s-set and s a state, returns the state s' which results from executing S[17], in s.

We now define recursively a function $\mathcal{R}$ to do the same on s-traces. Let t be a finite s-trace $S_0, S_1 ... S_n$.

$$\mathcal{R}(S_0)(s) = \rho(S_0)(s)$$

$$\mathcal{R}(S_i \circ S_{i+1})(s) = \mathcal{R}(S_i)(\mathcal{R}(S_{i+1})(s))$$

So, informally speaking, $\mathcal{R}(S)(s)$ returns a state s' that is reached by following the relevant part of trace S, starting from state s.

---

[15] So, informally, executing $\widetilde{[a]^{(i)}}$ means doing anything (any action) that does not involve doing a.

[16] The intuition being that executing some action a is, strictly speaking, to follow some path through the state-transition space. So, depending on which state we are in, executing a will take us to some state a' (or nowhere at all, if a was not an option to begin with).

[17] Being S an s-set, we may more precisely have to write "jointly executing all atomic actions in S".

We are ready to define the last meaning function we will need: $[\![.]\!]_R := Act \to (\Sigma \to \mathcal{P}(\Sigma))$, which is defined by:

$$[\![\alpha]\!]_R(\sigma) = \mathcal{R}(cut([\![\alpha]\!]))(\sigma)$$

The full action semantics is then defined by:

$$[\![\underline{a}]\!](\sigma) = \{S \mid a \in S\}^{(1)} \circ (\mathcal{P}^+(A)^{(0)})^\omega$$
$$[\![\alpha_1; \alpha_2]\!](\sigma) = cut([\![\alpha_1]\!](\sigma)) \circ [\![\alpha_2]\!]([\![\alpha_1]\!]_R(\sigma))$$
$$[\![\alpha_1 \underline{\cup} \alpha_2]\!](\sigma) = [\![\alpha_1]\!](\sigma) \cup [\![\alpha_2]\!](\sigma)$$
$$[\![\alpha_1 \,\&\, \alpha_2]\!](\sigma) = [\![\alpha_1]\!](\sigma) \cap [\![\alpha_2]\!](\sigma)$$
$$[\![\psi \to \alpha_1/\alpha_2]\!](\sigma) = \begin{cases} [\![\alpha_1]\!](\sigma) & \text{if } \sigma \models \psi; \\ [\![\alpha_2]\!](\sigma) & \text{if } \sigma \not\models \psi \end{cases}$$
$$[\![\bar{\alpha}_1]\!](\sigma) = \widetilde{[\![\alpha_1]\!]}(\sigma)$$
$$[\![\varnothing]\!](\sigma) = \varnothing$$
$$[\![\underline{U}]\!](\sigma) = \mathcal{P}^+(A)^1 \circ (\mathcal{P}^+(A)^{(0)})^\omega$$
$$[\![\alpha]\!]_R(\sigma) = \mathcal{R}(cut([\![\alpha]\!](\sigma)))(\sigma)$$

And finally, for $\tau \in \mathcal{P}(\Sigma)$ we define $[\![\alpha]\!]_R(\tau) = \bigcup_{\sigma \in \tau} [\![\alpha]\!]_R(\sigma)$.

Let $\sigma \in \Sigma$ and $\psi_1, \psi_2 \in Ass$; the boolean cases are defined in the usual way. If $\alpha \in Act$, we define

$$\sigma \models [\,\alpha\,]\psi \quad \text{iff } \forall \sigma' \in [\![\alpha]\!]_R(\sigma) : \sigma' \models \psi$$

$$\sigma \models \langle\alpha\rangle\psi \quad \text{iff } \exists \sigma' \in [\![\alpha]\!]_R(\sigma) : \sigma' \models \psi$$

Finally, truth in a model and validity can be defined as usual.

Appendix C: $d_e stit$

C.1 Syntax.

As usual, we give the syntax for $d_e stit$ in compact BNF:

$$\mathcal{L}_{d_e stit} : \varphi ::= p \mid \neg\varphi \mid \varphi \vee \varphi \mid \Box\varphi \mid F\varphi \mid P\varphi \mid [a\ dstit]\varphi \mid S\varphi$$

Where p ranges over a countable set of propositional variables Ψ and a over a countable set of agents Ag. We add to Ψ a special propositional atom V, whose intended meaning is, as we have already seen in the previously examined systems, 'there is a violation' or 'liability to punishment obtains', or something similar. Then, we introduce $\wedge, \supset p, \bot, \Diamond, \equiv$ as abbreviations, as usual.

I have been suggested to stress somewhere that $[a\ dstit]$ is *not* a normal modal operator, and this seems the right place to do it.

C.2 Semantics.

Essential to $d_e stit$[18] is an '*ockhamist*'[19] analysis of time; time is thus modelled as a set of *moments* endowed with a tree-like ordering by an earlier/later than relation $\leq$. Formally, the first building block of a $d_e stit$-frame is thus a tuple $\langle M, \leq \rangle$, where M is a set of moments (worlds) with typical elements $m, m', m''...$ and $\leq$ is an 'earlier than' relation defined on them. Features of the ordering are:

1. *no backwards branching*: $\forall m \forall m' \forall m''((m' \leq m \wedge m'' \leq m) \supset (m' \leq m'' \vee m'' \leq m'))$

2. *historical connection*: $\forall m \forall m' \exists m''(m'' \leq m' \wedge m'' \leq m)$

[18] And to *stit* in general, indeed. This whole section actually applies to *stit* generally, and not as much to $d_e stit$ specifically. However, since we are interested in the latter, we will keep using this talk to avoid confusions.

[19] The seminal work of Arthur Prior famously traced this analysis of time as a branching structure back to Ockham. See [Pri57].

An informal explanation for 1. and 2. could be the following: 1. makes sure that even though the future is 'open' (our choices in some moment can influence what will be the case at later moments), the past is 'closed': what has been can no longer change. 2. just says that any two moments have a meet. This avoids, for example, the existence of multiple temporal trees with no common past. Together, these two make sure that any two *histories* (maximal chains of moments) share an initial segment, then split up and never meet again.

Histories h, h' are said to be *undivided at* m (write $h \equiv_m h'$) iff $\exists m'(m < m' \wedge m' \in h \cap h')$. In words, two histories are undivided at some moment iff they share some *later* moment. Then we also define the set of histories that pass through a given moment as:

$$H_m := \{h \mid m \in h\}$$

Next, we introduce Ag, a primitive, which is a set of agents with typical elements $a, a', b....$

A *choice set* for an agent a in a moment m is a *partition* of the set of histories that pass through m; we write $\mathbf{Choice}_a(m)$ for such partition. Members of the partition are dubbed *possible choices* for a at m. One fundamental restriction we require is that there can be *no choice between undivided histories*. Formally, this amounts to:

$$\forall h \forall h' \forall H(h \equiv_m h' \wedge H \in \mathbf{Choice}_a(m) \rightarrow (h \in H \leftrightarrow h' \in H))$$

If h, h' belong to the same possible choice for a at m, we say that h, h' are choice-equivalent for a at m (or $\mathbf{Choice}_a(m)$-*equivalent*), and write $h \equiv^a_m h'$. This captures the intuition that such two histories cannot be told apart by a at moment m: no choice available to him at that moment can distinguish between them.

***d*$_e$*stit* Frames.** A $d_e stit$-frame, finally, is a 4-tuple $\langle M, \leq, Ag, \mathbf{Choice}\rangle$ with components satisfying the above conditions.

***d*$_e$*stit* Models.** Adding a valuation to the above is simply a matter of mapping the set Ψ of propositional variables to the powerset of the set $\{\langle m, h\rangle \mid m \in M, h \in \langle M, \leq\rangle\}$. This is done by an interpretation function I.

$$I := \Psi \mapsto \{\langle m, h\rangle \mid m \in M, h \in \langle M, \leq\rangle\}$$

Whenever $\langle m, h\rangle \in I(p)$, we informally say that "p is *true at* $\langle m, h\rangle$". That is, truth of propositions is evaluated against moment/history pairs $\langle m, h\rangle$.

A $d_e stit$-model is then a pair $\langle F, I\rangle$, where F is a $d_e stit$-frame, or equivalently a 5-tuple $\langle M, \leq, Ag, \mathbf{Choice}, I\rangle$.

Formally, we define an entailment relation $\models_{destit}$ (omitting the subscript whenever unambiguous):

$$\begin{aligned}
\mathcal{M}, m, h \models S\varphi \quad &\Longleftrightarrow_{df} \mathcal{M}, m, h' \models \varphi \text{ for all } h' \in H_m \\
\mathcal{M}, m, h \models \Box\varphi \quad &\Longleftrightarrow_{df} \mathcal{M}, m', h' \models \varphi \text{ for all } m' \text{ and all } h' \in H_{m'} \\
\mathcal{M}, m, h \models P\varphi \quad &\Longleftrightarrow_{df} \exists m' < m (\mathcal{M}, m', h \models \varphi) \\
\mathcal{M}, m, h \models F\varphi \quad &\Longleftrightarrow_{df} \exists m' > m (\mathcal{M}, m', h \models \varphi)
\end{aligned}$$

Semantics for the boolean cases is entirely classical, and so is the clause for truth in a model or truth in a frame.

We define $\mathcal{M}, m, h \models [a\ dstit]\varphi$ to hold iff the following two conditions are satisfied:

1. *Positive condition.* $\mathcal{M}, m, h' \models \varphi$ for all h' with $h' \equiv^a_m h$.

2. *Negative condition.* there exists some h'' s.t. $m \in h''$ and $\mathcal{M}, m, h'' \models \neg\varphi$. That is, $\mathcal{M}, m, h \not\models S\varphi$.

Defining O. Bartha defined O as

$$O[a\ dstit]\varphi \quad := \quad S(\neg[a\ dstit]\varphi \supset V) \tag{O}$$

However[20], we might want to define O in another way:

$$O'[a\ dstit]\varphi \quad := \quad S(\neg[a\ dstit]\varphi \supset [a\ dstit]V) \tag{O'}$$

(O') seems to have two main advantages over (O). The first is philosophical, the second is technical.

1. at a philosophical level, it seems reasonable that a violation (a wrongdoing) is carried out just like any action is, by the agent that did something he should not have done.

2. at a technical level, (O') makes sure that a particular agent can always be individuated as the 'source of violation'. A similar result would probably be obtained by indexing violation atoms to agents; a's wrongdoing is not the same as b's. Consequently, a world where a committed something false should be distinguishable from one where b did. Lacking past tense operators, using (O') is a cheap way to have that.

[20] As Broersen does in some of his works, such as [Bro09b].

If one wishes to adopt (O'), we would suggest to read V as '(the agent) makes himself liable to punishment'. This way, $[a\ stit]V$ would be granted a neater intuition.[21] For the purposes of this thesis the two definitions are interchangeable, and using (O') would just graphically complicate the formulae. For this reason, we will use (O).

[21] Though not motivating his choice, also Broersen adopts (O') as a definition of O: cf. [Bro09b, p. 15].

Appendix D: *xstit*

Here we will give the formal details of the logic following almost verbatim Broersen's [Bro11][22]. In the thesis however we use a slightly modified version (we add infinitely many V_φ violation atoms to the language). Having the differences no impact on completeness and soundness, there is no need to give such proofs for the 'new' logic.

We could as well have added, more simply, a set $V_1, V_2, V_3...$ of violation atoms and then render every norm (quite artificially) by using a 'fresh' atom from the list. This is the strategy employed in [Mey87]. Having a violation for each sentence, as we do, seems much more intuitive.

D.1 Syntax.

Let there be a countable set of propositions P, with p typical element and a finite set Ags of names of agents, with typical element a (and $A \subseteq Ags$). Unlike Broersen's XSTIT, we have a violation atom for each wff, including violation formulas themselves.[23] Also, in this thesis we use a single-agent sub-language where only *xstit* sentences of the form $[a\ xstit]\varphi$ can occur. So, the BNF for the full *xstit* language is:

$$\mathcal{L}_{xstit} : \varphi ::= p \mid V_\varphi \mid \neg\varphi \mid \varphi \vee \varphi \mid S\varphi \mid [a\ xstit]\varphi \mid X\varphi$$

Remark: *V is an operator.* However, since it is non-compositional (it is evaluated independently from the sentence it scopes over, and no axiom regulates its interaction with other operators; which makes it essentially alike atomic propositions), we will write V_φ instead of $V(\varphi)$ to distinguish it from 'standard' operators.

Finally we introduce P as a shortcut for $\neg S\neg$: $P\varphi := \neg S\neg\varphi$, the dual of historical necessity (settledness). Finally, we define a family of O_a operators for 'a is obliged to', as: $O_a\varphi := S(\neg[a\ xstit]\varphi \supset XV_\varphi)$.

[22] Also see [Bro09a, Bro09b].

[23] Else sentences such as V_{V_p} needed to model second-order obligations, wouldn't be wff.

Axiom system. An axiom system for *xstit* extends any one for propositional logic with the following axiom schemas:

S5 for S	(S)
KD for each $[A\ xstit]$	(Ax)
$\neg X\neg\varphi \supset X\varphi$	(Det)
$SX\varphi \equiv [\varnothing\ xstit]\varphi$	($\varnothing$-SettX)
$[Ags\ xstit]\varphi \equiv XS\varphi$	(Ags-XSett)
$[A\ xstit]\varphi \supset [A \cup B\ xstit]\varphi$	(C-Mon)
$(P[A\ xstit]\varphi \wedge P[B\ xstit]\psi) \supset P([A\ xstit]\varphi \wedge [B\ xstit]\psi)$ for each A, B s.t. $A \cap B = \varnothing$	(Indep-G)

If we want to implement the *no-pardon* variant discussed in §3.5, we shall add the following axiom:

$V_\varphi \supset XV_\varphi$	(x-np)

No other axiom beside (x-np) regulates the behaviour of V.

D.2 Semantics.

***xstit* frames** are tuples $\langle W, H, R_X, R_S, \{R_A \mid A \subseteq Ags\}\rangle$ such that:

- W is an infinite set of static states $s, s'...$
- $H \subseteq 2^{2^S \setminus \varnothing} \setminus \varnothing$ is a nonempty set of histories h, h'. Dynamic states are tuples $\langle s, h\rangle$ with $s \in h$.
- R_X is a 'next state' relation, serial, such that $\langle s, h\rangle R_X \langle s', h'\rangle \Rightarrow h = h'$.
- R_S is an 'historical necessity' relation with $\langle s, h\rangle R_S \langle s'h'\rangle \Longleftrightarrow s = s'$.
- the R_As are 'effectivity' relation over dynamic states such that:
  - $R_\varnothing = R_S \circ R_X$
  - $R_{Ags} = R_X \circ R_S$
  - $R_A \subseteq R_B$ for $B \subset A$
  - for $A \cap B = \varnothing$, if $\langle s_1, h_1\rangle R_S \langle s_2, h_2\rangle$ and $\langle s_1, h_1\rangle R_S \langle s_3, h_3\rangle$, then:
    1. $\exists s_4, h_4$ s.t. $\langle s_1, h_1\rangle R_S \langle s_4, h_4\rangle$
    2. if $\langle s_4, h_4\rangle R_A \langle s_5, h_5\rangle$ then $\langle s_2, h_2\rangle R_A \langle s_5, h_5\rangle$
    3. if $\langle s_4, h_4\rangle R_B \langle s_6, h_6\rangle$ then $\langle s_3, h_3\rangle R_B \langle s_6, h_6\rangle$

For an intuitive explanation of these conditions, we refer back to [Bro11, Bro09b, Bro09a].

***xstit* models.** The V_φ atoms are considered to be like atomic propositions; i.e. their truth-values depend solely on the model's evaluation function. Consequently given a set of atomic propositions P, let $P^+ := P \cup \{V_\varphi \mid V_\varphi \in \mathcal{L}_{xstit}\}$; then *xstit* models are *xstit* frames endowed with a valuation function

$$\pi : P^+ \mapsto 2^{W \times H}$$

which assigns to each atomic proposition (and each violation atom) the set of dynamic states in which they are true.

Truth definition. We define an entailment relation $\mathcal{M}, \langle s, h\rangle \models_{xstit} \varphi$, where $\mathcal{M}$ is a *xstit* model, $\langle s, h\rangle$ a dynamic state $\in \mathcal{M}$, $\varphi \in \mathcal{L}_{xstit}$, as follows (omitting the subscript whenever unambiguous):

$$\mathcal{M}, \langle s, h\rangle \models p \qquad \Longleftrightarrow_{df} \langle s, h\rangle \in \pi(p)$$

if $p \in P$ or $p = V_\varphi$ for some $\varphi \in \mathcal{L}_{xstit}$.[24] The boolean cases are defined in the usual way. The remaining interesting cases are:

$$\begin{aligned}
&\mathcal{M}, \langle s, h\rangle \models S\varphi && \Longleftrightarrow_{df} \langle s, h\rangle R_S \langle s', h'\rangle \Rightarrow \mathcal{M}, \langle s', h'\rangle \models \varphi \\
&\mathcal{M}, \langle s, h\rangle \models [\, A\ xstit\,]\varphi && \Longleftrightarrow_{df} \langle s, h\rangle R_A \langle s', h'\rangle \Rightarrow \mathcal{M}, \langle s', h'\rangle \models \varphi \\
&\mathcal{M}, \langle s, h\rangle \models X\varphi && \Longleftrightarrow_{df} \langle s, h\rangle R_X \langle s', h'\rangle \Rightarrow \mathcal{M}, \langle s', h'\rangle \models \varphi
\end{aligned}$$

[24] Again, this is an addition to Broersen's XSTIT.

Bibliography

[AB75] A.R. Anderson and N.D. Belnap. *Entailment: The Logic of Relevance and Necessity*, volume I. Princeton University Press, 1975.

[AB81] C.E. Alchourrón and E. Bulygin. The expressive conception of norms. In R. Hilpinen, editor, *NEW STUDIES IN DEONTIC LOGIC: Norms, Actions, and the Foundations of Ethics*. D. Reidel Publishing Company, 1981.

[ABvdT10] G. Aucher, G. Boella, and L. van der Torre. Prescriptive and descriptive obligations in dynamic epistemic deontic logic. In P. Casanovas et al., editor, *AICOL Workshops 2009*. Springer-Verlag, 2010.

[AHBG+13] T. Athan, H-Boley, G. Governatori, M. Palmirani, A. Paschke, and A. Wyner. OASIS legalruleml. In *International Conference on Artificial Intelligence and Law*, pages 3–12, 2013.

[AHK98] R. Alur, T.A. Henzinger, and O. Kupferman. Alternating-time temporal logic. *Technical Reports (CIS)*, 1998.

[And58] A.R. Anderson. A reduction of deontic logic to alethic modal logic. *Mind*, 1958.

[And67a] A.R. Anderson. The formal analysis of normative systems. In *The Logic of Decision and Action*. University of Pittsburgh Press, 1967.

[And67b] A.R. Anderson. Some nasty problems in the formalization of ethics. *Nous*, 1:345–360, 1967.

[Ang08] A.J.J. Anglberger. Dynamic deontic logic and its paradoxes. *Studia Logica*, 2008.

[Bar93] P. Bartha. Conditional obligation, deontic paradoxes, and the logic of agency. *Annals of Mathematics and Artificial Intelligence*, 1993.

[BdRV02] P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic*. Cambridge University Press, 2002.

[Bel91] N. Belnap. Backwards and forwards in the modal logic of agency. *Philosophy and Phenomenological Research*, 1991.

[BHar] J.M. Broersen and A. Herzig. Using stit theory to talk about strategies. In S. Ghosh J. van Benthem and R. Verbrugge, editors, *Modeling Strategic Reasoning*, Texts in Logic and Games. Springer, [to appear].

[BP90] N. Belnap and M. Perloff. Seeing to it that: A canonical form for agentives. In *Knowledge Representation and Defeasible Reasoning*. Springer, 1990.

[BPX01] N. Belnap, M. Perloff, and M. Xu. *Facing the future*. Oxford University Press, 2001.

[BRea12] J.F. Beltrán, G.B. Ratti, and et al. *The Logic of Legal Requirements: Essays on Defeasibility*. Oxford University Press, 2012.

[Bro03a] J.M. Broersen. *Modal Action Logics for Reasoning about Reactive Systems*. PhD thesis, University of Amsterdam, 2003.

[Bro03b] J.M. Broersen. Relativized action negation for dynamic logics. In F. Wolter P. Balbiani, N-Y. Suzuki and M. Zakharyaschev, editors, *Advances in Modal Logic*, volume 4, page 51–70. College Publications, 2003.

[Bro04a] J.M. Broersen. Action negation and alternative reductions for dynamic deontic logics. *Journal of applied logic*, 2004.

[Bro04b] J.M. Broersen. On the logic of 'being motivated to achieve ρ, before δ'. In J. Alferes and J. Leite, editors, *Proceedings Ninth European Conference on Logics in Artificial Intelligence (JELIA'04)*, volume 3229 of *Lecture Notes in Artificial Intelligence*, pages 334–346. Springer, 2004.

[Bro04c] M.A. Brown. Rich deontic logic: A preliminary study. *Journal of Applied Logic*, 2004.

[Bro06a] J.M. Broersen. Strategic deontic temporal logic as a reduction to atl, with an application to chisholm's scenario. In J-J.C. Meyer, editor, *DEON 2006, LNAI 4048*, pages 53–68, July 2006. Paper presented at DEON2006, Utrecht, The Netherlands.

[Bro06b] J.M. Broersen. Strategic deontic temporal logic as a reduction to atl, with an application to chisholm's scenario. In J-J.C. Meyer L. Goble, editor, *DEON 2006, LNAI 4048*, page 53–68, 2006.

[Bro09a] J.M. Broersen. A complete stit logic for knowledge and action, and some of its applications. In M. Baldoni et al., editor, *DALT 2008, LNAI 5397*, pages 47–59. Springer, 2009.

[Bro09b] J.M. Broersen. Deontic epistemic stit logic distinguishing modes of mens rea. *Journal of applied logic*, 2009.

[Bro10] J.M. Broersen. Problem 4: temporal deontic reasoning. In P. Casanovas et al., editor, *ESSLLI2010*. Springer-Verlag, 2010.

[Bro11] J.M. Broersen. Making a start with the stit logic analysis of intentional action. *Journal of Philosophical Logic*, 40(4):499–530, 2011.

[BvdT03] G. Boella and L. van der Torre. Permissions and obligations in hierarchical normative systems. *ICAIL '03: Proceedings of 9th international conference on Artificial Intelligence and law*, 2003.

[BvdT12] J.M. Broersen and L. van der Torre. Ten problems of deontic logic and normative reasoning in computer science. *Lecture Notes in Computer Science*, 7388:55–88, 2012.

[Cas81] H-N. Castañeda. The paradoxes of deontic logic: The simplest solution to all of them in one fell swoop. In *New Studies in Deontic Logic*. Synthese Library, 1981.

[CC86] N.C.A. Costa and W.A. Carnielli. Paraconsistent deontic logic. *Philosophia*, 16:293–305, 1986.

[Che69] B. Chellas. *The Logical Form of Imperatives*. PhD thesis, Stanford University, 1969.

[Che92] B. Chellas. Time and modality in the logic of agency. *Studia Logica*, 1992.

[Chi63] R.M. Chisholm. Contrary-to-duty imperatives and deontic logic. *Analysis*, 1963.

[CJ96] J. Carmo and A.J.I. Jones. Deontic database constraints, violation and recovery. *Studia Logica*, 57:139–165, 1996.

[CJ05] J. Carmo and A.J.I. Jones. Deontic logic and contrary-to-duties. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, volume 3, pages 203–279. Kluwer, 2 edition, 2005.

[CJ12] J. Carmo and A.J.I. Jones. Completeness and decidability results for a logic of contrary-to-duty conditionals. *Oxford University Press*, 23(3), 2012.

[CM09] P.F. Castro and T.S.E. Maibaum. Deontic logic, contrary to duty reasoning and fault tolerance. *Electronic Notes in Theoretical Computer Science*, 2009.

[CP09] M.E. Coniglio and N.M. Peron. A paraconsistentist approach to chisholm's paradox. *Principia*, 13, 2009.

[DMW96] R D'Altan, J.-J.Ch. Meyer, and R.J. Wieringa. An integrated framework for ought-to-be and ought-to-do constraints. *Artificial Intelligence and Law*, 1996.

[Esh14] A. Eshleman. Moral responsibility. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Summer 2014 edition, 2014.

[FH88] R. Fagin and J.Y. Halpern. Belief, awareness, and limited reasoning. *Artificial Intelligence*, 34:39–76, 1988.

[For84] J.W. Forrester. Gentle murder, or the adverbial samaritan. *Journal of Philosophy*, 1984.

[Gas95] R. Gaskin. *The Sea Battle and the Master Argument: Aristotle and Diodorus Cronus on the Metaphysics of the Future*. De Gruyter, 1995.

[Gob99] L. Goble. Deontic logic with relevance. In P. McNamara and H. Prakken, editors, *Norms, Logis and Information Systems*, pages 331–346. ISO Press, 1999.

[Gob00] L. Goble. Multiplex semantics for deontic logic. *Nordic Journal of Philosophical Logic*, 5(2):113–134, 2000.

[Gob03] L. Goble. Preference semantics for deontic logic part i — simple models. *Logique et Analyse*, 46(183-184), 2003.

[Gre39] K. Grelling. Zur logik der sollsätze. *Unity of Science Forum*, pages 44–47, 1939.

[Gre14] M. Green. Speech acts. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2014 edition, 2014.

[HB95] J.F. Horty and N. Belnap. The deliberative stit: A study of action, omission, ability, and obligation. *Journal of Philosophical Logic*, 24(6):583–644, Dec 1995.

[Heg75] G.W.F. Hegel. *Lectures on the philosophy of world history: introduction, reason in history*. Cambridge University Press, 1975.

[HF70] R. Hilpinen and D. Føllesdal. Deontic logic: An introduction. In *Deontic Logic: Introductory and Systematic Readings*. D. Reidel Publishing Company, 1970.

[HM39] A. Hofstadter and J.C.C. McKinsey. On the logic of imperatives. *Philosophy of Science*, pages 446–457, 1939.

[Hor89] J.F. Horty. An alternative stit operator. Technical report, Philosophy Department, University of Maryland, 1989.

[Hor93] J.F. Horty. Deontic logic as founded on nonmonotonic logic. *Annals of Mathematics and Artificial Intelligence*, pages 69–91, 1993.

[Hor01] J.F. Horty. *Agency and Deontic Logic*. Oxford University Press, 2001.

[Hum10] D. Hume. *A Treatise of Human Nature*. Project Gutenberg, 2010.

[Jø 8] J. Jørgensen. Imperatives and logic. *Erkenntnis*, pages 288–296, 1937-8.

[Kan71] S. Kanger. New foundations for ethical theory. In R. Hilpinen, editor, *Deontic Logic: Introductory and Systematic Readings*, page 36–58. D. Reidel Publishing Company, 1971. First published as a privately distributed pamphlet (1957).

[Kui12] L.B. Kuijer. Sanction semantics and contrary-to-duty obligations. In J.M. Broersen T. Agotnes and D. Elgesem, editors, *Deontic Logic in Computer Science: 11th International Conference, DEON 2012, Bergen, Norway, July 16-18, 2012, Proceedings*, 2012.

[LCR08] A. Lanteri, C. Chelini, and S. Rizzello. An experimental investigation of emotions and reasoning in the trolley problem. *Journal of Business Ethics*, 2008.

[Lok04] G-J.C. Lokhorst. Mally's deontic logic. *Grazer Philosophische Studien*, 2004.

[Lok06] G-J.C. Lokhorst. Andersonian deontic logic, propositional quantification, and mally. *Notre Dame Journal of Formal Logic*, 47(3):385–396, 2006.

[Lok13] G-J.C. Lokhorst. Mally's deontic logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2013 edition, 2013.

[LS07] S. Lindström and K. Segerberg. *Modal Logic and Philosophy*, chapter 21, pages 1149–1209. Elsevier, 2007.

[Mal26] E. Mally. *Grundgesetze des Sollens: Elemente der Logik des Willens*. Graz: Leuschner und Lubensky, Universitäts-Buchhandlung, 1926.

[Mar92] E.D. Mares. Andersonian deontic logic. *Theoria*, 58(1), 1992.

[Mar13] A. Marra. What should have been the case. a temporal update semantics for necessity deontic modals. Master's thesis, ILLC, 2013.

[McA81] R.P. McArthur. Anderson's deontic logic and relevant implication. *Notre Dame Journal of Formal Logic*, 22:145–154, 1981.

[McN96] P. McNamara. Making room for going beyond the call. *Mind*, pages 415–450, 1996.

[McN14] P. McNamara. Deontic logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2014 edition, 2014.

[MDW94] J-J.Ch. Meyer, F.M.P. Dignum, and R.J. Wieringa. The paradoxes of deontic logic revisited: A computer science perspective. Technical report, University of Utrecht, 1994.

[Men39] K. Menger. A logic of the doubtful: On optative and imperative logic. In *Reports of a Mathematical Colloquium*, pages pp. 53–64,. Indiana University Press, 1939.

[Mey86] J-J.Ch. Meyer. A sound and complete logic for propositional deontic reasoning. Technical report, Free University, 1986.

[Mey87] J-J.Ch. Meyer. A simple solution to the "deepest" paradox in deontic logic. *Logique et Analyse*, page 81–90, 1987.

[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*, 1988.

[Mey00] J-J.Ch. Meyer. Dynamic logic for reasoning about actions and agents. In J. Minker, editor, *Logic-Based Artificial Intelligence*, page 281–311. Kluwer, 2000.

[Mot73] P.L. Mott. On chisholm's paradox. *Journal of Philosophical Logic*, 1973.

[NR14] P.E. Navarro and J.L. Rodriguez. *Deontic Logic and Legal Systems*. Cambridge University Press, 2014.

[Pas92] J. Pasek. Prescriptive obligation and forrester's paradox. *Erkenntnis*, 1992.

[Pri54] A.N. Prior. The paradoxes of derived obligation. *Mind*, 1954.

[Pri57] A.N. Prior. *Time and Modality*. Clarendon Press, 1957.

[PS96] H. Prakken and M.J. Sergot. Contrary-to-duty obligations. *Studia Logica*, 1996.

[PS97] H. Prakken and M.J. Sergot. Dyadic deontic logic and contrary-to-duty obligations. In D. Nute, editor, *Defeasible Deontic Logic*, pages 223–262. Synthese, 1997.

[Ran39] R. Rand. Logik der forderungssiitze. *Revue internationale de la théorie du droit*, 1939.

[Ros41] A. Ross. Imperatives and logic. *Theoria*, pages 53–71, 1941.

[SA85] W. Sinnott-Armstrong. A solution to forrester's paradox of gentle murder. *The Journal of Philosophy*, 82(3):162–168, 1985.

[Sch35] E. Schrödinger. Die gegenwärtige situation in der quantenmechanik. *Naturwissenschaften*, 23(48):807–812, 1935.

[Ser90] M.J. Sergot. The representation of law in computer programs: A survey and comparison. In T. J. M. Bench-Capon, editor, *Knowledge Based Systems and Legal Applications*. Academic Press, 1990.

[SMK13] K. Segerberg, J-J. Meyer, and M. Kracht. The logic of action. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2013 edition, 2013.

[SW08] C. Semmling and H. Wansing. From bdi and stit to bdi-stit logic. *Logic and Logical Philosophy*, 2008.

[Tho81a] R.H. Thomason. Deontic logic and the role of freedom in moral deliberation. In R. Hilpinen, editor, *New Studies in Deontic Logic*, page 177–186. Reidel, 1981.

[Tho81b] R.H. Thomason. Deontic logic as founded on tense. In R. Hilpinen, editor, *New Studies in Deontic Logic: Norms, Actions and the Foundations of Ethics*. Reidel, 1981.

[Tho85] J.J. Thomson. The trolley problem. *The Yale Law Journal*, 1985.

[vBGL10] J. van Benthem, D. Grossi, and F. Liu. Deontics = betterness + priority. In G. Sartori G. Governatori, editor, *Deontic Logic in Computer Science, 10th International Conference, DEON 2010*, 2010.

[vBL10] J. van Benthem and F. Liu. Deontic logic and changing preferences. Technical report, ILLC, 2010.

[vBLG14] J. van Benthem, F. Liu, and D. Grossi. Priority structures in deontic logic. *Theoria*, 80(2):116–152, 2014.

[vdTH08] L. van der Torre and J. Hansen. Deontic logic in computer science. In *ESSLLI*, 2008. [Retrieved on 22-04-2015].

[vdTT98] L.W.N. van der Torre and Y-H. Tan. The temporal analysis of chisholm's paradox. In *AAAI-98 Proceedings*, 1998.

[vdTT99] L.W.N. van der Torre and Y-H. Tan. Contrary-to-duty reasoning with preference-based dyadic obligations. *Annals of Mathematics and Artificial Intelligence*, 1999.

[vK86] F. von Kutschera. Bewirken. *Erkenntnis*, 24:253–281, 1986.

[vR14] M. van Roojen. Moral cognitivism vs. non-cognitivism. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2014 edition, 2014.

[vW51] G.H. von Wright. Deontic logic. *Mind*, 1951.

[vW56] G.H. von Wright. A note on deontic logic and derived obligation. *Mind*, page 507–509, 1956.

[vW63] G.H. von Wright. *Norm and Action: a Logical Enquiry*. Routledge and Kegan Paul, 1963.

[Wan01] H. Wansing. Obligations, authorities and history dependence. In H. Wansing, editor, *Essays on Non-classical Logic*, pages 247–258. World Scientific, 2001.

[Wan04] H. Wansing. On the negation of action types: Constructive concurrent pdl. *Dresden Preprints in theoretical philosophy and philosophical logic*, 2004.

[Wik15a] Wikipedia. De broglie–bohm theory, 2015. [Retrieved on 25/05/2015].

[Wik15b] Wikipedia. Indeterminism, 2015. [Retrieved on 27/03/2015].

[Wik15c] Wikipedia. Trolley problem, 2015. [Retrieved on 20/02/2015].

[WM91] R.J. Wieringa and J-J.Ch. Meyer. Applications of deontic logic in computer science: A concise overview. Technical report, Computer Science Department, Free University, Amsterdam, The Netherlands, 1991. [Retrieved from http://eprints.eemcs.utwente.nl/ on date 20/02/2015].

[Xu15] M. Xu. Combinations of stit with know and ought. *Journal of Philosophical Logic*, 2015.

[Yam08] T. Yamada. Logical dynamics of some speech acts that affect obligations and preferences. *Synthese*, 165(2):295–315, 2008.

[Åq67] L. Åqvist. Good samaritans, contrary-to-duty imperatives, and epistemic obligations. *Nous*, 1967.