Alessio Lomuscio
Donald Nute (Eds.)

LNAI 3065

Deontic Logic in Computer Science

7th International Workshop on Deontic Logic in Computer Science, DEON 2004 Madeira, Portugal, May 2004, Proceedings

Springer

Lecture Notes in Artificial Intelligence 3065

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Springer
Berlin
Heidelberg
New York
Hong Kong
London
Milan
Paris
Tokyo

Alessio Lomuscio Donald Nute (Eds.)

Deontic Logic in Computer Science

7th International Workshop on
Deontic Logic in Computer Science, DEON 2004
Madeira, Portugal, May 26-28, 2004
Proceedings

Springer

eBook ISBN: 3-540-25927-9
Print ISBN: 3-540-22111-5

Created in the United States of America

Visit Springer's eBookstore at: http://ebooks.springerlink.com
and the Springer Global Website Online at: http://www.springeronline.com

Preface

This volume contains the workshop proceedings of DEON 2004, the Seventh International Workshop on Deontic Logic in Computer Science. The DEON workshop series aims at bringing together researchers interested in topics related to the use of deontic logic in computer science. It traditionally promotes research in the relationship between normative concepts and computer science, artificial intelligence, organisation theory, and law. In addition to these topics, DEON 2004 placed special emphasis on the relationship between deontic logic and multi-agent systems.

The workshop was held in Madeira, Portugal, on 26–28 May 2004. This volume includes all 15 papers presented at the workshop, as well as two abstracts from the two outstanding invited speakers we were privileged to host: Prof Mark Brown (Syracuse University, USA), and Prof Mike Wooldridge (University of Liverpool, UK). The reader will find that the topics covered span from theoretical investigations on deontic concepts and their formalisation in logic, to the use of deontic formalisms to verify and reason about multi-agent systems applications. We believe this makes it a well-balanced and interesting volume.

We wish to thank all those who contributed to this workshop, and especially the authors of the submitted papers and the referees. They were all forced to work on a very tight timescale to make this volume a reality.

April 2004

Alessio Lomuscio
Donald Nute

This page intentionally left blank

Workshop Organization

Organization Committee

Program Co-chairs:	Alessio Lomuscio, King's College London, UK
	Donald Nute, Syracuse University, USA
Local Organization Chair:	José Carmo, University of Madeira, Portugal
Local Organization Committee:	Eduardo Fermé, University of Madeira, Portugal
	Filipe Santos ICSCTE, Portugal

Program Committee Members

Laurence Cholvy	ONERA Toulouse
Marco Colombetti	Politecnico di Milano
Frederic Cuppens	ENST-Bretagne in Rennes
Robert Demolombe	ONERA Toulouse
Frank Dignum	University of Utrecht
Lou Goble	Willamette University
Guido Governatori	University of Queensland
Sven Ove Hansson	KTH Stockholm
John Horty	University of Maryland
Andrew Jones	King's College London
Steve Kimbrough	University of Pennsylvania
Lars Lindahl	University of Lund
Alessio Lomuscio	King's College London
Tom Maibaum	King's College London
Paul McNamara	University of New Hampshire
David Makinson	King's College London
Ron van der Meyden	University of New South Wales
John-Jules Meyer	University of Utrecht
Donald Nute	Syracuse University
Jeremy Pitt	Imperial College London
Henry Prakken	University of Utrecht
Giovanni Sartor	University of Bologna
Marek Sergot	Imperial College London
Leon van der Torre	CWI Amsterdam
Lennart Åqvist	Uppsala University

Sponsoring Institutions

Programa FACC (Fundo de apoio à Comunidade Científica)- FCT - Fundacao para a Ciencia e Tecnologia (Ministerio da Ciencia e Ensino Superior)
FLAD - Fundacao Luso-Americana para o Desenvolvimento
CITMA - Centro de Ciencia e Tecnologia da Madeira
AgentLink
Journal of Applied Logic

Table of Contents

Proceedings of DEON 2004

Abstracts of Invited Papers

Contributed Papers

Obligation, Contracts, and Negotiation

Mark A. Brown

Philosophy Department
Syracuse University
Syracuse, NY 13210, USA
mabrown@syr.edu

Many obligations can be seen as arising from contractual arrangements (or from situations resembling contractual arrangements) among agents. My obligation to repay you the $100 I borrowed is associated with a simple (quite possibly tacit and informal) contractual arrangement between us. My obligations as an employee of my university are associated with contractual arrangements with my university, which may be considered a collective agent. My university in turn has certain obligations to me. But some obligations change over time as a result of changing circumstances, and in at least some cases the changes that occur can be thought of as involving a renegotiation of a contract among the parties involved. When I pay back half the money I owe you, I have not fulfilled my original obligation; but neither does that original obligation to pay you $100 still stand. Instead, we may consider, we have renegotiated my contract with you so that my remaining obligation is to pay you $50 (or, depending on details of the negotiation, perhaps $50 plus interest or a late fee). Analogous, though usually more explicit, renegotiations of contracts are commonplace in the corporate world as well.

As we examine this way of looking at normative situations, we find a number of complications which must be considered, many of which we are accustomed to set aside in simpler treatments of deontic logic. We must consider the relationships among distinct agents, not just consider the normative positions of one agent at a time. We need to make room for corporate agents, i.e. agents which are organizations or groups of other agents. We need to consider that a single agent may be involved in multiple contractual arrangements, and thus may have a number of different normative roles simultaneously. As a result, we must make room for conflicting obligations. And we must allow for various kinds of modifications of contractual arrangements over time, including negotiation and renegotiation. Moreover, ultimately we must consider ways in which complex organizations are related to their changing roster of participant agents, whose roles within the organization alter over time.

In this paper, I will discuss a number of the issues which arise in any attempt to formalize a contractual model of our changing normative situations.

A. Lomuscio and D. Nute (Eds.): DEON 2004, LNAI 3065, p. 1, 2004.

Social Laws in Alternating Time

Michael Wooldridge

University of Liverpool

Since it was first proposed by Moses, Shoham, and Tennenholtz, the *social laws* paradigm has proved to be one of the most compelling approaches to the offline coordination of multiagent systems. In this paper, we make three key contributions to the theory and practice of social laws in multiagent systems. First, we show that the *Alternating-time Temporal Logic* of Alur, Henzinger, and Kupferman provides an elegant and powerful framework within which to express and understand social laws for multiagent systems. Second, we show that the *effectiveness, feasibility,* and *synthesis* problems for social laws may naturally be framed as ATL model checking problems, and that as a consequence, existing ATL model checkers may be applied to these problems. We illustrate the concepts and techniques developed by means of a running example.

(joint with Wiebe van der Hoek and Mark Roberts)

A. Lomuscio and D. Nute (Eds.): DEON 2004, LNAI 3065, p. 2, 2004.

Combinations of Tense and Deontic Modality

Lennart Åqvist

Department of Law, Uppsala University
P.O.Box 512, S-751 20 Uppsala, Sweden
lennart.aqvist@jur.uu.se

Abstract. We consider three infinite hierarchies of what I call "two-dimensional temporal logics with explicit realization operators", viz. (i) one without historical or deontic modalities, (ii) one with historical but without deontic modalities, and (iii) one with historical and with dyadic deontic modalities for conditional obligation and permission. Sound and complete axiomatizations are obtained for all three hierarchies relative to a simplified version of the finite co-ordinate-system semantics given for so-called T × W logic of historical necessity in Åqvist (1999).

Keywords: temporal realization operators, historical necessity, conditional obligation, finite two-dimensional co-ordinate system, frame constants.

1 Introduction

The purpose of this paper is to investigate some crucial properties of an infinite hierarchy of logics *combining* (i) a logic for the temporal realization operator $\boldsymbol{R}_t$ ["it is realized (true) at time *t* that"; see Rescher (1966), Rescher and Urquhart (1971)] *with* (ii) a modal logic for a temporally dependent necessity-modality, viz. "historical necessity" or "inevitability" [Åqvist (1999)], *and with* (iii) a dyadic deontic logic for conditional obligation [Åqvist (1997, 2000)].

In order to provide some necessary background to our present enterprise, let us briefly consider the recent contribution Carmo and Jones (2002): Section 7.1, where the authors make a number of useful observations concerning so-called *temporal approaches to the semantics of deontic notions* (like those of obligation and permission). The most important of these observations are, in my opinion, the following:

(I) The temporal approaches at issue are generally based on *tree-structures* representing branching time with the same past and open to the future.

(II) On top of these tree-structures, temporal deontic logics typically define one modal *necessity* operator, expressing some kind of *inevitability* or *historical* necessity, plus deontic *obligation* operators of either a monadic or dyadic kind (where the latter are to reflect notions of *conditional* obligation).

(III) A main difference appears in the way the temporal dimension is syntactically reflected in the formal language of the logics considered. One family of those logics *indexes* the modal and deontic operators with temporal *terms,* whereas another family introduces temporal *operators* that can be *iterated* and *combined with* the modal and deontic operators.

A. Lomuscio and D. Nute (Eds.): DEON 2004, LNAI 3065, pp. 3–28, 2004.

(IV) Leaving the "temporal-operator" family aside for the time being, we emphasize that a characteristic feature of the "indexed" temporal deontic logics is the presence in them of *time-indexed* modal and deontic operators. Carmo and Jones (2002) point out that the time-index could be "separated" from the modal / deontic operators so as to yield a uniform semantical and logical setting for analyzing the modal / deontic component of both types of temporal deontic logics, mentioned in (III) above. This, they say, can be achieved by means of the temporal realization operator R_t ["it is realized (true) at time *t* that"] of Rescher and Urquhart (1971). Let us add here that this means that, instead of writing, like van Eck (1981), Loewer and Belzer (1983), and many others,

N_tA for "it is necessary at time *t* that *A*",

O_tA for "it is obligatory at time *t* that A", and

p_t for "*p*-at-time-*t*"

we are to write, following Bailhache (1991, 1993) and myself in Åqvist (2002),

R_tNA,

R_tOA, and

R_tp

in order to express the corresponding notions, where the "separation" just spoken of is made perfectly clear and explicit.

In view of the above observations, the following problem naturally presents itself: What is the logic of the operators R_t, *N,* and *O,* considered (i) separately[1], and (ii) in combination with one another? As for the logic of the modal operator *N* of historical necessity (considered separately), the reader is referred to my earlier study Åqvist (1999) and, as for the logic of the dyadic deontic operator *O* (again considered separately), to my previous studies Åqvist (1997) and Åqvist (2000). As far as the informal, philosophical motivation for our concentrating on precisely the logics developed in those studies is concerned, we refer the reader to the introductions to those papers, where most relevant additional information can be found. However, some main points made in earlier work of mine deserve to be rehearsed here; this will be done at the end of the present introduction.

Two main novelties of the present paper are as follows.

(A) Inasmuch as we deal with the problem of *combining* the logic of R_t with that of *N,* we must be aware that the latter is represented as a special form of general two-dimensional modal logics in the sense of Segerberg (1973). *Two-dimensionality* means in the approach of Åqvist (1999) that, in the semantics for *N,* we work with frames considered as (finite) two-dimensional co-ordinate systems, where it is possible to distinguish between the *longitude* (i.e. *x*-value) and the *latitude* (i.e. *y*-value) of any point in such a co-ordinate system. Again, on that approach to the semantics[2]

[1] One should observe here that considering those logics "separately" does not preclude our basic two-dimensional temporal logic from containing *other* modal operators of great interest in their own right. See e.g. Section 2 below *in fine,* category (vi).

[2] We may notice here that our present semantics for *N* is simplified as compared to the one proposed in Åqvist (1999), where I worked simultaneously with two different semantical frameworks, a "relational" one and a "non-relational" one. As was correctly pointed out by the *J*PL-referee of that paper, this complication is unnecessary and can be overcome. However, the simplification achieved here is different from the one suggested by him / her.

for *N, times* were interpreted as longitudes, and *worlds,* or *histories,* as latitudes in such systems.

Now, facing the problem of combining the logic of *N* with a connective expressing temporal realization, we immediately see that the Rescher operator R_t, read as "it is realized at time *t* that", is *not* the one that comes most naturally to mind, because it is neither discriminative nor general enough. A more plausible and natural candidate for being combined with our *N* of historical necessity (inevitability) would instead seem to be R_{th}, read as "it is realized at time *t* in history *h* that". However, as will be seen in Sections 3-4 below, Rescher's R_t is readily definable[3] in terms of R_{th}, using the technique of so-called *systematic frame constants,* which was a characteristic feature of the Åqvist (1999) approach to the logic of historical necessity.

Upshot: the above combination problem will in this paper be re-formulated as one of combining the logic of the more general operator R_{th} with that of *N* and that of *O*.

(B) Let us next consider the question how to combine our logic of *N* with the one for the dyadic deontic operator *O,* proposed in Åqvist (1997, 2000), where we encounter a new application of the technique of systematic frame constants: they are, in the present deontic context, taken to represent different "levels of perfection" (as explained in those papers). It turns out that this problem admits of a fairly simple solution: (i) co-ordinate the various levels of perfection (denoted by our new frame constants) with the latitudes (representing worlds, or histories) in the semantics for *N,* and then (ii) re-interpret the modal operators for universal necessity and universal possibility used in Åqvist (1997, 2000) precisely as *historical* necessity and *historical* possibility in the sense of our present logic of *N* and *M*.

Having premised these observations, we can now outline the plan of this paper as follows.

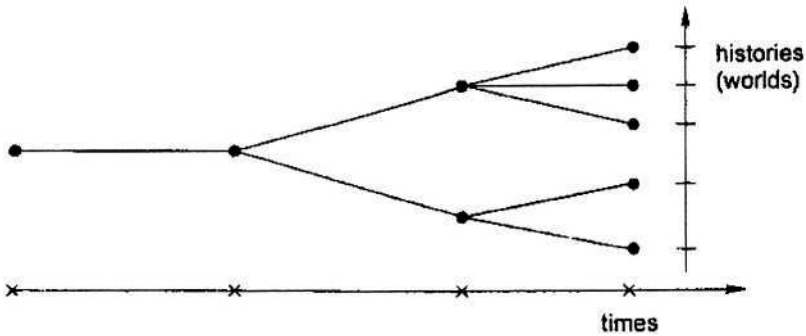
In Sections 2 – 4 we present the syntax, semantics and proof theory for an infinite hierarchy R*xy* (with *x, y* any natural numbers in ω) of two-dimensional temporal logics with explicit realization operators R_t and R_{th} *without* historical or deontic modalities. Section 5 establishes two fundamental results on so-called canonical R*xy*-structures (the proofs of which are given in the Appendix of an as yet unpublished paper, Åqvist (2004)); together they yield strong as well as weak completeness of the logics (axiomatic systems) R*xy* by means of the more or less standard argument given at the end of Section 5. Again, Sections 6 – 8 are devoted to the study of a new hierarchy HR*xy* ($x, y \in \omega$) of logics *with* the historical modalities *N* and *M* added to the vocabulary of the R*xy*, but still *without* deontic modalities: the semantics and proof theory of the systems R*xy* are extended to the HR*xy*, for which we obtain extended completeness results (main proofs being again relegated to the Appendix of Åqvist (2004)). Finally, in Sections 9 – 11, we achieve a desired extension of our R_t/R_{th} logics HR*xy* of historical necessity to a third infinite hierarchy DHR*xym* ($x, y, m \in \omega$) of dyadic deontic logics of conditional obligation and permission, for which similiar results are obtained in the same spirit.

---

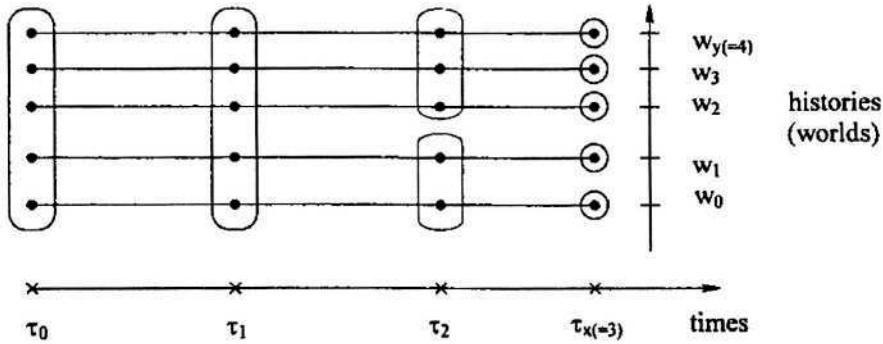
[3] See e.g. axiom schema A3(a) in Section 4 *infra.* Again, from axiom schemata A4(d)-(f) in the same section it appears that even the three operators N^{lg}, N^{lat} and $\Box$ (as well as their duals), introduced as primitives in Section 2 *infra,* are definable in terms of R_{th} and the frame constants a_i and b_j. This is easily verified by the reader.

We close this introduction by rehearsing some main points in earlier work of mine, the most important of which are as follows.

(i) In Åqvist (1999) we gave *two* types of semantics for our proposed infinite hierarchy of logics of historical necessity and illustrated how they differ in their treatment of that notion by means of two contrasting diagrams. According to the first type, we interpret sentences of the forms *NA* and *MA* (my present notation) relative to *tree*-structures like



by telling under what conditions such a sentence is *true at* a time *in* a history. On this approach, truth is thus relative to *two* indices. According to the second type of semantics, we proceed as follows: transform (convert) any tree-structure like the one shown above into a *rectangular grid*



where the encircled colonies of points represent equivalence classes under a certain equivalence relation on the 'longitudes' (heuristically *times*). The truth conditions for sentences of the forms *NA* and *MA* (my present notation) are then given in terms of this equivalence relation in the standard manner, i.e. truth is relative to just *one* index on this approach, viz. any point in the rectangular space.

We note that the two types of structure relative to which we interpret sentences of the forms *NA* and *MA* are to a certain extent analogous to the so-called T × W frames and *separated*

T × W frames in von Kutschera (1997). It turned out that our respective completeness proofs were facilitated by his use of *separated* T ×W frames and my use of *rec-*

tangular grids instead of the rival structures just mentioned, i.e. T × W frames and unconverted / unseparated trees.

(ii) An important difference between the Åqvist (1999) approach and the one advocated by von Kutschera (1997) is due to the fact that, on the former, we assumed time to be *discrete* and *finite* in the sense of having a beginning and an end. The main motivation for thus limiting our framework to a discrete and finite one goes back to my work on Causation in the Law together with Philip Mullock in our joint book Åqvist and Mullock (1989), which was judiciously reviewed by von Kutschera in the review von Kutschera (1996), dealing primarily with the philosophically relevant aspects of our enterprise. In Åqvist and Mullock (1989) we developed a detailed theory of causation by agents and the representation of causal issues in Tort and Criminal Law, which is based on a version of Games and Game Theory in Extensive Form and where legal cases (Anglo-American, Swedish, German) are examined and graphically represented by means of game trees. When embarking on our project we felt that using a discrete and finite framework was the natural starting point, mainly because of what we took to be *the fundamentally finitistic nature of legal reasoning*. In his (1996) review von Kutschera points out that this, of course, amounts to a partly serious limitation of our model.

However, he agrees with us on the need for a theory of liability and causation in the law that is, in the first place, intuitively clear in the sense of being sufficiently simple to apply. On the other hand he emphasizes that, since no simple model fits all the complex cases of human life, one has to find a compromise between simplicity and scope of applicability, whence he suggests that further refinements of our model be made later on according as needs for such refinements arise.

(iii) An interesting point intimately bound up with the foregoing one is this. The discrete and finite framework used in Åqvist and Mullock (1989), and later in Åqvist (2002a), turns out to be fruitful in enabling us to explicate and represent formally the useful distinction between (i) *basic action-sentences* asserting that such and such an act is performed / omitted by an agent, and (ii) *causative action-sentences* asserting that *by* performing / omitting a certain act, an agent *causes that* such and such a state-of-affairs is realized (e.g. comes about / ceases / remains / remains absent, - see von Wright (1983), p. 173 f.).As appears from Åqvist (2002a), the *discreteness* property of our framework is then seen to play an important role in the present context.

(iv) Again, the temporal setting of Åqvist and Hoepelman (1981) happened to be infinite – in the sense of requiring a denumerably infinite number of times and admitting a denumerably infinite number of histories, which were all taken to be of infinite length. However, its treatment of the Chisholm Contrary-to-Duty Paradox – a key problem in deontic logic – would, I suggest, be best understood in a finite framework like the one proposed here and in Åqvist (1999). Note also that Åqvist and Hoepelman (1981) represents an early attempt to combine temporal-logic-with-historical-necessity precisely with *dyadic* deontic logic of the sort studied in Åqvist (1997, 2000).

(v) The problem of combining the discrete and finite approach adopted in Åqvist (1999) and the present paper with the discrete and infinite one used in Åqvist and Hoepelman (1981) remains open.

2 Syntax of the Systems R*xy* of Two-Dimensional Temporal Logic with Explicit Realization Operators R_t ["It Is Realized at Time *t* That"] and R_{th} ["It Is Realized at Time *t* in History *h* That"]

The *vocabulary* (morphology, alphabet, language) of the systems R*xy* ($x, y \in \omega$) is a structure made up of the following disjoint basic syntactic categories:

(i) An at most denumerable set Prop of *propositional variables.*

(ii) *Prepositional constants,* subdivided into

(a) traditional: $\top$(*verum*) and $\bot$(*falsum*), and

(b) 'new': two families of *systematic frame constants,* viz.
$\{a_i\}_{i\in\omega}$, indicating positions on the *x*-axis ['longitudes'];
$\{b_j\}_{j\in\omega}$, indicating positions on the *y*-axis ['latitudes'].

(iii) A set $\mathbf{NT} = \{t_i\}_{i\in\omega}$ of *names of times (temporal names)* as well as a set $\mathbf{NH} = \{h_j\}_{j\in\omega}$ of *names of histories* (*'worlds'*).

(iv) The *Boolean sentential connectives* $\neg, \wedge, \vee, \rightarrow, \leftrightarrow$ with their usual readings.

(v) An indexed family $\{Rt_i\}_{i\in\omega}$ of one-place *temporal realization* operators, where Rt_i is read as "it is realized at time t_i that", as well as a doubly indexed family $\{Rt_ih_j\}_{i,j\in\omega}$ of one-place *time-history realization* operators, with Rt_ih_j read as "it is realized at time t_i in history (world) h_j that".

(vi) Three pairs (N^{lat}, M^{lat}), (N^{lg}, M^{lg}) and $(\Box,\Diamond)$ of one-place *modal operators* in two-dimensional temporal logic, the readings of which will be considered in a moment.

For any natural numbers *x, y,* we then define recursively the set Sent of well formed *sentences* of R*xy* in the straightforward manner, i.e. in such a way that all propositional variables and constants will be (atomic) sentences; moreover, Sent will be closed under every connective in the categories (iv) – (vi) *supra.* In particular, as far as the category (v) is concerned, we stipulate that if $t_i \in \mathbf{NT}$ and $A \in \mathbf{Sent}$, then $Rt_iA \in \mathbf{Sent}$; and that if in addition $h_j \in \mathbf{NH}$, then $Rt_ih_jA \in \mathbf{Sent}$ (this being the only non-standard clause in our recursive definition of Sent).

As to the readings of the one-place modalities in the category (vi), we tentatively invoke the following spatial metaphors:

$N^{lat}A$ – everywhere on this latitude, *A*
$M^{lat}A$ – somewhere on this latitude, *A*
$N^{lg}A$ – everywhere on this longitude, *A*
$M^{lg}A$ – somewhere on this longitude, *A*
$\Box A$ – everywhere, *A*
$\Diamond A$ – somewhere, A.

3 Semantics for R*xy:* Frames, Models and Truth Conditions

Rxy-Frames. For any pair of natural numbers (*x, y*), we mean by a R*xy-frame* an ordered quadruple

$$F = (U, (u_0, e_0, n_0), (\{\tau_i\}_{i \in \omega}, x), (\{w_j\}_{j \in \omega}, y))$$

where:

(i) $U \neq \emptyset$ [U is a non-empty, finite set of *points* in time].

(ii) u_0, e_0, n_0 are designated members of U [heuristically, u_0 is the *origin* in U, e_0 is the *eastern limit* of U, and n_0 is the *northern limit* of U].

(iii) $\{\tau_i\}_{i \in \omega}$ is an infinite sequence of subsets of U [heuristically *times*] and x is the first natural number under consideration.

Let $T = \{\tau_0, \tau_1, ..., \tau_x\}$. We then require the set T to be a *partition* of U in the familiar sense that

(iii:a) $\tau_0 \cup ... \cup \tau_x = U$.

(iii:b) For all $i, j \in \omega$ with $0 \leq i \neq j \leq x$: $\tau_i \cap \tau_j = \emptyset$.

(iii:c) For each $i \in \omega$ with $0 \leq i \leq x$: $\tau_i \neq \emptyset$.

(iii:d) For each $i \in \omega$ with $x < i$: $\tau_i = \emptyset$.

(iv) $\{w_j\}_{j \in \omega}$ is an infinite sequence of subsets of U [heuristically *histories, worlds*] and y is the second natural number under consideration.

Let $W = \{w_0, w_1, ..., w_y\}$. We require the set W to be a *partition* of U in the sense that

(iv:a) $w_0 \cup ... \cup w_y = U$.

(iv:b) For all $i, j \in \omega$ with $0 \leq i \neq j \leq y$: $w_i \cap w_j = \emptyset$.

(iv:c) For each $j \in \omega$ with $0 \leq j \leq y$: $w_j \neq \emptyset$.

(iv:d) For each $j \in \omega$ with $y < j$: $w_j = \emptyset$.

Furthermore, we require any R*xy*-frame to satisfy the following additional conditions:

(v) For each τ in T and each w in W there is exactly one u in U such that $\{u\} = \tau \cap w$.

(vi) $\{u_0\} = \tau_0 \cap w_0$, $\{e_0\} = \tau_x \cap w_0$, and $\{n_0\} = \tau_0 \cap w_y$.

Rxy-Models and Truth Conditions. Let $F = (U, (u_0, e_0, n_0), (\{\tau_i\}_{i \in \omega}, x), (\{w_j\}_{j \in \omega}, y))$ be any R*xy*-frame. By a *valuation on* such a frame we mean any function V which to each propositional variable p in Prop assigns a subset $V(p)$ of $T \times W$, i.e. a certain set of ordered pairs (τ, w) with $\tau \in T$ and $w \in W$.

By a R*xy-model* we then mean an ordered triple $M = (F, V, v)$ the first term of which is a R*xy*-frame, the second a valuation on that frame, and where v is the function defined on $\mathbf{NT} \cup \mathbf{NH}$ such that, for all $i \in \omega$:

$$v(t_i) = \begin{cases} \tau_i, & \text{if } 0 \leq i \leq x \\ \emptyset, & \text{otherwise, i.e. if } x < i \end{cases}$$

and such that, for all $j \in \omega$:

$$v(h_j) = \begin{cases} w_j, & \text{if } 0 \leq j \leq y \\ \emptyset, & \text{otherwise, i.e if } y < j \end{cases}$$

Let $\boldsymbol{M} = ((U, (u_0, e_0, n_0), (\{\tau_i\}_{i\in\omega}, x), (\{w_j\}_{j\in\omega}, y)), \boldsymbol{V}, \boldsymbol{v})$ be a R*xy*-model. We can now tell what it means for any sentence A to be *true at* a time $\tau \in T = \{\tau_0, \tau_1,...,\tau_x\}$ *in* a world $w \in W = \{w_0, w_1,...,w_y\}$ in $\boldsymbol{M}$ (in symbols: $\boldsymbol{M}, \tau, w \models A$) by the following recursion:

$\boldsymbol{M}, \tau, w \models p$ iff $(\tau, w) \in \boldsymbol{V}(p)$, for any p in Prop

$\boldsymbol{M}, \tau, w \models \top$

not: $\boldsymbol{M}, \tau, w \models \bot$

$\boldsymbol{M}, \tau, w \models a_i\ (i\in\omega)$ iff { $\tau = \tau_i$, if $0 \le i \le x$

$\tau \neq \tau$, otherwise, i.e. if $x < i$

$\boldsymbol{M}, \tau, w \models b_j\ (j\in\omega)$ iff { $w = w_j$, if $0 \le j \le y$

$w \neq w$, otherwise, i.e. if $y < j$

If A is a Boolean compound, the recursive definition goes on as usual. We then handle sentences having the characteristic one-place R*xy*-connectives as their main operator as follows:

$\boldsymbol{M}, \tau_i\ (= \boldsymbol{v}(t_i)), w_j\ (= \boldsymbol{v}(h_j)) \models B$, if $0 \le i \le x$ and $0 \le j \le y$

$\boldsymbol{M}, \tau, w \models Rt_ih_jB$ iff {

for all (τ', w') in $T \times W$ with $\tau' \neq \tau'$ and $w' \neq w'$:
$\boldsymbol{M}, \tau', w' \models B$, otherwise; i.e. if $x < i$ or $y < j$

$\boldsymbol{M}, \tau_i\ (= \boldsymbol{v}(t_i)), w \models B$, if $0 \le i \le x$

$\boldsymbol{M}, \tau, w \models Rt_iB$ iff {

for all τ' in T with $\tau' \neq \tau'$: $\boldsymbol{M}, \tau', w \models B$,
otherwise; i.e. if $x < i$

$\boldsymbol{M}, \tau, w \models N^{lat}B$ iff for all τ' in T: $\boldsymbol{M}, \tau', w \models B$

$\boldsymbol{M}, \tau, w \models N^{lg}B$ iff for all w' in W: $\boldsymbol{M}, \tau, w' \models B$

$\boldsymbol{M}, \tau, w \models \Box B$ iff for all τ' in T and all w' in W: $\boldsymbol{M}, \tau', w' \models B$.

The truth conditions for sentences having the possibility operators M^{lat}, M^{lg} and $\Diamond$ as their main connective are obtained in the dual way: just replace 'all' by 'some' to the right of the 'iff' in the last three conditions!

As usual, then, we say that sentence A is R*xy-valid* iff $\boldsymbol{M}, \tau, w \models A$ for all R*xy*-models $\boldsymbol{M}$, all τ in T and all w in W. And we say that a set Γ of sentences is R*xy-satisfiable* iff there exists a R*xy*-model $\boldsymbol{M}$ with members τ of T and w of W such that for all sentences A in Γ: $\boldsymbol{M}, \tau, w \models A$.

Remark. Consider the truth condition for sentences of the form Rt_iB, the case where $0 \le i \le x$. It looks simple and straightforward enough, but the impression of simplicity

is really deceptive, because the condition harbours a hidden complexity that can be spelled out as follows. Recall that w is to be a member of $W = \{w_0, w_1,...,w_y\}$ throughout the above truth definition. For the case we are considering this means that the truth condition for Rt_iB is to be split up into the following series of *conditional* truth conditions, whenever $0 \leq i \leq x$:

If $w = w_0$, then $M, \tau, w \models Rt_iB$ iff $M, \tau_i, w_0 \models B$ [iff $M, \tau, w \models Rt_ih_0B$]
If $w = w_1$, then $M, \tau, w \models Rt_iB$ iff $M, \tau_i, w_1 \models B$ [iff $M, \tau, w \models Rt_ih_1B$]
.
If $w = w_j$, then $M, \tau, w \models Rt_iB$ iff $M, \tau_i, w_j \models B$ [iff $M, \tau, w \models Rt_ih_jB$]
.
.
If $w = w_y$, then $M, \tau, w \models Rt_iB$ iff $M, \tau_i, w_y \models B$ [iff $M, \tau, w \models Rt_ih_yB$].

We see that the bracketed 'iff'-clauses are immediate by our truth condition for Rt_ih_jB, the case where $0 \leq i \leq x$ and $0 \leq j \leq y$. A further observation is now to the effect that the above series, or conjunction, of conditional truth conditions is in fact equivalent to the following *categorical* truth condition for Rt_iB, where the right member has the form of a *disjunction*:

$M, \tau, w \models Rt_iB$ iff either ($w = w_0$ and $M, \tau, w \models Rt_ih_0B$)
or ($w = w_1$ and $M, \tau, w \models Rt_ih_1B$)
or .....
or ($w = w_j$ and $M, \tau, w \models Rt_ih_jB$)
or .....
or ($w = w_y$ and $M, \tau, w \models Rt_ih_yB$).

Again, this categorical truth condition can be written more compactly as:

$$M, \tau, w \models Rt_iB \text{ iff } \vee_{0 \leq j \leq y}(w = w_j \text{ and } M, \tau, w \models Rt_ih_jB)$$

where the initial prefix in the right member of the equivalence may be replaced by an existential quantifier 'for some natural number j with $0 \leq j \leq y$', since we are dealing with a finite set.

The complexity just pointed out will reappear in the form of the "definitional" axiom schema A3(a) in Section 4 as well as in the proof of the so called Coincidence Lemma 5.3 in the Appendix *infra.* See also our discussion of the matter in the Introduction *supra.*

4 On the Proof Theory of Two-Dimensional Temporal Logic with the R_t and R_{th} Operators: The Axiomatic Systems Rxy

The infinite hierarchy of axiomatic systems Rxy ($x, y \in \omega$) is determined by one rule of inference (deduction), one rule of proof, and five groups of axiom schemata. They are as follows, where the letters 'i' and 'j' range throughout over the set ω of all the natural numbers, and where a notation like '$\vee_{0 \leq j \leq y}Rt_ih_jA$' ['$\wedge_{0 \leq i \leq x}Rt_ih_j$'] abbreviates

the finite disjunction '$Rt_ih_0A \vee Rt_ih_1A \vee ... \vee Rt_ih_y$' [the finite conjunction '$Rt_0h_jA \wedge Rt_1h_jA \wedge ... \wedge Rt_xh_jA$'], with j running from 0 to y [i from 0 to x]:

Rule of Inference

mp (modus ponens)

$$\frac{A,\ A \rightarrow B}{B}$$

Rule of Proof

Nec (necessitation for $\Box$)

$$\frac{\vdash A}{\vdash \Box A}$$

Axiom Schemata

A0. All truth-functional tautologies over Sent

A1. (a) $a_i \wedge b_j \wedge A \rightarrow Rt_ih_jA$, for all i, j such that $0 \leq i \leq x$ and $0 \leq j \leq y$

(b) $Rt_ih_j(a_i \wedge b_j)$, for all i, j such that $0 \leq i \leq x$ and $0 \leq j \leq y$

(c) $Rt_ih_j\bot$, if $x < i$ or $y < j$

(d) $a_0 \vee a_1 \vee ... \vee a_x$

(e) $b_0 \vee b_1 \vee ... \vee b_y$

(f) $a_i \rightarrow \neg a_j$, if $i \neq j$

(g) $b_i \rightarrow \neg b_j$, if $i \neq j$

(h) $\neg a_i$, if $x < i$

(k) $\neg b_j$, if $y < j$

(m) $a_i \rightarrow N^{lg}a_i$ (all $i \in \omega$)

(n) $b_j \rightarrow N^{lat}b_j$ (all $j \in \omega$).

A2. (a) $A \rightarrow Rt_ih_jA$, for all $i, j \in \omega$

(b) $Rt_ih_j(A \rightarrow B) \rightarrow (Rt_ih_jA \rightarrow Rt_ih_jB)$,

(c) $Rt_ih_jA \rightarrow \Box Rt_ih_jA$

(d) $\neg Rt_ih_j\neg A \rightarrow Rt_ih_jA$

(e) $Rt_ih_jA \rightarrow \neg Rt_ih_j\neg A$, for all i, j with $0 \leq i \leq x$ and $0 \leq j \leq y$

A3. (a) $Rt_iA \leftrightarrow \vee_{0 \leq j \leq y}(b_j \wedge Rt_ih_jA)$, if $0 \leq i \leq x$

(b) $Rt_i\bot$, *if* $x < i$

A4. (a) $A \rightarrow N^{lat}A$; $\Box A \rightarrow N^{lg}A$

(b) $A \leftrightarrow N^{lat}N^{lg}A$; $\Box A \leftrightarrow N^{lg}N^{lat}A$

(c) The modal logic **S5** for each pair of operators $(\Box,\Diamond)$, (N^{lat}, M^{lat}) and (N^{lg}, M^{lg})

(d) $a_i \rightarrow (N^{lg}A \leftrightarrow \wedge_{0 \leq j \leq y}Rt_ih_jA)$, for all $i \in \omega$

(e) $b_j \rightarrow (N^{lat}A \leftrightarrow \wedge_{0 \leq i \leq x}Rt_ih_jA)$, for all $j \in \omega$

(f) $\Box A \leftrightarrow \wedge_{0 \leq i \leq x}\wedge_{0 \leq j \leq y}Rt_ih_jA$

As usual, the above axiom schemata and rules determine syntactic notions of R*xy-provability* and R*xy-deducibility* as follows. We say that a sentence A is R*xy-provable* [in symbols: $\vdash_{Rxy} A$, or just $\vdash A$] iff A belongs to the smallest subset of Sent which (i) contains every instance of A0, A1(a)-(n),...,A4(a)-(f) as its member, and which (ii) is closed under the rule of inference mp and the rule of proof Nec. And we say that the sentence A is R*xy-deducible from* the set Γ ($\subseteq$ Sent) of *assumptions* [in symbols: $\Gamma \vdash_{Rxy} A$] iff there are sentences B_1,..., B_k in Γ, for some natural number $k \geq 0$, such that $\vdash_{Rxy} (B_1 \wedge ... \wedge B_k) \rightarrow A$ (i.e. the sentence $(B_1 \wedge ... \wedge B_k) \rightarrow A$ is to be R*xy*-provable in the sense of the preceding definition).

Again, letting $\Gamma \subseteq$ Sent, we say that Γ is R*xy-inconsistent* iff $\Gamma \vdash_{Rxy} \perp$, and R*xy-consistent* otherwise. Finally, we say that Γ is *maximal* R*xy-consistent* iff Γ is R*xy*-consistent and, for each A in Sent, either $A \in \Gamma$ or $\neg A \in \Gamma$; where this latter condition is known as requiring Γ to be *negation-complete.*

Soundness Theorem 4.1

Weak version: *Every* R*xy-provable sentence is* R*xy-valid.*
Strong version: *Every* R*xy-satisfiable set of sentences is* R*xy-consistent.*

Proof. As usual, we establish the weak version by showing (i) that every instance of the axiom schemata A0, A1(a)-(n),...,A4(a)-(f) is R*xy*-valid, and (ii) that the rules mp and Nec preserve R*xy*-validity. This is tedious, but entirely routine.

As to the strong version, it is easily obtained as a consequence of the weak one.

Again, we rehearse a few obvious results on our infinite hierarchy $\mathbf{Rxy}$ $(x, y \in \omega)$ in the following

Lemma 4.2. (Scott's Rule for R*xy*; Fresh Properties of Maximal R*xy*-Consistent Sets; Lindenbaum's Lemma for R*xy*). *Let* § *be any of the operators* $Rt_i h_j$ $(i, j \in \omega)$, N^{lat}, N^{lg}, $\Box$ *(all of which are 'necessity modalities' in a straightforward sense). Then*:

(I) *Let* Γ *be a set of sentences and let* A *be a sentence. If* $\Gamma \vdash_{Rxy} A$, *then*
$\{ \S B: B \in \Gamma \} \vdash_{Rxy} \S A$.

(II) *Let* Γ *be any maximal* R*xy-consistent set of sentences. Then it holds that:*
(i) $a_i \in \Gamma$, *for exactly one natural number* i *such that* $0 \leq i \leq x$.
(ii) $b_j \in \Gamma$, *for exactly one natural number* j *such that* $0 \leq j \leq y$.

(III) *For any* R*xy-consistent set* Γ *of sentences, define the* Lindenbaum extension Γ_ω *of* Γ *in the appropriate way. Then* Γ_ω is *maximal* R*xy-consistent.*

Proof. As for (I), this is familiar – see e.g. Makinson (1966) or Åqvist (1991), Lemma 6.5. As for (II), we argue as follows. By a well known property of maximal R*xy*-consistent sets, the disjunction A1(d) is in Γ. Hence, by another such property, at least one of its disjuncts must be in Γ. Hence the existence part of clause (i). The uniqueness of 'that' disjunct is then immediate by axiom A1(f) *supra.* The proof of clause (ii) is similar: just appeal to A1(e) and A1(g). Finally, the proof of (III) is familiar as well. ∎

Lemma 4.3 (Useful Properties of the Operators Rt_i, Rt_ih_j and N^{lat}, M^{lat}).

(I) *For all natural numbers x, y it holds that all instances of the following* theorem schemata *are* R*xy-provable*:

T0. $b_j \rightarrow (Rt_iA \leftrightarrow Rt_ih_jA)$ for all $i, j \in \omega$ such that $0 \leq i \leq x$ and $0 \leq j \leq y$

T1. $Rt_i(A \rightarrow B) \rightarrow (Rt_iA \rightarrow Rt_iB)$ for all i with $0 \leq i \leq x$

T2. $Rt_iA \leftrightarrow \neg Rt_i\neg A$

T3. $Rt_iA \leftrightarrow N^{lat}Rt_iA$ $(0 \leq i \leq x)$; $Rt_iA \leftrightarrow M^{lat}Rt_iA$ $(0 \leq i \leq x)$

T4. $b_j \rightarrow (N^{lat}A \rightarrow Rt_ih_jA)$ $(0 \leq i \leq x;\ j \in \omega)$

T5. $b_j \rightarrow (Rt_ih_jA \rightarrow M^{lat}A)$ $(0 \leq i \leq x;\ j \in \omega)$

T6. $b_j \rightarrow (M^{lat}A \leftrightarrow \vee_{0 \leq i \leq x} Rt_ih_jA)$ $(j \in \omega)$

T7. $Rt_iA \leftrightarrow Rt_kRt_iA$ for all $i, k \in \omega$ with $0 \leq i, k \leq x$.

(II) *In spite of the provability/validity in* R*xy of* T0, *there are instances of the schema*

$$Rt_ih_jA \rightarrow Rt_iA$$

which fail to be provable/valid in R*xy. (Similarly for the converse of that schema.)*

Proof. As to (I), the proofs in R*xy* of the theorem schemata T0 – T7 amount to useful exercises in the axiomatics for R*xy* that are left to the reader. As for (II), the task of constructing counterexamples to both directions in **T0-without-the-antecedent-b_j** can be left to the reader as well.

5 Canonical R*xy*-Structures and Semantic Completeness of the Logics R*xy*

Definition 5.1. For any natural numbers $x, y \in \omega$, let Ωxy be the set of all maximal R*xy*-consistent set of sentences. Let q be a fixed element of Ωxy. Furthermore, let ~lat/~lg/ be the binary relation on Ωxy such that for all u, v in Ωxy: u ~lat v /u ~lg v/ iff for each A in Sent, if $N^{lat}A$ /$N^{lg}A$/ $\in u$, then $A \in v$. Again, let ~ be the binary relation on Ωxy such that for all u, v in Ωxy: $u \sim v$ iff for each A in Sent, if $\Box A \in u$, then $A \in v$. Clearly, by the **S5**-properties of the operators N^{lat}, N^{lg} and $\Box$ [axiom schema A4(c) *supra*], the relations ~lat, ~lg, and ~ are equivalence relations on Ωxy.

We now define the *canonical* R*xy-structure generated by q* as the ordered sextuple

$$M^q = (U, (u_0, e_0, n_0), (\{\tau_i\}_{i \in \omega}, x), (\{w_j\}_{j \in \omega}, y), V, v)$$

where:

(i) $U = \{u \in \Omega xy$: for each sentence A, if $\Box A \in q$, then $A \in u\}$, i.e.
$U = \{u \in \Omega xy: q \sim u\} = [q]\sim$ (i.e. the ~-equivalence class of q in Ωxy).

(ii) $u_0 = \{A: Rt_0h_0A \in q\}$
$e_0 = \{A: Rt_xh_0A \in q\}$
$n_0 = \{A: Rt_0h_yA \in q\}$

$$\{u \in \Omega xy\colon \{A\colon Rt_i h_0 A \in q\} \sim\text{lg } u\}, \text{ if } 0 \leq i \leq x$$

(iii) For each $i \in \omega$: $\tau_i = \{$

$$\varnothing, \text{ otherwise, i.e. if } x < i$$

where x is the first natural number under consideration.

$$\{u \in \Omega xy\colon \{A\colon Rt_0 h_j A \in q\} \sim\text{lat } u\}, \text{ if } 0 \leq j \leq y$$

(iv) For each $j \in \omega$: $w_j = \{$

$$\varnothing, \text{ otherwise, i.e. if } y < j$$

where y is the second natural number under consideration.

(v) V = the function such that for all p in Prop: $V(p) = \{(\tau_i, w_j)\colon 0 \leq i \leq x, 0 \leq j \leq y,$ and there is exactly one u in U with $u \in \tau_i \cap w_j$ and $p \in u\}$.

(vi) v = the function on $\mathbf{NT} \cup \mathbf{NH}$ defined as in Section 3 *supra.*

Remark. Assume that $0 \leq i \leq x$ and $0 \leq j \leq y$. By condition (iii), $T = \{\tau_0, \tau_1, \ldots, \tau_x\}$ is identified with a certain set of ~lg-equivalence-classes of members of Ωxy, and, by condition (iv), $W = \{w_0, w_1, \ldots, w_y\}$ is identified with a certain set of ~lat-equivalence-classes of members of Ωxy.

We can now state two basic results concerning generated canonical R*xy*-structures.

Theorem 5.2. *As defined in Definition* 5.1 *above, the initial quadruple in* $\boldsymbol{M^q}$ *is a* R*xy-frame, and hence* $\boldsymbol{M^q}$ *as a whole is a* R*xy-model.*

Proof. See the Appendix of Åqvist (2004). ∎

Coincidence[4] Lemma 5.3. *Let q be any fixed maximal* R*xy-consistent set of sentences, and let* $\boldsymbol{M^q}$ *(as above) be the canonical Rxy-structure generated by q. Then, for each sentence A and each u in U,*

$$\boldsymbol{M^q}, [u]\sim\text{lg}, [u]\sim\text{lat} \models A \text{ iff } A \in u.$$

Here we use the following familiar definitional abbreviations: $[u]\sim\text{lg} = \{v \in U\colon u \sim\text{lg } v\}$ and $[u]\sim\text{lat} = \{v \in U\colon u \sim\text{lat } v\}$. Note also that the first set belongs to T and the second to W. (In order to verify that this is indeed so, just appeal to the fact [Theorem 5.2] that $\boldsymbol{M^q}$ satisfies the partition conditions (iii:a)-(iii:b) /(iv:a)-(iv:b)/ on R*xy*-frames in Section 3, to the definition of $\tau_i / w_j /$ in Definition 5.1 above, and to elementary properties of equivalence relations.)

Proof. By induction on the length of A. For details, see the Appendix of Åqvist (2004). ∎

Completeness Theorem 5.4.

Weak version: *Every* R*xy-valid sentence is* R*xy-provable.*
Strong version: *Every* R*xy-consistent set of sentences is* R*xy-satisfiable.*

---

[4] The word "Coincidence" used in the name of this Lemma is meant to suggest that, as applied to any sentences (of R*xy*), the notions of *truth* and *membership* coincide, or are co-extensive, with respect to the points in generated canonical R*xy*-structures.

Proof. As the weak version is immediate from the strong one, let us concentrate on the latter. Let Γ be any R*xy*-consistent set of sentences. Form the Lindenbaum extension Γ_ω of Γ: by Lemma 4.2 (III), Γ_ω is maximal R*xy*-consistent. Again, form the canonical R*xy*-structure generated by Γ_ω, i.e. the structure $\boldsymbol{M}^{\Gamma\omega}$ as defined in Definition 5.1 *supra*: by the basic Theorem 5.2, then, $\boldsymbol{M}^{\Gamma\omega}$ is a R*xy*-model. By the Coincidence Lemma 5.3 for generated canonical R*xy*-structures, we obtain in particular that for each sentence *A*:

$$\boldsymbol{M}^{\Gamma\omega},\ [\Gamma_\omega]\sim\mathrm{lg},\ [\Gamma_\omega]\sim\mathrm{lat}\ \models A\ \text{ iff }\ A \in \Gamma_\omega$$

since Γ_ω is known to belong to the universe *U* of $\boldsymbol{M}^{\Gamma\omega}$. Hence, since $\Gamma \subseteq \Gamma_\omega$, we have $\boldsymbol{M}^{\Gamma\omega},\ [\Gamma_\omega]\sim\mathrm{lg},\ [\Gamma_\omega]\sim\mathrm{lat}\ \models A$ for every $A \in \Gamma$. In other words, assuming Γ to be any R*xy*-consistent set of sentences, we have constructed a R*xy*-model, viz. $\boldsymbol{M}^{\Gamma\omega}$, with members τ of *T* and *w* of *W,* viz. $[\Gamma_\omega]\sim\mathrm{lg}$ and $[\Gamma_\omega]\sim\mathrm{lat}$, such that for all sentences *A* in Γ: $\boldsymbol{M}^{\Gamma\omega}, \tau, w \models A$. Hence, we have shown Γ to be R*xy*-satisfiable, as desired.

6 Introducing Historical Modalities: The Systems HR*xy* and Their Semantics

Syntax of HRxy. Let us add to the vocabulary of the systems R*xy* a pair of primitive one-place modal operators, *N* and *M,* to be read respectively as

'it is *necessary* on the basis of the *past* and the *present* that', and
'it is *possible* on the basis of the *past* and the *present* that'.

Call the resulting language that of the systems HR*xy* (of two-dimensional temporal logic with the *Rt* and *Rth* operators *and with* historical necessity), where, as usual, *x, y* are any members of the set ω of natural numbers. The definition of the set Sent of well formed *sentences* of HR*xy* is then obvious: Sent will be closed under the two new one-place connectives as well.

Semantics for HRxy. Consider any R*xy*-frame as described in Section 3 *supra.* We lay down a few preliminary definitions. First of all, observe that for each *u* in *U* there is, by conditions (iii:a) and (iii:b), *exactly one* τ in *T* such that $u \in \tau$, as well as, by (iv:a) and (iv:b), *exactly one w* in *W* such that $u \in w$. Define for all *u, v* in *U*:

$\mathrm{lg}(u) = \underline{\text{the}}\ \tau \in T:\ u \in \tau$
$\mathrm{lat}(u) = \underline{\text{the}}\ w \in W:\ u \in w$
$u \sim\mathrm{lg}\ v$ iff $\mathrm{lg}(u) = \mathrm{lg}(v)$
$u \sim\mathrm{lat}\ v$ iff $\mathrm{lat}(u) = \mathrm{lat}(v)$

where the function lg/lat/ is to mean *the longitude /latitude/ of,* and where the binary relation ~lg / ~lat/ means *is on the same longitude /latitude/ as.* (On the present general definitions of the relations ~lg and ~lat – 'general' in the sense of applying to all R*xy*-frames, not just to the initial quadruple in $\boldsymbol{M}^q$ of the preceding section – these two relations are equivalence relations on *U,* so we are still entitled to use the standard notation for equivalence classes, i.e. '[ ]~lg' and '[ ]~lat'.)

Again, since any R*xy*-frame satisfies the condition (v) in Section 3, we can define the following function f from $T \times W$ into U. For each τ in T and each w in W:

$$f(\tau, w) = \underline{\text{the}}\ u \text{ in } U\text{: } \{u\} = \tau \cap w$$

i.e. $f(\tau, w)$ is identical to the sole member of the intersection $\tau \cap w$.

HRxy-Frames. We now take a HR*xy-frame* to be the result of adding to any R*xy*-frame a binary equivalence relation $\approx$ on U satisfying the following conditions, for all u, v in U:

(C1≈) If $u \approx v$, then u ~lg v.

(C2≈) If u_0 ~lg u, then $u_0 \approx u$.

Moreover, $\approx$ is to satisfy, for all integers i, k such that $0 \leq k < i \leq x$ and all integers j, m with $0 \leq j, m \leq y$:

(C3≈) If $f(\tau_i, w_j) \approx f(\tau_i, w_m)$, then $f(\tau_k, w_j) \approx f(\tau_k, w_m)$.

Remark. The intuitive import of the relation $\approx$ is this: we have $u \approx v$ iff (i) $\lg(u) = \lg(v)$, i.e. the time of u is identical to the time of v, and, moreover, (ii) $\text{lat}(u)$ and $\text{lat}(v)$, i.e.the histories to which u, v respectively belong, share the same past and present up to and including the time which is common to u and v; in other words, $\text{lat}(u)$ and $\text{lat}(v)$ are to 'coincide' at the time $\lg(u)$ [$=\lg(v)$] and at all times in T previous to $\lg(u)$. Hence, (C1≈) requires the equivalence relation $\approx$ to be, for any u in U, a relation on the ~lg-equivalence class $[u]_{\sim\text{lg}}$ (= $\{v \in U\text{: } u \sim\text{lg } v\}$). Furthermore, (C2≈) guarantees that the ~lg-equivalence class $[u_0]_{\sim\text{lg}}$ can be taken as the *origin* of the finite *tree* which is definable on any HR*xy*-frame by means of $\approx$. Finally, (C3≈) is a characteristic condition on $\approx$, leading to the representation of time by such a tree. In the suggestive terminology of Zanardo (1985), we may say that (C3≈) requires $f(\tau_i, w_j) \approx f(\tau_i, w_m)$ to hold only if the (strict) pasts of $f(\tau_i, w_j)$ and $f(\tau_i, w_m)$ 'coincide modulo $\approx$'.

HRxy-Models and Truth Conditions. We now take a *valuation on* a HR*xy*-frame still to be any function V from Prop into the power-set of $T \times W$. In accordance with our informal characterization of $\approx$ just given above, we then require V to satisfy the following condition, for all p in Prop and all u, v in U:

(C4≈) If $u \approx v$, then $([u]_{\sim\text{lg}}, [u]_{\sim\text{lat}}) \in V(p)$ iff $([v]_{\sim\text{lg}}, [v]_{\sim\text{lat}}) \in V(p)$.

As usual, then, we mean by a HR*xy-model* any ordered triple

$$M = (F, V, v)$$

where the first term is a HR*xy*-frame, the second a valuation on that frame, and where v is as in Section 3 *supra*. In the extended truth definition relative to HR*xy*-models we now have the following truth condition for sentences of the form *NB*:

$$M, \tau, w \models NB \text{ iff for all } w' \text{ in } W \text{ such that } f(\tau, w) \approx f(\tau, w')\text{: } M, \tau, w' \models B;$$

and dually for *MB*.

Finally, the notions of HR*xy-validity* and HR*xy-satisfiability* are straightforward.

7 Proof Theory of Two-Dimensional Temporal Logic with the R_t and R_{th} Operators and Historical Necessity: The Axiomatic Systems HR*xy*

Each axiomatic system in the infinite hierarchy HR*xy* $(x, y \in \omega)$ still has mp as its sole primitive rule of inference and Nec (for $\Box$) as its sole primitive rule of proof, but, in addition to the five groups A0, A1(a)-(n),...,A4(a)-(f) of axiom schemata, it has a sixth group of axioms governing the new modalities *N* and *M*, viz. the following:

A5. (a) $N^{lg}A \rightarrow NA;\ MA \rightarrow M^{lg}A$

(b) $Rt_0h_0NA \rightarrow Rt_0h_0N^{lg}A$

(c) The modal logic **S5** for *N* and *M*

(d) $p \rightarrow Np$, for all *p* in Prop

(e) $wNA \rightarrow NwA$.

(f) $Rt_{i-n}NA \rightarrow Rt_iNRt_{i-n}A$, for all $i, n \in \omega$ with $1 \leq n \leq i \leq x$

Remark. In the next to last axiom, A5(e), there occurs a one-place operator *w* ["at the last point west of here on this latitude", "yesterday"] used already in Åqvist (1999): Section 11, which can be *defined* in our present framework in terms of the systematic frame constants a_i and b_j together with the realization operators Rt_ih_j as follows:

Def.*w*. $wA \leftrightarrow_{df} \vee_{0 \leq j \leq y}(b_j \wedge (a_0 \vee \vee_{0 < i \leq x}(a_i \wedge Rt_{i-1}h_jA)))$

where the $\vee$-notations denote certain finite disjunctions in the familiar way.

The definitions of the notions of *provability, deducibility, consistency* and *maximal consistency* for HR*xy* are straightforward.

Soundness Theorem 7.1 and Lemma 7.2

Both versions of the Soundness Theorem 4.1 are readily extended to the logics HR*xy*. In like manner, Lemma 4.2 is easily extended so as to apply to the fresh hierarchy of systems HR*xy*.

For instance, in the validation of Scott's Rule for HR*xy* [clause (I)] we allow for the case that § = *N*. Further details are left to the reader.

Let us list some useful properties of the defined operator *w* in the following

Lemma 7.3. (i) *As defined by* Def.*w*, *this is a derived rule of proof both in* HR*xy* + Def.*w and already in* R*xy* + Def.*w*, *where, for the sake of expository simplicity, we write just* '$\vdash$' *to indicate provability in the relevant system*:

w-necessitation: *from* $\vdash A$ *to infer* $\vdash wA$.

Moreover, (ii) *all instances of the following theorem schemata are provable in* R*xy* + Def.*w as well as in* HR*xy* + Def.*w*:

T1. $N^{lat}A \rightarrow wA$

T2. $a_i \rightarrow (wA \rightarrow \neg w\neg A)$, for all integers *i* such that $1 \leq i < \omega$

T3. $w(A \rightarrow B) \rightarrow (wA \rightarrow wB)$.

Proof. *Ad w*-necessitation. To derive this rule of proof in the system at issue, use the primitive rule of proof Nec (= necessitation for $\Box$) together with axiom schema A4(a) and the fresh theorem schema T1 just stated.

Ad T1. Use various axioms under A1, A2 and A4 [notably A4(e)] together with Def.*w*!

Ad T2 and T3. Similarly. ∎

Lemma 7.4 (Properties of the New Operators *N, M*).

(I) *For all natural numbers x, y it holds that all instances of the following* theorem schemata *are* HR*xy-provable*:

T4. $Rt_iNA \leftrightarrow Rt_iNRt_iA$ for all $i \in \omega$ such that $0 \leq i \leq x$

T5. $Rt_{i-n}NRt_iA \rightarrow Rt_iNRt_iA$ for all $i, n \in \omega$ with $1 \leq n \leq i \leq x$

T6. $Rt_ip \rightarrow Rt_iNRt_ip$ for all $i \in \omega$ with $0 \leq i \leq x$ and all p in Prop

(II) *None of the following* sentence schemata *are* HR*xy-provable or* HR*xy-valid*:

(a) $b_j \rightarrow Nb_j$ $(0 \leq j \leq y)$

(b) $Rt_iA \rightarrow NRt_iA$ $(0 \leq i \leq x)$ [in contrast to T3 in Lemma 4.3 (I)]

(c) $NA \rightarrow N^{lg}A$

(d) $NA \rightarrow N^{lat}A$

(e) $NA \rightarrow \Box A$

(f) $N^{lat}A \rightarrow NN^{lat}A$

(g) $M^{lat}A \rightarrow NM^{lat}A$

(h) $Rt_{i-n}NA \rightarrow Rt_iNA$, for all $i, n \in \omega$ with $1 \leq n \leq i \leq x$

(k) $Rt_{i-n}A \rightarrow Rt_iNRt_{i-n}A$, for all $i, n \in \omega$ with $1 \leq n \leq i \leq x$

(II) *In the spirit of Åqvist & Hoepelman (1981), Section 12*: *Theorem 2, axiom* A5(d) [$p \rightarrow Np$, for p in Prop] *can be generalized so as to yield the following* theorem schema *of* HR*xy:*

T7. $A \rightarrow NA$, provided that A contains no occurrences of the operators Rt_i, N^{lat}, or M^{lat} or of any frame constant b_j $(0 \leq j < \omega)$.

Proof. As to (I), the proofs in HR*xy* of the theorem schemata T4 – T6 amount to exercises left to the reader, somewhat tedious in the first case. [We observe that analogues of T4 and T5 are taken as axioms in Bailhache (1991), Ch.IV, p. 74 f., and that an analogue of T6 is discussed by the author in the very same context.] As for (II), we leave to the reader the task of constructing appropriate counterexamples to the HR*xy*-validity of the schemata (a) – (k). [Note that the schemata (h) and (k) are analogues of van Eck's Th1 and Th2 on p. 280 in the *Logique et Analyse* **25** (1982) version of van Eck (1981).] As for (III), the proof of T7 is by an easy induction on the length of A. In the basis we appeal to A5(d), A5(a) and A1(m). For the interesting cases in the induction step, we use *inter alia* A4(a), A4(c), A5(a), A5(c) and A2(c). ∎

8 Semantic Completeness of the Logics HR*xy*

Preliminaries: Generated Canonical HRxy-structures. We begin by extending Definition 5.1. So the *canonical* HR*xy-structure generated by* any fixed maximal HR*xy*-consistent set of sentences q will be the ordered septuple

$$M^q = (U, (u_0, e_0, n_0), (\{\tau_i\}_{i \in \omega}, x), (\{w_j\}_{j \in \omega}, y), \approx, V, v)$$

Where

(i) $U = \{u \in \Omega_{HR}xy$: for each A in Sent, if $\Box A \in q$, then $A \in u\}$

(with $\Omega_{HR}xy$ being the set of all maximal HRxy-consistent sets of sentences)
and where the remaining conditions (ii)-(vi) now apply to U in this new sense and to the set Sent of sentences in our expanded language of HRxy. Moreover, there will be the following fresh condition governing our new equivalence relation $\approx$:

(iv $\approx$) $\approx$ is the binary relation on U such that for all u, v in U: $u \approx v$ iff for each A in Sent, if $NA \in u$, then $A \in v$.

Theorem 8.1. *As defined in the extended Definition* 5.1 *just presented, the initial quintuple in* M^q *is a* HRxy*-frame and* M^q *as a whole is a* HRxy*-model.*

Proof. See the Appendix of Åqvist (2004). ∎

Furthermore, for q and M^q as above and for all A in Sent and u in U, we have the following extended

Coincidence Lemma 8.2. M^q, $[u]$~lg, $[u]$~lat $\models A$ iff $A \in u$.

Proof. See again the Appendix of Åqvist (2004). ∎

Finally, as a consequence of the last two results, we obtain the following

Completeness Theorem 8.3 for HRxy. Both versions of the Completeness Theorem 5.4 are extended so as to apply to the new hierarchy of logics HRxy. The pattern of argument remains the same as in the case of the Rxy: in addition to the results 8.1 and 8.2 we use Lemma 7.2 (= Lemma 4.2 as extended to HRxy) in the proof. ∎

9 Extending the R_t/R_{th} Logics of Historical Necessity to Dyadic Deontic Logics of Conditional Obligation and Permission: Syntax and Semantics for the Systems DHRxym

Syntax of the Systems DHRxym. In this and the following sections we deal with an infinite hierarchy DHRxym of logics combining *dyadic deontic modalities* with the temporal ones so far studied in this essay, where, as usual, x, y are any members of the set ω of natural numbers, and where m is any positive integer with $1 \leq m \leq y+1$. Those logics are based on a common formal language obtained by our adding to the vocabulary of the earlier systems HRxy the following items:

(i) A third infinite family of *systematic frame constants, viz* $\{Q_k\}_{k=1,2....}$ indicating various 'levels of perfection'; as well as

(ii) a pair of *dyadic deontic modalities,* O (for conditional obligation) and P (for conditional permission), the readings of which will be considered in a moment.

The definition of the set Sent of well formed *sentences* of DHR*xym* is then straightforward: all the new frame constants will be (atomic) sentences; moreover, whenever A, B are sentences, so are O_BA and P_BA.

Remark on Notation for Dyadic Deontic Operators. We write O_BA [P_BA] in order to render the ordinary language locution "if *B,* then it ought to be that *A"* ["if *B,* then it is permitted (permissible) that A"]. We prefer this style of notation to the current one *O(A/B)* [*P(A/B)*], because (i) it is paranthesis-free, and (ii) the reading goes from left to right, and not the other way around.

Semantics for DHRxym: A General Remark. Our presentation of the semantics for DHR*xym* will differ from the one given in the cases of R*xy* and HR*xy* in the following crucial respect. In those earlier cases we started out by defining (1) the notion of a *frame,* then (2) that of a *model* (using the concept of a *valuation on* a frame), whereupon we gave (3) a recursive definition of the notion of *truth* relative to a model, in terms of which, finally, (4) we could characterize the notions of *validity* and *satisfiability.* For reasons that will hopefully appear as we go along, we have to change the terminology a bit and adopt another order of progression in the present case of DHR*xym:* we start out by defining (1) the concept of a *structure,* which already includes that of a *valuation* (on a frame) and which enables us to give (2) a recursive definition of the notion of *truth* relative to a *structure,* whereupon (3) we introduce the concept of a *model* as a special kind of structure, in terms of which (4) we characterize the notions of *validity* and *satisfiability.*

Semantics for DHRxym: DHRxym-structures. Let $x, y \in \omega$ and let m be any positive integer with $1 \leq m \leq y+1$. By a DHR*xym-structure* we shall mean a sequence

$$\boldsymbol{M} = ((U, (u_0, e_0, n_0), (\{\tau_i\}_{i \in \omega}, x), (\{w_j\}_{j \in \omega}, y), \approx, \boldsymbol{V}, v), (\{\mathrm{opt}_k\}_{k=1,2,\ldots,} m), \{\mathrm{R}_B\}_{B \in \mathrm{Sent}})$$

where:

(i) The initial septuple is a HR*xy*-model.

(ii) $\{\mathrm{opt}_k\}_{k=1,2,\ldots}$ is an infinite sequence of subsets of U [to be thought of as representing *levels of perfection*], and m is the positive integer $\leq y+1$ under consideration.

(iii) $\{\mathrm{R}_B\}_{B \in Sent}$ is a family, indexed by Sent, of binary relations on U.

We can now tell what it means for any sentence A to be *true at* a time $\tau \in T = \{\tau_0, \tau_1, \ldots, \tau_x\}$ *in* a history $w \in W = \{w_0, w_1, \ldots, w_y\}$ relative to any DHR*xym*-structure $\boldsymbol{M}$. Our extended truth-definition will contain two fresh clauses, one governing atomic sentences Q_k, and one governing dyadic deontic sentences of the forms O_BC and P_BC:

$$\boldsymbol{M}, \tau, w \models Q_k\,(k = 1,2,\ldots.) \text{ iff } \begin{cases} f(\tau, w) \in \mathrm{opt}_k, & \text{if } 1 \leq k \leq m \\ \tau \neq \tau, & \text{otherwise, i.e. if } m < k \end{cases}$$

$\boldsymbol{M}, \tau, w \models O_BC$ iff for all w' in W such that $f(\tau, w)\ \mathrm{R}_B\ f(\tau, w')$: $\boldsymbol{M}, \tau, w' \models C$

$\boldsymbol{M}, \tau, w \models P_BC$ iff for some w' in W with $f(\tau, w)\ \mathrm{R}_B\ f(\tau, w')$: $\boldsymbol{M}, \tau, w' \models C$.

Semantics for DHRxym: DHRxym-Models

We now focus our attention on a special kind of DHR*xym*-structures called 'DHR*xym-models*'. So by a DHR*xym-model* we shall mean any DHR*xym*-structure $\boldsymbol{M}$, where $\{\mathrm{opt}_k\}$, m and $\{\mathrm{R}_B\}$ satisfy the following three additional conditions:

δ1 [Exactly m Non-Empty Levels of Perfection]. This condition requires the set $\{\mathrm{opt}_1, \mathrm{opt}_2, ..., \mathrm{opt}_m\}$ to be a *partition* of U in the sense that

(a) $\mathrm{opt}_i \cap \mathrm{opt}_j = \varnothing$, for all positive integers i,j with $1 \leq i \neq j \leq m$

(b) $\mathrm{opt}_1 \cup .... \cup \mathrm{opt}_m = U$

(c) $\mathrm{opt}_k \neq \varnothing$, for each positive integer k with $1 \leq k \leq m$

(d) $\mathrm{opt}_k = \varnothing$, for each positive integer k with $m < k < \omega$.

The second condition on DHR*xym*-models requires $\boldsymbol{M}$ to be such that for all u, v in U and any integer k with $1 \leq k \leq m$:

δ2 [Closure of the opt_k under ~lat]. If $u \in \mathrm{opt}_k$ and u ~lat v, then $v \in \mathrm{opt}_k$.

Clearly, **δ2** requires each opt_k with $1 \leq k \leq m$ to be *closed* under the relation ~lat of *being on the same latitude as.*

Our third, crucial condition on DHR*xym*-models pertains to the indexed family $\{\mathrm{R}_B\}_{B \in \mathrm{Sent}}$; it requires any u, v in U and any sentence B to be such that:

δ3 [Import of the relations R_B]. $u \, \mathrm{R}_B \, v$ iff $u \approx v$ and $\boldsymbol{M}$, [v]~lg, $[v]\sim\mathrm{lat} \models B$ and for each z in U with $u \approx z$ and $\boldsymbol{M}$, [z]~lg, $[z]\sim\mathrm{lat} \models B$ it holds that $v \geq z$.

Here, the weak *preference* relation $\geq$, "is at least as good (ideal) as", is to be understood as follows. First of all, by clauses (a) and (b) in the condition **δ1** [Exactly m Non-empty Levels of Perfection], we have that for each u in U there is *exactly one* positive integer k with $1 \leq k \leq m$ such that $u \in \mathrm{opt}_k$. We then define a 'ranking' function r from U into the closed interval [1, m] of integers by setting, for each u *in U:*

$$\mathrm{r}(u) = \underline{\text{the}}\ k, \text{ with } 1 \leq k \leq m, \text{ such that } u \in \mathrm{opt}_k.$$

Finally, define $\geq$ as the binary relation on U such that for all u, v in U:

$$u \geq v \quad \text{iff} \quad \mathrm{r}(u) \leq \mathrm{r}(v).$$

Validity and Satisfiability in DHRxym. Armed with the notion of a DHR*xym*-model, we then introduce the notions of DHR*xym-validity* and DHR*xym-satisfiability* in the same way as we defined the corresponding notions for the logics R*xy* and HR*xy*. See Section 3 *supra.*

10 Proof Theory for the R_t/R_{th} Logics of Historical Necessity with Conditional Obligation and Permission: The Axiomatic Systems DHR*xym*

Each axiomatic system in the infinite hierarchy DHR*xym* ($x, y, m \in \omega$, $1 \leq m \leq y+1$) still has modus ponens (mp) as its sole primitive rule of inference and Nec (for $\Box$) as its sole primitive rule of proof. In addition to the six groups of axiom schemata A0, A1(a)-(n),..., A4(a)-(f) [Section 4 *supra*] and A5(a)-(e) [Section 7 *supra*], DHR*xym* has a seventh group of axiom schemata governing the new frame constants Q_k (with k=1,2,....) and the new dyadic deontic modalities O and P, viz. the following:

A6.
(a) $Q_1 \vee Q_2 \vee .... \vee Q_m$
(b) $Q_i \rightarrow \neg Q_j$, for all i, j in ω with $1 \leq i \neq j < \omega$
(c) $M^{lg}Q_1 \wedge M^{lg}Q_2 \wedge .... \wedge M^{lg}Q_m$
(d) $Q_k \rightarrow N^{lat}Q_k$, for all k in ω with $1 \leq k \leq m$
(e) $P_BA \leftrightarrow \neg O_B\neg A$
(f) $O_B(A \rightarrow C) \rightarrow (O_BA \rightarrow O_BC)$
(g) $O_BA \rightarrow NO_BA$
(h) $NA \rightarrow O_BA$
(i) $N(A \leftrightarrow B) \rightarrow (O_AC \leftrightarrow O_BC)$
(j) O_AA
(k) $O_{A \wedge B}C \rightarrow O_A(B \rightarrow C)$
(l) $MA \rightarrow (O_AB \rightarrow P_AB)$
(m) $P_AB \rightarrow (O_A(B \rightarrow C) \rightarrow O_{A \wedge B}C)$
(n) $P_AQ_k \rightarrow (Q_j \rightarrow \neg A)$, for all j, k in ω with $1 \leq j < k \leq m$
(o) $Q_1 \rightarrow (O_BA \rightarrow (B \rightarrow A)$
(p) $(Q_k \wedge O_BA \wedge B \wedge \neg A) \rightarrow P_B(Q_1 \vee .... \vee Q_{k-1})$, for all k in ω with $1 < k < m$.
(q) $Rt_{i-n}O_TA \rightarrow Rt_iO_TRt_{i-n}A$, for all $i, n \in \omega$ with $1 \leq n \leq i \leq x$

The definitions of the notions of *provability, deducibility, consistency* and *maximal consistency* for DHR*xym* are then straightforward.

Soundness Theorem 10.1
The Soundness Theorems 4.1 and 7.1 (both versions) are again readily extended to the logicsDHR*xym*. The detailed proof is left to the reader. ∎

Lemma 10.2
In like manner, Lemma 4.2 on our basic two-dimensional temporal logics R*xy* with the R_t and R_{th} operators is extended so as to apply to the new hierarchy DHR*xym*. Thus, in the validation of Scott's Rule for DHR*xym* [clause (I)] we allow for the cases where § = N, O_B or P_B. Moreover, in the extended Lemma 4.2 [clause (II)], we have

the following fresh subclause governing the constants Q_k, where Γ is any maximal DHR*xym*-consistent set of sentences:

(iii) $Q_k \in \Gamma$, for exactly one positive integer k with $1 \leq k \leq m$.

In the proof of this subclause, we appeal to the axiom A6(a) in order to establish *existence,* and to axiom A6(b) in order to get *uniqueness.*

Lemma 10.3 (Properties of the New Constants/Operators Q_k, O_B and P_B). *All instances of the following* theorem schemata *are provable in* DHR*xym*:

T1. $\neg Q_k$, for all positive integers k with $m < k < \omega$

T2. $P_B A \rightarrow NP_B A$

T3. $P_B(A \vee C) \leftrightarrow (P_B A \vee P_B C)$

T4. $Rt_i O_B A \leftrightarrow Rt_i O_B Rt_i A \qquad (0 \leq i \leq x)$

T5. $Rt_{i\text{-}n} O_T(Rt_{i\text{-}n} O_T Rt_i A \rightarrow Rt_i O_T Rt_i A) \qquad (1 \leq n \leq i \leq x)$

Proof. *Ad* T1. Suppose Q_k for some k such that $m < k < \omega$. Then, by axiom schema A6(b), we obtain $\neg Q_1$, $\neg Q_2$ and ... and $\neg Q_m$, whence $\neg(Q_1 \vee ... \vee Q_m)$, contrary to axiom A6(a). Hence, $\vdash Q_k \rightarrow \bot$ and T1, as desired.

Ad T2. We first obtain $MP_B A \rightarrow P_B A$ by contraposing A6(g), and then $NMP_B A \rightarrow NP_B A$ by familiar **S5**-principles [axiom schema A5(c) in Section 7]. Since by A5(c) we also get $P_B A \rightarrow NMP_B A$, the desired result T2 is immediate.

Ad T3. Immediate by axioms A6(e), (f) and (h) together with the underlying logic of N.

Ad T4. Exercise (easy, though somewhat tedious). Use *inter alia* A4(d)!

Ad T5. Assuming $Rt_{i\text{-}n} O_T Rt_i A$, we obtain by A6(q) $Rt_i O_T Rt_{i\text{-}n} Rt_i A$, and by T7 in Lemma 4.3 *supra* [enabling us to reduce the compound $Rt_{i\text{-}n} Rt_i$ to Rt_i] $Rt_i O_T Rt_i A$. So we get $\vdash Rt_{i\text{-}n} O_T Rt_i A \rightarrow Rt_i O_T Rt_i A$ by the Deduction Theorem for DHR*xym*. Applying the easily derived rule of proof:

$$from \vdash A \ to \ infer \vdash Rt_{i\text{-}n} O_T A \qquad (1 \leq n \leq i \leq x)$$

to this last result, the desired conclusion T5 is immediate.

Remark 1. The above theorem schemata T1-T3 will be explicitly appealed to in the proof of Theorem 11.2 *infra,* which is an essential ingredient in our Completeness Theorem 11.3 for the systems DHR*xym;* see the Appendix below. More precisely, we use them in showing that the canonical DHR*xym*-structure M^q [Section 11 *infra*] is indeed a DHR*xym*-model in the sense of satisfying the characteristic conditions **δ1** – **δ3** in the semantics for those systems.

Remark 2. The theorem schemata T4 and T5 are DHR*xym*-analogues of the axioms AOt1 and AOt2, respectively, in Bailhache (1991), Ch.IV, p.81. Note that neither T5

nor A6(q) can be strengthened by replacing O_T by O_B (with arbitrary formula *B*) in those schemata.

Remark 3. We observe that the axiom A6(c) can be – *prima facie* – weakened as follows:

A6(c') $\Diamond Q_1 \wedge \Diamond Q_2 \wedge ....\wedge \Diamond Q_m$

However, using axiom schemata A4(b)-(c) [Section 4 *supra*] together with A6(d), we easily derive our original formulation A6(c) from A6(c'). The derivation is left to the reader as an exercise. Thus, the two formulations are in effect equivalent in the axiomatics for DHR*xym*.

Lemma 10.4 (DHR*xym*-Analogues of Results in Åqvist & Hoepelman (1981), Sections 15-16). *All instances of the following* theorem schemata *are provable in* DHR*xym*:

T6. $A \rightarrow NA$; $A \rightarrow O_B A$, provided that *A* contains no occurrences of the operators R_{ti}, N^{lat}, or M^{lat}, or of any frame constants b_j $(0 \leq j < \omega)$ or Q_k $(1 \leq k < \omega)$.

T7. $(NA \vee N\neg A) \rightarrow (O_T A \leftrightarrow A)$

T8. $NA \vee N\neg A$, where *A* satisfies the proviso of T6

T9. $O_T A \leftrightarrow A$, where *A* satisfies the proviso of T6

T10. $O_B A \leftrightarrow (B \rightarrow O_T A)$, where *B* satisfies the same proviso

T11. $O_B A \leftrightarrow O_T(B \rightarrow A)$, where *B* satisfies the same proviso

T12. $(B \rightarrow O_T A) \leftrightarrow O_T(B \rightarrow A)$, where *B* satisfies the same proviso.

Proof. *Ad* T6. Easily handled in the spirit of the proof of T7 in Lemma 7.4 (III) *supra.*

Ad T7. For $NA \rightarrow (O_T A \leftrightarrow A)$, use A0, A5(c) and A6(h). For $N\neg A \rightarrow (O_T A \leftrightarrow A)$, use A5(c), A6(h), A6(l) [relying on $\vdash\!\!\!- MT$] and A6(e).

A*d* T8. Use AO and T6 *supra.*

Ad T9. Immediate from T7 and T8.

Ad T10. For the *left-to-right* direction, assume $O_B A$ and *B,* where *B* satisfies the proviso of T6 [containing no occurrences etc.]. Then *NB* by T6 as well as $N(B \leftrightarrow T)$ by A5(c). Hence, $O_B A \leftrightarrow O_T A$ by A6(i), so $O_T A$ and $B \rightarrow O_T A$ by A0. For the *right-to-left* direction, make the counterassumption that $(B \rightarrow O_T A)$ together with $P_B \neg A$, where *B* still satisfies the proviso at issue. We then leave to the reader the task of showing that this counterassumption implies the contradiction that $N\neg B \wedge MB$.

Ad T11. For the *left-to-right* direction, make the counterassumption that $O_B A \wedge P_T(B \wedge \neg A)$. Show that *MB* follows from the second conjunct, and that $O_T A$ follows from the first one [the proviso gives us $\vdash\!\!\!- MB \rightarrow B$ and $\vdash\!\!\!- B \rightarrow (O_T A \leftrightarrow O_B A)$],

whence $O_T(B \rightarrow A)$ contrary to the second conjunct. The opposite direction is handled in the same spirit. Note that the axiom schema A6(j) is used in the proof of both directions!

Ad T12. Immediate from T10 and T11.

11 Semantic Completeness of the Logics DHR*xym*

Preliminaries: Generated Canonical DHRxym-Structures. We begin by extending Definition 5.1. For any natural numbers $x, y, m \in \omega$ with $1 \leq m \leq y+1$, let $\Omega_{\mathrm{DHR}}xym$ be the set of all maximal DHR*xym*-consistent sets of sentences. Let q be a fixed element of $\Omega_{\mathrm{DHR}}xym$. Define the *canonical* DHR*xym-structure generated by* q as the sequence

$$M^q = ((U, (u_0, e_0, n_0), (\{\tau_i\}_{i\in\omega}, x), (\{w_j\}_{j\in\omega}, y), \approx, V, v), (\{\mathrm{opt}_k\}_{k=1,2,\ldots,m}), \{\mathrm{R}_B\}_{B\in \mathrm{Sent}})$$

where

(i) $U = \{u \in \Omega_{\mathrm{DHR}}xym$: for each A in Sent, if $\Box A \in q$, then $A \in u\}$

and where the remaining conditions (ii)-(vi) in Definition 5.1 now apply to U in this new sense and to the richer set Sent of sentences in our expanded language of DHR*xym*. Similarly for the condition (iv $\approx$) stated in the *Preliminaries* to Section 8 above, which condition is still assumed to govern the equivalence relation $\approx$. Furthermore, we must define the remaining items in M^q:

(vii) $\mathrm{opt}_k = \{u \in U: Q_k \in u\}$ $(k=1,2,\ldots)$

(viii) m is the third natural number under consideration.

Finally, as to the indexed family $\{\mathrm{R}_B\}$, we require each B in Sent to satisfy:

(ix) R_B = the binary relation on U such that for all u, v in U:
$u \mathrm{R}_B v$ iff for all A in Sent: if $O_B A \in u$, then $A \in v$.

Moreover, for canonical DHR*xym*-structures (generated by $q \in U$) as just defined by conditions (i)-(iv),(iv $\approx$),(v)-(ix), we introduce the ranking function r from U into the closed interval $[1, m]$ of integers by setting, for each u in U:

$\mathrm{r}(u) =$ the k, with $1 \leq k \leq m$, such that $Q_k \in u$.

This definition of r is clearly justified by our fresh subclause (iii) in Lemma 10.2 (II).

Again, $\geq$ is the binary relation on U such that for all u, v in U: $u \geq v$ iff $\mathrm{r}(u) \leq \mathrm{r}(v)$.

Having gone through these preliminaries, we now state two basic results on generated canonical DHRxym-structures. However, the order of presentation will be reversed as compared with the one adopted in Sections 5 and 8 *supra,* and similarly for their proofs given in the Appendix of Åqvist (2004). The reason for this reversal will again become apparent in that Appendix.

Coincidence Lemma 11.1. *Let q be any fixed maximal* DHR*xym-consistent set of sentences, and let* $\mathfrak{M}^q$, *as just defined, be the canonical* DHR*xym-structure generated by q. Then, for each sentence A and each u in U:*

$$\mathfrak{M}^q, [u]\sim\text{lg}, [u]\sim\text{lat} \models A \text{ iff } A \in u.$$

Proof. By induction on the length of *A*. For details, see the Appendix of Åqvist (2004).

∎

Theorem 11.2. *As just defined,* $\mathfrak{M}^q$ *is a* DHR*xym-model.*

Proof. See again the Appendix of Åqvist (2004).

As a consequence of the two basic results just stated, we obtain the desired

Completeness Theorem 11.3 for DHR*xym*. Both versions of the Completeness Theorems 5.4 and 8.3 are extended so as to apply to the new hierarchy of logics DHR*xym*.

Proof. The pattern of argument remains the same as in the case of the R*xy* and the HR*xy*: just use right Lemmata/Theorems!

∎

References

Åqvist, L. (1991): Discrete tense logic with beginning and ending time: An infinite hierarchy of complete axiomatic systems, *Logique et Analyse* **34** (1991), 359 – 401. Åqvist, L. (1997): Systematic frame constants in defeasible deontic logic: A new form of Andersonian reduction, in D. Nute (ed.), *Defeasible Deontic Logic,* Kluwer, Dordrecht / Boston / London, pp. 59–77.

Åqvist, L. (1999): The logic of historical necessity as founded on two-dimensional modal tense logic, *J. Philos. Logic* **28** (1999), 329 – 369. Missing list of References appeared in *J. Philos. Logic* **29** (2000), 541–542.

Åqvist, L. (2000): Three characterizability problems in deontic logic, *Nordic Journal of Philosophical Logic* **5** (2000), 65–82.

Åqvist, L. (2002): Conditionaliy and branching time in deontic logic: Further remarks on the Alchourrón & Bulygin example, in J. Horty and A.J.I. Jones (eds.), *ΔEON'02: Sixth International Workshop on Deontic Logic in Computer Science, Imperial College, London, May 2002,* pp. 323–343.

Åqvist, L. (2002a): Old Foundations for the Logic of Agency and Action, *Studio Logica* **72** (2002), 313–338.

Åqvist, L. (2004): On the R_t Approach to Temporal Logic with Historical Necessity and Conditional Obligation. Manuscript 39 pp. Forthcoming.

Åqvist, L. and Hoepelman, J. (1981): Some theorems about a "tree" system of deontic logic, in R.Hilpinen (ed.), *New Studies in Deontic Logic,* Reidel, Dordrecht Boston / London, pp. 187–221.

Åqvist, L. and Mullock, Ph. (1989): *Causing Harm: A Logico-Legal Study,* W. de Gruyter, Berlin.

Bailhache, P. (1991): *Essai de logique déontique.* LIBRAIRIE PHILOSOPHIQUE J. VRIN, Paris. Chapitre quatre: Normes et temps, pp. 62– 87.

Bailhache, P. (1993): The deontic branching time: Two related conceptions, *Logique et Analyse* **36** (1993), 159–175.

Carmo, J. and Jones, A.J.I. (2002): Deontic logic and contrary-to-duties, in D.M.Gabbay andF.Guenthner (eds.), *Handbook of Philosophical Logic, 2nd Edition, Volume 8,* Kluwer, Dordrecht / Boston / London, pp. 265–343.

van Eck, Job A. (1981): *A System of Temporally Relative Modal and Deontic Predicate Logic and its Philosophical Applications.* Rijksuniversiteit te Groningen: Department of Philosophy. Doctoral dissertation, 1981.Also in *Logique et Analyse* **25** (1982), 249 – 290 and 339–381.

von Kutschera, F. (1996): Rezension von (Review of) Åqvist and Mullock (1989), *Erkenntnis* **44**(1996), 113–118.

von Kutschera, F. (1997): T × W completeness, *J. Philos. Logic* 26 (1997), 241–250.

Loewer, B. and Belzer, M. (1983): Dyadic deontic detachment, *Synthese* **54** (1983), 295–318.

Makinson, D. (1966): On some completeness theorems in modal logic, *Zeitschrift für mathematische Logik und Grundlagen der Mathematik* **12** (1966), 379–384.

Rescher, N. (1966): On the logic of chronological propositions, *Mind* **75** (1966), 75–96.

Rescher, N. and Urquhart, A. (1971): *Temporal Logic,* Springer, Wien / New York.

Segerberg, K. (1973): Two-dimensional modal logic, *J. Philos. Logic* **2** (1973), 77–96.

von Wright, G. H. (1983): *Practical Reason,* Blackwell, Oxford.

Zanardo, A. (1985): A finite axiomatization of the set of strongly valid Ockhamist formulas, *J. Philos. Logic* 14 (1985), 447–68.

Δ: The Social Delegation Cycle

Guido Boella[1] and Leendert van der Torre[2,3,*]

[1] Dipartimento di Informatica, Università di Torino, Italy
guido@di.unito.it
[2] CWI Amsterdam, The Netherlands
torre@cwi.nl
[3] Delft University of Technology, The Netherlands

Abstract. In this paper we consider the relation between desires and obligations in normative multiagent systems. We introduce a model of their relation based on what we call the social delegation cycle, which explains the creation of norms from agent desires in three steps. First individual agent desires generate group goals, then a group goal is individualized in a social norm, and finally the norm is accepted by the agents when it leads to the fulfilment of the desires the cycle started with. We formalize the social delegation cycle by formalizing goal generation as a merging process of the individual agent desires, we formalize norm creation as a planning process for both the obligation and the associated sanctions or rewards, and we formalize the acceptance relation as both a belief of agents that the fulfilment of the norm leads to achievement of their desires, and the belief that other agents will act according to the norm.

1 Introduction

The relation between obligations and actions is a classical field of study in deontic logic. However, when we consider actions of agents with beliefs and desires, then some questions arise which are traditionally not studied in this area. In agent theory, the relation between desires and obligations has been formalized in BOID agent architectures as a combination of BDI agent architectures [10] and normative (BO) agent architectures, and more generally in normative multiagent systems (NMAS) as a combination of multiagent systems (MAS) and normative systems (NS) for applications like virtual communities [5]. Whereas BDI and NMAS conceptualize the agent's decision making behavior in terms of goals and desires, BO and NS conceptualize the agent's behavior in terms of obligations and permissions.

$$BOID = BDI + BO \qquad NMAS = MAS + NS$$

However, the proposed BOID architectures and normative multiagent systems do not explain how desires and obligations are related. Consequently, the formalization of desires and obligations has raised many questions. For example, is the logic of desire different from the logic of obligation [20]? How do agents deal with conflicts between desires and obligations in their decision making [11]? Why do agents often respect obligations even if they know that their violations are not or cannot be sanctioned? What

* Supported by the ArchiMate research project.

A. Lomuscio and D. Nute (Eds.): DEON 2004, LNAI 3065, pp. 29–42, 2004.

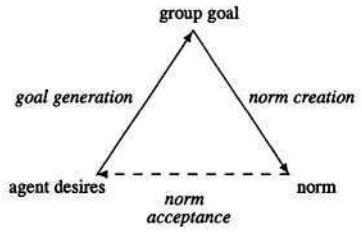


Fig. 1. Δ: the social delegation cycle.

does this imply for the rational creation of norms in such systems, and which mechanisms do not work properly without a normative system? How are social constructions like normative systems constructed from multiagent systems [27]? When is a separation of powers as in trias politica a necessary precondition for norm creation to be efficient?

In this paper we introduce the social delegation cycle, which explains the creation of norms from desires from a rational (e.g., Kantian) perspective. We assume that norms are only accepted if they are respected by the other agents, and therefore sometimes sanctions are needed. Informally, it consists of three steps visualized in Figure 1. Individual agents have desires, which turn into group (or joint, or social) goals. A group goal is individualized by a social norm. The individual agents accept the norm, together with its associated sanctions and rewards, because they recognize that it serves to achieve their desires the cycle started with.

We study the social delegation cycle in a formal framework. The research questions of this paper are:

1. How to balance goal generation, norm creation, and acceptance?
2. How to formalize joint goal generation? We formalize goal generation as a merger of individual desires.
3. How to formalize norm generation? We formalize norm creation as a planning problem, distinguishing between creation of the obligation and creation of the associated sanctions and rewards;
4. How to formalize the acceptance relation? We formalize the acceptance relation by distinguishing between the fulfilment of the agents' desires, and the belief that other agents will fulfill the norm.

The conceptual model we use to study and formalize the social delegation cycle is based on a formal characterization of normative multiagent systems we have developed elsewhere [5,8,9], which is based on rule based systems and input/output logics. Moreover, this other work is based on the assumption that the normative system can be modelled as an agent. This paper is not based on this assumption, but it is related to it, as we explain in detail in Section 9.

The layout of this paper is as follows. In Section 2 we discuss the balance between goal generation, norm creation and acceptance. In Section 3 we define the conceptual model in which we study and formalize the social delegation cycle, and in Section 4 we define the logic of rules. In Section 5 we formalize goal generation, in Section 6 we formalize norm creation, and in Section 7 we formalize the acceptance relation.

2 Social Delegation Cycle

When developing a formal model for the social delegation cycle, we have to make two fundamental choices.

- We may define a general model of the social delegation cycle, defining a range of possibilities, or we may define an actual procedure. The two are not exclusive, since we can first define a general theory of social delegation cycle, thereafter desirable properties within this framework, and finally procedures within the framework that satisfy some or all of the desirable properties.
- We have to define how the elements of the social delegation cycle, i.e., goal generation, norm creation and acceptance, are balanced. For example, strictly defined norm creation procedures only create norms that will always be accepted, and analogously strictly defined goal generation procedures generate only goals for which a norm can be created that is accepted.

In this paper, we propose a fairly general formal model of the social delegation cycle, which delimits the kind of norms that can be created, but that does not give an actual procedure to create norms. The reason is that we aim to capture the fundamental properties of the social delegation cycle, which later can be used to design actual procedures. However, compared to informal characterizations of the construction of social reality, such as in the work of Searle [27], our model is fairly limited as we do not introduce for example beliefs or institutions. This issue is discussed in Section 10.

Concerning the balance between the elements of the cycle, we do not aim to define strict goal generation and norm creation procedures. The reason is that we believe that our setting is more realistic and may cover a wider range of social delegation cycles. Moreover, it facilitates the use of formal theories developed elsewhere, such as merging theories for joint goal generation, planning theories for norm creation, and game theories for acceptance. We consider the definition of strict mechanisms more relevant for the design of mechanisms of norm creation.

Our model builds on several existing formal theories and formalizes the three steps as follows:

Goal generation generates a set of goals based on merging operators, which have been proposed as generalizations of belief revision operators inspired by social choice theory.

Norm creation creates for each goal a set of norms (or revisions of existing norms) based on planning theories as used in most theories in artificial intelligence.

Acceptance relation accepts or rejects a norm based on game theories. We assume that norms are thus only accepted if they are respected, and we formalize the acceptance relation by distinguishing between fulfilment of the agents desire, and the belief that other agents will fulfill the norm.

Before we present our formalizations, we define in the following two sections the conceptual framework we use based on rule based systems, and the logic of rules based on input/output logics.

3 Conceptual Model

The conceptual model is visualized in Figure 2, in which we distinguish the multiagent system (normal lines) and additions for the normative system (thick lines). Following the usual conventions of for example class diagrams in the unified modelling language (UML), □ is a concept or set, — and → are associations between concepts, and —▷ is the "is-a" or subset relation. The logical structure of the associations is detailed in the definitions below.

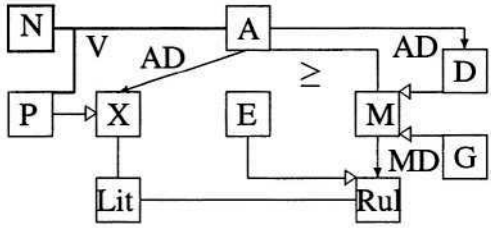


Fig. 2. Conceptual model of normative multiagent system.

The model consists of a set of agents *(A),* which are described *(AD)* by a set of boolean variables *(X)* including *decision variables* it can perform and desires *(D)* guiding its decision making. The motivational state of the group *(G)* is composed of its goals. Desire rules can be conflicting, and the way the agent resolves its conflicts is described by a priority relation $(\geq)$ that expresses its agent characteristics [11]. The priority relation is defined on the powerset of the motivations such that a wide range of characteristics can be described, including social agents that take the desires or goals of other agents into account. The priority relation contains at least the subset-relation which expresses a kind of independence between the motivations. Variables which are not decision variables are called parameters *(P).*

Definition 1 (AS). *An agent set is a tuple* $\langle A, X, D, G, AD, \geq \rangle$*, where:*

- *the agents A, variables X, agent desires D and group goals G are four finite disjoint sets. We write* $M = D \cup G$ *for the motivations defined as the union of the desires and goals.*
- *an agent description* $AD : A \rightarrow 2^{X \cup D}$ *is a complete function that maps each agent to sets of variables (its decision variables) and desires, but that does not necessarily assign each variable to at least one agent. For each agent* $a \in A$*, we write* X_a *for* $X \cap AD(a)$*, and* D_a *for* $D \cap AD(a)$*. We write parameters* $P = X \setminus \cup_{a \in A} X_a$*.*
- *a priority relation* $\geq$*:* $A \rightarrow 2^M \times 2^M$ *is a function from agents to a transitive and reflexive relation on the powerset of the motivations containing at least the subset relation. We write* $\geq_a$ *for* $\geq (a)$*.*

Desires and goals are abstract concepts which are described by – though conceptually not identified with – rules *(Rul)* built from literals *(Lit).* They are therefore not

represented by prepositional formulas, as in some other approaches to agency [13,25]. Agents may share decision variables, or desires, though this complication is not used in this paper. Background knowledge is formalized by a set of effect rules *(E)*.

Definition 2 (MAS). *A multiagent system is a tuple* $\langle A, X, D, G, AD, E, MD, \geq \rangle$, *where* $\langle A, X, D, G, AD, \geq \rangle$ *is an agent set, and:*

- *the set of literals built from X, written as Lit(X), is* $X \cup \{\neg x \mid x \in X\}$, *and the set of rules built from X, written as* $\mathbf{Rul}(X) = 2^{Lit(X)} \times \mathbf{Lit}(X)$, *be the set of pairs of a set of literals built from X and a literal built from X, written as* $\{l_1, \ldots, l_n\} \rightarrow l$. *We also write* $l_1 \wedge \ldots \wedge l_n \rightarrow l$ *and when* $n = 0$ *we write* $\top \rightarrow l$. *Moreover, for* $x \in X$ *we write* $\sim x$ *for* $\neg x$ *and* $\sim(\neg x)$ *for* x.
- *the set of effects* $E \subseteq \mathbf{Rul}(X)$ *is a set of rules built from X.*
- *the motivational description* $MD : M \rightarrow \mathbf{Rul}(X)$ *is a complete function from the sets of desires and goals to the set of rules built from X. For a set of motivations* $S \subseteq M$, *we write* $MD(S) = \{MD(s) \mid s \in S\}$.

We now extend the multiagent system to a normative multiagent system to take norm generation into account. To describe the normative system, we introduce a set of norms *(N)* and a norm description that associates violations with variables (V).

Definition 3 (NMAS). *A normative multiagent system NMAS is a tuple*

$$\langle A, X, D, G, AD, E, MD, \geq, N, V \rangle$$

where $MAS = \langle A, X, D, G, AD, E, MD, \geq \rangle$ *is our multiagent system, and moreover:*

- *the norms N is a set disjoint from A, X, D, and G.*
- *the norm description* $V : N \times A \rightarrow P$ *is a complete function that maps each pair of a norm and an agent to the parameters, where* $V(n, a)$ *represents the parameter that counts as a violation by agent* a *of the norm* n.

We define sanction and reward-based obligations in the normative multiagent system using an extension of Anderson's well-known reduction [2], like Meyer [24] also does: violations and sanctions are the consequences of not fulfilling a norm. It covers a kind of ought-to-do and a kind of ought-to-be obligations. Moreover, we can also have that x is obligatory for agent a while it is a decision variable of another agent b. The logic of obligations, sanctions and rewards satisfies only replacements by logical equivalents.

Definition 4 (Obligation). *Let* $NMAS = \langle A, X, D, G, AD, E, MD, \geq, N, V \rangle$. *We say that:*

- x *is obligatory for agent* a *in NMAS iff* $\exists n \in N$ *with* $\sim x \rightarrow V(n, a) \in E$,
- s *is a sanction for agent* a *in NMAS iff* $\exists n \in N$ *with* $V(n, a) \rightarrow s \in E$, *and*
- r *is a reward for agent* a *in NMAS iff* $\exists n \in N$ *with* $\neg V(n, a) \rightarrow r \in E$.

The obligation for x *is called ought-to-do when* $x \in \mathbf{Lit}(X \setminus P)$ *and it is called ought-to-be when* $x \in \mathbf{Lit}(P)$.

This is clearly a very weak notion of obligation, and more sophisticated notions within this kind of framework are developed elsewhere [5]. In this paper we now turn to the representation of the desires and goals.

4 Logic of Rules

We use a simplified version of the input/output logics introduced in [21,22]. A rule base is a set of rules, i.e., a set of ordered pairs $p \rightarrow q$. For each such pair, the body p is thought of as an input, representing some condition or situation, and the head q is thought of as an output, representing what the norm tells us to be desirable, obligatory or whatever in that situation. We use input/output logics since they do not necessarily satisfy the identity rule. Makinson and van der Torre write (p, q) to distinguish input/output rules from conditionals defined in other logics, to emphasize the property that input/output logic does not necessarily obey the identity rule. In this paper we do not follow this convention.

In this paper, input and output are respectively a set of literals and a literal. We use a simplified version of input/output logics, since it keeps the formal exposition simple and it is sufficient for our purposes here. In Makinson and van der Torre's input/output logics, the input and output can be arbitrary prepositional formulas, not just sets of literals and literal as we do here. Consequently, in input/output logic there are additional rules for conjunction of outputs and for weakening outputs.

Definition 5 (Input/output logic [21]). *Let a rule base B be a set of rules $\{p_1 \rightarrow q_1, \ldots, p_n \rightarrow q_n\}$, read as 'if input p_1 then output q_1', etc., and consider the following proof rules, strengthening of the input (SI), disjunction of the input (OR), and cumulative transitivity (CT) defined as follows:*

$$\frac{p \rightarrow r}{p \wedge q \rightarrow r} SI \qquad \frac{p \wedge q \rightarrow r, p \wedge \neg q \rightarrow r}{p \rightarrow r} OR \qquad \frac{p \rightarrow q, p \wedge q \rightarrow r}{p \rightarrow r} CT$$

The following four output operators are defined as closure operators on the set B using the rules above:

out_1: SI (simple-minded output) *out_3: SI+CT (simple-minded reus. output)*
out_2: SI+OR (basic output) *out_4: SI+OR+CT (basic reusable output)*

We write out(B) for any of these output operations and $B \vdash_{iol} p \rightarrow q$ iff $p \rightarrow q \in out(B)$, and we write $B \vdash_{iol} B'$ iff $B \vdash_{iol} p \rightarrow q$ for all $p \rightarrow q \in B'$.

The following definition of the so-called input-output and output constraints checks whether the derived conditional goals are consistent with the input.

Definition 6 (Constraints [22]). *Let B be a set of rules, and C a set of literals. B is consistent with C, written as $cons(B \mid C)$, iff there do not exist two contradictory literals p and $\neg p$ in $C \cup \{l \mid B \vdash_{iol} C \rightarrow l\}$. We write cons(B) for $cons(B \mid \emptyset)$.*

Due to space limitations we have to be brief on technical details with respect to input/output logics, see [21,22] for their semantics, further details on their proof theory, the extension with the identity rule, alternative constraints, and examples.

5 Joint Goal Generation by Merging Agent Desires

We characterize the goal generation process as a merger or fusion of the desires of the agents, which may be seen as a particular kind of social choice process [19]. In

this paper, we use the merging operators for merging desires into goals in the context of beliefs, defined in [15]. We adapt these operators in two ways. First we simplify the operators, because we do not use beliefs. Secondly, and most importantly, we make them more complex, because we extend the operators defined on propositional formulas to merge rules.

Definition 7. *A rule base B is a set of rules, a rule set S is a multi-set of rule bases. Two rule sets S_1 and S_2 are equivalent, noted $S_1 \leftrightarrow S_2$, iff there exists a bijection f from $S_1 = \{B_1^1, \ldots, B_1^n\}$ to $S_2 = \{B_2^1, \ldots, B_2^n\}$ such that $out(f(B)) = out(B)$. We write $\bigwedge S$ for the union of all rules in S, and $\sqcup$ for union with multi-sets.*

Most of these postulates are generalizations of belief revision postulates [1, 16, 18]. (R0) states that the result of merging complies with the integrity constraints. (R1) ensures that, when the integrity constraints are consistent we always manage to extract a coherent piece of information from the knowledge set. (R2) says that, if possible, the result of the merging is simply the conjunction of the knowledge bases of the knowledge set with the integrity constraints, (R3) is the principle of irrelevance of syntax. The purely 'merging' postulates are (R4), (R5) and (R6). (R4) is what is called the fairness postulate. It ensures that when merging two knowledge bases, the operator cannot give full preference to one of them. (R5) and (R6) correspond to Pareto's conditions in social choice theory [3] and were proposed in [26] to model fitting operators. Finally (R7) and (R8) state conditions on the conjunction of integrity constraints and make sure that 'closeness' is well-behaved [18]. See the above mentioned papers for further details and motivations. In the following definition, as well as in all following definitions, we assume that a logic of rules has been fixed.

Definition 8. *Let $\vdash_{iol}$ be an output operation, S be a rule set, E a rule base, and ∇ an operator that assigns to each rule set S and rule base E a rule base $\nabla_E(S)$. ∇ is a rule merging operator if and only if it satisfies the following properties:*

R0 *If not cons(E), then* $\nabla_E(S) \leftrightarrow E$
R1 *If cons(E), then* $cons(\nabla_E(S))$
R2a $\bigwedge S \vdash_{iol} \nabla_E(S)$
R2b *If* $cons(\bigwedge S \cup E)$, *then* $\nabla_E(S) \vdash_{iol} \bigwedge S$
R3 *If* $S_1 \leftrightarrow S_2$ *and* $E_1 \leftrightarrow E_2$, *then* $\nabla_{E_1}(S_1) \leftrightarrow \nabla_{E_2}(S_2)$
R4 *If* $B \vdash_{iol} E$, $B' \vdash_{iol} E$, *and* $cons(\nabla_E(\{B\} \sqcup \{B'\}) \cup B \cup E)$, *then* $cons(\nabla_E(\{B\} \sqcup \{B'\}) \cup B' \cup E)$
R5 $\nabla_E(S_1) \cup \nabla_E(S_2) \vdash_{iol} \nabla_E(S_1 \sqcup S_2)$
R6 *If* $cons(\nabla_E(S_1) \cup \nabla_E(S_2) \cup E)$, *then* $\nabla_E(S_1 \sqcup S_2) \vdash_{iol} \nabla_E(S_1) \cup \nabla_E(S_2)$
R7 *If* $cons(E_1 \cup E_2)$, *then* $\nabla_{E_1}(S) \vdash_{iol} \nabla_{E_1 \cup E_2}(S)$
R8 *If* $cons(\nabla_{E_1}(S) \cup E_1 \cup E_2)$, *then* $\nabla_{E_1 \cup E_2}(S) \vdash_{iol} \nabla_{E_1}(S)$

Additional properties can be accepted [19], but due to space limitations we do not discuss them. For the same reason we do not discuss the semantics of merging operators. The merging operator is illustrated in the following example.

Example 1. Let $\vdash_{iol}$ be out_3, and consider four rule bases each consisting of a single rule $S = \{\{\top \rightarrow p\}, \{\top \rightarrow q\}, \{p \rightarrow r\}, \{q \rightarrow \neg r\}\}$. Now $cons(\bigwedge E)$ does not

hold, so due to R1 we cannot have $\nabla_\emptyset(S) = \bigwedge S$. They can be merged into a maximal subset of these rules, for example we may have $\nabla_\emptyset(S) = \{\top \rightarrow p, \top \rightarrow q, p \rightarrow r\}$ or $\nabla_\emptyset(S) = \{\top \rightarrow q, p \rightarrow r, q \rightarrow \neg r\}$. Note that the latter merger selects a maximal consistent subset of *S,* but it does not select a set of rules that maximizes the output $\{x \mid \nabla_\emptyset(S) \vdash_{iol} \top \rightarrow x\}$ (a distinction discussed in [22]). If we assume $\vdash_{iol}$ be out_1, then $cons(\bigwedge E)$, and due to R1 we have $\nabla_\emptyset(S) = \bigwedge S$.

In our conceptual model, goals are a subset of the merger of desires of the agents.

Definition 9 (Goal generation). *There is a rule merging operator* ∇ *such that* $MD(G) \subseteq \nabla_E(MD(D_x) \mid x \in A)$.

In the latter definition we use variable *x* to refer to agents. We use variables also in many other places, e.g., in the following example, but these variables are just used to shorten the presentation and are not part of the logical language. It is just some *syntactic sugar.* For example, quantification over rules means that it is schema: there is a set of rules, one for each agent involved. Since the set of agents *A* is finite, we are still in prepositional logic. Joint goal generation is illustrated by the following example.

Example 2. Let $NMAS = \langle A, X, D, G, AD, E, MD, \geq, N, V\rangle$ with the following ingredients:

variables in *X:*
$\{\neg collision(x_1, x_2), accident, drive\_right(x), drive\_left(x) \mid x_1, x_2, x \in A\}$.
Moreover, each agent can decide to drive on left or right side of street, e.g., $X_a = \{drive\_right(a), drive\_left(a)\}$,

effect rules *E:*
$\{drive\_right(x_1) \wedge drive\_left(x_2) \rightarrow collision(x_1, x_2) \mid x_1, x_2 \in A\}$
$\cup \{(\bigwedge_{x_1,x_2 \in A} \neg collision(x_1, x_2)) \rightarrow \neg accident\}$
$\cup \{collision(x_1, x_2) \rightarrow accident \mid x_1, x_2 \in A\}$
$\cup \{collision(x_1, x_2) \rightarrow collision(x_2, x_1) \mid x_1, x_2 \in A\}$
If two agents do not drive on same side then they collide, and if there are no collisions then there is no accident

desires: $D_x = \{\top \rightarrow \neg collision(x, y) \mid y \in A\}$ for each agent $x \in A$. Agents desire not to be part of a collision.

goal $G = \{\top \rightarrow \neg accident\}$.

The system generates a joint goal of *NMAS* for absence of accidents.

Goals can be generated using negotiation processes. Alternatively, the process can be facilitated by an agent playing the role of legislator. Here we do not further consider the construction of goals.

6 Norm Creation

We formalize norm creation as a planning problem, distinguishing between the creation of the obligation and the creation of the associated sanctions and rewards. In some cases sanctions must be associated with the norms to ensure that some agent fulfills the norm, and therefore to ensure that the other agents accept the norm, but in some other cases this is not necessary. Here are two prototypical examples.

- Agents do not want to crash into each other, and the norm to drive on the right side of the road (or the left side, for that matter) is accepted by all members. In this case, no sanction is necessary and the norm may be called a convention. Other examples of this kind can be found in coordination games in game theory.
- Agents want to cooperate in a prisoner's dilemma, so the norm to cooperate is accepted by all members. In this case, a sanction must be associated with the norm, because otherwise the agent will defect (as game theory shows).

The two elements of norm creation are formalized as two sequential steps: first determining the obligation, and thereafter determining the associated sanctions or rewards. The first step is essentially a planning problem: the obligations of the agents must imply the joint goal $Y \rightarrow g$. We represent a norm n by an obligation for all agents in the multiagent system, that is, for every agent a we introduce an obligation $\sim x \rightarrow V(n,a)$. Moreover, since goals can only be in force in a context, e.g., Y, we introduce in context Y an obligation $Y \wedge \sim x \rightarrow V(n,a)$. Roughly, the condition is that all obligations x imply the goal g.

However, to determine whether the obligations imply the goal, we have to take the existing normative system into account. We assume that the normative system only creates obligations that can be fulfilled together with the already existing obligations. Moreover, for the test that the goal g will be achieved, we propose the following condition: if every agent fulfills its newly introduced obligation, and it fulfills all its other obligations, then g is achieved. We define a global violation constant $\mathbf{V}$ as the disjunction of all indexed violation constants like $V(n,a)$, i.e., $\mathbf{V} = \bigvee_{n \in N, a \in A} V(n,a)$.

Definition 10 (Norm creation). *Let* $NMAS = \langle A, X, D, G, AD, E, MD, \geq, N, V \rangle$ *with* $Y \rightarrow g \in MD(G)$. *The parameters contain the global violation constant* $\mathbf{V} \in P$ *and E contains the following set of rules:*
$\{V(n,a) \rightarrow \mathbf{V} \mid n \in N, a \in A\} \cup \{\neg\mathbf{V} \rightarrow \neg V(n,a) \mid n \in N, a \in A\}$
The creation of norm n' *to achieve joint goal* $Y \rightarrow g$ *leads to the updated normative multiagent system* $\langle A, X, D, G, AD, E \cup E', MD, \geq, N \cup \{n'\}, V \rangle$ *such that:*

1. *The norm* n' *is not already part of N;*
2. *A set of rules* $E' = \{Y \wedge x \rightarrow V(n', a) \mid a \in A, x \in Lit(X)\}$ *is a set of obligations for each* $a \in A$ *such that* $E \cup E' \vdash_{iol} \neg\mathbf{V} \wedge Y \rightarrow g$, *if all norms are fulfilled, then the joint goal is satisfied;*
3. $cons(E \mid Y \wedge \neg\mathbf{V})$, *it is possible that no norm is violated.*

The creation of norms is illustrated by the following example.

Example 3. Let $NMAS = \langle A, X, D, G, AD, E, MD, \geq, N, V \rangle$ as defined in Example 2. Assume that the normative system creates a norm n' with the following obligations: $\forall a \in A : \neg right\_side(a) \rightarrow V(n', a)$: $\neg right\_side(a)$ counts as a violation of norm n' by agent a.

The second step is adding sanctions and rewards. The condition of this second step is that sanctions are disliked, and rewards are desired.

Definition 11 (Norm creation with sanctions and rewards). *Let NMAS =be a normative multiagent system* $\langle A, X, D, G, AD, E, MD, \geq, N, V \rangle$ *with* $Y \rightarrow g \in MD(G)$.

The creation of norm n' with sanctions and rewards to achieve joint goal $Y \rightarrow g$ leads to the updated system $\langle A, X, D, G, AD, E \cup E' \cup E'', MD, \geq, N \cup \{n'\}, V\rangle$ with:

1. *The creation of norm n' to achieve joint goal $Y \rightarrow g$ leads to updated system $\langle A, X, D, G, AD, E \cup E', MD, \geq, N \cup \{n'\}, V\rangle$ and*
2. *The set of rules $E'' = \{Y \wedge V(n', a) \rightarrow s \mid a \in A, s \in Lit(X)\} \cup \{Y \wedge \neg V(n', a) \rightarrow r \mid a \in A, r \in Lit(X)\}$ is a set of sanctions and rewards for each $a \in A$ such that for all such s and r we have $D_a \vdash_{iol} Y \rightarrow \neg s$ or $D \vdash_{iol} Y \rightarrow r$: sanctions are undesired and rewards are desired.*

Sanctions are illustrated by the following example.

Example 4. Let $NMAS = \langle A, X, D, G, AD, E, MD, \geq, N, V\rangle$ with the following ingredients:

agents A: $\{a, b\}$;
variables in X:
$\{c(a), c(b)$, *cooperation*, $s(a), s(b)\}$ with $X_a = \{c(a)\}$, $X_b = \{c(b)\}$, each agent can cooperate (e.g., $c(a)$) or not, each agent can be sanctioned (e.g., $s(a)$) or not.
effect rules E.
$\{c(a) \wedge c(b) \rightarrow cooperation\}$, there is cooperation if both agents cooperate.
desires D: $D_a = \{\top \rightarrow \neg c(a), \top \rightarrow cooperation, \top \rightarrow \neg s(a)\}$,
$D_b = \{\top \rightarrow \neg c(b), \top \rightarrow cooperation, \top \rightarrow \neg s(b)\}$.
Agents desire to defect (e.g., $\neg c(a)$), but they also desire cooperation, and they desire not to be sanctioned.
goal $G = \{\top \rightarrow cooperation\}$, the system has generated a joint goal of *NMAS* for cooperation.

Assume that the normative system creates a norm n' with the following obligations: $E' = \{\neg c(a) \rightarrow V(n', a) \mid a \in A\}$: $\neg c(a)$ counts as a violation of norm n' by agent a. Moreover, it adds the following sanctions: $E'' = \{V(n', a) \rightarrow s(a) \mid a \in A\}$.

There may be a third step that adds controls to the obligations, sanctions and rewards. We do not consider this extension in this paper.

7 Norm Acceptance

An agent accepts a norm when the obligation implies the desires the cycle started with, and moreover, it believes that the other agents will fulfill their obligations. We propose the following games: agent a plays a game with arbitrary agent b and accepts the norm if agent b fulfills the norm *given that all other agents fulfill the norm,* and this fulfillment leads to fulfillment of its personal desire the cycle started with. This implies that fulfillment of the goal g is kind of normative equilibrium.

Definition 12 (Decision). *Let $NMAS = \langle A, X, D, G, AD, E, MD, \geq, N, V\rangle$. The optimal decision of agent $b \in A$ given a set of literals C is defined as follows.*

- *The set of decisions is the set of subsets of $Lit(X_b)$ that do not contain a variable and its negation. A decision δ is complete if it contains, for each variable in X_b, either this variable or its negation.*

- *The unfulfilled desires of decision δ for agent $b \in A$ are the desires whose body is part of the decision, but whose head is not.*
  $U(\delta, b) = \{d \in D_b \mid MD(d) = L \rightarrow l, E \vdash_{iol} C \cup \delta \rightarrow l'$ *for* $l' \in L$ *and* $E \nvdash_{iol} C \cup \delta \rightarrow l\}$.
- *A decision δ is* optimal *for agent b if and only if there is no decision δ' such that* $U(\delta, b) >_b U(\delta', b)$.

We use the definition of optimal decision to define the acceptance relation. We define a variant $\mathbf{V}_{\sim b}$ of the global violation constant $\mathbf{V}$ as the disjunction of the violation constants of all agents except agent b. We assume here that the agents only consider typical cases. In reality there are always exceptions to the norm, but we do not take this into account.

Definition 13 (Acceptance). *Let* $NMAS = \langle A, X, D, G, AD, E, MD, \geq, N, V\rangle$, *and let* $NMAS' = \langle A, X, D, G, AD, E \cup E' \cup E'', MD, \geq, N \cup \{n'\}, V\rangle$ *be the system after the creation of a norm and its associated sanctions and rewards. The parameters contain the global violation constants* $\mathbf{V}_{\sim b} \in P$ *and E contains the following rules:*
$\{V(n, x) \rightarrow \mathbf{V}_{\sim b} \mid n \in N, x \in A \setminus \{b\}\} \cup \{\neg\mathbf{V}_{\sim b} \rightarrow \neg V(n, a) \mid n \in N, x \in A \setminus \{b\}\}$
An agent $a \in A$ accepts the norm if:

1. *There is a desire in D which is not satisfied in NMAS, but it is satisfied in $NMAS'$.*
2. *For all other agents $b \in A$, we have that the optimal decision of agent b assuming $\neg\mathbf{V}_{\sim b}$ implies $\neg\mathbf{V}$.*

Norms do not always need to be accepted in order to be fulfilled, since the sanction provides a motivation to the agents. However, for a norm to be really effective must be respected due to its acceptance, and not only due to fear of sanctions.

It can easily be shown that in the two running examples, both norms are accepted.

8 Further Research

8.1 Trust

For more realistic but also more complex social trust, we have to enrich the model with beliefs. We have to extend the merging operators to merging in the context of beliefs, see [15]. Consequently, we have to introduce beliefs in norm creation, and we have to make the acceptance relation relative to beliefs.

8.2 The Creation of Permissive Norms

It is not directly clear how the social delegation cycle can explain the creation of permissive norms. One way to proceed is to define permissions as exceptions within hierarchical normative systems [12].

8.3 Social Institutions and the Creation of Constitutive Norms

How to take social institutions into account in the social delegation cycle? Based on Searle's construction of social reality, we may introduce besides the obligations or regulative norms also constitutive norms, which are definitions of the normative system based on a counts-as conditional [9].

9 Related Work

9.1 Other Work

The relation between 'desires' or internal motivations and 'obligations' or external motivations has been studied in many areas, for example:

Religion. The Golden Rule or the ethic of reciprocity is found in the scriptures of nearly every religion. It is often regarded as the most concise and general principle of ethics. It is a condensation in one principle of all longer lists of ordinances such as the Decalogue.

Ethics. Kant's categorical imperative [17] expresses the moral law as ultimately enacted by reason and demanding obedience from mere respect for reason.

Political theory. Marx (ideology) [23]: the ruling class forms a theory (obligations) justifying itself (its desires).

Social theory. Norms (obligations) are only accepted if the legislator does not make them only for his own interests (desires) ([14]).

Agent theory. Your wish is my command: the desires of the master are the obligations of the slave.

Within formal and semi-formal agent theory, there has been some work by Castelfranchi, Conte and colleagues on norm adoption and norm acceptance [14].

9.2 Normative System as an Agent

In other work we discuss applications of normative multiagent systems [5], of which the formal machinery based on rule based systems and input/output logics has been developed in various papers. In those papers the agents consider the normative system as an agent, and they attribute mental attitudes to it, because the agents are playing games with the normative system to determine whether to fulfill or violate norms. We refer to this use of the agent metaphor as "your wish is my command": the goals of the normative agent are the obligations of the normal agents. In the present paper, however, the agents play games with other agents, and the attribution of mental attitudes to normative system is not a necessary assumption. In our other work, we have informally discussed the notion of the social delegation cycle in a short paper [4]. In that short paper we have suggested that the social delegation cycle can be used to explain the agent metaphor "your wish is my command", because the group goal from which the norm is created, may be interpreted as the goal of the normative system, and the normative system is doing a kind of planning.

In this framework, we have not discussed the merging of desires into group goals, but we have mentioned the notion of rational norm creation in a second short paper [7]. In that paper we do not present a formalization of norm creation, and we do not consider norm creation within the context of the social delegation cycle. Finally, we introduce an extension of our formal model with constitutive norms in [9] and we observe that constitutive norms play an important role in norm creation, but we do not formally study it. The creation of permissions in this framework has been mentioned in [6].

Finally, in none of our other work we have discussed the acceptance relation, and we have not discussed games between ordinary agents.

10 Summary

In this paper we consider the relation between desires and obligations in normative multiagent systems. We introduce a model of their relation based on what we call the social delegation cycle, which explains the creation of norms from agent desires in three steps. First individual agent desires generate group goals, then a group goal is individualized in a social norm, and finally the norm is accepted by the agents when it leads to the fulfillment of their initial desires. The social delegation cycle may be seen as a generalization of single agent decision making, which can also be defined as a combination of goal generation and planning. Additional issues in the social delegation cycle are the role of sanctions and rewards, the acceptance relation, and the implicit assumption of fairness in goal generation. Moreover, in the social delegation cycle institutions may play a role.

We formalize the social delegation cycle combining theories developed in a generalization of belief revision called merging operators, planning and game theory. First, we formalize joint goal generation as a merging process of the individual agent desires, for which we extend existing merging operators to deal with rules. Second, we formalize norm creation as a planning process for both the obligation and the associated sanctions or rewards. Third, we formalize the acceptance relation as both a belief of agents that the norm leads to achievement of their desires, and the belief that other agents will act according to the norm, introducing a notion of normative equilibrium which states that agents fulfill norms when other agents do so.

There are two main directions for further research. First, the theories have to be extended with beliefs and institutions to cover social delegation cycles based on trust and norm creation by institutions. Second, for the social delegation cycle efficient mechanisms should be designed which can be employed in actual implementations of normative multiagent systems. Desirable properties may be soundness (compliance with our framework), completeness (for each possible goal there is a goal generated), conciseness of goals and norms generated, generality of goals and norms generated, strictness of goal generation and norm creation, *et cetera.*

References

1. C. Alchourron, P. Gärdenfors, and D. Makinson. On the logic of theory change. *Journal of Symbolic Logic,* 50(2):510–530, 1985.
2. A. Anderson. A reduction of deontic logic to alethic modal logic. *Mind,* 67:100–103, 1958.
3. K.J. Arrow. *Social choice and individual values.* Wiley, New York, second edition, 1963.
4. G. Boella and L. van der Torre. Attributing mental attitudes to normative systems. In *Procs. of AAMAS'03,* pages 942–943. ACM Press, 2003.
5. G. Boella and L. van der Torre. Local policies for the control of virtual communities. In *Procs. of IEEE/WIC Web Intelligence Conference,* pages 161–167. IEEE Press, 2003.
6. G. Boella and L. van der Torre. Permissions and obligations in hierarchical normative systems. In *Procs. of ICAIL'03,* pages 109–118, Edinburgh, 2003. ACM Press.
7. G. Boella and L. van der Torre. Rational norm creation: Attributing mental attitudes to normative systems, part 2. In *Procs. of ICAIL'03,* pages 81–82, Edinburgh, 2003. ACM Press.
8. G. Boella and L. van der Torre. Contracts as legal institutions in organizations of autonomous agents. In *Procs. of AAMAS'04,* New York, 2004.

9. G. Boella and L. van der Torre. Regulative and constitutive norms in normative multiagent systems. In *Procs. of 9th International Conference on the Principles of Knowledge Representation and Reasoning,* Whistler (CA), 2004.
10. M.E. Bratman. *Intentions, plans, and practical reason.* Harvard University Press, Harvard (Massachusetts), 1987.
11. J. Broersen, M. Dastani, J. Hulstijn, and L. van der Torre. Goal generation in the BOID architecture. *Cognitive Science Quarterly,* 2(3-4):428–447, 2002.
12. E. Bulygin. Permissive norms and normative systems. In A. Martino and F. Socci Natali, editors, *Automated Analysis of Legal Texts,* pages 211–218. Publishing Company, Amsterdam, 1986.
13. P. R. Cohen and H. J. Levesque. Intention is choice with commitment. *Artificial Intelligence,* 42:213–261, 1990.
14. R. Conte, C. Castelfranchi, and F. Dignum. Autonomous norm-acceptance. In *Intelligent Agents V(ATAL'98),* volume 1555 of *LNAI,* pages 319–333. Springer, 1999.
15. M. Dastani and L. van der Torre. Specifying the merging of desires into goals in the context of beliefs. In *Procs. of The First Eurasian Conference on Advances in Information and Communication Technology (EurAsia ICT'02),* volume 2510 of *LNCS,* pages 824–831. Springer, 2002.
16. P. Gärdenfors. *Knowledge in flux.* MIT Press, Cambridge (Massachusetts), 1988.
17. I. Kant. *Kritik der praktischen Vernunft.* Johann Friedrich Hartnoch, Riga, 1788.
18. H. Katsuno and A.O. Mendelzon. Propositional knowledge base revision and minimal *change. Artificial Intelligence,* 52:263–294, 1991.
19. S. Konieczny and R.P. Pérez. On the frontier between arbitration and majority. In *Procs. of 8th International Conference on the Principles of Knowledge Representation and Reasoning (KR'02),* pages 109–120, Toulouse, 2002.
20. J. Lang. Conditional desires and utilities - an alternative approach to qualitative decision theory. In *Procs. of ECAI'96,* pages 318–322, Budapest, 1996.
21. D. Makinson and L. van der Torre. Input-output logics. *Journal of Philosophical Logic,* 29:383–408, 2000.
22. D. Makinson and L. van der Torre. Constraints for input-output logics. *Journal of Philosophical Logic,* 30(2): 155–185, 2001.
23. K. Marx. *Das Kapital.* Verlag von Otto Meissner, Hamburg, 1867.
24. J. J. Ch. Meyer. A different approach to deontic logic: Deontic logic viewed as a variant of dynamic logic. *Notre Dame Journal of Formal Logic,* 29(1): 109–136, 1988.
25. A. S. Rao and M. Georgeff. The semantics of intention maintenance for rational agents. In *Procs. of 14th International Joint Conference on Artificial Intelligence (IJCAI'95),* pages 704–710, Montreal, 1995.
26. P.Z. Revesz. On the semantics of arbitration. *International Journal of Algebra and Computation,* 7(2): 133–160, 1997.
27. J.R. Searle. *The Construction of Social Reality.* The Free Press, New York, 1995.

Designing a Deontic Logic of Deadlines

Jan Broersen, Frank Dignum, Virginia Dignum, and John-Jules Ch. Meyer

Institute of Information and Computing Sciences
Utrecht University,
P.O.Box 80.089, 3508 TB Utrecht, The Netherlands
{broersen,dignum,Virginia,jj}@cs.uu.nl

Abstract. This paper studies the logic of a dyadic modal operator for being obliged to meet a condition ρ before a condition δ becomes true. Starting from basic intuitions we arrive at a simple semantics for deadline obligations in terms of branching time models. We show that this notion of deadline obligation can be characterized in the branching time logic CTL. The defined operator obeys intuitive logic properties, like monotony w.r.t. ρ and anti-monotony w.r.t. δ, and avoids some counter-intuitive properties like agglomeration w.r.t ρ and 'weak agglomeration' w.r.t. δ. However, obligations of this type are implied by the actual achievement of ρ before the deadline. We argue that this problem is caused by the fact that we model the obligation only from the point of view of its violation conditions. We show that the property might be eliminated by considering success conditions also.

1 Introduction

This paper studies the logic of a dyadic modal operator, denoted $O(\rho \leq \delta)$, for being obliged to meet a condition ρ before a condition δ becomes true. To satisfy the obligation, it suffices to satisfy the condition ρ only once, at a time of ones choosing, as long as it is *before* (or, ultimately, *at*) the point where the condition δ occurs. We refer to the operator $O(\rho \leq \delta)$ as a 'deontic deadline' operator. We do not claim that all deadlines have a deontic aspect. For instance, in the field of 'scheduling' [1], deadlines are hard constraints that have to be satisfied under all circumstances. However, in more realistic situations, where agents may choose to violate deadlines imposed on them by other agents, it is much harder to deny the deontic aspect[1].

Conceptually, deontic deadlines are interactions between two dimensions: a deontic (normative) dimension and a temporal dimension. So, to study deadlines, it makes sense to take a standard temporal logic, say CTL [2–4], and a standard deontic logic, say SDL [5], and combine the two in one system. This type of approach is for instance taken in [6]. The problem then is how to account for the interactions. Conceptually, we have to keep in mind that in the combined system we can express that the normative content of the deontic operators can

[1] If deadlines are not due to commitments towards other agents, but the result of personal decisions based on personal desires, it is more adequate to talk about 'deadline intentions'.

A. Lomuscio and D. Nute (Eds.): DEON 2004, LNAI 3065, pp. 43–56, 2004.

be temporal (e.g, being obliged to be polite *always*), but also that obligations can have some (non-)dynamical behavior over time (e.g., *always* being obliged to be polite). It is easy to mix up these essentially different propositions. The same kind of confusion threatens the study of deadlines. Is a deadline (1) an obligation *at* a certain point in time to achieve something *before* another point in time, or (2) is a deadline simply an obligation that persists in time until a deadline is reached, or (3) is it both? In natural language it is actually quite hard to be precise about this distinction. Therefore, for now, we rely on in informal understanding of the branching time temporal logic CTL, and the standard deontic logic SDL, to discuss the distinction using formulas. In CTL, the symbols E and A denote an existential and a universal quantifier respectively, ranging over possible future courses (branches) of time. Within the scope of these quantifiers, CTL uses the linear time operators (LTL [3]) $\varphi U \psi$ (strong Until, i.e., ψ will occur, and φ holds up until then), $\varphi U_w \psi$ (weak Until, i.e., if ψ will occur, then φ holds up until then) to talk about individual future courses of time (from now on simply called 'possible futures'). From SDL we use the operator O, for obligation.

In a language that combines CTL and SDL we can talk about both the temporal and deontic dimensions independently. For instance, we can talk about a certain obligation being preserved over time: $A(O\rho U_w \delta)$, which says 'the obligation to achieve ρ persists until δ, and if δ does not occur, it persists forever'. Or we can talk about the obligation that a certain condition ρ has to be achieved before a condition δ occurs: $O(\neg E(\neg \rho U \delta))$. This says 'it is obliged that on no possible future ρ is avoided until δ becomes true' (alternatively we may read this as 'it is forbidden that on some possible future ρ is avoided up until δ becomes true').

However, in this paper we do not use a language where we can talk about both dimensions independently. We see the deontic dimension as 'embedded' in the temporal dimension[2], the only difference being that it is considered exclusively with certain violation [7] and ideality [8] constants that hold or do not hold at certain points in time.

The advantage of this approach is that we study deadlines entirely in CTL supplemented with violation[3] and ideality constants. The disadvantage is that the language is not expressive enough to talk about the deontic and temporal dimensions independently. In particular, we cannot talk about the dynamics of obligations. So, a background assumption of our study will be that agents do not get new obligations, or are explicitly discharged of some of their obligations, when time evolves. In yet other words: there are no explicit 'deontic updates'. This implies that if in a next state the deadline δ or the achievement ρ is not realized, the deadline obligation persists. In sections 4 and 5, we present formulas that correspond to this background assumption for two different version of the deadline operator.

[2] Technically this corresponds with the deontic accessibility relation being enclosed by the temporal accessibility relation.

[3] The idea of expressing the semantics of deontic deadlines by characterizing violation conditions in CTL supplemented with violation constants [7], was first explored in [9].

We model the deadlines themselves as propositions. This seems a reasonable choice given that we do not want to model a deadline in a logic of explicit time (real time). Our view is more abstract, and a deadline is simply a condition δ true at some point in time. A consequence of this abstract view is that we have to deal with the possibility that δ actually never occurs. Note that for a theory of deadlines that uses an explicit notion of time, this would never be a problem. In particular, the point 'two hours from now' will always occur, while meeting a condition 'δ' may be impossible or extremely unlikely. However, our abstract view contributes to the relevance of the present research for other logical systems. For instance, Rao and Georgeff's commitment strategies [10] are actually a sort of deadlines: an agent commits to an intention until the action is performed or believed not to be feasible any longer.

The choice in this paper for the temporal logic CTL is a pragmatic one. We believe the theory applies equally well, and maybe better, to linear time temporal logic (LTL [3]). However, CTL has nice properties (P-complete complexity of the model checking problem for CTL, versus PSPACE-complete complexity for LTL [11]), and is popular in agent theory [12].

2 Preliminaries: CTL

A well-formed formula φ of the temporal language $\mathcal{L}_{\mathrm{CTL}}$ is defined by:

$$\varphi, \psi, \ldots := p \mid \neg\varphi \mid \varphi \wedge \psi \mid E(\varphi U^{ee} \psi) \mid A(\varphi U^{ee} \psi)$$

where φ, ψ represent arbitrary well-formed formulas, and where the p are elements from an infinite set of propositional symbols $\mathcal{P}$. We use the superscript 'ee' to denote that this is the version of the 'until' where φ is not required to hold for the present, nor for the point where ψ, i.e., the present and the point where ϕ are *both* excluded. This gives us the following informal meanings of the until operators:

$E(\varphi U^{ee} \psi)$: there is a future for which eventually, at some point m, the condition ψ will hold, while φ holds from the next moment until the moment before m

$A(\varphi U^{ee} \psi)$: for all futures, eventually, at some point m, the condition ψ will hold, while φ holds from the next moment until the moment before m

We define all other CTL-operators as abbreviations[4]. Although we do not use all of the LTL operators *X, F,* and *G* in this paper, we give their abbreviations

[4] Often, the CTL-operators $EG\varphi$ and $E(\psi U \varphi)$ are taken as the basic ones, and other operators are defined in terms of them. The advantage of that approach is that we do not have to use the notion of 'full path', that is crucial for the truth condition of $A(\varphi U^{ee} \psi)$. However, that approach is not applicable here, since we cannot define the 'exclusive' versions of the operators in terms of them. And, even if we take $EG\varphi$ and $E(\psi U^{ee} \varphi)$ as basic, we can still not define the for our purposes important operator $A(\psi U^{e} \varphi)$ as an abbreviation.

(in combination with the path quantifiers E and A) in terms of the defined operators for the sake of completeness. We also assume the standard propositional abbreviations.

$$
\begin{array}{ll}
EX\varphi \equiv_{def} E(\bot U^{ee}\varphi) & AX\varphi \equiv_{def} \neg EX\neg\varphi \\
EF\varphi \equiv_{def} \varphi \vee E(\top U^{ee}\varphi) & AG\varphi \equiv_{def} \neg EF\neg\varphi \\
AF\varphi \equiv_{def} \varphi \vee A(\top U^{ee}\varphi) & EG\varphi \equiv_{def} \neg AF\neg\varphi \\
A(\varphi U^{e}\psi) \equiv_{def} \varphi \wedge A(\varphi U^{ee}\psi) & E(\varphi U^{e}\psi) \equiv_{def} \varphi \wedge E(\varphi U^{ee}\psi) \\
A(\varphi U\psi) \equiv_{def} A(\varphi U^{e}(\varphi \wedge \psi)) & E(\varphi U\psi) \equiv_{def} E(\varphi U^{e}(\varphi \wedge \psi)) \\
A(\varphi U_w\psi) \equiv_{def} \neg E(\neg\psi U\neg\varphi) & E(\varphi U_w\psi) \equiv_{def} \neg A(\neg\psi U\neg\varphi)
\end{array}
$$

The informal meanings of the formulas with a universal path quantifier are as follows (the informal meanings for the versions with an existential path quantifier follow trivially):

$A(\varphi U^{e}\psi)$: for all futures, eventually, at some point m, the condition ψ will hold, while φ holds from now until the moment before m
$A(\varphi U\psi)$: for all futures, eventually, at some point the condition ψ will hold, while φ holds from now until then
$A(\varphi U_w\psi)$: for all possible futures, if eventually ψ will hold, then φ holds from now until then, or forever otherwise
$AX\varphi$: at any next moment φ will hold
$AF\varphi$: for all futures, eventually φ will hold
$AG\varphi$: for all possible futures φ holds globally

A CTL model $\mathcal{M} = (S, \mathcal{R}, \pi)$, consists of a non-empty set S of states, an accessibility relation $\mathcal{R}$, and an interpretation function π for propositional atoms. A full path σ in M is a sequence $\sigma = s_0, s_1, s_2, \ldots$ such that for every $i \geq 0$, s_i is an element of S and $s_i \mathcal{R} s_{i+1}$, and if σ is finite with s_n its final situation, then there is no situation s_{n+1} in S such that $s_n \mathcal{R} s_{n+1}$. We say that the full path σ starts at s if and only if $s_0 = s$. We denote the state s_i of a full path $\sigma = s_0, s_1, s_2, \ldots$ in $\mathcal{M}$ by σ_i. Validity $\mathcal{M}, s \models \varphi$, of a CTL-formula φ in a world s of a model $\mathcal{M} = (S, \mathcal{R}, \pi)$ is defined as:

$\mathcal{M}, s \models p \quad \Leftrightarrow s \in \pi(p)$
$\mathcal{M}, s \models \neg\varphi \quad \Leftrightarrow$ not $\mathcal{M}, s \models \varphi$
$\mathcal{M}, s \models \varphi \wedge \psi \quad \Leftrightarrow \mathcal{M}, s \models \varphi$ and $\mathcal{M}, s \models \psi$
$\mathcal{M}, s \models E(\varphi U^{ee}\psi) \Leftrightarrow \exists\sigma$ in $\mathcal{M}$ with $\sigma_0 = s$, and $\exists n > 0$ such that:
(1) $\mathcal{M}, \sigma_n \models \psi$ and
(2) $\forall i$ with $0 < i < n$ it holds that $\mathcal{M}, \sigma_i \models \varphi$
$\mathcal{M}, s \models A(\varphi U^{ee}\psi) \Leftrightarrow \forall\sigma$ in $\mathcal{M}$ such that $\sigma_0 = s$, it holds that $\exists n > 0$ such that
(1) $\mathcal{M}, \sigma_n \models \psi$ and
(2) $\forall i$ with $0 < i < n$ it holds that $\mathcal{M}, \sigma_i \models \varphi$

Validity on a CTL model $\mathcal{M}$ is defined as validity in all states of the model. If φ is valid on a CTL model $\mathcal{M}$, we say that $\mathcal{M}$ is a model for φ. General validity of a formula φ is defined as validity on all CTL models. The logic CTL is the set of all general validities of $\mathcal{L}_{\mathrm{CTL}}$ over the class of CTL models.

3 A Dyadic Deontic Deadline Operator

We minimally extend the language $\mathcal{L}_{\mathrm{CTL}}$ by extending the set of propositional atoms with a violation constant of the form *Viol*[5]. Furthermore, the formal interpretation of the atom *Viol* is treated like that of all other atomic propositions. So, we can view the propositional constant *Viol* also as a special element of $\mathcal{P}$: a 'special purpose' proposition solely used to interpret deontic formulas in a temporal dimension.

Let $\mathcal{M}$ be a CTL model, s a state, and σ a full path starting at s. A straightforward modal semantics for the operator $O^V(\rho \leq \delta)$, where the V is only a label to emphasize that this operator is defined in terms of *V*iolations, is then defined as follows:

$$\begin{aligned}\mathcal{M}, s \models O^V(\rho \leq \delta) \Leftrightarrow\ & \forall \sigma \text{ with } \sigma_0 = s, \forall j : \\ & \text{if} \\ & \mathcal{M}, \sigma_j \models \delta \text{ and } \forall 0 \leq i \leq j : \mathcal{M}, \sigma_i \models \neg\rho \\ & \text{then} \\ & \mathcal{M}, \sigma_j \models Viol\end{aligned}$$

This says: if at some future point the deadline occurs, and until then the result has not yet been achieved, then we have a violation at that point. This semantic definition is equivalent to the following definition as a reduction to CTL:

$$O^V(\rho \leq \delta) \equiv_{def} \neg E(\neg\rho U(\delta \wedge \neg Viol))$$

This formula simply 'negates' the situation that should be excluded when a deontic deadline is in force[6]. In natural language this *negative* situation is: 'δ becomes true at a certain point, the achievement has not been met until then, and there is *no* violation at δ'. Therefore this CTL formula exactly characterizes the truth condition for the above defined deontic deadline operator: the semantic conditions are true is some state if and only if the the CTL formula is true in that state.

4 Logical Properties

What logical properties of the operator $O^V(\rho \leq \delta)$ does this bring us? We first discus the property that corresponds to out background assumption that there are no deontic updates. It holds that:

$$\models O^V(\rho \leq \delta) \rightarrow A(O^V(\rho \leq \delta) U_w \rho)$$

To see that this holds[7], it is easiest to fall back on the semantics of the operator. The semantics says that on futures (branches of time) where δ occurs

[5] For reasoning in a multi-agent context we may provide violation constants of the form $Viol(a)$ where $a \in \mathcal{A}$, and $\mathcal{A}$ an infinite set of agent identifiers.

[6] Alternatively this definition can be given using the weak until: $O^V(\rho \leq \delta) \equiv_{def} A((\neg\delta \vee Viol) U_w \rho)$. But for the version with the strong until it is much easier to see that it corresponds with the semantic truth conditions defined above.

[7] Alternatively we may write this as $\models O(\rho \leq \delta) \rightarrow \neg E(\neg\rho U \neg O(\rho \leq \delta))$. But in our opinion, here the version with the weak until is easier to understand.

at some point t, while until then ρ has not been done once, there is a violation at t. Now, if we follow such a branch for some time-steps in the future, and we do not meet a ρ, then, the deadline conditions do still apply: still it holds that if δ will occur later on, we get a violation if we do not meet a ρ until then. An important observation is that even if we have passed one or more δ-states, the obligation still applies; only if we meet a ρ, the conditions are no longer guaranteed. Thus, the defined notion of deadline persists, even if we have passed a deadline. This might be considered counter-intuitive, since it seems correct to assume that the deadline obligation itself is discharged by passing the deadline. Therefore, in section 5 we show how to define a version that does not have this property. However, we consider the present notion of deadline not as counter-intuitive, but merely as a variant. Persistence of the obligation at the passing of a deadline is not a priori counter-intuitive. An example is the following: you are obliged to repair the roof of your house before it will rain (or otherwise you and your interior get wet). This obligation is only discarded by the act of repairing the roof, and not by the event of raining.

We now turn to other properties of the operator of section 3. First of all we get monotonicity with respect to ρ (other terminology: validity of the operator is closed under weakening of ρ)[8]. Monotonicity says that we have as validities[9]:

$$\models O^V((\rho \wedge \chi) \leq \delta) \rightarrow O^V(\rho \leq \delta)$$

$$\models O^V(\rho \leq \delta) \rightarrow O^V((\rho \vee \chi) \leq \delta)$$

This is in accordance with intuition: if ρ is made logically weaker, it is easier to satisfy. So, if the stronger condition has to be accomplished before δ occurs, then certainly also the weaker condition has to be accomplished before δ occurs.

A property we do not have is agglomeration with respect to ρ, i.e.:

$$\not\models O^V(\rho \leq \delta) \wedge O^V(\chi \leq \delta) \rightarrow O^V((\rho \wedge \chi) \leq \delta)$$

This shows that $O^V(\rho \leq \delta)$, is monotonic with respect to ρ, but is *not* a normal modal operator with respect to ρ. This means that it is a *strictly* monotonic modal operator with respect to ρ. Exactly this same logic behavior is known from intentions [13]: an intention for p and an intention for q do not necessarily give an intention for $p \wedge q$, because we may intend p for another point in the future then the point for which we intend q. That the behavior of deadline obligations is similar to that of future directed intentions is not unlikely, given the intuition that intentions can be seen as a kind of obligations to oneself. A consequence of the absence of agglomeration is that it is consistent to have

[8] In section 6 we will see a simple way to prove weakening and strengthening for the defined operators. However, all other verifications are left to the reader.

[9] To express the property we call 'monotonicity' it suffices to give just one of these theorems, because they can be derived from each other using only the rules of uniform substitution and substitution by logical equivalents. However, to check the intuitiveness of monotony, especially for deontic operators, it is wise to consider both these 'appearances' of monotony.

$O^V(p \leq \delta) \wedge O^V(\neg p \leq \delta)$. Consistency of obligations of the form Op and $O\neg p$ is heavily debated in deontic logic. Here we have consistency simply because we are free to choose our time of compliance, as long as it is before the deadline.

Also we get that the operator is anti-monotonic with respect to δ (other terminology: validity of the operator is closed under strengthening of δ):

$$\models O^V(\rho \leq \delta) \rightarrow O^V(\rho \leq (\delta \wedge \gamma))$$

$$\models O^V(\rho \leq (\delta \vee \gamma)) \rightarrow O^V(\rho \leq \delta)$$

For this version of the operator, this is in accordance with intuition: if δ is made logically stronger, it is harder to satisfy. And if ρ already has to be accomplished before the weaker condition occurs, it will certainly have to be accomplished before the stronger condition occurs. This property does not go through for the version of the deadline we discuss in section 5. As said, in that variant, the obligation is discharged by the first condition δ we meet. Then, by strengthening δ, it is not necessarily the case that we preserve the obligation.

A property we do not have for $O^V(\rho \leq \delta)$ is 'weak agglomeration' with respect to ρ, i.e.:

$$\not\models O^V(\rho \leq \delta) \wedge O^V(\rho \leq \gamma) \rightarrow O^V(\rho \leq (\delta \vee \gamma))$$

This means that the deontic deadline operator $O^V(\rho \leq \delta)$, is strictly anti-monotonic with respect to δ. If it would also obey weak agglomeration, it would have been a window operator [14,15], which means that it would have been a normal modal operator with respect to $\neg\delta$ [16,17]. However, the operator is *strictly* anti-monotonic. This is intuitive. Weak agglomeration should not hold, because having to achieve something before tomorrow and having to achieve the same thing before the end of the day does not imply that I have the choice to do it before tomorrow *or* before the end of the day: it simply gives me no other choice than to do it before the end of the day.

The combination of monotony for ρ and anti-monotony for δ gives us the following transitivity property for the deontic deadline operator $O^V(\rho \leq \delta)$ [10]:

$$\models O^V(\rho \leq \delta) \wedge O^V(\delta \leq \gamma) \rightarrow O^V(\rho \leq \gamma)$$

Also this property is intuitive: if an agent is obliged to brush his teeth before going to bed, and take a medicine before he brushes his teeth, then he is certainly obliged to take has medicine before going to bed.

Clearly, the deadline operator should not be symmetric. Indeed we have:

$$\not\models O^V(\rho \leq \delta) \rightarrow O^V(\delta \leq \rho)$$

Another property we do obey is reflexivity:

$$\models O^V(\gamma \leq \gamma)$$

[10] Taking advantage of the definability in CTL, it can be shown that we actually obey a stronger version of this property: $\models O^V(\rho \leq \delta) \wedge AG(\delta \rightarrow \chi) \wedge O^V(\chi \leq \gamma) \rightarrow O^V(\rho \leq \gamma)$

This is exactly the reason why we use the symbol '$\leq$' and not the symbol '$<$', in the denotation for the operator. If we achieve the obliged condition *at* the point of the deadline, we are still in time. In particular, if the deadline condition itself coincides with the condition we are obliged to achieve, whatever this condition is, we are always 'just in time' to meet the deadline. However, some would say that it is counter-intuitive to actually always be *obliged* to achieve any γ up and until γ.

We might argue that the situation is comparable to the axiom $O\top$ of standard deontic logic. The common denominator of these properties is that they concern an obligation for something that actually can never be violated. The point is that although it seems strange that our logic validates obligations for things that cannot be violated, it is not harmful either. No agent will ever let his decision making be influenced by obligations for things that are true inevitably and always. In other words, such obligations are void. However, we will see in section 7 a solution to a related, but more serious problem will discard this property, which means that we no longer have to defend it by saying that it is counter-intuitive but harmless.

Let us now consider the related issue of having a tautology or contradiction as deadline, or as a condition to achieve. We first consider the case where ρ equals $\top$. We have that:

$$\models O^V(\top \leq \delta)$$

This is related to the monotony with respect to ρ; we can weaken ρ up until it coincides with $\top$. This situation is similar to standard deontic logic's $O\top$, which we already discussed.

Just like we can weaken ρ up until $\top$, we can strengthen δ up until $\bot$ (from the anti-monotony with respect to δ).

$$\models O^V(\rho \leq \bot)$$

Clearly, $\bot$ is a condition that will be never met. So, an obligation to perform something before the (absent) point that $\bot$, can never be violated. We can postpone the obligation forever, without ever falling pray to a violation. In our view, such obligations are void. Therefore, also this case is similar to standard deontic logic's $O\top$.

Another issues is the case where ρ equals $\bot$ or δ equals $\top$. These conditions deserve extra attention. First we discuss the case where ρ equals $\bot$. This concerns the question whether something general holds for obligations for conditions that under no circumstance can be achieved. One view is that obligations of the form $O(\bot \leq \delta)$ are impossible or inconsistent. After all, it seems reasonable to take the position that one can never be obliged to achieve the impossible. This view would demand that we validate $\neg O^V(\bot \leq \delta)$, which is similar to standard deontic logic's D-axiom $\neg O\bot$. However, it is clear that we do not validate $\neg O^V(\bot \leq \delta)$. In our semantics, this would mean that we validate $EF(\delta \wedge \neg Viol)$, which directly contradicts our intuitions: it is not the case that any condition δ will be met eventually. But this does not answer the question whether we *should* obey $\neg O^V(\bot \leq \delta)$. We belief we should not. Note first that our setting is weaker than

that of standard deontic logic. In particular, since we do not have agglomeration, we can satisfy $O^V(p \leq \delta) \wedge O^V(\neg p \leq \delta)$. This simply says that before δ, we have to satisfy p at some point, and we have to satisfy $\neg p$ at some point. That this cannot be the same point does not exclude the conjunction. However, this does not yet explain why it is not excluded that we satisfy $O^V(\bot \leq \delta)$. This is because this is no ordinary obligation but a deadline obligation. As we already discussed, we can have that the deadline itself is a condition that can never occur. And we argued that for that situation, the obligation is trivially met. But then we can also satisfy the formula $O^V(\bot \leq \delta)$ by choosing $\bot$ for δ. We get $O^V(\bot \leq \bot)$, which is not only satisfiable, but also valid. So, obligations of the form $O^V(\bot \leq \delta)$ are not inconsistent; in particular they can be met if δ never occurs. Intuitively: an agent can consistently meet up to the obligation to do something impossible before δ just in case that δ will never occur. Analogously, we can discuss the case where δ equals $\top$. Now the agent is obliged to achieve ρ *now*. In our semantics this is possible. Therefore $O^V(\rho \leq \top)$ is satisfiable. Similar to the above case, we may even choose ρ to be $\top$ to get the valid formula $O^V(\top \leq \top)$, which says: an agent is obliged to obey a tautology now.

However, from the above discussion, it follows that there is a deadline obligation that really should be inconsistent: $O^V(\bot \leq \top)$: agents cannot achieve the impossible now, since, by definition, the present state is not an impossibility. And indeed, we have the following property:

$$\models \neg O^V(\bot \leq \top)$$

5 A Variant without Strengthening of the Deadline Condition

The deadlines as discussed in section 3 are not discarded by meeting the deadline: as long as the condition ρ is not yet achieved, we have a violation at every point where the deadline condition δ holds. In other words: the obligation is not discarded by having failed a deadline. Here we drop this property. Thus the obligation is dropped the first time we meet the deadline condition, irrespective of whether we have achieved the goal or not. In the definition of section 3, we need to add that only the first δ occurring, is relevant.

$$\begin{aligned}\mathcal{M}, s \models O'^V(\rho \leq \delta) \Leftrightarrow \forall \sigma \text{ with } \sigma_0 = s, \forall j : \\ &\text{if} \\ &\mathcal{M}, \sigma_j \models \neg\rho \wedge \delta \text{ and } \forall 0 \leq i < j : \mathcal{M}, \sigma_i \models \neg\rho \wedge \neg\delta \\ &\text{then} \\ &\mathcal{M}, \sigma_j \models Viol\end{aligned}$$

This says: if at some future point the deadline occurs for the first time, and until then the result has not yet been achieved, then we have a violation at that point. For this notion of deadline it is a slightly harder to give a CTL characterization. We need to use the notion of until that talks about the states until the last state before φ (i.e., $\psi U^e \varphi$).

$$O'^V(\rho \leq \delta) \equiv_{def} \neg E((\neg\rho \wedge \neg\delta) U^e (\delta \wedge \neg\rho \wedge \neg Viol))$$

The main point of this variant is thus that it has a different dynamical behavior. In particular, it is discarded by the first δ, even if the achievement has not been met. Therefore, the following preservation property holds:

$$\models O'^{V}(\rho \leq \delta) \rightarrow A(O'^{V}(\rho \leq \delta) U_w (\rho \vee \delta))$$

For this variant all logical properties of the variant of section 3 hold, except strengthening. Thus:

$$\not\models O'^{V}(\rho \leq \delta) \rightarrow O'^{V}(\rho \leq (\delta \wedge \gamma))$$
$$\not\models O'^{V}(\rho \leq (\delta \vee \gamma)) \rightarrow O'^{V}(\rho \leq \delta)$$

It is clear that the following holds for the relation between the two variants:

$$\models O^{V}(\rho \leq \delta) \rightarrow O'^{V}(\rho \leq \delta)$$

6 A Counter-Intuitive Logical Property

The operators defined in sections 3 and 5 obey intuitive properties. However, there is a property, or more precise, a class of properties, which are satisfied by it, but whose intuitiveness is disputable. These possibly counter-intuitive properties are caused by the definition of a deadline from the viewpoint of its violation conditions only. The idea behind the definitions was 'give an exact temporal characterization of the conditions under which the deadline is *violated*'. This idea is correct as long as we are interested in the temporal conditions *implied by* a deontic deadline. But what about the temporal conditions that *give rise to* a deontic deadline? It turns out that here something might be missing. For instance, we have the following property (From now on we will only consider the first version of the operator. The discussion for the other version is analogous.):

$$\models \rho \rightarrow O^{V}(\rho \leq \delta)$$

It says that the deadline obligation of section 3 is implied by the actual achievement of ρ in the current state. Moreover, this property is only an instance of a more general, stronger property that holds for the deontic deadline operator of section 3. The obligation is valid in any state where it is sure that the deadline will be met. In particular:

$$\models \neg E(\neg \rho U \delta) \rightarrow O^{V}(\rho \leq \delta)$$

This can be verified by substituting the CTL characterization of the deadline obligation: $\neg E(\neg \rho U \delta) \rightarrow \neg E(\neg \rho U (\delta \wedge \neg Viol))$. We may see this as the strengthening of δ to $\delta \wedge \neg Viol$ in the schema $\neg E(\neg \rho U \delta)$. It is quite easy to see that this strengthening property holds. We start with the fact that validity of the the schema $E(\varphi U \psi)$ is closed under *weakening* with respect to φ and with respect to ψ, that is, if at some point in a model we satisfy $E(\varphi U \psi)$, we also satisfy both

$E((\varphi \vee \gamma)U\psi)$ and $E(\varphi U(\psi \vee \gamma))$. But this means[11] that the schema $\neg E(\varphi U\psi)$ is closed under *strengthening* with respect to ψ, which is what we needed to show (with $\neg\rho$ substituted for φ, and δ for ψ).

Now the question rises whether we cannot defend intuitiveness of this property in the same way as we defended intuitiveness of, for instance $\models O^V(\gamma \leq \gamma)$ and $\models O^V(\top \leq \delta)$ and $\models O^V(\rho \leq \bot)$. We might argue that if ρ is unavoidable, in particular, if it is true now, then the deadline $O^V(\rho \leq \delta)$ is void, because it concerns an achievement that is met anyway.

However, we consider the issue whether or not $\rho \rightarrow O^V(\rho \leq \delta)$ to be different from, for instance, the issue whether or not $O^V(\top \leq \delta)$. Whereas the second obligation is void because the obligation concerns a tautology, i.e., something that is considered to be true inevitably and always, the first obligation results from a condition that can be considered to be only *occasionally* true. Therefore, we would like to have a mechanism that enables us to avoid this property while retaining the good properties.

7 A Solution

We argue that this problem is caused by the fact that we model the obligation only from the point of view of its violation conditions. We show that the undesired property is eliminated by considering success conditions also. The solution we arrive at, preserves the good properties. First we investigate how we can define a deadline operator $O^S(\rho \leq \delta)$ using success conditions (propositional 'ideality' constants) only. We show that if we look at the operator from this more positive angle, we arrive at similar logical properties. However, also this approach has a (quite obvious) counter-intuitive consequence. We show that to eliminate all counter-intuitive properties we may combine both failure and success conditions.

We extend the language $\mathcal{L}_{\mathrm{CTL}}$ with an ideality constant [8] *Idl*. Let $\mathcal{M}$ be a CTL model, s a state, and σ a full path starting at s. We can now define a success condition based semantics for a deontic deadline operator $O^S(\rho \leq \delta)$, where the S stands for Success, as follows:

$$\begin{aligned} \mathcal{M}, s \models O^S(\rho \leq \delta) \Leftrightarrow \forall\sigma \text{ with } \sigma_0 = s, \forall j : \\ \text{if} \\ \mathcal{M}, \sigma_j \models \delta \\ \text{then} \\ \exists 0 \leq i \leq j : \mathcal{M}, \sigma_i \models \rho \wedge Idl \end{aligned}$$

This says: for all possible futures it holds that if at some point the deadline occurs, then until then, there has at least been one ideal state where ρ has been achieved. Note that it would not be correct to define that *all* ρ-states before δ

[11] We actually use some background theory here about how logical properties of defined operators can be determined by looking at the way they are constructed from simpler operators. In particular, a negation in the definition flips closure under strengthening to closure under weakening and vice versa. This is why any modal operator $M\varphi$ is closed under weakening (strengthening) if and only if its dual $\neg M\neg\varphi$ is closed under weakening (strengthening).

are ideal; if a ρ is met, the obligation is discharged, and no ideal states should occur anymore[12].

The above semantic definition is equivalent to the following definition as a reduction to CTL:

$$O^S(\rho \leq \delta) \equiv_{def} \neg E(\neg(\rho \wedge Idl)U\delta)$$

Note that due to its form, this definition also obeys all the logical properties discussed in section 4. To be more precise, also the operator $O^S(\rho \leq \delta)$ is a monotonic operator with respect to ρ (i.e., closed under weakening with respect to ρ), and an anti-monotonic operator with respect to δ (i.e., closed under strengthening with respect to δ). And, in addition, it does *not* obey the counter-intuitive $\rho \rightarrow O^S(\rho \leq \delta)$, because now it requires the presence of an ideal state to have an obligation of the form $O^S(\rho \leq \delta)$. To be more precise, we have that:

$$\not\models \neg E(\neg\rho U\delta) \rightarrow O^S(\rho \leq \delta)$$

This follows, because, as we argued in section 4, the construct $\neg E(\neg\rho U\delta)$, is not closed under agglomeration with respect to ρ, which implies that it is certainly not anti-monotonic (closed under strengthening) with respect to ρ. So ρ cannot be strengthened to $\rho \wedge Idl$ while preserving truth.

However, obviously, also with this operator something is wrong. We have that:

$$\models O^S(\rho \leq \delta) \rightarrow \neg E(\neg\rho U\delta)$$

That is, deadline obligations $O^S(\rho \leq \delta)$ cannot be violated; success is guaranteed. Before giving the remedy, let us first explain why the above property is valid for the success based deadline definition. Valididty of the schema $\neg E(\neg(\rho \wedge Idl)U\delta)$ is closed under weakening with respect to $\rho \wedge Idl$, so weakening $\rho \wedge Idl$ to ρ, does not destroy truth.

Now how can we combine the intuitions from the present section with the ones of the previous sections, to arrive at a deadline operator that excludes all counterintuitive properties? We will not give the semantic truth-conditions of this final operator we define, and leave it to a characterization as a CTL formula (the semantic truth-conditions can easily be obtained by combining the conditions for the earlier defined operators):

$$O(\rho \leq \delta) \equiv_{def} \neg E(\neg(\rho \wedge Idl)U(\delta \wedge \neg Viol))$$

First of all, it is clear that this operator preserves the good properties. Due to its form we have monotonicity with respect to ρ, anti-monotonicity with respect to δ, etc. But we also avoid the counter-intuitive property $\neg E(\neg\rho U\delta) \rightarrow O(\rho \leq \delta)$, because we have strengthened ρ to $\rho \wedge Idl$. And we avoid the counter-intuitive $O(\rho \leq \delta) \rightarrow \neg E(\neg\rho U\delta)$, because we have strengthened δ to $\delta \wedge \neg Viol$ (which

[12] This actually implies that an ideal state can only be the *first* ρ-state encountered before the deadline δ. The consequences of introducing this stronger condition will be investigated on another occasion.

means that δ is weaker than $\delta \wedge \neg Viol$). Informally, the formula says that there is a deadline obligation only if there is a violation if the achievement is not met at the deadline, or there is success if the achievement is accomplished before the deadline.

A positive side-effect of this operator is that we now have that $\not\models O(\gamma \leq \gamma)$ and $\not\models O(\top \leq \delta)$. So, some of the properties we considered to be intuitively unattractive, but harmless, are no longer valid. But, we do still have that $\models O(\rho \leq \bot)$.

8 Discussion and Conclusion

Given that obligation concerns action, that action involves change, and that change presupposes time, deontic and temporal aspects have very strong conceptual connections. Therefore, any contribution to the study of such connections is welcome.

In this paper we discussed intuitions concerning the notion of 'being obliged to obey a condition ρ before a condition δ occurs'. We made a simplifying assumption that enabled us to study this notion in the logic CTL, minimally extended with violation constants. We defined two dyadic modal operators for the mentioned notion, and showed that they obey several intuitive logical properties. Finally, to prevent the operators from obeying some counter-intuitive property also, we proposed to consider success conditions.

It would be interesting to test the logic by means of a CTL-theorem prover. There are no such implemented theorem provers available. However,they can be written by using the results in either [18] or [19]. We plan to do this in the near future.

There are many directions for future research. For instance, we want to investigate whether the semantics also applies to other temporal formalisms (in particular LTL). Another point concerns abandoning the background assumption that there are no deontic updates. How much of the theory can be preserved if we do allow updates? Also we want to study the notion of permission in this setting (a simple definition is $P(\rho \leq \delta) \equiv_{def} \neg O(\neg\rho \leq \delta)$).

Finally we note that the combination of failure and success conditions was used before in deontic formalisms [20]. However, to our knowledge, the idea to evaluate failure and success conditions at *different* points in time for defining the semantics of a deontic concept, is new.

We thank Leendert van der Torre, Joris Hulstijn, Mehdi Dastani and Henry Prakken for lively and illuminating discussions on this subject. We thank the anonymous referees for valuable suggestions and references.

References

1. Dean, T., Firby, R., Miller, D.: Hierarchical planning involving deadlines, travel time, and resources. Computational Intelligence **6:1** (1990) 381–398
2. Clarke, E., Emerson, E., Sistla, A.: Automatic verification of finite-state concurrent systems using temporal logic specifications. ACM Transactions on Programming Languages and Systems **8** (1986)

3. Emerson, E.: Temporal and modal logic. In Leeuwen, J.v., ed.: Handbook of Theoretical Computer Science, volume B: Formal Models and Semantics. Elsevier Science (1990) 996–1072
4. Clarke, E., Grumberg, O., Long, D.: Verification tools for finite-state concurrent systems. In: A decade of concurrency. Volume 803 of Lecture Notes in Computer Science. Springer (1993) 124–175
5. Wright, G.v.: Deontic logic. Mind **60** (1951) 1–15
6. Broersen, J., Dastani, M., Torre, L.v.d.: BDIO-CTL: Obligations and the specification of agent behavior. In: Proceedings of Eighteenth International Joint Conference on Artificial Intelligence (IJCAI2003). (2003)
7. Anderson, A.: A reduction of deontic logic to alethic modal logic. Mind **67** (1958) 100–103
8. Kanger, S.: New foundations for ethical theory. In Hilpinen, R., ed.: Deontic Logic: Introductory and Systematic Readings. D. Reidel Publishing Company (1971) 36–58
9. Dignum, F., Kuiper, R.: Combining dynamic deontic logic and temporal logic for the specification of deadlines. In Jr., R.S., ed.: Proceedings of thirtieth HICSS. (1997)
10. Rao, A., Georgeff, M.: Modeling rational agents within a BDI-architecture. In Allen, J., Fikes, R., Sandewall, E., eds.: Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning (KR'91), Morgan Kaufmann Publishers (1991) 473–484
11. Schnoebelen, P.: The complexity of temporal logic model checking. In Balbiani, P., Suzuki, N.Y., Wolter, F., Zakharyaschev, M., eds.: Advances in Modal Logic. Volume 4. (2003) 393–436
12. Schild, K.: On the relationship between BDI-logics and standard logics of concurrency. Autonomous agents and multi-agent systems **3** (2000) 259–283
13. Cohen, P., Levesque, H.: Intention is choice with commitment. Artificial Intelligence **42** (1990) 213–261
14. Benthem, J.v.: Minimal deontic logics. Bulletin of the Section of Logic **8** (1979) 36–42
15. Lutz, C., Sattler, U.: The complexity of reasoning with boolean modal logic. In: Advances in Modal Logic. Volume 3., World Scientific (2002) 365–387
16. Gasquet, O., Herzig, A.: Translating non-normal modal logics into normal modal logics. In Jomes, A., Sergot, M., eds.: Proceedings International Workshop on Deontic Logic, TANO, Oslo (1993)
17. Gasquet, O., Herzig, A.: (From classical to normal modal logics)
18. Bolotov, A., Fisher, M.: A clausal resolution method for CTL branching-time temporal logic. Journal of Experimental and Theoretical Artificial Intelligence **11** (1999) 77–93
19. Shankar, N.: Machine-assisted verification using theorem proving and model checking. In Broy, M., ed.: Mathematical Methods in Program Development. Springer (1997)
20. Torre, L.v.d., Tan, Y.: Diagnosis and decision making in normative reasoning. Artificial Intelligence and Law **7** (1999) 51–67

Obligation Change in Dependence Logic and Situation Calculus

Robert Demolombe[1] and Andreas Herzig[2]

[1] ONERA Toulouse
France
Robert.Demolombe@cert.fr
[2] IRIT-Université Paul Sabatier
France
herzig@irit.fr

Abstract. Obligation change raises the "frame problem" which is to characterise what obligations remain unchanged after an action has been performed. Many general solutions have been proposed but even if they are attractive from a thoretical point of view they have practical drawbacks.
In this paper simple solutions are proposed thanks to the restriction to obligations that take the form of modal literals. These solutions are presented in the framework of dependence logic and of situation calculus, and it is shown that they are based on the same intuitive idea. This idea is to express that we have a complete representation of actions and circumstances that can change an obligation.

1 Introduction

The problem of the characterisation of what remains to be true after the performance of an action is recognised as a difficult problem in the field of Articial Intelligence. This problem is usually called the "frame problem".

The same problem arises in the field of deontic logic if we want to characterise the set of obligations that persist after an action. It has some connections with deontic defeasibility but it is not the same problem (see [10,1,12,9,17]).

An interesting solution to the frame problem has been proposed by Reiter [11] in the framework of situation calculus for modelling the evolution of the world. Later on this solution has been extended to the evolution of beliefs about the world by Scherl and Levesque [16,14]. This work has been extended to revision by Shapiro et al. in [15]. In [4] Demolombe has adapted their intuitive ideas to the evolution of obligations. However, this solution has practical drawbacks because it requires to assign to all the ideal situations an ideality level in the same way as Scherl and Levesque require the assignement of a plausibility level.

In this paper we investigate solutions to the frame problem which are less general, in the sense that we only consider facts that can be represented by literals, but are simpler to formalise and much easier to use for practical applications.

A. Lomuscio and D. Nute (Eds.): DEON 2004, LNAI 3065, pp. 57–73, 2004.

The first idea is to consider the dependence logic, which has been defined by Castilho, Herzig et al. in [2,3] (section 2) and to extend it to obligations (section 3). The second idea is to extend the simple idea of successor state axioms in situation calculus to obligations about literals (section 4). At the end of the paper the two formalisations are compared and it is shown that they are based on the same intuitive ideas (section 5).

2 Dependence Logic

The dependence logic is a propositional modal logic with the two modal operators $\Box$ and $[\alpha]$. Sentences of the form $\Box(p)$ are read: p is true after any sequence of actions, and sentences of the form $[\alpha](p)$ are read: p is true after the action α.

For modelling an application domain the effects of the actions are defined by properties of the form:

$$\Box(q \rightarrow [\alpha]p)$$

For instance, in the typical example of the Yale Shooting Scenario we have:

$$\Box(Loaded \rightarrow [shoot]\neg Alive)$$

This intuitively means that after any sequence of actions, if the gun is loaded then after shooting the man is not alive.

In addition to the definition of the action effects we have a set of frame axioms of the form:

$$\Box(\neg C \rightarrow (L \rightarrow [\alpha]L))$$

where L is a literal and C is a formula of classical propositional logic.

In a metalanguage this axiom says that if we are not in the context C the truth value of L is independent of the action α. That is, there is a ternary relation between α, L and C, and the frame axioms could be represented in the metalanguage by this independence relation.

The problem is that for almost every applications the set of frame axioms is very large, because after an action most of the literals have the same truth value.

Then, it is easier to represent the dependence relation, which is the complement of the independence relation, than the independence relation itself. Let us call D the dependence relation, we suppose that D is finite. The fact that the tuple $\langle \alpha, P, C \rangle$ is in D means that in the context C the truth value of the atom P may be changed by the action α. It is assumed that the dependence relation is **complete** in the sense that α may change the truth value of P only if there is a tuple $\langle \alpha, P, C \rangle$ in D.

The logic which is based on this dependence relation is called LAPD[1]. It is formally defined as follows.

ATM is the set of atomic formulas of the language. We have $ATM = \{P, Q, \ldots\}$. LIT is the set of literals. ACT is the set of actions. We have $ACT = \{\alpha, \beta, \ldots\}$. $PFOR$ denotes the set of formulas of classical propositional logic.

The dependence relation is such that $D \subseteq ACT \times ATM \times PFOR$.

[1] LAPD abreviates Logic for Action and Plan with Dependence relation.

Semantics

A model for the logic LAPD is a structure μ such that:
$\mu =< W, \{R_\alpha : \alpha \in ACT\}, R_\Box, \tau >$.

In μ:

- W is a set of possible worlds,
- $R_\Box$ and R_α are two accessibility relations which interpret $\Box$ and $[\alpha]$,
- τ is a function from ATM to 2^W; τ is extended as usual to the logical connectives.

The following constraints are imposed on μ:

- $R_\Box$ is reflexive and transitive,
- $R_\alpha \subseteq R_\Box$,
- if $wR_\alpha w'$ then
$$\forall P \in ATM \text{ if } \forall C \in PFOR(\langle \alpha, P, C\rangle \in D \Rightarrow w \not\models C) \text{ then}$$
$$w \in \tau(P) \text{ iff } w' \in \tau(P).$$

The intuition of the last constraint is that if all the contexts C where α may influence P are false in w, then P has the same truth value in w and w'.

We adopt the notation:

$$Pre_D(\alpha, P) = \bigvee_{\langle \alpha, P, C\rangle \in D} C$$

It is assumed that $Pre_D(\alpha, P) = \bot$ if there is no tuple in D of the form: $\langle \alpha, P, C\rangle$.

Let us denote by $|L|$ the atom of the literal L. We have the property:

$$\models_{LAPD} \Box(\neg Pre_D(\alpha, |L|) \rightarrow (L \rightarrow [\alpha]L))$$

From the relation D we obtain the formula $Pre_D(\alpha, |L|)$, and from this property we have the corresponding frame axioms.

For instance, if $Pre_D(\alpha, P) = C$ we have the frame axioms:

$$\models_{LAPD} \Box(\neg C \rightarrow (P \rightarrow [\alpha]P))$$
$$\models_{LAPD} \Box(\neg C \rightarrow (\neg P \rightarrow [\alpha]\neg P))$$

Axiomatics

The axiomatics of the LAPD logic is defined as follows:

- all the tautologies of the classical propositional logic,
- $[\alpha]$ obeys the schema K,
- $\Box$ obey the schemas K, T and 4,
- *(I)* $\Box p \rightarrow [\alpha]p$,
- *(Persist)* $\neg Pre_D(\alpha, |L|) \rightarrow (L \rightarrow [\alpha]L)$,
- Modus Ponens and Necessitation for $\Box$ and $[\alpha]$.

It has been proved (see in the Annex) that this axiomatics is valid and complete.

Example

We can see now how this logic can be applied to the Yale Shooting Scenario. The dependence relation D is $D = \{d1, d2, d3\}$, where we have:

($d1$) $\langle load, Loaded, \top\rangle$
($d2$) $\langle shoot, Loaded, \top\rangle$
($d3$) $\langle shoot, Alive, \top\rangle$

The set of effect laws is LAW = {1,2,3,4}, where we have:

(1) $\Box[load]Loaded$
(2) $\Box[shoot]\neg Loaded$
(3) $\Box(Loaded \rightarrow [shoot]\neg Alive)$
(4) $\Box(\neg Loaded \wedge Alive \rightarrow [shoot]Alive)$

Let us assume that the current situation is represented by KB = {$\neg Loaded$, $Alive$}.

Since there is no tuple in D of the form $\langle load, Alive, C\rangle$ we have $Pre_D(load, |Alive|) = \bot$. Then, from the schema *(Persist)* we have the frame axiom:

(5) $Alive \rightarrow [load]Alive$

From (5) and KB we have:

(6) $[load]Alive$

From (1) and (T) we have:

(7) $[load]Loaded$

From the schema (I) and (3) we have:

(8) $[load](Loaded \rightarrow [shoot]\neg Alive)$

And from (7) and (8) we have:

(9) $[load]\ [shoot]\neg Alive$

From (4) and (T) we also have:

(10) $\neg Loaded \wedge Alive \rightarrow [shoot]Alive$

Then, from KB we have:

(11) $[shoot]Alive$

If we apply the Necessitation rule to the frame axiom (5) we have:

(12) $[shoot](Alive \rightarrow [load]Alive)$

From (11) and (12) we have:

(13) $[shoot][load]Alive$

Finally we have:

$$\vdash_{LAPD} KB \wedge LAW \rightarrow [load][shoot]\neg Alive$$
$$\vdash_{LAPD} KB \wedge LAW \rightarrow [shoot][load]Alive$$

It is interesting to see how the property *Alive* persists after the actions *shoot* and *load.*

3 Extension of Dependence Logic to Obligations

In the previous section we have seen how dependence logic provides us with a simple solution to the frame problem. This simplicity comes from the fact that the evolution of the world is described in terms of evolution of classical literals.

Here this approach is extended to the evolution of obligations, where this evolution is described in terms of modal literals.

We introduce the new modality *Obg* and sentences of the form $Obg(p)$ are read: it is obligatory that p. The new dependence logic extended to obligations is called *LAPDO*.

A modal literal has the form: *Obg(P), ¬Obg(P), Obg(¬P)* or *¬Obg(¬P)*, where $P \in ATM$. The set of modal literals is denoted by *LITM*.

If $LM \in LITM$ we denote by $|LM|$ the classical atom in *LM*. For instance, we have $|Obg(\neg P)| = P$.

To characterise the modal literals whose truth values may change after an action we define the dependence relation *DO* such that $DO \subseteq ACT \times ATM \times PFOR$.

The fact that a tuple $\langle \alpha, P, C \rangle$ is in *DO* means that if *C* holds the action α may change the truth value of *Obg(P), ¬Obg(P), Obg(-P)* or *¬Obg(¬P)*.

Semantics

A model of the logic *LAPDO* is a structure μ such that:
$\mu = \langle W, \{R_\alpha : \alpha \in ACT\}, R_\Box, R_{Obg}, \tau \rangle$.

In μ:

- W, $R_\Box$, R_α and τ are defined like in *LAPD*[2].
- R_{Obg} is an accessibility relation which interprets *Obg* and is reflexive.

The following constraints are imposed to μ:

- $R_\alpha \subseteq R_\Box$ and $R_{Obg} \subseteq R_\Box$,
- if $wR_\alpha w'$ then
  $\forall P \in ATM$ if $\forall C \in PFOR(\langle \alpha, P, C \rangle \in D \Rightarrow w \not\models C)$ then
  $w \in \tau(P)$ iff $w' \in \tau(P)$.
- if $wR_\alpha w'$ then
  $\forall P \in ATM$ if $\forall C \in PFOR(\langle \alpha, P, C \rangle \in DO \Rightarrow w \not\models C)$ then
  $w \in \tau(Obg(P))$ iff $w' \in \tau(Obg(P))$ and
  $w \in \tau(Obg(\neg P))$ iff $w' \in \tau(Obg(\neg P))$.

The last constraint on *LAPDO* models means that if we are not in a context where the action α may change the truth values of the modal literals formed with *P*, then their truth values remain unchanged after α.

The constraint $R_{Obg} \subseteq R_\Box$ requires some comments. Indeed, a consequence of this constraint is that we have $\models \Box(p) \rightarrow Obg(p)$. Then, for example, from $\Box(Loaded \rightarrow [shoot]\neg Alive)$ we can infer $Obg(Loaded \rightarrow [shoot]\neg Alive)$. This consequence may seem to be odd in a first approach.

In fact this is acceptable if the intuitive meaning of $Obg(p)$ is: p is true in all the ideal worlds, and if we accept that ideal worlds are a subset of the "real worlds". Here we call real world a world which satisfies all the properties that are

[2] The function τ is extended to obligations in a natural way. We have $\tau(Obg(p)) = \{ w : w \models Obg(p) \}$, and $w \models Obg(p)$ iff $wR_{Obg}w'$ implies $w' \models p$.

necessarily true in a given application domain. In particular all the properties that define the effects of the actions must hold in a real world.

Why should we impose that the ideal worlds are a subset of the real worlds? Suppose, on the contrary, that there is an ideal world w which is not a real world. That means that in w there is a property of the domain which is not satisfied.

Let us consider, for example, the property: a person cannot be at two different places at the same time. Then, in w it could be the case that the same person is at two different places at the same time, and, from a normative point of view, it would be permitted for a person to be at two different places at the same time. It would be very odd to define a regulation with such a permission. That is why it is imposed that ideal worlds are real worlds.

We adopt the notation:

$$Pre_{DO}(\alpha, P) = \bigvee_{\langle \alpha, P, C \rangle \in DO} C$$

If $LM \in LITM$ we have the property:

$$\models_{LAPDO} \Box(\neg Pre_{DO}(\alpha, |LM|) \rightarrow (LM \rightarrow [\alpha]LM))$$

For example, if $Pre_{DO}(\alpha, P) = C$ we have:

$$\models_{LAPDO} \Box(\neg C \rightarrow (Obg(P) \rightarrow [\alpha]Obg(P)))$$
$$\models_{LAPDO} \Box(\neg C \rightarrow (\neg Obg(P) \rightarrow [\alpha]\neg Obg(P)))$$
$$\models_{LAPDO} \Box(\neg C \rightarrow (Obg(\neg P) \rightarrow [\alpha]Obg(\neg P)))$$
$$\models_{LAPDO} \Box(\neg C \rightarrow (\neg Obg(\neg P) \rightarrow [\alpha]\neg Obg(\neg P)))$$

Axiomatics

The axiomatics of the *LAPDO* logic is defined as follows:

- all the tautologies of the classical prepositional logic,
- $[\alpha]$ obeys the schema K,
- $\Box$ obeys the schemas K, T and 4,
- *Obg* obeys the schemas K and D,
- *(I)* $\Box p \rightarrow [\alpha]p$,
- *(O)* $\Box p \rightarrow Obg(p)$,
- *(Persist)* $\neg Pre_D(\alpha, |L|) \rightarrow (L \rightarrow [\alpha]L)$, if $L \in LIT$,
- $(Persist_O)$ $\neg Pre_{DO}(\alpha, |LM|) \rightarrow (LM \rightarrow [\alpha]LM)$, if $LM \in LITM$,
- Modus Ponens and Necessitation for $\Box$, $[\alpha]$ and *Obg*.

It has been proved that this logic is valid and complete (see in the Annex).

Example

Let us take the example of the traffic lights to show how obligation change is formalised in *LAPDO*. We use the following notations[3]:

[3] As a matter of simplification we have ignored the case where the light is orange.

Red: the light is red.
Green: the light is green.
InCrossing: the car is crossing the crossroads.

For the actions we use the notations:

red: to switch the light to red.
green: to switch the light to green.
start.cr: to start to cross the crossroads.
end.cr: to end to cross the crossroads.

The relation D is $D = \{d1, d2, d3, d4, d5, d6\}$ where:
($d1$) $\langle red, Red, \top\rangle$
($d2$) $\langle red, Green, \top\rangle$
($d3$) $\langle green, Red, \top\rangle$
($d4$) $\langle green, Green, \top\rangle$
($d5$) $\langle start.cr, InCrossing, \top\rangle$
($d6$) $\langle end.cr, InCrossing, \top\rangle$

The relation DO is $DO = \{do1, do2\}$ where;
($do1$) $\langle red, InCrossing, \top\rangle$
($do2$) $\langle green, InCrossing, \top\rangle$

Note that *start.cr* and *end.cr* have no influence on obligations.

The set of effects laws is $LAW = \{l1, l2, l3, l4, l5, l6, l7\}$ where:

($l1$) $\Box[red]Red$
($l2$) $\Box[green]Green$
($l3$) $\Box[start.cr]InCrossing$
($l4$) $\Box[end.cr]\neg InCrossing$
($l5$) $\Box[red]Obg(\neg InCrossing)$
($l6$) $\Box[green]Perm(InCrossing)$
($l7$) $\Box\neg(Red \wedge Green)$

As usual $Perm(p)$ is an abreviation for $\neg Obg(\neg p)$. The current situation is represented by $KB = \{\neg InCrossing, Green, Perm(InCrossing)\}$.

From (l1) and (T) we have:
(1) $[red]Red$

From (Persist) we have:
$\neg InCrossing \rightarrow [red]\neg InCrossing$

Then, from KB we have:
(2) $[red]\neg InCrossing$

From (l5) and (T) we have:
(3) $[red]Obg(\neg InCrossing)$

Therefore from (1), (2) and (3) we have:
$\vdash_{LAPDO} LAW \wedge KB \rightarrow [red](Red \wedge \neg InCrossing \wedge Obg(\neg InCrossing))$

It is worth noting that $(Persist_O)$ does **not** allow to infer:

$$Perm(InCrossing) \rightarrow [red]Perm(InCrossing)$$

because we have the tuple $\langle red, InCrossing, \top \rangle$ in *DO*. Then the permission *Perm(InCrossing)* does not persist after the action *red*.

From (Persist) we have:

$Red \rightarrow [start.cr]Red$

By Necessitation we have:

$[red](Red \rightarrow [start.cr]Red)$

And from (1) we have:

(4) $[red][start.cr]Red$

From (13) and (I) we have:

(5) $[red][start.cr]InCrossing$

From $(Persist_O)$ we have:

$Obg(\neg InCrossing) \rightarrow [start.cr]Obg(\neg InCrossing)$

And by Necessitation we have:

$[red](Obg(\neg InCrossing) \rightarrow [start.cr]Obg(\neg InCrossing))$

Then, from (3) we have:

(6) $[red][start.cr]Obg(\neg InCrossing)$

Therefore from (4), (5) and (6) we have

$\vdash_{LAPDO} LAW \wedge KB \rightarrow [red][start.cr]$
$(Red \wedge InCrossing \wedge Obg(\neg InCrossing))$

It can be shown in a similar way that we have:

$\vdash_{LAPDO} LAW \wedge KB \rightarrow [red][green][start.cr]$
$(Green \wedge InCrossing \wedge Perm(InCrossing))$

This example shows how the obligations about the fact *InCrossing* are updated when the actions *red* and *green* are performed.

4 A Simple Extension of Situation Calculus to Obligation Change

The situation calculus is a typed first order classical logic (except some limited fragments that are in the second order.). The characteristic feature of this logic is that dynamic aspects are represented by the notion of situation, which can be quantified, and each predicate whose truth value may change when actions are performed has an argument of the type situation. These predicates are called fluents.

For instance, the fact that the light is red in the situation s is represented by $Red(s)$. A situation may be the initial situation S_0, or the situation obtained after performance of the action a from the situation s. This situation is represented by the term $do(a, s)$.

For example, the situation $do(start.cr, S_0)$ represents the situation where the car has crossed the crossroads, and $do(red, do(start.cr, S_0))$ represents the situation where the light has switched to red after the car has crossed the crossroads.

To solve the frame problem in a given application domain we have to define for each fluent the complete list of the actions and circumstances that cause the fluent to be true or that cause the fluent to be false.

For example, the action *red* causes the light to be red and the action *green* causes the light not to be red. This is formally represented by:

$(S1)\ \ \forall s \forall a(a = red \rightarrow Red(do(a, s)))$

$(S2)\ \ \forall s \forall a(a = green \rightarrow \neg Red(do(a, s)))$

To represent the fact that there are no other action that cause *Red* or $\neg Red$ we have to add the properties:

$(C1)\ \ \forall s \forall a(\neg Red(s) \wedge Red(do(a, s)) \rightarrow a = red)$

$(C2)\ \ \forall s \forall a(Red(s) \wedge \neg Red(do(a, s)) \rightarrow a = green)$

It can be shown that (S1), (S2), (C1) and (C2) are logically equivalent to (SS1).

$(SS1)\ \ \forall s \forall a(Red(do(a, s)) \leftrightarrow a = red \vee Red(s) \wedge \neg(a = green))$

In the same way we have:

$(SS2)\ \ \forall s \forall a(Green(do(a, s)) \leftrightarrow a = green \vee Green(s) \wedge \neg(a = red))$

$(SS3)\ \ \forall s \forall a(InCrossing(do(a, s)) \leftrightarrow a = start.cr \vee InCrossing(s) \wedge \neg(a = end.cr))$

Notice that from (SS1) and (SS2) it can be easily proved by induction that $\neg(Green(S_0) \wedge Red(S_0)) \rightarrow \forall s(\neg(Green(s) \wedge Red(s)))$.

If we assume that each action has a unique name, from (SS1) we have:

$\forall s(Red(do(start.cr, s)) \leftrightarrow Red(s))$

Its intuitive meaning is that the action *start.cr* does not change the fact that the color of the light is red. In other terms the status of *Red* persists after any action other than *red* and *green*. That gives a very simple solution to the frame problem.

In general, for each fluent we have to define a successor state axiom of the form:

$$\forall s \forall a(p(do(a, s)) \leftrightarrow \Gamma^{+}(a, s) \vee p(s) \wedge \neg\Gamma^{-}(a, s))$$

To avoid inconsistencies we have to impose the constraint:

$$\neg\exists s \exists a(\Gamma^{+}(a, s) \wedge \Gamma^{-}(a, s))$$

The solution to the frame problem is based on two key ideas: we define the evolution of the world by defining the evolution of each literal, and we assume that we have a complete knowledge of the causes of their evolution. The same ideas will be applied to the evolution of the obligations in the same way as Demolombe and Pozos did for the evolution of beliefs [6].

In a first step we define obligations in the same way as Scherl and Levesque have defined beliefs in the situation calculus.

We adopt the definition:

$$Obg(p, s) \stackrel{\text{def}}{=} \forall s'(O(s', s) \rightarrow p[s'])$$

where the arguments of the type situation have been removed in p, and they have been replaced by s' in $p[s']$. $O(s', s)$ is a classical predicate that plays the same role as an accessibility relation.

To define the successor state axioms for obligations the only difference is that modal literals correspond to four truth values, while classical literals correspond to two truth values.

For example, to define the evolution of the four modal literals formed with the atom *InCrossing* we have the properties:

$(OS1)\ \ \forall s \forall a(\bot \rightarrow Obg(InCrossing, do(a, s)))$
$(OS2)\ \ \forall s \forall a(a = red \rightarrow \neg Obg(InCrossing, do(a, s)))$
$(OS3)\ \ \forall s \forall a(a = red \rightarrow Obg(\neg InCrossing, do(a, s)))$
$(OS4)\ \ \forall s \forall a(a = green \rightarrow \neg Obg(\neg InCrossing, do(a, s)))$

And we have four properties to represent the fact that the causes of change are complete:

$(OC1)\ \ \forall s \forall a(\neg Obg(InCrossing, s) \wedge Obg(InCrossing, do(a, s)) \rightarrow \bot)$
$(OC2)\ \ \forall s \forall a(Obg(InCrossing, s) \wedge \neg Obg(InCrossing, do(a, s)) \rightarrow a = red)$
$(OC3)\ \ \forall s \forall a(\neg Obg(\neg InCrossing, s) \wedge Obg(\neg InCrossing, do(a, s)) \rightarrow a = red)$
$(OC4)\ \ \forall s \forall a(Obg(\neg InCrossing(s)) \wedge \neg Obg(\neg InCrossing, do(a, s)) \rightarrow a = green)$

It can be shown that (OS1)-(OS4) and (OC1)-(OC4) are logically equivalent to (OSS1) and (OSS2).

$(OSS1)\ \ \forall s \forall a(Obg(InCrossing, do(a, s)) \leftrightarrow Obg(InCrossing, s) \wedge \neg(a = red))$
$(OSS2)\ \ \forall s \forall a(Obg(\neg InCrossing, do(a, s)) \leftrightarrow a = red \vee Obg(\neg InCrossing, s) \wedge \neg(a = green))$

Let us consider an initial situation defined by $KB = \{\neg InCrossing(S_0), Green(S_0), Perm(InCrossing, S_0)\}$.

From (SS1) we have:

(1) $Red(do(red, S_0))$

From (SS3) and KB we have:

(2) $\neg InCrossing(do(red, S_0))$

From (OSS2) we have:

(3) $Obg(\neg InCrossing, do(red, S_0))$

If we denote by AS the set of properties $AS = \{SS1, SS2, SS3, OSS1, OSS2\}$ we have:

$\vdash AS \wedge KB \rightarrow Red(do(red, S_0)) \wedge \neg InCrossing(do(red, S_0)) \wedge Obg(\neg InCrossing, do(red, S_0))$

Notice that in S_0 it is permitted to cross the crossroads while in $do(red, S_0)$ it is forbidden to cross. This shows that the action *red* requires obligation updating. We can also notice that the fact $\neg InCrossing$ persists after the action *red.*

From (SS1) and (1) we also have:

(4) $Red(do([red, start.cr], S_0))$[4]

From (SS3) we have:

(5) $InCrossing(do([red, start.cr], S_0))$

From (OSS2) and (3) we have:

(6) $Obg(\neg InCrossing, do([red, start.cr], S_0))$

Therefore we have:

$\vdash AS \wedge KB \rightarrow Red(do([red, start.cr], S_0)) \wedge InCrossing(do([red, start.cr], S_0)) \wedge Obg(\neg InCrossing, do([red, start.cr], S_0))$

[4] $do([red, start.cr], S_0)$ is an abreviation for $do(start.cr, do(red, S_0))$.

It can be shown in a similar way that we have:

$\vdash AS \wedge KB \rightarrow Green(do([red, green, start.cr], S_0)) \wedge InCrossing(do([red, green, start.cr], S_0)) \wedge Perm(InCrossing, do([red, green, start.cr], S_0))$

In general for each normative fluent we must define two successor state axioms for obligations of the form:

$\forall s \forall a(Obg(p, do(a, s)) \leftrightarrow \Gamma_1^+(a, s) \vee Obg(p, s) \wedge \neg\Gamma_1^-(a, s))$

$\forall s \forall a(Obg(\neg p, do(a, s)) \leftrightarrow \Gamma_2^+(a, s) \vee Obg(\neg p, s) \wedge \neg\Gamma_2^-(a, s))$

To guarantee the consistency of obligations we impose the constraints:

$\neg\exists s \exists a(\Gamma_1^+(a, s) \wedge \Gamma_1^-(a, s))$

$\neg\exists s \exists a(\Gamma_2^+(a, s) \wedge \Gamma_2^-(a, s))$

To satisfy the schema (D) we impose the constraint:

$\neg\exists s \exists a(\Gamma_1^+(a, s) \wedge \Gamma_2^+(a, s))$

Moreover, from (D) we have: $Obg(p, do(a, s)) \rightarrow \neg Obg(\neg p, do(a, s))$. In addition we have: $\Gamma_1^+(a, s) \rightarrow Obg(p, do(a, s))$. Then, we can infer: $\Gamma_1^+(a, s) \rightarrow \neg Obg(\neg p, do(a, s))$. Since $\Gamma_2^-(a, s))$ represent all the circumstances that cause $\neg Obg(\neg p, do(a, s))$ we must impose the constraint:

$\forall s \forall a(\Gamma_1^+(a, s) \rightarrow \Gamma_2^-(a, s))$

For a similar reason we impose the constraint:

$\forall s \forall a(\Gamma_2^+(a, s) \rightarrow \Gamma_1^-(a, s))$

5 Comparison between Situation Calculus and Dependence Logic

To analyse the links between the evolution of obligations expressed in the situation calculus or in the dependence logic, we shall consider a translation from situation calculus to a dynamic logic, and from this dynamic logic to dependence logic. Then, it is shown that consequences derived in dynamic logic correspond to the consequences derived in dependence logic.

5.1 From Situation Calculus to Dynamic Logic

In [5] Demolombe has presented a general method to translate situation calculus formulas into formulas of a dynamic logic.

As a matter of simplification we only consider here the translation of the successor state axioms for the obligations.

Without loss of generality it can be assumed that the Γ_is have the following form.

$\Gamma_1^+(a, s) \stackrel{\text{def}}{=} a = \alpha \wedge C_1^+(s)$

$\Gamma_1^-(a, s) \stackrel{\text{def}}{=} (a = \beta \wedge C_1^-(s)) \vee (a = \gamma \wedge C_2^+(s))$

$\Gamma_2^+(a, s) \stackrel{\text{def}}{=} a = \gamma \wedge C_2^+(s)$

$\Gamma_2^-(a, s) \stackrel{\text{def}}{=} (a = \delta \wedge C_2^-(s)) \vee (a = \alpha \wedge C_1^+(s))$

It is assumed that the C_is contain no symbol of the type action.

In Γ_1^- we have the subformula $a = \gamma \wedge C_2^+(s)$ because $a = \gamma \wedge C_2^+(s)$ implies $Obg(\neg p, do(a, s))$, and, since obligations should obey (D), $Obg(\neg p, do(a, s))$ implies $\neg Obg(p, do(a, s))$. Then, $a = \gamma \wedge C_2^+(s)$ causes $\neg Obg(p, do(a, s))$. We have $a = \alpha \wedge C_1^+(s)$ in Γ_2^- for a similar reason.

All the results would be the same if instead of Γ_1^+ we had:

$\Gamma_1^+(a, s) \stackrel{\text{def}}{=} (a = \alpha_1 \wedge C_{1,1}^+(s)) \vee \ldots \vee (a = \alpha_n \wedge C_{1,n}^+(s))$

The same comment holds for the other Γ_is.

Thanks to the unique name axioms we can easily check that the Γ_is satisfy all the constraints mentioned in the previous section.

Then, the properties that define the effects of the actions on the obligations, and the completion properties are:

$(C1)$ $\forall s \forall a(a = \alpha \wedge C_1^+(s) \rightarrow Obg(p, do(a, s)))$
$(C2)$ $\forall s \forall a((a = \beta \wedge C_1^-(s)) \vee (a = \gamma \wedge C_2^+(s)) \rightarrow \neg Obg(p, do(a, s)))$
$(C3)$ $\forall s \forall a(a = \gamma \wedge C_2^+(s) \rightarrow Obg(\neg p, do(a, s)))$
$(C4)$ $\forall s \forall a((a = \delta \wedge C_2^-(s)) \vee (a = \alpha \wedge C_1^+(s)) \rightarrow \neg Obg(\neg p, do(a, s)))$
$(C5)$ $\forall s \forall a(\neg Obg(p, s) \wedge Obg(p, do(a, s)) \rightarrow a = \alpha \wedge C_1^+(s))$
$(C6)$ $\forall s \forall a(Obg(p, s) \wedge \neg Obg(p, do(a, s)) \rightarrow (a = \beta \wedge C_1^-(s)) \vee (a = \gamma \wedge C_2^+(s)))$
$(C7)$ $\forall s \forall a(\neg Obg(\neg p, s) \wedge Obg(\neg p, do(a, s)) \rightarrow a = \gamma \wedge C_2^+(s)$
$(C8)$ $\forall s \forall a(Obg(\neg p, s) \wedge \neg Obg(\neg p, do(a, s)) \rightarrow (a = \delta \wedge C_2^-(s)) \vee (a = \alpha \wedge C_1^+(s)))$

It is worth noting that the set of formulas (C1)-(C8) is logically equivalent to (C9) and (C10).

$(C9)$ $\forall s \forall a(Obg(p, do(a, s)) \leftrightarrow (a = \alpha \wedge C_1^+(s)) \vee Obg(p, s) \wedge \neg((a = \beta \wedge C_1^-(s)) \vee (a = \gamma \wedge C_2^+(s))))$

$(C10)$ $\forall s \forall a(Obg(\neg p, do(a, s)) \leftrightarrow (a = \gamma \wedge C_2^+(s)) \vee Obg(\neg p, s) \wedge \neg((a = \delta \wedge C_2^-(s)) \vee (a = \alpha \wedge C_1^+(s))))$

The translation of these properties into dynamic logic is based on the following property:

$\vdash Obg(p, do(a, s)) \leftrightarrow \forall s''(s'' = do(a, s) \rightarrow \forall s'(O(s', s'') \rightarrow p[s']))$

This property justifies the translation of $Obg(p, do(a, s))$ into $[a]Obg(p)$.

Formulas of the form $\forall s F(s)$ are translated in dynamic logic into $\Box F$, where all the arguments of the type situation have been removed from the fluents that occur in $F(s)$.

To translate formulas of the form $\forall a G(a)$ it is assumed that the quantification domain for the actions is the set of actions that occur in some formula to be translated, plus another distinct action ϵ.

Then, we assume that we have the following domain closure axiom for the actions:

$\forall a(a = \alpha \vee a = \beta \vee a = \gamma \vee a = \delta \vee a = \epsilon)$

From this axiom the translation of $\forall a G(a)$ is $G(\alpha) \wedge G(\beta) \wedge G(\gamma) \wedge G(\delta) \wedge G(\epsilon)$, which is equivalent to the set of formulas: $G(\alpha)$, $G(\beta)$, $G(\gamma)$, $G(\delta)$, $G(\epsilon)$.

Notice that it is not necessary to have several distinct actions $\epsilon_1, \ldots, \epsilon_n$ like ϵ. Indeed, in the evaluation of the conditions of the form: $\epsilon_i = \alpha$, $\epsilon_i = \beta$, $\epsilon_i = \gamma$ and $\epsilon_i = \delta$, we always get the result $\perp$ for every ϵ_i. Then, every ϵ_i would lead to a translated formula of the same form.

Finally the translation of the set of properties (C1)-(C8) leads to:

$(D1)$ $\Box(C_1^+ \rightarrow [\alpha]Obg(p))$
$(D2)$ $\Box(C_1^- \rightarrow [\beta]\neg Obg(p))$
$(D2')$ $\Box(C_2^+ \rightarrow [\gamma]\neg Obg(p))$
$(D3)$ $\Box(C_2^+ \rightarrow [\gamma]Obg(\neg p))$
$(D4)$ $\Box(C_2^- \rightarrow [\delta]\neg Obg(\neg p))$
$(D4')$ $\Box(C_1^+ \rightarrow [\alpha]\neg Obg(\neg p))$
$(D5)$ $\Box(\neg Obg(p) \wedge [\alpha]Obg(p) \rightarrow C_1^+)$
$(D6)$ $\Box(Obg(p) \wedge \neg[\beta]Obg(p) \rightarrow C_1^-)$
$(D6')$ $\Box(Obg(p) \wedge \neg[\gamma]Obg(p) \rightarrow C_2^+)$
$(D7)$ $\Box(\neg Obg(\neg p) \wedge [\gamma]Obg(\neg p) \rightarrow C_2^+)$
$(D8)$ $\Box(Obg(\neg p) \wedge [\delta]\neg Obg(\neg p) \rightarrow C_2^-)$
$(D8')$ $\Box(Obg(\neg p) \wedge [\alpha]\neg Obg(\neg p) \rightarrow C_1^+)$

Since in the situation calculus the actions are deterministic, in dynamic logic we must have the schema $\neg[\alpha]\neg p \rightarrow [\alpha]p$. Moreover, in the situation calculus every situation has a successor for any action. Then, we must also have the schema: $[\alpha]p \rightarrow \neg[\alpha]\neg p$. To sum it up, in the dynamic logic we must have the axiom schema (DD).

(DD) $[\alpha]p \leftrightarrow \neg[\alpha]\neg p$

5.2 From Dynamic Logic to Dependence Logic

We consider a propositional dynamic logic with the axiom schema (DD).

The effects of the actions are represented by the properties (D1-(D8').

From the properties (D1)-(D4') we know that the following tuples are in the dependence relation for obligations *DO*.

$(o1)$ $\langle \alpha, p, C_1^+ \rangle$
$(o2)$ $\langle \beta, p, C_1^- \rangle$
$(o3)$ $\langle \gamma, p, C_2^+ \rangle$
$(o4)$ $\langle \delta, p, C_2^- \rangle$

From the completion properties (D5)-(D8') we know that there is no other tuple in *DO*. Therefore we have: $DO = \{o1, o2, o3, o4\}$.

The set of formulas in *LAW* is (D1)-(D4').

To have the same properties as in the dynamic logic we add to the dependence logic the axiom schema (DD).

5.3 From Dependence Logic to Dynamic Logic

Let us consider a dependence logic with the axiom schema (DD).

Let us assume that in this dependence logic the effects of the actions are defined by the set of sentences in *LAW,* and the dependence relation for obligation is *DO,* and *LAW* and *DO* are the same as in the previous section.

We can show that in this dependence logic obligations change in the same way as in the previous dynamic logic.

Let us denote by Mp a modal literal of the form: $Obg(p)$, $\neg Obg(p)$, $Obg(\neg p)$ or $\neg Obg(\neg p)$. From the axiom schema $(Persist_O)$, and from the relation DO, we have the following frame axioms.

(*f1*) $\neg C_1^+ \rightarrow (Mp \rightarrow [\alpha]Mp)$
(*f2*) $\neg C_1^- \rightarrow (Mp \rightarrow [\beta]Mp)$
(*f3*) $\neg C_2^+ \rightarrow (Mp \rightarrow [\gamma]Mp)$
(*f4*) $\neg C_2^- \rightarrow (Mp \rightarrow [\delta]Mp)$

We can prove that in the dependence logic from (f1)-(f4) we can infer (D5)-(D8'). Since (D1)-(D4') are in the dependence logic and in the dynamic logic, in both logics we have (D1)-(D8'), and the evolution of obligations is the same.

For example, we can prove that (f1) implies (D5). Indeed, if (f1) is transformed in clausal form and if we apply Necessitation for the operator $\Box$ we get: $\Box(\neg Mp \vee [\alpha]Mp \vee C_1^+)$. Then, for $Mp = \neg Obg(p)$ we have:$\Box(Obg(p) \vee [\alpha]\neg Obg(p) \vee C_1^+)$. Moreover, from the schema (DD) we have: $\neg[\alpha]Obg(p) \leftrightarrow [\alpha]\neg Obg(p)$; then, we have: $\Box(Obg(p) \vee \neg[\alpha]Obg(p) \vee C_1^+)$, which is the clausal form of (D5).

In a similar way we can prove that (f2) implies (D6). Indeed, if (f2) is transformed in clausal form and if we apply Necessitation for the operator $\Box$ we have: $\Box(\neg Mp \vee [\beta]Mp \vee C_1^-)$. Then, for $Mp = Obg(p)$ we have: $\Box(\neg Obg(p) \vee [\beta]Obg(p) \vee C_1^-)$ which is the clausal form of (D6).

6 Conclusion

Two simple solutions to the frame problem for obligations have been presented in the framework of dependence logic and of situation calculus. These solutions are restricted to obligations that apply to classical literals, and obligations are given the semantics of standard deontic logic. As we can see by the traffic light example the solutions work for iterated obligation changes, too.

It has been shown that both frameworks lead to the same consequences for obligation change. At the intuitive level the two solutions are based on the same ideas. A technical difference is that dependence logic requires some kind of meta reasoning, while situation calculus deals with classical logic but modalities have to be represented by a predicate that plays the role of an accessibility relation. The similarity between intuitive ideas can be shown as follows.

Let us assume that the action ϵ has no influence on the obligations about the atom p. That means in the dependence logic that there is no tuple of the form $\langle \epsilon, p, C \rangle$ in the dependence relation DO, and by meta reasoning we can infer $Pre_{DO}(\epsilon, p) = \bot$. Then, from $(Persist_O)$ we have the four frame axioms:

$Obg(p) \rightarrow [\epsilon]Obg(p)$
$\neg Obg(p) \rightarrow [\epsilon]\neg Obg(p)$
$Obg(\neg p) \rightarrow [\epsilon]Obg(\neg p)$
$\neg Obg(\neg p) \rightarrow [\epsilon]\neg Obg(\neg p)$

From the schema (DD) we have $\neg[\epsilon]\neg\phi \leftrightarrow [\epsilon]\phi$. Then the four frame axioms are equivalent to the two frame axioms:

$[\epsilon]Obg(p) \leftrightarrow Obg(p)$
$[\epsilon]Obg(\neg p) \leftrightarrow Obg(\neg p)$

In the situation calculus, since ϵ does not influence the obligations about p, ϵ is different from α, β, γ and δ. Then, from (C9) and (C10) we have:

$\forall s(Obg(p, do(\epsilon, s)) \leftrightarrow Obg(p, s))$

$\forall s(Obg(\neg p, do(\epsilon, s)) \leftrightarrow Obg(\neg p, s))$

We see that we obtain frame axioms that have the same semantics in both frameworks, the difference is just technical.

An important issue that deserves more work is the ramification problem, that is to integrate in these frameworks invariant constraints between obligations like, for example, in the situation calclus $\forall s(Obg(p, s) \rightarrow Obg(q, s))$. Solutions proposed by Lin and Reiter in [7] and McIllraith in [8] could be adapted to the case of modal literals.

References

1. A. Artosi, G. Governatori, and G. Sartor. Towards a computational treatment of deontic defeasibility. In M. A. Brown and J. Carmo, editors, *Deontic Logic, Agency and Normative Systems,* pages 27–46. Springer, 1996.
2. M. A. Castilho, O. Gasquet, and A. Herzig. Formalizing action and change in a modal logic I:the frame problem. *Journal of Logic and Computation,* 9(5), 1999.
3. M. A. Castilho, A. Herzig, and I. J.Varzinczak. It depends on the context! A decidable logic of actions and plans based on a ternary dependence relation. In *9th Int. Workshop on Non Monotonic Reasoning,* 2002.
4. R. Demolombe. From belief change to obligation change in the Situation Calculus. A preliminary study. In J. Horty and A.J.I. Jones, editor, *Six International Workshop on Deontic Logic in Computer Science,* 2002.
5. R. Demolombe. Belief change: from Situation Calculus to Modal Logic. In G. Brewka and P. Peppas, editor, *Proc. ofthe Workshop on Nonmonotonic Reasoning, Action and Change,* 2003.
6. R. Demolombe and M. P. Pozos-Parra. A simple and tractable extension of situation calculus to epistemic logic. In Z. W. Ras and S. Ohsuga, editors, *Proc. of 12th International Symposium ISMIS 2000.* Springer. LNAI 1932, 2000.
7. F. Lin and R. Reiter. State constraints revisited. *Journal ofLogic and Computation,* 4:655–678, 1994.
8. S. McIllraith. A closed-form solution to the ramification problem (sometimes). In *Proc. of the IJCAI'97. Workshop on Non Monotonic Reasoning Action and Change,* 1997.
9. D. Nute. Norms, priorities and defeasibility. In P. McNamara and H. Prakken, editors, *Norms, Logics and Information Systems,* pages 201–218. IOS Press, 1999.
10. H. Prakken. Two approaches of defeasible reasoning. In A.J.I. Jones and M. Sergot, editors, *2d International Workshop on Deontic Logic in Computer Science,* pages 281–295. Tano A.S., 1994.
11. R. Reiter. The frame problem in the situation calculus: a simple solution (sometimes) and a completeness result for goal regression. In V. Lifschitz, editor, *Artificial Intelligence and Mathematical Theory of Computation: Papers in Honor of John McCarthy,* pages 359–380. Academic Press, 1991.
12. L. Royakkers and F. Dignum. Defeasible reasoning with legal rules. In M. A. Brown and J. Carmo, editors, *Deontic Logic, Agency and Normative Systems,* pages 174–193. Springer, 1996.

13. Sahlqvist, H. Completeness and correspondence in the first and second order semantics for modal logics. In Stig Kanger, editor, *Proc. 3rd Scandinavian Logic Symposium 1973,* number 82 in Studies in Logic. North Holland, 1975.
14. R. Scherl and H.J. Levesque. Knowledge, action and the frame problem. *Artificial Intelligence,* 144:1–39, 2003.
15. S. Shapiro, M. Pagnuco, Y. Lespérance, and H. Levesque. Iterated belief change in the situation calculus. In *Proc. of the 7th Conference on Principles on Knowledge Representation and Reasoning (KR2000).* Morgan Kaufman Publishers, 2000.
16. W. Snyder and C. Lynch. Goal directed strategies for paramodulation. In *Proc. 14th Int. Conf. on Rewriting Techniques and Applications (LNCS 488).* Springer-Verlag, 1991.
17. L. W. N. van der Torre and Y. H. Tan. An update semantics for deontic reasoning. In P. McNamara and H. Prakken, editors, *Norms, Logics and Information Systems,* pages 73–92. IOS Press, 1999.

Annex

Soundness and Completeness

Soundness of *LAPDO* can be proved as usual by proving that all the theorems are valid, and that the inference rules preserve validity. We prove completeness in several steps.

First we define the set of all instances of axioms *(Persist)* and $(Persist_O)$:

$$Indep(D) = \{\neg Pre_D(\alpha, |L|) \rightarrow (L \rightarrow [\alpha]L) : \alpha \in ACT \text{ and } L \in LIT\}$$

$$Indep(DO) =$$

$$\{\neg Pre_{DO}(\alpha, |LM|) \rightarrow (LM \rightarrow [\alpha]LM) : \alpha \in ACT \text{ and } LM \in LITM\}$$

We abbreviate $Indep(D, DO) = Indep(D) \cup Indep(DO)$.

Let $LAPDO_0$ be the basic logic of dependence and obligations such that

$$D_0 = DO_0 = \{\langle \alpha, P, \top \rangle : \alpha \in ACT \text{ and } P \in ATM\}$$

Lemma. If $\models_{LAPDO} p$ then $Indep(D, DO) \models_{LAPDO_0} p$.

This follows from the fact that the class of models of *LAPDO* is just the same as the class of those models μ of $LAPDO_0$ where *Indep(D, DO)* is true in μ. (A set of formulas is true in μ iff each of its elements is true in every possible world of μ.)

Now we restrict *Indep(D)* and *Indep(DO)* to the language of p:

$$Indep(D, DO, p) = Indep(D, DO) \cap lang(p)$$

where $lang(p)$ is the language of p, i.e. the set of formulas built from the actions and atoms appearing in p.

Lemma. If $Indep(D, DO) \models_{LAPDO_0} p$ then $Indep(D, DO, p) \models_{LAPDO_0} p$.

As $Indep(D, DO, p)$ is finite we can formulate the following.

Lemma. If $Indep(D, DO, p) \models_{LAPDO_0} p$ then

$$\models_{LAPDO_0} (\Box \bigwedge Indep(D, DO, p)) \rightarrow p.$$

This follows from the fact that $R_{\Box}$ contains the reflexive and transitive closure of the union of R_{Obg} and all the accessibility relations R_{α}.

In logic $LAPDO_0$ we have that $Pre_D(\alpha, P) = Pre_{DO}(\alpha, P) = \top$ for every α and p. Therefore axioms *(Persist)* and $(Persist_O)$ are redundant in that logic and can be dropped. As the remaining axioms are standard ones, the following is guaranteed by Sahlqvist's completeness theorem [13].

Lemma. If $\models_{LAPDO_0} (\Box \bigwedge Indep(D, DO, p)) \rightarrow p$ then
$\vdash_{LAPDO_0} (\Box \bigwedge Indep(D, DO, p)) \rightarrow p$.

Finally we have:

Lemma. If $\vdash_{LAPDO_0} (\Box \bigwedge Indep(D, DO, p)) \rightarrow p$ then
$Indep(D, DO) \vdash_{LAPDO_0} \rightarrow p$.

and

Lemma. If $Indep(D, DO) \vdash_{LAPDO_0} \rightarrow p$ then $\vdash_{LAPDO} p$.

The latter is because the axioms of *LAPDO* are those of $LAPDO_0$ plus axioms *(Persist)* and $(Persist_O)$. The set *Indep(D, DO)* collects all instances of the latter axioms.

Putting the preceding lemmas together we obtain that $\models_{LAPDO} p$ implies $\vdash_{LAPDO} p$. Hence our logic *LAPDO* is complete.

It follows a fortiori that LAPD is complete, too.

A Proposal for Dealing with Deontic Dilemmas

Lou Goble

Willamette University, Salem, Oregon 97301, USA
lgoble@willamette.edu

Abstract. In this paper I propose a simple modification of standard deontic logic that will enable the system to accommodate deontic dilemmas without inconsistency and without deontic explosion, while at the same time preserving the range of genuinely valid inferences. The proposal applies both to monadic deontic logic and to a dyadic logic of conditional obligation. In the Appendix these systems are proved to be sound and complete with respect to an appropriate semantics and also to be decidable.

In what follows I want to discuss deontic dilemmas and the proper way to treat them in deontic logic. In doing so, I shall follow a fairly simple, modest, even conservative, course and treat deontic logic as an elementary modal logic of a quite ordinary sort[1]. It is my intention to show that such a logic can accommodate deontic dilemmas in a reasonable way despite some objections that have been raised, especially by Horty in a number of works, [19], [20], [21].

In Section 1 below I will describe what I mean by 'deontic dilemma' more precisely and the problem that such dilemmas pose for deontic logic. After that, in Section 2, I consider briefly some common approaches to this problem that appear not entirely adequate. Drawing on lessons seen there, I will present my new proposal in Section 3; it should fare better. The discussion to that point will concern monadic deontic logic only, since that is the easiest framework to

[1] Hence, throughout this discussion, I suppose a propositional language with formulas A in the usual vocabulary, and a single monadic modal operator O to represent 'it ought to be that …' with OA well-formed whenever A is. (Later I will extend the language to include a dyadic operator $O(-/-)$ for conditional oughts.) I thus use the idiom of ought-to-be, which some would distinguish from ought-to-do. I do not discriminate between the two since analogous issues arise for both locutions. I also abstract from considerations of agency and action, and from such issues as the time of an obligation, the authority that institutes it or the person to whom it might be directed, if any. These are all significant factors of normative discourse and so deserve to be treated in deontic logic. But it is reasonable to suppose that they do not have particular bearing with respect to the question of deontic dilemmas before us. The same issues should arise with any further sophistication of the logic, and should probably be treated in much the same way. Thus, what I present here might be taken as a blueprint for a more detailed treatment in richer contexts. In a similar vein, I do not distinguish between so-called *prima facie* oughts and *actual,* or all-things-considered, oughts, since here too similar problems should arise for both.

A. Lomuscio and D. Nute (Eds.): DEON 2004, LNAI 3065, pp. 74–113, 2004.

work with, and it suffices to bring out the central issues that the prospect of deontic dilemmas raises. Section 4, however, extends these considerations to cover conditional obligation. There we will also see how a similar maneuver can answer another, separate problem that Horty has raised against ordinary dyadic deontic logics. Section 5 concludes with some comparison between the approach taken here and Horty's way of developing deontic logic since Horty put forward the objections that have motivated this proposal the most. I reserve for the Appendix the full formal presentation of the deontic logics that result from my account, their proof theory and semantics, and there establish that the systems are sound and complete, and as a side benefit of that demonstration, that they have the finite model property and are decidable.

1 The Problem of Deontic Dilemmas

By a 'deontic dilemma' I mean a situation in which, in a univocal sense of 'ought', some state of affairs, *A,* both ought to be and ought not to be, in which, that is, both OA and $O\neg A$ are true. More broadly, a deontic dilemma would be a situation in which there are inconsistent states of affairs, *A* and *B,* both of which ought to be, that is, a case where $\vdash A \rightarrow \neg B$ and yet both OA and OB are true. More broadly still, a deontic dilemma would be a situation in which it is impossible for both *A* and *B* to be realized even though both ought to be, where the sense of impossibility could be anything appropriate to the context of discourse, from some metaphysical impossibility to the most mundane practical incompatibility. More generally too, along another dimension, deontic dilemmas could be conditional. These would be situations in which both it ought to be that *A* on a condition *B,* and also it ought to be that not-*A* on the same condition, i.e., where both $O(A/B)$ and $O(\neg A/B)$ are true, and similarly for the other senses of incompatible requirements.

Of course, the first sense of deontic dilemma is easily seen to be a case of the others. In addition, given seemingly innocent assumptions, like the inheritance rule,

RM) If $\vdash A \rightarrow B$ then $\vdash OA \rightarrow OB$

or natural variants of it, then the second broader sense of dilemma reduces to the first; i.e., any case of the second will imply a case of the first. Hence, any deontic logic that eschews the first sort of dilemma, must eschew the second. Similarly, if the sense of impossibility in the third description is such that the logic contains anything like

NM) $\vdash \neg\Diamond(A \wedge \neg B) \rightarrow (OA \rightarrow OB)$

(with $\Diamond$ for the appropriate sense of possibility), then the yet broader sense of dilemma also reduces to the first, and any logic that eschews the one must eschew the other. Hence we can focus our attention primarily on the first sort of dilemma, since it is easiest to discuss, though natural examples might take the form of the second or third.

It is plausible that there are deontic dilemmas, indeed that they are very common[2]. I shall not argue that here, however. Instead, I will simply take it for granted that there are, or could be, such cases in order to investigate how deontic logic should accommodate them.

Any deontic logic that accommodates deontic dilemmas must, of course, not contain the principle

D) $\quad \vdash OA \to \neg O \neg A$

lest it license contradictions. (D) is central to standard deontic logic, SDL, in its many variants, and so SDL can be thought of as denying the possibility of dilemmas. Or, one might think of commitment to (D) rather as defining the range of application of the logic. One might think that, while deontic dilemmas might be possible, standard deontic logic only applies to the logic of normative systems that are in fact consistent or dilemma-free[3]. Such an approach has severe drawbacks, however, and cannot be maintained. Nevertheless, it does suggest a measure of adequacy that we might apply to a logic that does accommodate dilemmas, namely that it should be equivalent to SDL in case there aren't any. That is, for purposes of this inquiry, we should make minimal changes to SDL in order to tolerate deontic dilemmas. One way to put this is to say

(∗) A deontic logic for dilemmas should be such that the result of adding (D) as an axiom to it is equivalent to SDL.

I think this is a worthwhile criterion, but I do not insist on it. Most proposals to accommodate deontic dilemmas do not meet this condition; the one I propose below does.

[2] This is especially so when one considers the norms that apply to multiple agents, for one agent might be required to do one thing while another agent is required to do something else that is incompatible with the first. Think of two players of a game, or competitors of any sort. Or think of two people who have each promised something which precludes the other's fulfilling his promise, etc. Of course, since its inception, standard deontic logic has denied the possibility of conflicts of obligation, and there is a long standing philosophical tradition that argues against the possibility at least of *moral* dilemmas. See Horty [21] for an examination of several such arguments, and also Forrester [6] for more sources. Many in that tradition nowadays maintain that what look like cases of dilemmas, where it looks as though a person ought to do something A and ought to do something else B but can't do both, are not really dilemmas; rather they are situations where it is not the case that the person ought to do A and not the case that the person ought to do B, but only the case that the person ought to do (A-or-B). I am unconvinced, and all the more so when we consider that the agents of the obligations might be distinct. See Routley and Plumwood [32] for more discussion of how normative conflicts pervade our lives.

[3] This is analogous to the way one might preserve the inference

(A) All S are P
$\therefore$ (I) Some S are P

by saying that the logic that contains it applies only to that part of the language in which all terms S have existential import. Cf. Lambert [24], p. 261f. for discussion of this, and the reasons why this sort of maneuver should be rejected. I develop this theme further below.

Any deontic logic that accommodates dilemmas must also not contain principles of 'deontic explosion', such as

DEX) $\vdash (OA \wedge O\neg A) \rightarrow OB$

which says that if there is any instance of a deontic dilemma then *everything* is obligatory. (And similarly for the other broader senses of dilemma.) It is plausible that there are deontic dilemmas; it is not plausible that everything ought to be the case. Hence, (DEX) must be rejected. We might make this too a condition of adequacy for a deontic logic that can accommodate dilemmas:

(**) A deontic logic for dilemmas should not contain (DEX), or anything like it.

(In what follows I will be more concerned with (**) than (*).)

The *problem* that deontic dilemmas present for deontic logic is simply the question of how to avoid deontic explosion, and (D), while at the same time accounting for the full range of inferences that do seem valid for normative concepts. Any logic that contains the rule (RM) mentioned above and the aggregation principle

AND) $\vdash (OA \wedge OB) \rightarrow O(A \wedge B)$

and the principle of *ex falso quodlibet*

EFQ) $\vdash (A \wedge \neg A) \rightarrow B$

will *ipso facto* contain (DEX). Hence, to be adequate for deontic dilemmas, the logic must reject or restrict at least one of the principles (RM), (AND) and (EFQ). The question is, What is the best way to do that[4]?

[4] Thus, the problem to be addressed here is the problem of what principles should, and should not, be contained in a logic that allows for deontic dilemmas. This should be distinguished from the question of how such dilemmas should be resolved, or how one should decide to act in the face of such a dilemma. It should also be distinguished from the kind of case often envisaged in the literature of defeasible reasoning, whereby one might have information that, from a classical point of view, seems to lead to inconsistent conclusions, but where some of that information defeats the application of other information, so that no conflict is in fact generated. (Birds fly; emus are birds; emus don't fly; Edward is an emu. One concludes that Edward doesn't fly, because he's an emu; one doesn't conclude that Edward flies, in spite of his being a bird.) Similarly, one might have information that points to the conclusion that one ought to do *A,* but further information that points to the conclusion that one ought to do *B,* when *A* and *B* are incompatible, where the latter information overrides, or defeats, the former reasoning. (One might think, for example, of sets of regulations some of which enjoin *A* and others forbid *A;* there might be mechanisms of priority that make only one injunction operative in a particular case, so that the prohibition of *A* defeats the injunction for *A*.) A deontic dilemma is a case where neither claim of obligation, *OA* and *OB* for incompatible *A* and *B,* is defeated; both are true. How a deontic logic should accomodate that sort of situation is what concerns me here.

It is convenient to note that the inheritance rule (RM) is equivalent to the converse of (AND) and also to the principle (OR), namely

M) $\vdash O(A \wedge B) \rightarrow (OA \wedge OB)$
OR) $\vdash OA \rightarrow O(A \vee B)$

given the rule of replacement for equivalents

RE) If $\vdash A \leftrightarrow B$ then $\vdash OA \leftrightarrow OB$

which seems a *sine qua non* for any reasonable deontic logic. That is to say, given (RM) then both (M) and (OR) are derivable (and, of course (RE)), and given either (M) or (OR), with (RE), then (RM) is derivable. Hence (M) and (OR) are equally implicated in the derivation of (DEX).

2 Some Proposed Solutions

The problem posed by deontic dilemmas is really two-faced. On the one hand, one wants a logic that is not too strong; it must avoid (D), which is easy, and it must avoid deontic explosion (DEX), which is also easy, though less so. On the other hand, it must not be too weak; it must capture all the inferences one wants for the operator *O*. That second side will become more clear in the discussion in Sect. 2.3 below. It is this that makes the problem of deontic dilemmas a challenge.

As noted above, to avoid (DEX), at least one of (EFQ), (RM) and (AND) must be rejected or restricted. This suggests three ways one might try to weaken standard deontic logic. Let us consider them briefly in turn, with emphasis on (AND) since that might be the most common strategy[5].

2.1 Reject (EFQ)

Perhaps the most direct way to avoid the derivation of (DEX) is to deny the principle of *ex falso quodlibet*. This means basing one's deontic logic on a paraconsistent logic, rather than the classical propositional calulus, PC, that is usually assumed. A natural, and well-developed, alternative is the relevant logic **R**. Routley and Plumwood recommend this in [32], and in my own [7] and [8], I proposed similar systems for this purpose. I find relevant logic attractive, and think that (EFQ) is indeed the real culprit behind deontic explosion. Nevertheless, this is a fairly radical departure from standard deontic logics, and requires defending

[5] This will not be an exhaustive review of the proposals that have been made. Van der Torre and Tan [35], for example, present an interesting proposal that I do not discuss. Theirs is a hybrid that would modify two of the principles, (RM) and (AND), with a 'two-phase' deontic logic that prevents deontic explosion by not only restricting (AND) but also controlling the order of the application of the rules; in effect, this distinguishes two senses of 'ought' where some standard principles apply to one and other principles to the other. While intriguing, this kind of approach goes in a rather different direction from the proposal I want to offer, although it may bear some similarity to the system I call **DPM.2** below. I have not investigated those connections, however.

an approach to logic in general that goes well beyond deontic considerations. (Amongst other things, it requires abandoning such intuitive principles as the Disjunctive Syllogism, $\vdash ((A \vee B) \wedge \neg A) \rightarrow B$.) Rather than enter those battles, for present purposes I will simply set this approach aside[6].

2.2 Reject Modal Inheritance (RM)

Keeping all of classical PC, including (EFQ), but denying the rule of monotonicity or inheritance for *O,* the rule (RM), and its partners (M) and (OR), will also clearly block the derivation of (DEX). Various authors have, for various reasons, called this rule into question, e.g., Jackson [23], Hansson [16], [17], pp. 141ff., and myself in [9], [10], [11], [13]. Generally speaking, however, the reasons for questioning (RM) have little to do with the question of deontic dilemmas, and more to do with other paradoxes of deontic logic. Indeed, my own proposals along these lines contained (D) and so are incompatible with accepting the possibility of deontic dilemmas.

While I continue to be suspicious of (RM), I will not pursue its wholesale rejection here. In a discussion, not of deontic dilemmas, but of the other deontic paradoxes, Nute and Yu ([31], p.5) comment on my rejection of (RM), saying,

> But the principle of inheritance of obligations is one of the most fundamental principles of SDL and has strong intuitive appeal. It requires the agent to take moral responsibility for the logical consequences of what he/she has committed to do. The rejection of the principle, therefore, seems to be contrary to one of our basic moral reasoning patterns.

Certainly (RM) does have strong intuitive appeal. Nonetheless, to anticipate later discussion, I shall propose modifying it. This will not be a wholesale rejection of the principle, as in the works cited above, but rather a limitation on it that should take the intuitive appeal of the rule into account. That is the subject of Section 3 below[7].

[6] Other paraconsistent deontic logics are found in da Costa [3], da Costa and Carnielli [4], and Loparic and Puga [27]. Casey McGinnnis, in work as yet unpublished [28], [29], presents a variation on this theme with what he calls 'semiparaconsistent' deontic logic. In this, the actual world is construed classically, and validity is determined with respect to that world. Hence all of PC is valid, including disjunctive syllogism, which gives rise to *ex falso.* At the same time, however, deontically alternative points are construed paraconsistently, in a 3- or 4-valued way. The deontic logic that results contains the principles (AND) and (K) of standard deontic logic, but not (D) and not (DEX) because it lacks deontic disjunctive syllogism, $(O(A \vee B) \wedge O\neg A) \rightarrow OB$. As a result, this proposal is vulnerable to an objection raised by Horty that is discussed in Sect. 2.3 below. (The same is true for deontic logics based on **R**.) McGinnis's approach also has some peculiar consequences, such as lacking a full replacement theorem (rule (RE)) since $(A \rightarrow B) \leftrightarrow (\neg A \vee B)$ is valid but $O(A \rightarrow B) \leftrightarrow O(\neg A \vee B)$ is not. How much this vitiates the logic is a worthy question.

[7] That proposal does not address the general issues of the standard deontic paradoxes. My aim in this paper is to focus entirely on the question of deontic dilemmas and deviate as little as possible from standard deontic logic in order to accommodate them.

2.3 Reject Aggregation (AND)

Given the strong appeal of (RM) (and complacent attachment to PC, including (EFQ)), perhaps the most natural suggestion for avoiding deontic explosion is to reject the aggregation principle (AND). In [12], [14] and [15], I recommend such a logic precisely for this purpose. I called this logic **P**. It is axiomatized by PC, with closure under *modus ponens,* the inheritance principle (RM) and two minimal axioms, (N) $\vdash O\top$ and (P) $\vdash \neg O\bot$, where $\top$ is any tautology and $\bot$ is $\neg\top$. Since **P** lacks (AND), neither (D) nor (DEX) is derivable. (Adding (AND) to **P** yields full SDL, but adding (D) alone does not. Hence **P** fails condition $(*)$.)[8]

P is very well-behaved. It has a natural interpretation in terms of neighborhood semantics, after Segerberg [34] or Chellas [2], and in terms of preference-based models, [12], [14] and [15], as well as in an extension of Kripke-models, [12], [33]. Nevertheless, it is a very weak deontic logic, perhaps too weak. This is the concern to which I alluded as the second face of the problem posed by deontic dilemmas. It is this that motivates the present discussion.

Van Fraassen [37] and after him Horty [19], [20], [21] have argued that systems like **P** fail to account for patterns of inference that seem unobjectionable and that seem to require the principle of aggregation (AND). Horty frequently gives the example of a person, perhaps a conscientious objector, who recognizes the obligations for someone, Smith,

i) Smith ought to fight in the army or perform alternative service to his country — $O(F \vee S)$
ii) Smith ought not to fight in the army — $O\neg F$

and who then reasons to the conclusion

iii) Smith ought to perform alternative service to his country — OS

Whether this is a case of Smith deliberating for himself what he should do, or someone else describing the situation that pertains to Smith, the inference from (i) and (ii) to (iii) seems valid. Given the principle of aggregation, that is easy to explain. By (AND), (i) and (ii) entail $O((F \vee S) \wedge \neg F)$. Since $\vdash ((F \vee S) \wedge \neg F) \rightarrow S$, $\vdash O((F \vee S) \wedge \neg F) \rightarrow OS$, by (RM). So, given $O((F \vee S) \wedge \neg F)$, (iii) OS follows. Nothing in **P** licenses this inference, however, and this seems a significant shortcoming of the system and others like it.

It is considerations like this that make the problem of deontic dilemmas a difficult problem. How can one steer a middle course between a normal logic like SDL, or even **K**, which are clearly too strong, and a minimal logic like

[8] Others have also proposed this, or a very similar system, for the same purpose. For example, Schotch and Jennings [33] likewise introduced the same system in order to allow for deontic dilemmas. **P** is very like the first system van Fraassen proposed in [37], p. 16, though he backed away from it for reasons we will discuss below. **P** differs from van Fraassen's in that **P** contains (N) while his does not, but I take this to be an insignificant difference; (N) fails only in models in which nothing is obligatory. Chellas [2], p. 202 proposes the same system as van Fraassen's first as a minimal deontic logic; this too in order to permit deontic dilemmas.

P, which appears to be too weak? Since aggregation (AND) seems to be what distinguishes **P** from SDL, perhaps there is a way to restrict (AND) without rejecting it altogether.

2.4 Restrict Aggregation

Here we look at three ways to limit the aggregation principle; at least two of them won't work.

2.4.1 Consistent Aggregation. A first natural suggestion for a way to accommodate both situations in which there are deontic dilemmas and the cases of innocent inferences using aggregation is to adopt a principle that allows $OA \wedge OB$ to entail $O(A \wedge B)$ except when that would get one into trouble, as when A and B are incompatible, which, as we have seen, would lead to deontic explosion, not to mention a violation of principle (P). Hence, it seems plausible simply to restrict aggregation to those cases where A and B are consistent (or jointly possible). Call this the principle of *Consistent Aggregation* or

ConAND) If $\nvdash A \rightarrow \neg B$ then $\vdash (OA \wedge OB) \rightarrow O(A \wedge B)$ [9]

While this move might seem natural[10], (ConAND) is still too strong and will not serve as it is supposed to. Here is a counterexample (adapted from Horty [21] p. 581). Suppose a situation in which someone, Jones, ought to visit his daughter Abby at a certain time — OV_a. It is plausible that in this situation he should notify her he is coming and then visit her — $O(V_a \wedge N_a)$. But it could also be that in the very same situation Jones ought also to visit his daughter Beth at that same time, and indeed that he should notify her he is coming and then visit her — $O(V_b \wedge N_b)$. Because of circumstances, however, such as that Abby and Beth live on opposite sides of the country, it is impossible for Jones to visit both at that time. Thus he faces a deontic dilemma. Both $O(V_a \wedge N_a)$ and $O(V_b \wedge N_b)$ are true, though presumably $O((V_a \wedge N_a) \wedge (V_b \wedge N_b))$ is not. But from $O(V_a \wedge N_a)$, ON_a follows by (RM), and similarly ON_b follows from $O(V_b \wedge N_b)$. N_a and N_b are consistent; hence they are candidates for (ConAND). Since both ON_a and ON_b are true and N_a and N_b are consistent, it follows by (ConAND) that $O(N_a \wedge N_b)$ is true, that Jones ought to notify both his daughters he is coming to visit, and indeed that he ought to notify both that he is coming even if he only goes to see one of them. That seems going too far.

[9] As with the principle (NM), mentioned in Sect. 1, if the language has alethic modalities, this rule might be replaced with a stronger postulate $\vdash \Diamond(A \wedge B) \rightarrow ((OA \wedge OB) \rightarrow O(A \wedge B))$, with $\Diamond$ for any appropriate sense of possibility. All the remarks to follow would apply *mutatis mutandis* to this as well.

[10] I know of no published source that adopts this rule, and for good reason. Nevertheless, it is the sort of proposal that comes to mind first when considering how to handle deontic dilemmas. At least, it has come up often in conversations. (Van der Torre and Tan [35], p. 411 attribute this principle, with this name, to van Fraassen [37], but I do not find it there.)

Jörg Hansen presented a similar problem for a version of (ConAND) proposed by Paul McNamara. As McNamara [30] presents the example, the story goes like this (with free use of alethic modalities): Suppose a person ought to do something A that necessitates something else C, and likewise ought to do something B that necessitates D, when A and B are incompatible but C and D are not. With (ConAND) one then infers $O(C \wedge D)$, which has no support in the situation. For example, Jones ought to keep an appointment in Montreal on Monday morning (OA) and Jones ought to keep an appointment in London on Monday afternoon (OB), where we may assume it is impossible for Jones to do both, given the distances ($\neg\Diamond(A \wedge B)$). Hence there is a dilemma. To keep the appointment in Montreal necessitates traveling to Montreal in the morning ($\Box(A \to C)$), while keeping the appointment in London necessitates departing for London in the morning ($\Box(B \to D)$). It is, however, possible to travel to Montreal in the morning and depart from there for London, ($\Diamond(C \wedge D)$). With (RM) one can then infer first both OC and OD, thence $O(C \wedge D)$ by (ConAND). But that seems contrary to the facts of the case.

Examples like this should make one suspicious of (ConAND). Moreover, we can make a stronger, more general case against this rule[11]. Consider a case of a deontic dilemma where we suppose OA and $O\neg A$ to hold, and let B be any consistent proposition, so that $\nvdash B$. We show that OB. B must be consistent with either A or ¬A; suppose it is $\neg A$, so that $\nvdash B \to A$, and argue:

i)	OA	hyp
ii)	$O\neg A$	hyp
iii)	$\nvdash B$	hyp
iv)	$\nvdash B \to A$	hyp
v)	$O(A \vee B)$	i, PC, RM
vi)	$\nvdash (A \vee B) \to \neg\neg A$	iv, PC
vii)	$O((A \vee B) \wedge \neg A)$	ii, v, vi, ConAND
viii)	$\vdash ((A \vee B) \wedge \neg A) \to B$	PC
ix)	$\vdash O((A \vee B) \wedge \neg A) \to OB$	viii, RM
x)	OB	vii, ix, PC

In case B is consistent with A the argument is similar, and so we may discharge the hypothesis at (iv). Thus we conclude that if there is any deontic dilemma, then anything consistent is obligatory. Call this rule:

DEX-1) If $\nvdash B$ then $\vdash (OA \wedge O\neg A) \to OB$

(DEX-1) does not go quite as far as full deontic explosion (DEX) that follows from full aggregation, where B could be anything at all, but it is still absurd. It still means the collapse of normative distinctions in plausible circumstances. Moreover, B could easily be taken to be something specified to be normatively neutral, and then there would be a direct contradiction. Clearly then, an adequate deontic logic must reject (DEX-1) no less than (DEX). Hence, consistent aggregation, (ConAND), is far too strong, not much better than complete aggregation, (AND), itself.

[11] Van der Torre and Tan [35], p. 412, observe much the same.

2.4.2 Weakened Consistent Aggregation. The logic that Horty [21] presents to countenance deontic, or moral, dilemmas does not satisfy consistent aggregation. Instead it supports a weaker rule he calls 'consistent consequent agglomeration'. Just what this is, is difficult to describe without more machinery than we have available. Very roughly, Horty's account goes like this (but see [21] for details). First, he distinguishes two sorts of 'ought'; one, represented by formulas $!(A)$, is for *prima facie* oughts, perhaps derived directly from imperatives. The other, represented by formulas $\bigcirc(A)$, is for the all-things-considered ought[12].

Horty's question is how such all-things-considered oughts are derived from sets of prima facie oughts, especially in the face of moral conflicts. His answer, very roughly, is that $\bigcirc(A)$ follows from a set $\mathcal{I}$ of *prima facie* oughts just in case A is a logical consequence of a maximal consistent subset of the applicable binding *prima facie* oughts in $\mathcal{I}$. The rule of consistent consequent agglomeration (CCA) can now be (roughly) stated: Suppose a number of oughts $\bigcirc(B_1), \ldots, \bigcirc(B_n)$ are consequent on a set of *prima facie* oughts $\mathcal{I}$, then the aggregate $\bigcirc(B_1 \wedge \cdots \wedge B_n)$ is consequent on $\mathcal{I}$ just in case the set $\{B_1, \ldots, B_n\}$ is both (i) consistent and (ii) a subset of the set of propositions enjoined by the binding members of $\mathcal{I}$. (Cf. [21] p. 580.)

Condition (i) of (CCA) is like the rule of consistent aggregation; condition (ii) lets (CCA) escape the problems that confronted that rule. In the example of Jones visiting his daughters, we can suppose that the setup is such that the relevant set of *prima facie* oughts $\mathcal{I}$ is $\{!(V_a \wedge N_a), !(V_b \wedge N_b)\}$ and that both are binding on Jones. From this set, both $\bigcirc(V_a \wedge N_a)$ and $\bigcirc(V_b \wedge N_b)$ follow, though $\bigcirc((V_a \wedge N_a) \wedge (V_b \wedge N_b))$ does not, just as we should want. And neither does $\bigcirc(N_a \wedge N_b)$, for the set $\{N_a, N_b\}$, though consistent, is not a subset of the propositions enjoined by the members of $\mathcal{I}$ since neither $!N_a$ nor $!N_b$ is in $\mathcal{I}$. (Both $\bigcirc(N_a)$ and $\bigcirc(N_b)$ would, however, follow from $\mathcal{I}$, but their aggregate $\bigcirc(N_a \wedge N_b)$ does not, which is the key point now.) Hansen's example would be treated similarly, as would the more general problem that led to (DEX-1). The crucial step there is step (vii), but this cannot be inferred from (ii) and (v) by (CCA) even given (vi) since $\{A \vee B, \neg A\}$ is not a subset of the propositions enjoined by the operative set of background oughts, which we can take now to be just $!(A)$ and $!(\neg A)$. By contrast, when aggregation is wanted, it is available. Thus in the example of Smith and his service to his country, let us suppose the operative set of *prima facie* oughts $\mathcal{I}$ is $\{!(F \vee S), !(\neg F)\}$ and that both are binding. Then $\bigcirc(F \vee S)$ and $\bigcirc(\neg F)$ follow from $\mathcal{I}$, and so too does $\bigcirc((F \vee S) \wedge \neg F)$ since $\{(F \vee S), \neg F\}$ is a consistent subset of propositions enjoined by the binding oughts in $\mathcal{I}$. From this $\bigcirc S$ follows since (RM) holds for $\bigcirc$ in Horty's system.

Thus the rule (CCA) seems to do what is asked of it. Moreover, if there are no deontic dilemmas, i.e., if the background set $\mathcal{I}$ of *prima facie* obligations is

[12] Horty [21] actually sets everything up for conditional oughts, of both kinds, to have formulas $!(B/A)$ and $\bigcirc(B/A)$, each to say in its sense that under conditions A, it ought to be that B. B here is the 'consequent', which explains the name of Horty's rule. The present monadic simplification will suffice for our purposes. As is customary, $!(A)$ and $\bigcirc(A)$ are defined as $!(A/\top)$ and $\bigcirc(A/\top)$, respectively.

conflict free, then Horty's system agrees with SDL; this seems a desirable feature, not found in systems that deny (EFQ), or (RM) or (AND)[13].

We might find further philosophical support for this sort of approach as follows. Consider familiar accounts of the Kripke-style semantics for standard deontic logic. There *OA* is said to hold at a possible world just in case *A* holds at all the 'ideal' or deontically 'best' possible worlds (accessible from the given world). A possible world is typically considered ideal insofar as it is a world where all obligations are fulfilled (cf. Hilpinen [18], p. 163). If there are deontic dilemmas, however, there can be no ideal worlds in this sense (so long as we keep to a classical view that possible worlds are entirely consistent). But we might still think that a world is ideal, not perhaps when *all* obligations are fulfilled, but when all that can consistently be fulfilled are, when the world is as good as it *can* be. That is, we might think of a world as ideal just when a maximal consistent set of obligations are fulfilled in it. This might be thought of as a semantic counterpart to Horty's picture.

Unfortunately, there does not seem any way to realize this picture in a straight-forward possible-worlds type model theory for deontic logic. Although we explain the idea of an ideal world in terms of the obligations that obtain in a given world, to define truth conditions for formulas *OA* we must take the notion of ideal world as primitive. If we were to say that *OA* is true at a possible world just in case *A* is true at all ideal worlds, we are back in the original fix that either there are no ideal worlds, in which case deontic explosion is validated, or else there are, in which case (D) is validated. If we were to say that *OA* is true just in case *A* is true in some ideal world, we lose the validity of the inference concerning Smith's obligations to his country, inferences with deontic disjunctive syllogism when there is no conflict of obligation involved.

Horty himself does not try to present his system in this sort of possible-world/model-theoretic terms, and indeed it is considerations like this that lead him away from thinking of deontic logic in the framework of traditional modal logics. His proposals require a rather radical rethinking of the foundations for deontic logic. Not being able to implement the modified semantic picture of ideal worlds suggested above, does, however, point to a limitation of the approach Horty has taken. He himself remarked that condition (ii) of his rule of consistent consequent agglomeration may seem "peculiar, or at least excessively syntactic" ([21] p. 580). This limits aggregation to cases determined by the particular specification of the *prima facie* oughts in $\mathcal{I}$. It means, amongst other things, that $\mathcal{I}$ cannot be considered closed under logical consequence. If it were, then given $!(V_a \wedge N_a) \in \mathcal{I}$ and given $!(V_b \wedge N_b) \in \mathcal{I}$, we should have $!N_a \in \mathcal{I}$ and $!N_b \in \mathcal{I}$, and then $\bigcirc(N_a \wedge N_b)$ would follow from $\mathcal{I}$. Similarly, if $!(A \vee B)$ followed from $!(A)$, then the problem that gave rise to (DEX-1) would reappear. This suggests that, although Horty has given an account of how all-things-considered oughts follow from specific sets of *prima facie* oughts, this account leaves no room for a logic of *prima facie* oughts themselves.

[13] Cf. the criterion of adequacy $(*)$ of Sect. 1, though we won't go so far as to say that Horty's system plus (D) is equivalent to SDL because they are too different in their fundamentals.

2.4.3 Permitted Aggregation. The rules of consistent aggregation (ConAND) and consistent consequent agglomeration (CCA) screen candidates for combination by conjunction through the logical property of consistency (or more broadly possibility). In place of that one might consider screening for *normative* consistency (or normative possibility). This is simply the notion of (joint) permissibility, which is already available in the language. Then one could say that aggregation is permitted, as it were, when the aggregate is itself permitted. This would be the rule

PAND) $\vdash P(A \wedge B) \rightarrow ((OA \wedge OB) \rightarrow O(A \wedge B))$

where, as usual, $PA =_{df} \neg O \neg A$. This rule could then be added to the logic **P** to form the system **PA**[14].

Consider how **PA** handles the previous examples. For the case of Smith and his service to his country, one wants to infer $O((F \vee S) \wedge \neg F)$ from $O(F \vee S)$ and $O\neg F$ in order to conclude OS. Implicit in the example is that it is all right, i.e., permitted, that Smith perform alternate service to his country and not fight in the army; without that, the example has no intuitive appeal. Thus we can take $P(\neg F \wedge S)$ as an implicit premise in the setup. So we have given

i)	$O(F \vee S)$	hyp
ii)	$O\neg F$	hyp
iii)	$P(\neg F \wedge S)$	hyp

and reason as follows: $\neg F \wedge S$ is logically equivalent to $(F \wedge \neg F) \vee (\neg F \wedge S)$, which is logically equivalent to $(F \vee S) \wedge \neg F$, hence

iv)	$P((F \vee S) \wedge \neg F)$	iii, PC, RM
v)	$O((F \vee S) \wedge \neg F)$	i, ii, iv, PAND
vi)	$\vdash (F \vee S) \wedge \neg F) \rightarrow S$	PC
vii)	$\vdash O((F \vee S) \wedge \neg F) \rightarrow OS$	vi, RM
viii)	OS	v, vii, PC

which seems to be just the argument one has in mind with examples like this.

In the case of Jones notifying his daughters he is coming to visit, for the example to count against consistent aggregation (ConAND) it must be assumed that there is something wrong with his notifying them both (when he will visit at most one). That is, implicit in the setup is the proposition that $O\neg(N_a \wedge N_b)$. If that is so, then we are given $\neg P(N_a \wedge N_b)$, and so we cannot have the condition necessary to apply (PAND) to infer $O(N_a \wedge N_b)$ from ON_a and ON_b under pain of having a contradictory premise set.

A similar point applies to the general problem that gave rise to (DEX-1) in Sect. 2.4.1. Suppose we are given OA and $O\neg A$ and we take some proposition B that is supposed to be not only logically consistent (compossible) with $\neg A$ (or A as the case might be) but normatively consistent with it. That is, let us assume

[14] This system is new. I once thought that, since (PAND) is properly weaker than (ConAND), it would be an adequate way to accommodate deontic dilemmas, but, as we shall see, it is not. Nevertheless, I present it here in part as a cautionary tale, and also as a step to the new proposal I will develop in the next section.

$P(B \wedge \neg A)$, and try to run the argument as before, with this assumption as line (iv) ((iii) drops out), and the counterpart of line (vi) following by the logical equivalence of $B \wedge \neg A$ and $(A \vee B) \wedge \neg A$. Then line (vii) seems justified by the new rule (PAND). Thus:

i)	OA	hyp
ii)	$O(\neg A)$	hyp
iv)	$P(B \wedge \neg A)$	hyp
v)	$O(A \vee B)$	i, PC, RM
vi)	$P((A \vee B) \wedge \neg A)$	iv, PC, RE
vii)	$O((A \vee B) \wedge \neg A)$	ii, v, vi, PAND
viii)	$\vdash ((A \vee B) \wedge \neg A) \rightarrow B$	PC
ix)	$\vdash O((A \vee B) \wedge \neg A) \rightarrow OB$	viii, RM
x)	OB	vi, ix, PC

(The argument would be similar if B were co-permissible with A instead of $\neg A$, i.e., with $P(B \wedge A)$ in place of $P(B \wedge \neg A)$ at line (iv).)

At first sight then (PAND) would seem to suffer the same sort of failing as (ConAND): if there is any deontic dilemma then anything co-permissible with one of the obligations is itself obligatory. This is not so, however. For as we look more closely at what we are given in (i), (ii) and (iv), we see that (i) OA entails $O(B \rightarrow A)$, by (PC) and (RM); hence it entails the logically equivalent $O\neg(B \wedge \neg A)$, which is to say, $\neg P(B \wedge \neg A)$. Thus (i) and (iv) are contradictory assumptions. It is no wonder then that untoward consequences, e.g., (x), follow from them. So long as the premise set of the argument is consistent, no problem should arise. Or so it might appear. So **PA** might seem an appropriate logic for deontic dilemmas.

Unfortunately, this proposal really fares no better than the rule (ConAND) above, for like (ConAND), (PAND) also yields a form of deontic explosion, namely that if there is any case of a deontic dilemma, then anything that is permitted will be obligatory, which seems absurd.

The argument for this consequence is much like the argument against the principle of Consistent Aggregation (ConAND) that was given above. Suppose a deontic dilemma, OA and $O\neg A$, and a proposition B such that PB.

i)	OA	hyp
ii)	$O\text{-}A$	hyp
iii)	PB	hyp
iv)	$O(A \vee B)$	i, PC, RM
v)	$O(\neg A \vee B)$	ii, PC, RM
vi)	$\vdash B \leftrightarrow ((A \vee B) \wedge (\neg A \vee B))$	PC
vii)	$P((A \vee B) \wedge (\neg A \vee B))$	iii, vi, RE
viii)	$O((A \vee B) \wedge (\neg A \vee B))$	iv, v, vii, PAND
ix)	OB	vi, viii, RE
x)	$\vdash (OA \wedge O\neg A) \rightarrow (PB \rightarrow OB)$	i–ix, Conditional Proof.

Call the principle (x) here

DEX-2) $\vdash (OA \wedge O\neg A) \rightarrow (PB \rightarrow OB)$

Though weaker than the original (DEX) and (DEX-1), this too seems absurd, and contrary to the spirit of accepting deontic dilemmas. It is enough to bar (PAND) and so the system **PA.** Moreover, the pattern of this argument would seem to generalize. As applied to the principle of Consistent Aggregation, it showed that if there is a deontic dilemma then any proposition that is consistent (or possible) is obligatory. As applied here, it shows that if there is a deontic dilemma then any proposition that is permitted is obligatory. Consider then any proposed restriction on the principle aggregation, a principle that if *OA* and *OB* and *Cond*(*A, B*), then $O(A \wedge B)$, where *Cond*(*A, B*) is some condition to be met by *A* and *B* together to reflect the limitation of aggregation, such as mutual consistency, compossibility, or co-permissibility, etc. If *Cond*(*A, B*) is such that it is appropriate to speak of its applying to a single proposition, *B,* or that *Cond*(*B, B*) could hold, especially if this is given as a modality on the conjunction of *A* and *B,* and if this condition or modality is preserved under replacement for logical equivalents, then the preceding sort of argument would seem to apply, regardless of the particular condition. If the condition is such that it is plausible that a proposition could meet it without being obligatory, then the restricted principle of aggregation based on it will be in trouble. This should cast doubt on any attempt to accommodate deontic dilemmas by limiting, but not excluding, aggregation (AND)[15].

3 Another Proposal: Permitted Inheritance

In this section I present my new proposal, drawing on the discussion of Permitted Aggregation, but redirecting its device. That is, instead of limiting the aggregation rule (AND), I propose to limit the inheritance principle (RM). The idea behind (ConAND), (CCA) and (PAND) was that aggregation should be allowed except in cases where it gets one into trouble, by producing deontic explosion, (DEX) or its variants. The same idea can be applied to the inheritance rule. Thus, I propose to replace the rule (RM) with

RPM) if $\vdash A \rightarrow B$ then $\vdash PA \rightarrow (OA \rightarrow OB)$

where, as before, $PA =_{df} \neg O\neg A$. Thus, if *A* entails *B* then if one ought to do *A* then one ought to do *B, provided that A* is permitted or consistent with the normative code. (Or, since $\vdash A \leftrightarrow (A \wedge B)$ when $\vdash A \rightarrow B$, this could be phrased in terms of the conjoint permissibility of *A* and *B,* as in (PAND), but that is not necessary.)[16].

[15] This remark does not apply to Horty's rule (CCA), which operates in a significantly different framework.

[16] As with (NM) and (ConAND), adapted in Footnote 9, this rule too can be strengthened to $\vdash \Box(A \rightarrow B) \rightarrow (PA \rightarrow (OA \rightarrow OB))$ if the language contains alethic modalities with $\Box$ for an appropriate necessity.

Since this rule is weaker than (RM), we cannot simply add it to the weak logic **P**, as with rules like (ConAND) or (PAND). Instead, let us build a system from scratch. There are two plausible variations for how this could be done. I present them both. Let **DPM.1,** the first Deontic logic with Permitted Inheritance, be given by adding to classical PC, with closure under *modus ponens,*

RE) if $\vdash A \leftrightarrow B$ then $\vdash OA \leftrightarrow OB$
RPM) if $\vdash A \rightarrow B$ then $\vdash PA \rightarrow (OA \rightarrow OB)$
N) $\vdash O\top$
AND) $\vdash (OA \wedge OB) \rightarrow O(A \wedge B)$

(RE) is simply a replacement rule for logical equivalents; I consider it a prerequisite for any plausible deontic logic, regardless of the question of deontic dilemmas. To do without it, or to restrict it in the manner of (RPM), would mean that assertions of obligation would depend for their truth value on the particular syntactic structure of their embedded formulas. In systems with unrestricted (RM), (RE) is derivable; here it must be postulated separately. (N) is included here primarily so that the logic will approximate SDL; given (N) and (RE), the rule form of necessitation (RN), If $\vdash A$ then $\vdash OA$, is derivable. One might dispense with either form, but given (RPM), and PC, alone, $\vdash (OA \wedge PA) \rightarrow O\top$ would be derivable. Thus, if there were anything that was both obligatory and permitted, i.e., any case of a non-conflicted obligation, as no doubt there is, then the tautology $\top$ would be obligatory. So one gains very little by not including (N) as stated. **RPM.1** has an unrestricted principle of aggregation (AND), yet it will still avoid (D) and especially deontic explosion (DEX) (as well as (DEX-1) and (DEX-2)). Because it has (AND) without restriction, **RPM.1** must then not posit (P), $\vdash \neg O\bot$, since otherwise (D), $\vdash \neg(OA \wedge O\neg A)$ would be derivable, contrary to our desire to allow for deontic dilemmas.

If one wanted (P), in order to maintain that although there could be *conflicts* of obligation, there could be no obligatory *contradictions,* 'ought' implies 'can' and all that, then one could restrict (AND) along the lines of (PAND) in Sect. 2.4.3 above. This yields the second variation **DPM.2,** given by

RE) if $\vdash A \leftrightarrow B$ then $\vdash OA \leftrightarrow OB$
RPM) if $\vdash A \rightarrow B$ then $\vdash PA \rightarrow (OA \rightarrow OB)$
N) $\vdash O\top$
P) $\vdash \neg O\bot$
PAND) $\vdash P(A \wedge B) \rightarrow ((OA \wedge OB) \rightarrow O(A \wedge B))$

In this system too, neither (D) nor (DEX), or its variants, is derivable. Hence both are candidates for a logic that admits the possibility of deontic dilemmas. I will use the term **DPM** when remarks apply equally to both versions. (One might also weaken **DPM.2** by not including (P) while keeping the restricted (PAND); the remarks below will apply equally to this minor variation.)

Neither version of **DPM** contains the distribution principle (K), $O(A \rightarrow B) \rightarrow (OA \rightarrow OB)$. If it did, then the unrestricted inheritance rule (RM) would be derivable, and then (DEX) (or (DEX-2)) would reoccur, and **DPM.2** would contain (D) and be equivalent to SDL. That (RM) is derivable given (K), is very quick from (N) and (RE):

i)	$\vdash A \to B$	hyp
ii)	$\vdash (A \to B) \leftrightarrow \top$	i,PC
iii)	$\vdash O\top$	N
iv)	$\vdash O(A \to B)$	iii, RE
v)	$\vdash OA \to OB$	iv, K

But even without (N), unrestricted (K) would yield the rule: If $\vdash A \to B$ then $\vdash (OC \wedge PC) \to (OA \to OB)$, and from this another form of deontic explosion would follow, namely

DEX-3) $\quad \vdash (OC \wedge PC) \to ((OA \wedge O\neg A) \to (PB \to OB))$

This says that if there were anything that was both obligatory and permitted, any non-conflicted obligation, as there surely is, then if there were any deontic dilemma, then whatever is permitted is obligatory. (Derivation is left to the reader for fun.) While less than full (DEX), (DEX-3) is still unacceptable. Hence (K) too is unacceptable in this context.

Although (K) is unacceptable, and not derivable in **DPM**, this restricted form, 'permitted (K)', is derivable:

PK) $\quad \vdash P(A \wedge B) \to (O(A \to B) \to (OA \to OB))$

This follows with (RPM) and either unrestricted aggregation (AND) or permitted aggregation, (PAND). Likewise, given (PK), along with (N) and (RE), the rule (RPM) can be derived. Hence, either might be taken as primitive. (Derivations are left to the reader.)

I will present **DPM** in full formal dress in the Appendix, where I will also demonstrate the claim that indeed neither (D) nor (DEX), including the several variations described above, is derivable. There I will prove that the systems are sound and complete with respect to an appropriate semantics, and also that they are decidable. In the meantime, let us consider informally how **DPM** responds to the concerns raised for the other systems.

With the failure of (D) and (DEX) in its several forms, one primary issue is clearly resolved. Both versions of **DPM** are able to tolerate deontic dilemmas. Moreover, they escape the problematic cases from Sect. 2.4.1 that arise for systems with Consistent Aggregation (ConAND), such as the scenario of Jones visiting his daughters and Hansen's example of traveling to Montreal and London.

For logics with the rule (RM) if there is a case of a deontic dilemma where OA and OB are both true but A is incompatible with B, it follows that OA and $O\neg A$ are both true, and so there is a dilemma in the narrow sense. This result does not quite hold for **DPM**, but something similar does, namely that if OA and OB are both true and A and B are incompatible, then either OA and $O\neg A$ are both true or else OB and $O\neg B$ are both true. Hence, if there is a deontic dilemma, then at least one of the conflicting obligations is itself conflicted; either OA and $\neg PA$ holds or else OB and $\neg PB$ holds. This will suffice to block the application of (RPM). Thus, with Jones and his daughters, we are given that $O(V_a \wedge N_a)$ and $O(V_b \wedge N_b)$ when it is impossible to have both $V_a \wedge N_a$ and $V_b \wedge N_b$ (because it is impossible to have both V_a and V_b). This posed a problem

for (ConAND) because, with (RM) one could infer both ON_a and ON_b, and then the undesirable $O(N_a \wedge N_b)$ follows since N_a and N_b are jointly possible. With (RPM) in place of (RM), however, one cannot infer both ON_a and ON_b since, by the above, one will not have both of the initial conditions $P(V_a \wedge N_a)$ and $P(V_b \wedge N_b)$ that are required for the two applications of the rule. Hence, even with the unrestricted aggregation rule (AND) of **DPM.1,** the unwanted $O(N_a \wedge N_b)$ is not derivable.

Similar remarks apply to Hansen's example. Granted that Jones ought to keep the appointment in Montreal and also ought to keep the appointment in London $(OA \wedge OB)$ though it is impossible to do both $(\neg\Diamond(A \wedge B))$, we know, by the remark above, that either $\neg PA$ or $\neg PB$. Hence, even though keeping each appointment necessitates the described travel ($\Box(A \rightarrow C)$ and $\Box(B \rightarrow D)$), the inference to either OC or OD by (RPM) will be blocked since the requisite clause concerning permission will be missing. Thus there can be no inference to the unwanted $O(C \wedge D)$ from the initial premises.

Thus the logics **DPM** avoid not only deontic explosion, but also the other untoward cases that troubled earlier proposals. So it seems these logics are not too strong. The real question, though, is whether they are too weak, as, for example, the system **P** seemed to be. Since these systems restrict (RM), and since (RM) has strong intuitive appeal (cf. the comment of Nute and Yu in Sect. 2.2), won't **DPM** automatically fail to capture all the inferences one expects? Further, is **DPM** adequate to represent the sort of inference that was brought against the system **P**, which originally raised this concern?

Regarding the intuitive force of (RM), this seems drawn from considering cases in which the antecedent A of the entailment will be considered (normatively) consistent, and thus the intuitive support for a rule of inheritance applies to (RPM) rather than the unrestricted (RM). I strongly suspect we have no clear intuitions about inheritance in conflict cases; at any rate, I have none. And thus the limitation on the rule does not automatically mean the system is too weak; more argument would be required.

Here is an analogy to illustrate what the restriction on the inheritance principle accomplishes. Consider free logic[17]. In classical first-order logic, the rule

UI) a) $\forall x Ax$
$\therefore$ b) At

is valid for all individual constants t. And it certainly has strong intuitive appeal. Nevertheless, the free logician maintains that (UI) is not valid; it fails in cases where the individual constant t does not refer to anything that exists. For example, the argument

a) For all objects, x, there is an object, y, identical to x. —
$\forall x \exists y(y = x)$
$\therefore$ b) There is an object y identical to the planet Vulcan. —
$\exists y(y = \text{Vulcan})$

[17] See, e.g., [24] for a useful introduction to this kind of logical system and its motivations.

has a true premise and a false conclusion. (The quantifiers here are construed classicially, as ranging over existent entities.) The plausibility, the intuitive appeal, of (UI) derives from the presupposition that the singular term t refers to an existent. By adopting this rule, classical logic, in effect, limits its range of application to languages containing no terms that fail to refer in this way. This is a severe limitation.

In its place, the free logician recommends letting the language contain singular terms that lack existential import, but build the presupposition required for the application of (UI) into the rule itself. Thus, although free logic rejects (UI), it accepts the restricted rule

RUI) a) $\forall x Ax$
c) t exists
$\therefore$ b) At

With the logics **DPM 1** recommend something analogous. We accept inferences based on deontic inheritance (RM) (and perhaps aggregation (AND)), we find that they have strong intuitive appeal, under the presupposition that the situations described are conflict free. Standard deontic logic, with its unrestricted rule (RM) and (AND), in effect limits itself to reasoning with normative structures that exclude deontic dilemmas. This is a severe limitation. In its place I propose that we recognize the possibility of such conflicts, and then build the presupposition required for the application of the rule into the rule itself. That is what (RPM) does. Given that *A* entails *B,* we accept that *OA* entails *OB,* provided that the obligation that *A* is not itself conflicted.

That is just how **DPM** treats arguments like that of Smith's obligation to serve his country. This follows the pattern described in Sect. 2.4.3 for **PA** to illustrate the restricted rule of permitted aggregation (PAND). We are given that Smith ought to fight in the army or perform alternate service — $O(F \vee S)$ — and that he ought not to fight in the army — $O\neg F$ — and we take it as an implicit premise that it really is all right, i.e., permitted, that Smith not fight but perform alternate service — $P(\neg F \wedge S)$. We then argue that Smith ought to perform alternate service (OS) as follows, in **DPM.1:**

i)	$O(F \vee S)$	hyp
ii)	O-F	hyp
iii)	$P(\neg F \wedge S)$	hyp
iv)	$\vdash (\neg F \wedge S) \leftrightarrow ((F \vee S) \wedge \neg F)$	PC
v)	$P((F \vee S) \wedge \neg F)$	iii, iv, RE
vi)	$O((F \vee S) \wedge \neg F)$	i, ii, AND
vii)	$\vdash (F \vee S) \wedge \neg F) \rightarrow S$	PC
viii)	$\vdash P((F \vee S) \wedge \neg F) \rightarrow (O((F \vee S) \wedge \neg F) \rightarrow OS)$	vii, RPM
ix)	OS	v, vi, viii, PC

In **DPM.2** the argument would insert a step by (PAND)

v)′ $P((F \vee S) \wedge \neg F) \rightarrow O((F \vee S) \wedge \neg F)$ i, ii, PAND

between (v) and (vi) and then conclude (vi) by *modus ponens* from (v).

Thus these systems seem well equipped to handle arguments like this, provided one is prepared to accept the implicit premise (iii). I will return to this in the Conclusion below.

Finally, it is worth noting that in case there were no deontic dilemmas, no violations of (D), then **DPM.1** would agree wholly with SDL. That is, if (D) were added as an axiom to **DPM.1,** the result is equivalent to SDL. (Criterion of adequacy (∗).) Oviously all of **DPM.1** + (D) is contained in SDL. The converse follows from the fact that (RM) is derivable given (D) with (RPM). Thus:

i)	$\vdash A \to B$	hyp
ii)	$\vdash PA \to (OA \to OB)$	i, RPM
iii)	$\vdash OA \to PA$	D
iv)	$\vdash OA \to (OA \to OB)$	ii, iii, PC
v)	$\vdash OA \to OB$	iv, PC

Hence, **DPM.1** + (D) contains (RM), (N), (D), and (AND), which are adequate for SDL. Therefore **DPM.1** + (D) is equivalent to SDL.

This is not so for **DPM.2.** Although the derivation of (RM) from (RPM) and (D) still holds, the full proposition (AND) does not follow just given (PAND), along with (D) and (RM), etc. Thus, in a sense, **DPM.2** does not correspond to SDL in a dilemma free universe. But even so, I suspect, though will not argue, that **DPM.2,** with its limited principle of aggregation, would still appear adequate in such a case.

4 Conditional Obligation

So far I have only discussed monadic deontic logic because that is simplest, and suffices to raise the question of the proper way to accommodate deontic dilemmas. Nevertheless, since so much of normative discourse seems to require a notion of conditional oughts, it is worthwhile to extend the previous considerations to apply to them. Thus, when speaking of deontic dilemmas, we should include cases in which some state of affairs B is enjoined under a condition A, $O(B/A)$, and so is $\neg B$, $O(\neg B/A)$, or more generally situations in which $O(B/A)$ and $O(C/A)$ when B and C are jointly impossible, given A. Formulas $O(B/A)$ may be read 'it ought to be that B under the condition A.

Given an operation of conditional obligation $O(-/-)$, the monadic operator $O(-)$ can be introduced by definition so that $OA =_{df} O(A/\top)$, and then one should expect the logic of such a defined monadic 'ought' to be the same as one has originally settled on. Indeed, the logic for $O(-/A)$ should also be the same for any condition A held constant, (or perhaps any consistent condition; there are subtleties that can arise when A is inconsistent). These considerations suggest that a proper logic of conditional obligation that will accommodate (conditional) deontic dilemmas will contain at least the principles, corresponding to **DPM.1:**

RCE) If $\vdash A \leftrightarrow B$ then $\vdash O(C/A) \leftrightarrow O(C/B)$
CRE) If $\vdash B \leftrightarrow C$ then $\vdash O(B/A) \leftrightarrow O(C/A)$
CRPM) If $\vdash B \rightarrow C$ then $\vdash P(B/A) \rightarrow (O(B/A) \rightarrow O(C/A))$
CN) $\vdash O(\top/\top)$
CAND) $\vdash (O(B/A) \wedge O(C/A)) \rightarrow O(B \wedge C/A)$

where $P(B/A) =_{df} \neg O(\neg B/A)$. Call the result of adding these to PC, and closing under *modus ponens,* **CDPM.1.** The first rule is the rule of replacement of equivalents in the antecedent position. The remaining postulates correspond directly to those of **DPM.1.** For **CDPM.2,** the counterpart to **DPM.2,** add

CP) $\vdash \neg O(\bot/A)$

and replace (CAND) with

CPAND) $\vdash (O(A/C) \wedge O(B/C) \wedge P((A \wedge B)/C)) \rightarrow O((A \wedge B)/C)$

Beyond these minimal principles there is opportunity for a lot of variation for the logic of conditional obligation, especially in the way it manipulates antecedents. For present purposes, however, let us consider just these postulates[18].

The purpose of bringing conditional obligation in now, aside from the intrinsic virtues of such a notion, if any, is not so much to raise issues about deontic dilemmas, for these should play out the same in the dyadic context as in the

[18] Van Fraassen [36] proposed another fairly minimal axiom

$$\vdash O(B/A) \rightarrow O(A \wedge B/A)$$

One might also consider a general axiom of reflexivity $O(A/A)$, which with conditional aggregation would render van Fraassen's redundant. Van Fraassen's own first proposal for a logic of conditional obligation with conflicts [37] p. 17 included (RCE), a rule corresponding to (RM) for inheritance in the consequent,

CRM) If $\vdash B \rightarrow C$ then $\vdash O(B/A) \rightarrow O(C/A)$

as well as the above, but neither (CN) nor (CP). Chellas [2] §10.2 proposed a system he called *CD* (not to be confused with van Fraassen's **CD** of [36]) for a minimal conditional deontic logic. It is given by adding to (PC) just (RCE), the unrestricted inheritance rule (CRM), and a weaker version of (CP), namely, $\Diamond A \rightarrow \neg O(\bot/A)$, where the $\Diamond$ represents alethic possibility such as given by **S5**. Thus the language of *CD* requires alethic modalities; in their absence, this principle would have to be strengthened to (CP). Chellas prefers the weaker version in order to allow models in which $O(\bot/\bot)$ is true. In [14] and [15] I presented a conditional analog to the system **P** described above in Sect. 2.3; I called this **DP.** It contains (RCE), (RCM), (CN), (CP), and also a principle of transitivity for (weak) preference

$$(A \geq B \wedge B \geq C) \rightarrow A \geq C$$

where $A \geq B =_{df} \neg O(\neg A/A \vee B)$.This system **DP** corresponds to the dyadic deontic logic of van Fraassen of [36], his **CD,** and of David Lewis in [25] Ch. 6, there called **VN**, and in [26], much as the logic **P** stands to SDL. **DP** will accept conditional deontic dilemmas, but does not account for conditional versions of arguments like that of Smith's service to his country.

monadic, and thus not require anything new, as rather to suggest that a move similar to that which let **DPM** accommodate deontic dilemmas can be applied to another problem that arises specifically in the context of dyadic deontic logic, and which is independent of questions of deontic dilemmas or normative conflict. This is a problem broached, but not settled, by Horty in a number of places, e.g., [19], [20], to question the treatment of conditional obligation within the framework of traditional modal or conditional logic, such as we see in standard dyadic deontic logics, like van Fraassen's **CD** or their weaker counterparts like **DP** mentioned in footnote 18.

In logics of conditional obligation, the principle of 'strengthening the antecedent' (SA)

$$O(A/B) \rightarrow O(A/B \wedge C)$$

is not valid. This is the virtue of these systems, in contrast to attempts to define conditional obligation in terms of a monadic ought-operator and ordinary conditional, e.g., as $O(A/B) = O(A \rightarrow B)$ or $A \rightarrow OB$. Nevertheless, Horty argues, there seem to be cases where strengthening the antecedent seems appropriate, and by its wholesale rejection systems like **CD** or **DP** are unable to account for these cases. This is a lot like the situation with the argument about Smith's service to his country, which challenged the wholesale rejection of Aggregation, though the form of the case is different.

The example Horty uses to make this point draws on rules of etiquette, but it is easy to imagine counterparts in other normative domains. We are given these rules:

i) You ought not to eat with your fingers — $O(\neg F/\top)$
ii) You ought to put your napkin on your lap — $O(N/\top)$
iii) If you are served asparagus, you ought to eat it with your fingers — $O(F/A)$

Here the third rule might be said to override the first, so that we should not be able to conditionalize (i) to A. That is, one should not be able to strengthen its antecedent to

iv) If you are served asparagus, you ought not to eat with your fingers — $O(\neg F/A)$

inferred from (i). And indeed in the usual logics one cannot. Horty's concern, however, is with rule (ii). This does not seem overriden by (iii), and it seems plausible to infer from (ii) that

v) If you are served asparagus, you ought to put your napkin on your lap — $O(N/A)$

Yet this inference is not valid the usual logics of conditional obligation.

Thus the situation seems much like that that arose with respect to aggregation as discussed in Sect. 2.3. One wants to exclude some applications of the rule, but not all. I suggest a similar solution. One wants to block the rule in cases of conflict, as in the case of (i), but allow it when there is no conflict (ii). 'Conflict' here certainly means logical conflict, logical inconsistency, but

also, I suggest, deontic dilemma, impermissibility. Hence one might include this restricted principle of permitted strengthening of the antecedent (PSA)

PSA) $\vdash (O(B/A) \wedge P(B/A \wedge C)) \rightarrow O(B/A \wedge C)$

in one's logic of conditional obligation.

This blocks the inference from (i) to (iv) in the presence of (iii), for (iii) entails $\neg P(\neg F/A)$ (by definition of $P(-/-)$), and so the conjunct necessary for the antecedent of (PSA) cannot be included in the premise set without contradiction. On the other hand, we can suppose that $P(N/A)$ is implicitly present, and with it (ii) yields (v) by (PSA).

In this way, just as with the rule of permitted inheritance (RPM), a system of deontic logic within the traditional framework of modal or conditional logic is able to steer a middle course between accepting all inferences of the original pattern and accepting none.

The principle (PSA) is interesting for it corresponds quite closely to a rule discussed in the literature of nonmonotonic, or defeasible, reasoning. This is the rule known [1] as 'determinacy preservation' (DP),

DP) If $A \mathrel{|\!\sim} B$ and $A \wedge C \not\mathrel{|\!\sim} \neg B$ then $A \wedge C \mathrel{|\!\sim} B$

where the sign '$\mathrel{|\!\sim}$' represents a nonmonotonic inference relation, that B (normally) follows from A. Although their interpretations are quite different, there is a strong formal analogy between conditional assertions $A \mathrel{|\!\sim} B$ and assertions of (first-degree) conditional obligation $O(B/A)$. With that in mind, we can see that (DP) maps directly to (PSA) above.

Given the other principles of a preferential nonmonotonic inference relation (DP) turns out to be equivalent to a principle of 'rational transitivity' (RT),

RT) If $A \mathrel{|\!\sim} B$ and $B \mathrel{|\!\sim} C$ and $A \not\mathrel{|\!\sim} \neg C$ then $A \mathrel{|\!\sim} C$

(Cf. [1].) This is quite a strong rule, and it, or rather its analog for conditional obligation

$$\vdash (O(B/A) \wedge O(C/B) \wedge \neg O(\neg C/A)) \rightarrow O(C/A)$$

probably takes one farther than one would want to go for a logic of conditional obligation. (In a preference based semantics for the deontic logic it would require the preference relation between possible worlds be quasi-linear.)

Horty's problem can, however, be answered with a principle weaker than (PSA) or (DP). Consider this

RSA) $\vdash (O(B/A) \wedge P(C/A)) \rightarrow O(B/A \wedge C)$

In other words, one can strengthen an antecedent when the added condition, C, is permitted under the main condition, A. With this, and the implicit premise $P(N/A)$ as above, then the inference from (ii) to (v) still goes through, even while the inference from (i) to (iv) is blocked.

This rule (RSA) of 'restricted strengthening the antecedent', or 'rational SA', corresponds to the rule of 'rational monotonicity' (RatMono) in nonmonotonic logic,

RatMono) If $A \mathrel{|\!\sim} B$ and $A \not\mathrel{|\!\sim} \neg C$ then $A \wedge C \mathrel{|\!\sim} B$

This rule defines the class of preferential nonmonotonic inference relations that are called 'rational'. Semantically it corresponds to the modularity of the preference ordering on models (equivalently, the transitivity of the complement of the converse of the strict ordering). Its deontic analog, (RSA) is *already* present in standard systems of dyadic deontic logic, and thus nothing new is needed to accommodate Horty's concern. In a preference-based semantics for such a deontic logic its validity is guaranteed by the transitivity of the weak preference ordering of alternative worlds. With some other basic assumptions of dyadic deontic logic, it is equivalent to the principle of transitivity mentioned in footnote 18[19].

As noted above, this particular problem is independent of the question of deontic dilemmas and how a deontic logic should accommodate them. Indeed, the principle (RSA) is not contained in the minimal systems **CDPM** as initially set out, though it could be added. The purpose of these last remarks was to indicate that just as one might want to steer a middle ground between rejecting a rule altogether, like deontic (RM) or (AND), and accepting it wholesale, and that one can do this by qualifying it through a clause relating to permission, so the same sort of maneuver will apply to this other kind of case. One wants to find a middle way between unrestricted strengthening of the antecedent and none at all, and that too can be done by paying attention to permissions, or in the framework of nonmonotonic inference relations, by bringing non-Horn premises into the rules.

5 Conclusion

Most of the discussion of this paper is motivated by problems Horty presented for deontic logic that wants to allow for the possibility of deontic dilemmas. While it is easy to design a logic that accepts that possibility while keeping to a classical base for deontic logic, e.g., the logic **P** of Sect. 2.3, it is not so easy to design a logic that also accounts for the further inferences that Horty puts forward as unproblematic. In Sect. 2.4 I looked at a couple of proposals, which, however, turn out to be inadequate. In Sect. 3 I presented a different approach that fares better, and so I recommend it as a basic monadic deontic logic to accomplish both purposes, to accommodate deontic dilemmas while also accounting for Horty's examples. In Sect. 4,I sketched how this approach can be extended to the logic of conditional obligation, and suggest how similar maneuvers can also respond

[19] Delgrande [5] argues that a rule like the present (RatMono) is too strong; it resolves Horty's original problem, but then lets in inferences that should not be accepted. Analogous cases would confront the standard systems of dyadic deontic logic, such as van Fraassen's **CD** of [36], which I call SDDL, [14]. Delgrande's cases are blocked, however, in the weaker dyadic logic **DP** of [14], which corresponds to the weak monadic logic **P** described in Sect. 2.3 above, and likewise in its nonmonotonic counterpart. That is because this system lacks (conditional) aggregation (CAND), and also principles corresponding to CUT and OR for the antecedent.

to other problems Horty has raised for this way of doing deontic logic, or the logic of nonmonotonic reasoning.

Nevertheless, there is a fundamental difference between the way I have proposed looking at these kinds of situations and the way Horty would look at them. Horty sees deontic logic belonging to the domain of nonmonotonic logic, that is, a logic in which an argument with premises Γ and conclusion A might be valid even while an argument with premises Γ' and conclusion A is not valid, even though $\Gamma \subseteq \Gamma'$. In classical logic validity is never lost with the addition of new premises. The logics **DPM,** and **CDPM,** and their enrichments with principles like (RSA) described above are all classical in this respect.

The difference in approach can be seen most clearly with the example concerning strengthening the antecedent and how to eat asparagus, though much the same could be said with regard to the example of Smith's service to his country. With F for 'you eat with your fingers' and A for 'you are eating asparagus', the nonmonotonic approach would regard the argument

I) a) $O(\neg F/\top)$
$\therefore$ c) $O(\neg F/A)$

to be valid, but the argument

II) a) $O(\neg F/\top)$
b) $O(F/A)$
$\therefore$ c) $O(\neg F/A)$

not to be valid. This illustrates the nonmonotonicity of the consequence relation. (Cf. [20], p. 35.)

By contrast, the approach I proposed would say that, strictly speaking, argument (I) is not valid. If it appears to be, that is because there is an additional premise implicit in the context, and that (I) should be considered enthymematic for

I)′ a) $O(\neg F/\top)$
d) $P(\neg F/A)$
$\therefore$ c) $O(\neg F/A)$

Then (c) follows from (a) and (d) by the rule (RSA). And if (II) seems not to be valid, that is because with the addition of premise (b) the implicit premise (d) is withdrawn, in which case (RSA) does not apply. (If (d) is maintained, so that the argument is really

II)′ a) $O(\neg F/\top)$
b) $O(F/A)$
d) $P(\neg F/A)$
$\therefore$ c) $O(\neg F/A)$

then one is confronted with an inconsistent premise set since (d) is equivalent to (d′) $\neg O(F/A)$, which contradicts the new premise (b). Although (II)′ is classically valid, it should be rejected for that inconsistency.)

Similarly, in the example concerning Smith's service, from the nonmonotonic point of view, the argument

III) a) $O(F \vee S)$
b) $O(\neg F)$
$\therefore$ c) OS

is regarded to be valid as it stands, whereas **DPM** would say that it is, strictly speaking, not valid, but it might appear to be because of a tacit premise, and that what is really valid is the argument

III)′ a) $O(F \vee S)$
b) $O(\neg F)$
d) $P(\neg F \wedge S)$
$\therefore$ c) OS

where the conclusion (c) follows from (a), (b) and (d) by (RPM) and (AND), or (PAND) as described at the end of Sect. 3.

Much like the free logician described in Sect. 3,I take it that arguments like (I) and (III), insofar as they are to be considered valid, rest on presuppositions, tacit premises that are made explicit in (I)′ and (III)′. Because the latter forms are valid in **DPM,** the apparent validity of (I) and (III) is explained. At the same time, the non-validity of argument (II) is also explained since the additional premise requires cancelling the presuppositions of the first argument.

Thus the difference between Horty's approach and what I have recommended concerns what concept of validity is being ascribed when evaluating arguments, and also the identification of the precise argument that is evaluated (does it include the tacit premise, or not?). It is plausible that both approaches have their proper roles to play in analyzing normative discourse. At any rate, the issues raised by the contrast go far beyond the purposes of the present discussion, and do not need to be decided here.

Appendix

In this Appendix I dress the logics **DPM** and **CDPM** in more formal clothing, and demonstrate that they are sound and complete with respect to an appropriate semantics. Because these are non-normal, but still classical modal logics the neighborhood semantics familiar from Segerberg [34] or Chellas [2] Ch. 7–9 is readily adapted to them. Nevertheless, the completeness theorems given below are a bit tricky, and so, perhaps, more interesting. In the course of completing these proofs I also establish that these logics have the finite model property and hence, because they are finitely axiomatizable, they are decidable.

Let us take up the monadic logics **DPM** first; results for **CDPM** will then follow quickly by similar procedures. The language, $\mathcal{L}$, for **DPM** is a prepositional language adequate for classical prepositional logic plus the monadic deontic operator O such that OA is well-formed whenever A is. '*A*', '*B*', '*C*', etc. are variables for arbitrary formulas of $\mathcal{L}$. $A \rightarrow B$ is understood to be equivalent to $\neg A \vee B$ and to $\neg(A \wedge \neg B)$. $A \leftrightarrow B$ is $(A \rightarrow B) \wedge (B \rightarrow A)$. As usual, PA abbreviates $\neg O \neg A$. $\top$ is any classical tautology and $\bot$ is $\neg\top$.

DPM.1 is the least set of formulas containing all classical tautologies of formulas of $\mathcal{L}$, plus all instances of

N) $O\top$
AND) $(OA \wedge OB) \rightarrow O(A \wedge B)$

and closed under the rules

MP) if $\vdash A \rightarrow B$ and $\vdash A$ then $\vdash B$
RE) if $\vdash A \leftrightarrow B$ then $\vdash OA \leftrightarrow OB$
RPM) if $\vdash A \rightarrow B$ then $\vdash PA \rightarrow (OA \rightarrow OB)$

where $\vdash$ indicates membership in **DPM.1.**

DPM.2 is the least the set of formulas containing all classical tautologies of formulas of $\mathcal{L}$, plus (N) as above and also all instances of

P) $\neg O\bot$
PAND) $(OA \wedge OB \wedge P(A \wedge B)) \rightarrow O(A \wedge B)$

and closed under the rules (MP), (RE) and (RPM) (with $\vdash$ for membership in **DMP.2;** henceforth, we shall take it to be clear what $\vdash$ signifies in context).

For the semantics for these logics, formulas of $\mathcal{L}$ are interpreted with respect to *neighborhood frames.* The key idea here is to take obligatoriness, or normative requirement, to be a property or attribute of propositions. A formula OA is then true just in case the proposition expressed by A has this property. More precisely, consider a proposition to be a set of possible worlds, and accordingly the proposition expressed by A to be the set of worlds at which A is true; designate that set $|A|$. Consider a property of such propositions extensionally, as a set of propositions, and thus a set of sets of possible worlds. Each possible world a has associated with it a set $\mathcal{O}_a$ of propositions; these are the propositions that are obligatory (from the point of view of a). If $|A|$ is the proposition expressed by A (on a model), i.e., the set of possible worlds where A is true (on the model), then OA is true (at a on the model) just in case $|A|$ is a member of $\mathcal{O}_a$.

More formally, define a neighborhood frame, F, to be a pair $\langle W, \mathcal{O}\rangle$ in which W is a non-empty set of points, e.g., possible worlds, and $\mathcal{O}$ is a function assigning every $a \in W$ a set, $\mathcal{O}_a$, of subsets of W; i.e., $\mathcal{O}_a \subseteq \wp W$. A *model, M,* is a pair $\langle F, v\rangle$ where F is a neighborhood frame $\langle W, \mathcal{O}\rangle$, and v is a function assigning every atomic formula p of $\mathcal{L}$ a subset of W, i.e., $v(p) \subseteq W$. A satisfaction relation $\models$ is defined as usual, so that for any model $M = \langle F, v\rangle$ on a frame $F = \langle W, \mathcal{O}\rangle$, for any $a \in W$,

Tp) $M, a \models p$ iff $a \in v(p)$
T$\neg$) $M, a \models \neg A$ iff $M, a \not\models A$
T$\wedge$) $M, a \models A \wedge B$ iff $M, a \models A$ and $M, a \models B$
T$\vee$) $M, a \models A \vee B$ iff $M, a \models A$ or $M, a \models B$

and in particular

TO) $M, a \models OA$ iff $|A|_M \in \mathcal{O}_a$

where $|A|_M = \{a \in W : M, a \models A\}$. $|A|_M$ is the proposition expressed by A on the model M.

It is helpful to note

Proposition 1. *For any model, M, (i)* $|A \wedge B|_M = |A|_M \cap |B|_M$*; (ii)* $|A \vee B|_M = |A|_M \cup |B|_M$*;(iii)* $|\neg A|_M = -|A|_M$*; (iv)* $|\top|_M = W$*; and (v)* $|\bot|_M = \emptyset$.

where $-X$ is the complement of X with respect to W, i.e., $-X = W - X = \{a \in W : a \notin X\}$. Generally speaking, the relativization to W should be understood in context; in later arguments when two models, one a submodel of the other, are being discussed together, it will be necessary to be careful to distinguish one complement from the other.

As usual, a model M satisfies A, or A holds on M — $M \models A$ —iff $M, a \models A$ for every $a \in W$ when $M = \langle F, v\rangle$ and $F = \langle W, \mathcal{O}\rangle$. A is *valid on* a frame F — $F \models A$ — iff $M \models A$ for every model $M = \langle F, v\rangle$ on F, and A is *valid in* a class of neighborhood frames $\mathcal{F}$ — $\mathcal{F} \models A$ — iff $F \models A$ for every $F \in \mathcal{F}$. A set of formulas **S** is *sound* with respect to a class of frames $\mathcal{F}$ iff $\mathcal{F} \models A$ for every $A \in \mathbf{S}$. **S** is *complete* with respect to $\mathcal{F}$ iff for every A that $\mathcal{F} \models A$, $A \in \mathbf{S}$. Similarly, **S** is sound with respect to a class of *models,* $\mathcal{M}$, iff $M \models A$ for every model $M \in \mathcal{M}$, and **S** is complete with respect to $\mathcal{M}$ iff for every A such that for all $M \in \mathcal{M}$, $M \models A$, $A \in \mathbf{S}$. The contrast between frame-completeness and model-completeness will be useful below.

For later reference, it is also useful to note

Proposition 2. *For any model, M, (i)* $M \models A \rightarrow B$ *iff* $|A|_M \subseteq |B|_M$*, and (ii)* $M \models A \leftrightarrow B$ *iff* $|A|_M = |B|_M$.

The set of formulas that comprise **DPM.1** is both sound and complete with respect to the class of neighborhood frames $F = \langle W, \mathcal{O}\rangle$ that meet the following three conditions: For all $X, Y \subseteq W$ and all $a \in W$,

a) $W \in \mathcal{O}_a$
b) If $X \in \mathcal{O}_a$ and $Y \in \mathcal{O}_a$ then $X \cap Y \in \mathcal{O}_a$
c) If $X \subseteq Y$ and $X \in \mathcal{O}_a$ and $-X \notin \mathcal{O}_a$ then $Y \in \mathcal{O}_a$

Condition (a) validates axiom (N), condition (b) validates the aggregation axiom (AND) and condition (c) validates (RPM). ((RE), and later (RCE) and (CRE), come for free.)

The set of formulas that comprise **DPM.2** is both sound and complete with respect to the class of neighborhood frames $F = \langle W, \mathcal{O}\rangle$ that meet the conditions (a) and (c) above for all $X, Y \subseteq W$ and all $a \in W$, but with condition (b) modified to

b)′ If $X \in \mathcal{O}_a$ and $Y \in \mathcal{O}_a$ and $-(X \cap Y) \notin \mathcal{O}_a$ then $X \cap Y \in \mathcal{O}_a$

and also the additional condition

d) $\emptyset \notin \mathcal{O}_a$

(b)′ validates the weakened aggregation principle (PAND) while (d) validates axiom (P).

Theorem 1. *(i)* **DPM.1** *is sound with respect to the class of frames that satisfy conditions (a)–(c); (ii)* **DPM.2** *is sound with respect to the class of frames that satisfy conditions (a), (b)′, (c) and (d).*

Proof. As usual this is simply a matter of showing that the axioms, (N) and (AND) of **DPM.1** and (N), (PAND) and (P) of **DPM.2,** are valid in the respective classes of frames and that the rules preserve validity. These are easy enough to leave to the reader, but here is the argument for the rule (RPM) since this is new to the literature. With $\mathcal{F}$ a class of frames that satisfies condition (c), suppose that $F \models A \rightarrow B$, and show that $F \models PA \rightarrow (OA \rightarrow OB)$. For that suppose a frame $F = \langle W, \mathcal{O} \rangle \in \mathcal{F}$ and a model $M = \langle F, v \rangle$ and an $a \in W$ such that $M, a \models PA$ and $M, a \models OA$. By Proposition 2.i, $|A|_M \subseteq |B|_M$. Since $M, a \models PA$, $M, a \not\models O\neg A$, so that $|\neg A|_M \notin \mathcal{O}_a$. By Proposition 1.iii, $-|A|_M \notin \mathcal{O}_a$. Since $M, a \models OA$, $|A|_M \in \mathcal{O}_a$. Hence, since F satisfies condition (c), $|B|_M \in \mathcal{O}_a$, and therefore $M, a \models OB$, as required. (The argument for the validity of permitted aggregation (PAND), given condition (b)′, is similar.) □

In light of later arguments, it is useful to mention this immediate corollary to Theorem 1,

Corollary 1. DPM *is sound with respect to the class of finite frames that satisfy conditions (a), (b), (c), or (a), (b)′, (c), (d), as appropriate.*

Before taking up the question of completeness, let us first fulfill a promise made in the main text, to demonstrate that neither (D) nor deontic explosion, in its various forms, is derivable in **DPM.** That (D) is not, is manifest from the fact that **DPM** is a subsystem of the normal modal logic **K**, and (D) is well known not to be derivable in **K**. Not only is deontic explosion in the form (DEX), $(OA \wedge O\neg A) \rightarrow OB$, not derivable in **DPM,** but neither is (DEX-2), $(OA \wedge O\neg A) \rightarrow (PB \rightarrow OB)$, which undermined the logic **PA** described in Sect. 2.4.3, nor is (DEX-3), $(OC \wedge PC) \rightarrow ((OA \wedge O\neg A) \rightarrow (PB \rightarrow OB))$, which would vitiate **DPM** enriched with an unrestricted (K) principle, discussed in Sect. 3. To prove this, here is a model for each version of **DPM** that will falsify the latter, and so the others, including also (DEX-1). By Soundness of the systems, Theorem 1, it follows that these schemas are not derivable in **DPM.**

For that model for **DPM.1,** let $F = \langle W, \mathcal{O} \rangle$ with $W = \{a, b, c\}$, and $\mathcal{O}_a = \{W, \emptyset, \{a, b\}\}$, and $\mathcal{O}_b = \mathcal{O}_c = \{W, \emptyset\}$. Given F, let $M = \langle F, v \rangle$, with $v(p) = W$, $v(q) = \{a, b\}$ and $v(r) = \{a\}$, (v for any other atomic formula could be anything.) With these specifications, it is not difficult to show that F satisfies the conditions (a), (b) and (c) for a frame for **DPM.1,** and also that $M, a \models Op$, $M, a \models O\neg p$, $M, a \models Oq$, $M, a \models Pq$, $M, a \models Pr$ and $M, a \not\models Or$. (Verification may be left to the reader.) This suffices to falsify this instance of (DEX-3), $(Oq \wedge Pq) \rightarrow ((Op \wedge O\neg p) \rightarrow (Pr \rightarrow Or))$, as promised.

For **DPM.2** we require a model on a frame that meets condition (d) and (b)′ rather than (b), as well as (a) and (c). For this take $W = \{a, b, c\}$, as before, but let $\mathcal{O}_a = \{W, \{a\}, \{b, c\}\}$ (and $\mathcal{O}_b = \mathcal{O}_c = \{W\}$). It is then easy to show (and so left to the reader) that $F = \langle W, \mathcal{O} \rangle$ meets the requisite conditions. For M on F, let $v(p) = \{a\}$, $v(q) = W$ and $v(r) = \{b\}$. This too will falsify the instance of (DEX-3) (verification left to the reader).

Turning now to the completeness of the logics **DPM,** as is so often the case, this is more difficult to establish than soundness, and indeed, for these it is more complicated than one might have predicted at first. Let us begin with the familiar. Define a canonical model $M^c = \langle F^c, v^c \rangle$ on a frame $F^c = \langle W^c, \mathcal{O}^c \rangle$, as follows: W^c is the set of all maximal consistent extensions of **DPM** (either version as appropriate). Let $[A] = \{a \in W^c : A \in a\}$. For each $a \in W^c$, let

$$\mathcal{O}^c_a = \{X \subseteq W^c : \exists A(X = [A] \text{ and } OA \in a)\}$$

$\mathcal{O}^c$ assigns $\mathcal{O}^c_a$ to a. Let v^c be such that for every atomic formula p

$$v^c(p) = \{a \in W^c : p \in a\}$$

As a syntactical counterpart to Proposition 1, it is helpful to note

Proposition 3. *For any A and B, (i)* $[A \wedge B] = [A] \cap [B]$*; (ii)* $[A \vee B] = [A] \cup [B]$*; (iii)* $[\neg A] = -[A]$*; (iv)* $[\top] = W^c$*; (v)* $[\bot] = \emptyset$.

These may be left to the reader to verify. Also, corresponding to Proposition 2,

Proposition 4. *For any A and B, (i)* $[A] \subseteq [B]$ *iff* $\vdash A \rightarrow B$*, and (ii)* $[A] = [B]$ *iff* $\vdash A \leftrightarrow B$.

Proof. For (i), suppose $[A] \subseteq [B]$ but $\nvdash A \rightarrow B$. Then $\{A, \neg B\}$ is consistent, and so has a maximal consistent extension, b. $A \in b$ so $b \in [A]$. Hence $b \in [B]$, which is to say $B \in b$, contrary to the consistency of b since $\neg B \in b$. Therefore, $\vdash A \rightarrow B$. Further, if $\vdash A \rightarrow B$, then since maximal consistent extensions are closed under provable implications, it is automatic that for any $a \in [A]$, $a \in [B]$, or $[A] \subseteq [B]$. Part (ii) follows immediately from (i). □

Lemma 1. *For all A and all* $a \in W^c$, $M^c, a \models A$ *iff* $A \in a$ *(or,* $|A|_{M^c} = [A]$*).*

Proof. As usual, by induction on A. This is immediate from the definition of v^c when $A = p$, and it is routine when $A = B \wedge C$, $A = B \vee C$ or $A = \neg B$. We consider the case when $A = OB$ under the assumption that $|B|_{M^c} = [B]$. (a) Suppose $OB \in a$. By the inductive hypothesis $|B|_{M^c} = [B]$, hence there is a C such that $|B|_{M^c} = [C]$ and $OC \in a$. By definition $|B|_{M^c} \in \mathcal{O}^c_a$, which suffices immediately for $M^c, a \models OB$. (b) Suppose $M^c, a \models OB$, so that $|B|_{M^c} \in \mathcal{O}^c_a$. Then $[B] \in \mathcal{O}^c_a$, by the inductive hypothesis. Consequently, there is a C such that $[B] = [C]$ and $OC \in a$. For such a C, since $[C] = [B]$, $\vdash C \leftrightarrow B$, by Proposition 4.ii. Hence, $\vdash OC \rightarrow OB$ by (RE), so $OB \in a$, as required. □

Ordinarily that would suffice to establish completeness for the system, since if $\nvdash A$, $\{\neg A\}$ is consistent, and so has a maximal consistent extension b. By Lemma 1, $M^c, b \models \neg A$, and so $M^c, b \not\models A$, and then $M^c \not\models A$. Hence A could not be valid. Thus, by contraposition, if A were valid, it would have to be provable in **DPM.** What is missing, however, is that M^c is a model on a frame F^c that satisfies all of conditions (a)–(c) (for **DMP.1**) or (a), (b)′, (c), (d) (for **DMP.2**). And we do not have that. Condition (c) fails, as does (b)′.

To see how (c) fails, consider an $X, Y \subseteq W^c$ such that $X \subseteq Y$ and $X \in \mathcal{O}_a^c$ and $-X \notin \mathcal{O}_a^c$. Hence, there is a C such that $X = [C]$ and $OC \in a$, and also, with some manipulation, $O\neg C \notin a$, whence, by maximality, $\neg O\neg C \in a$ or $PC \in a$. If there were a formula D such that $Y = [D]$, then (c) would follow, since then $[C] \subseteq [D]$, so $\vdash C \rightarrow D$, by Proposition 4, and so $\vdash PC \rightarrow (OC \rightarrow OD)$, by (RPM), and so $OD \in a$, by (MP) twice, which would suffice for $Y \in \mathcal{O}_a^c$. But there is no guarantee that there will be any such D. If we were to modify the definition of $\mathcal{O}_a^c$, e.g., to have, for example,

$$\mathcal{O}_a^c = \{X \subseteq W^c : \exists A([A] \subseteq X \text{ and } OA \in a)\}$$

or equivalently if we take the supplementation of F^c to be the canonical frame, then (c) would be satisfied, indeed, F^c would be a frame for full inheritance (RM), but then the key Lemma 1 would be lost. A similar problem infects (b)′ for **DPM.2.** Hence we must find another model that will do the job of falsifying any non-theorem while being based on a frame in the class $\mathcal{F}$ of frames that meet the requisite conditions.

Nevertheless, this stumbling block does suggest that from Lemma 1 we can conclude at least the model-completeness of **DPM.** From that, as we shall see, it will be possible to establish full-blooded frame-completeness.

For model-completeness, consider any model $M = \langle F, v\rangle$ on a neighborhood frame $F = \langle W, \mathcal{O}\rangle$. Let $\mathcal{E}_M$ be the set of *expressible* propositions on M. That is,

$$\mathcal{E}_M = \{X \subseteq W : \exists B(X = |B|_M)\}$$

We can then modify the frame conditions above to form conditions on models, by restricting X and Y to expressible propositions. Thus

a)^m $W \in \mathcal{O}_a$

b)^m For all $X, Y \in \mathcal{E}_M$ and $a \in W$, if $X \in \mathcal{O}_a$ and $Y \in \mathcal{O}_a$ then $X \cap Y \in \mathcal{O}_a$

c)^m For all $X, Y \in \mathcal{E}_M$ and $a \in W$, if $X \subseteq Y$ and $X \in \mathcal{O}_a$ and $-X \notin \mathcal{O}_a$ then $Y \in \mathcal{O}_a$

b)′^{m} For all $X, Y \in \mathcal{E}_M$ and $a \in W$, if $X \in \mathcal{O}_a$ and $Y \in \mathcal{O}_a$ and $-(X \cap Y) \notin \mathcal{O}_a$ then $X \cap Y \in \mathcal{O}_a$

d)^m $\emptyset \notin \mathcal{O}_a$

Theorem 2. *(i)* **DPM.1** *is sound and complete with respect to the class of models that satisfy conditions* $(a)^m$, $(b)^m$, *and* $(c)^m$; *(ii)* **DMP.2** *is sound and complete with respect to the class of models that satisfy* $(a)^m$, $(b)'^m$, $(c)^m$ *and* $(d)^m$.

Proof. As usual, soundness is routine; the argument for Theorem 1 applies *mutatis mutandis.* For completeness, given Lemma 1, it suffices to establish that M^c as specified above meets the requisite conditions.

For $(a)^m$, since $W^c = [\top]$ (Proposition 3.iv) and since $\vdash O\top$ (axiom (N)), so that $O\top \in a$, it follows that there is a B such that $W^c = [B]$ and $OB \in a$, which yields $W^c \in \mathcal{O}_a^c$.

For $(b)^m$, consider any $X, Y \in \mathcal{E}_{M^c}$ and $a \in W^c$, and suppose $X \in \mathcal{O}_a^c$ and $Y \in \mathcal{O}_a^c$. Hence there are B and C such that $X = |B|_{M^c}$ and $OB \in a$ and $Y = |C|_{M^c}$ and $OC \in a$. By Lemma 1, $|B|_{M^c} = [B]$ and $|C|_{M^c} = [C]$. Since $(OB \wedge OC) \rightarrow O(B \wedge C)$ is a theorem of **DMP.1,** $O(B \wedge C) \in a$. Hence $[B \wedge C] \in \mathcal{O}_a^c$. By Proposition 3.i, and Lemma 1, $[B \wedge C] = [B] \cap [C] = |B|_{M^c} \cap |C|_{M^c} = X \cap Y$. Hence $X \cap Y \in \mathcal{O}_a^c$, as required.

For $(c)^m$, consider any $X, Y \in \mathcal{E}_{M^c}$ and $a \in W^c$, and suppose $X \subseteq Y$ and $X \in \mathcal{O}_a$ and $-X \notin \mathcal{O}_a^c$. Hence, there is a B and a C such that $X = |B|_{M^c}$ and $Y = |C|_{M^c}$ and there is a D such that $X = [D]$ and $OD \in a$. By Lemma 1, $|B|_{M^c} = [B] = [D]$, hence $OB \in \mathcal{O}_a^c$ by (RE) and Proposition 4.ii. Since $-|B|_{M^c} \notin \mathcal{O}_a^c$, for all D such that $-X = [D]$, $OD \notin \mathcal{O}_a^c$. $-X = -|B|_{M^c} = |\neg B|_{M^c}$, by Proposition 1.iii, $= [\neg B]$, by Lemma 1. Hence $O\neg B \notin a$, and by maximality, $\neg O\neg B \in a$, or $PB \in a$. Since $X \subseteq Y$, $|B|_{M^c} \subseteq |C|_{M^c}$, so by Lemma 1, $[B] \subseteq [C]$, whence by Proposition 4.i, $\vdash B \rightarrow C$. Therefore, by (RPM), $\vdash PB \rightarrow (OB \rightarrow OC)$. Hence, by (MP) twice, $OC \in a$. That suffices for $Y \in \mathcal{O}_a^c$, as required for this case.

For $(b)'^m$, consider any $X, Y \in \mathcal{E}_{M^c}$, and suppose $X \in \mathcal{O}_a$ and $Y \in \mathcal{O}_a$ and $-(X \cap Y) \notin \mathcal{O}_a$, and show that $X \cap Y \in \mathcal{O}_a$. Given that $X = |B|_{M^c}$ and $Y = |C|_{M^c}$, for some B and C, *so* that $X = [B]$ and $Y = [C]$, by Lemma 1, then $-(X \cap Y) = [\neg(B \wedge C)]$, by Proposition 3. Since $[B] \in \mathcal{O}_a$ and $[C] \in \mathcal{O}_a$, there are D and E such that $[B] = [D]$ and $[C] = [E]$ and $OD \in a$ and $OE \in a$. $\vdash B \leftrightarrow D$ and $\vdash C \leftrightarrow E$, by Proposition 4. Hence $OD \rightarrow OB$ and $OE \rightarrow OC$, by (RE), so that $OB \in a$ and $OC \in a$. Further, $O\neg(B \wedge C) \notin a$, for if it were, then $-(X \cap Y) \in \mathcal{O}_a$, contrary to the supposition. Therefore, $\neg O\neg(B \wedge C) \in a$, or $P(B \wedge C) \in a$. By the postulate (PAND) of **DMP. 2**, it follows that $O(B \wedge C) \in a$, which suffices for $X \cap Y \in \mathcal{O}_a$, as required.

For $(d)^m$, suppose $\emptyset \in \mathcal{O}_a$. Then there is a formula B such that $\emptyset = [B]$ and $OB \in a$. $\emptyset = [\bot]$ (Proposition 3.v). Hence $[\bot] = [B]$, and so $\vdash \bot \leftrightarrow B$, by Proposition 4.ii.. Therefore, $O\bot \in a$ by (RE) . But$\neg O\bot$ is a theorem of **DPM.2,** and so $\neg O\bot \in a$, contrary to the consistency of a. Thus, $\emptyset \notin \mathcal{O}_a$.

This suffices for the theorem since if A is a formula not provable in **DPM,** then $\{\neg A\}$ is consistent and so has a maximal consistent extension $a \in W^c$. By Lemma 1, $M^c, a \models \neg A$, and so $M^c, a \not\models A$, and thus $M^c \not\models A$. Since M^c satisfies the conditions $(a)^m$–$(d)^m$, there is thus a model meeting these conditions on which A does not hold. Hence, by contraposition, if A is a formula that holds of every model that meets these conditions, it must be provable in **DPM.** □

We can extract frame-completeness from Theorem 2 if, for any model M meeting the conditions $(a)^m$, $(b)^m$, $(c)^m$ for **DPM.1** or $(a)^m$, $(b)'^m$, $(c)^m$, $(d)^m$ for **DPM.2** that falsifies a non-theorem A of the system, there is a corresponding model on a frame that meets the original frame conditions (a), (b), (c), or (a), $(b)'$, (c), (d), that also falsifies A. We obtain this by taking a filtration of M, and applying a little more manipulation. This will yield a model on a frame in which every proposition is expressible, and so the satisfaction of conditions $(a)^m$, $(b)^m$, $(c)^m$ ($(a)^m$, $(b)'^m$, $(c)^m$, $(d)^m$) will imply the satisfaction of (a), (b), (c) ((a), $(b)'$, (c), (d)). The resulting model will be a model on a finite frame, and so as a spin-off benefit, we shall establish that each version of **DPM** is complete

with respect to the class of finite frames, and so has the finite model property, from which it follows that **DPM** is decidable since it is finitely axiomatizable.

Given a model $M = \langle F, v\rangle$ on a frame $F = \langle W, \mathcal{O}\rangle$ with M satisfying conditions $(a)^m$, $(b)^m$, $(c)^m$, or $(a)^m$, $(b)'^m$, $(c)^m$, $(d)^m$), as above, construct an alternative model $M^* = \langle F^*, v^*\rangle$ as follows: Let ψ be a finite set of formulas closed under subformulas, i.e., if $A \in \psi$ and B is a subformula of A then $B \in \psi$. Let p_ψ be a particular atomic formula in ψ and let $\top = p_\psi \to p_\psi$ and $\bot = \neg\top$. Let Ψ be the closure of ψ under truth-functions, i.e., Ψ is the smallest set of formulas such that $\psi \subseteq \Psi$ and if $A, B \in \Psi$, then $A \wedge B \in \Psi$, $A \vee B \in \Psi$ and $\neg A \in \Psi$. We note that $\top \in \Psi$ and $\bot \in \Psi$, and that Ψ itself is closed under subformulas.

Given M and such a ψ, define an equivalence relation $\equiv_\psi$ on W such that, for all $a, b \in W$,

$$a \equiv_\psi b \text{ iff } \forall B(\text{if } B \in \psi \text{ then } (M, a \models B \text{ iff } M, b \models B))$$

This generalizes to Ψ, thus

Proposition 5. *For all $a, b \in W$, if $a \equiv_\psi b$, then for all $B \in \Psi$, $M, a \models B$ iff $M, b \models B$.*

Proof. Suppose $a \equiv_\psi b$ and $B \in \Psi$. Proof is by induction on B. If B is atomic or of the form OC, then $B \in \psi$, and so $M, a \models B$ iff $M, b \models B$ follows from the definition of $\equiv_\psi$. If $B = C \wedge D$ or $C \vee D$ or $B = \neg C$, for some C and D, then the result follows directly from the inductive hypothesis. □

The relation $\equiv_\psi$ partitions W into finitely many equivalence classes $[a]$ for $a \in W$,

$$[a] = \{b \in W : b \equiv_\psi a\}$$

This is familiar from standard modal logic. Now, however, we do something a little different; we make what might be called a 'thin' filtration. This is required in order to assure that the frame F^* that is derived from F will meet condition (c), and also $(b)'$ as appropriate.

For each of the equivalence classes $[a]$, take exactly one member; call it a^*. (a^* need not be a itself.) Let W^* be the set of all such selected a^*. Then the following facts obtain.

Proposition 6. *(i) $W^* \subseteq W$; (ii) W^* is finite; (iii) for all $b \in W$ there is an $a^* \in W^*$ such that $b \equiv_\psi a^*$; (iv) for all $a^*, b^* \in W^*$, if $a^* \neq b^*$ then it is not the case that $a^* \equiv_\psi b^*$.*

These all follow directly from the definitions (and the finitude of ψ).

For some convenient notation, for every $X \subseteq W$, let $X\downarrow = X \cap W^*$. Also we continue to write $|A|_M$ for $\{b \in W : M, b \models A\}$, but since M is given by the context of this discussion, let us drop the subscript, except when it is really required for disambiguation (e.g., in the proof of Lemma 6 below to distinguish $|A|_M$ from $|A|_{M^*}$).

Given $M = \langle F, v\rangle$ with $F = \langle W, \mathcal{O}\rangle$, we can now specify the alternative model $M^* = \langle F^*, v^*\rangle$. Let $F^* = \langle W^*, \mathcal{O}^*\rangle$ with W^* as specified ($W^* = W\downarrow$), and $\mathcal{O}^*$ such that for all $a^* \in W^*$ and all $X^* \subseteq W^*$,

$$X^* \in \mathcal{O}^*_{a^*} \text{ iff } \exists B(B \in \Psi \text{ and } X^* = |B|\downarrow \text{ and } |B| \in \mathcal{O}_{a^*})$$

And for all atomic formulas p,

$$v^*(p) = v(p)\downarrow$$

The principle task now is to show (1) that if M satisfies conditions $(a)^m$, $(b)^m$, $(c)^m$, or $(a)^m$, $(b)'^m$, $(c)^m$, $(d)^m$, as appropriate, then F^* satisfies (a), (b), (c), or (a), (b)′, (c), (d), and (2) that M and M^* are equivalent modulo ψ. To this end some preliminary lemmas are required.

Lemma 2. *For all $a^* \in W^*$, there is a formula $B \in \Psi$ such that $|B|\downarrow = \{a^*\}$.*

Proof. This is established by a sort of diagonal argument[20]. Note first that for all $a^*, b^* \in W^*$, if $a^* \neq b^*$ then there is a formula C such that $C \in \Psi$ and $a^* \in |C|$ and $b^* \notin |C|$. For suppose otherwise. Suppose $a^* \neq b^*$ but for every $C \in \Psi$ if $a^* \in |C|$ then $b^* \in |C|$. Then $a^* \equiv_\psi b^*$, for consider any $D \in \psi$, hence $D \in \Psi$. If $M, a^* \models D$, then $a^* \in |D|$, so by the supposition $b^* \in |D|$, and thus $M, b^* \models D$. Suppose then that $M, b^* \models D$, i.e., $b^* \in |D|$, but that it is not the case that $M, a^* \models D$. Then $M, a^* \models \neg D$ and $a \in |\neg D|$. $\neg D \in \Psi$; so, by the supposition, $b^* \in |\neg D|$, or $M, b^* \models \neg D$. That means $M, b^* \not\models D$, a contradiction. Hence, if $M, b^* \models D$, then $M, a^* \models D$, and so $M, a^* \models D$ iff $M, b^* \models D$, which suffices for $a^* \equiv_\psi b^*$. But if $a^* \neq b^*$ then it is not the case that $a^* \equiv_\psi b^*$, by Proposition 6.iv, a contradiction. Therefore, it must be the case that if $a^* \neq b^*$, there is a $C \in \Psi$ such that $a^* \in |C|$ and $b^* \notin |C|$. For each b^* such that $b^* \neq a^*$, select one such formula, and call it C_{b^*}. Let α be the set of all such formulas C_{b^*}. α is finite since W^* is finite. Let B_{a^*} be a conjunction of all the members of α. $B_{a^*} \in \Psi$ since each conjunct $C_{b^*} \in \Psi$ and Ψ is closed under truth-functions. We now show that $|B_{a^*}|\downarrow = \{a^*\}$.

(i) Suppose $x \in |B_{a^*}|\downarrow$. So $x \in |B_{a^*}|$ and $x \in W^*$. Suppose $x \neq a^*$. Then there is a formula $C_x \in \alpha$ such that $a^* \in |C_x|$ and $x \notin |C_x|$. Because B_{a^*} is a conjunction of all members of α, $\vdash B_{a^*} \rightarrow C_x$. Hence, by soundness, Theorem 1, $M \models B_{a^*} \rightarrow C_x$, and so $|B_{a^*}| \subseteq |C_x|$, by Proposition 2.i. Since $x \in |B_{a^*}|$, $x \in |C_x|$, a contradiction. Therefore, if $x \in |B_{a^*}|$, $x = a^*$ and so $x \in \{a^*\}$. Thus, $|B_{a^*}|\downarrow \subseteq \{a^*\}$.

(ii) Suppose $x \in \{a^*\}$, i.e., $x = a^*$. Thus $x \in W^*$. For all $D \in \alpha$, $x \in |D|$. Hence if $\alpha = \{D_1, \ldots, D_n\}$, $M, x \models D_1$ and … and $M, x \models D_n$. Consequently, $M, x \models D_1 \wedge \cdots \wedge D_n$. But $D_1 \wedge \cdots \wedge D_n = B_{a^*}$, so that $M, x \models B_{a^*}$. That is to say, $x \in |B_{a^*}|$, and therefore $x \in |B_{a^*}|\downarrow$. Thus, $\{a^*\} \subseteq |B_{a^*}|\downarrow$. Therefore, by (i) and (ii) together, $|B_{a^*}|\downarrow = \{a^*\}$, as required for the lemma. □

Lemma 3. *For all $X^* \subseteq W^*$, there is a formula B such that $B \in \Psi$ and $X^* = |B|\downarrow$.*

[20] This argument draws from Hughes and Cresswell [22], p. 166, which draws in turn from Segerberg [34], pp. 31–33.

Proof. Suppose $X^* \subseteq W^*$. X^* is finite, because W^* is. Let $X^* = \{a_1^*, \ldots, a_n^*\}$. For every $a_i^* \in X^*$ there is a formula B such that $B \in \Psi$ and $|B_{a_i^*}| = \{a_i^*\}$, Lemma 2. For each such $a_i^* \in X^*$, select one such formula, and call it $B_{a_i^*}$. Let $B_{X^*} = B_{a_1^*} \vee \cdots \vee B_{a_n^*}$. $B_{X^*} \in \Psi$ since each disjunction is in Ψ. We show that $X^* = |B_{X^*}| \downarrow$.

(i) Suppose $x \in X^*$. Then $x \in W^*$. $x = a_i^*$ for some $1 \leq i \leq n$. $A_i^* \in |B_{a_i^*}|$, by specification, so $x \in |B_{a_i^*}|$, which is to say, $M, x \models B_{a_i^*}$. But then $M, x \models B_{a_1^*} \vee \cdots \vee B_{a_n^*}$, i.e., $M, x \models B_{X^*}$. Thus $x \in |B_{X*}|$, and so $x \in |B_{X^*}| \cap W^*$, which is to say $x \in |B_{X^*}| \downarrow$. Hence, $X^* \subseteq |B_{X^*}| \downarrow$.

(ii) Suppose $x \in |B_{X^*}| \downarrow$, i.e., $x \in |B_{X^*}| \cap W^*$. So $x \in W^*$, and $M, x \models B_{X^*}$, i.e., $M, x \models B_{a_1^*} \vee \cdots \vee B_{a_n^*}$ Hence, $M, x \models B_{a_1^*}$ or ... or $M, x \models B_{a_n^*}$. Suppose $M, x \models B_{a_i^*}$. Then $x \in |B_{a_i^*}|$, and since $|B_{a_i^*}| = \{a_i^*\}$, by specification, $x = a_i^*$. Since $A_i^* \in X^*$, $x \in X^*$. Hence, $|B_{X^*}| \downarrow \subseteq X^*$. Putting (i) and (ii) together, $X^* = |B_{X^*}| \downarrow$, as required. □

Lemma 4. For *all* $B, C \in \Psi$, *(i) if* $|B| \downarrow \subseteq |C| \downarrow$, *then* $|B| \subseteq |C|$; *(ii) if* $|B| \downarrow = |C| \downarrow$, *then* $|B| = |C|$.

Proof. For (i), suppose some $B, C \in \Psi$ such that $|B| \downarrow \subseteq |C| \downarrow$, and consider any $b \in |B|$. $b \in W$. Hence, there is an $a^* \in W^*$ such that $b \equiv_\psi a^*$, by Proposition 6.iii. Since $M, b \models B$, and $B \in \Psi$, $M, a^* \models B$, by Proposition 5, so $a^* \in |B|$, whence $a^* \in |B| \downarrow$. So, $a^* \in |C| \downarrow$ and then $a^* \in |C|$, or $M, a^* \models C$. Since $C \in \Psi$ and $b \equiv_\psi a^*$, $M, b \models C$, by Proposition 5. Thus, $b \in |C|$, as required for $|B| \subseteq |C|$. For (ii), the argument is just the same. □

We are now in a position to establish that, given a model $M = \langle F, v \rangle$ meeting model conditions $(a)^m$, $(b)^m$, $(c)^m$, or $(a)^m$, $(b)'^m$, $(c)^m$, $(d)^m$, as appropriate, then the derived frame F^* meets the corresponding frame conditions.

Lemma 5. *(i) If* $M = \langle F, v \rangle$ *satisfies conditions* $(a)^m$, $(b)^m$, $(c)^m$, *then* F^* *satisfies (a), (b), (c), and (ii) if* M *satisfies* $(a)^m$, $(b)'^m$, $(c)^m$, $(d)^m$, *then* F^* *satisfies (a), (b)', (c), (d).*

Proof. We consider the two parts together. Suppose M meets the described model conditions, as appropriate.

For (a), since M meets $(a)^m$, $W \in \mathcal{O}_{a^*}$. $W = |\top|$, so $|\top| \in \mathcal{O}_{a^*}$. $\top \in \Psi$. $W^* = |\top| \cap W^* = |\top| \downarrow$. Hence there is a $B \in \Psi$ such that $W^* = |B| \downarrow$ and $|B| \in \mathcal{O}_{a^*}$. That suffices for $W^* \in \mathcal{O}^*_{a^*}$, as required for F^* to meet condition (a).

For (b), consider some $X, Y \subseteq W^*$ and $a^* \in W^*$ such that $X \in \mathcal{O}^*_{a^*}$ and $Y \in \mathcal{O}^*_{a^*}$ and show that $X \cap Y \in \mathcal{O}^*_{a^*}$. By hypothesis, there are B and C such that $B \in \Psi$ and $X = |B| \downarrow$ and $|B| \in \mathcal{O}_{a^*}$ and $C \in \Psi$ and $Y = |C| \downarrow$ and $|C| \in \mathcal{O}_{a^*}$. $B \wedge C \in \Psi$ since Ψ is closed under truth-functions. Because M meets condition $(b)^m$, $|B| \cap |C| \in \mathcal{O}_{a^*}$; hence, $|B \wedge C| \in \mathcal{O}_{a^*}$, by Proposition 1.i. $X \cap Y = |B| \downarrow \cap |C| \downarrow = (|B| \cap W^*) \cap (|C| \cap W^*) = (|B| \cap |C|) \cap W^* = (|B| \cap |C|) \downarrow = |B \wedge C| \downarrow$ (the last by Proposition 1.i again). Since $|B \wedge C| \in \mathcal{O}_{a^*}$, $X \cap Y \in \mathcal{O}^*_{a^*}$, as required.

For (c), consider $X, Y \subseteq W^*$ and $a^* \in W^*$, and suppose that $X \subseteq Y$ and $X \in \mathcal{O}^*_{a^*}$ and $-X \notin \mathcal{O}^*_{a^*}$, where $-$ is complementation with respect to

W^*, that is, suppose $W^* - X \notin \mathcal{O}^*_{a^*}$. We show that $Y \in \mathcal{O}^*_{a^*}$. By Lemma 3, there are formulas B and C such that $B \in \Psi$ and $X = |B|\downarrow$ and $C \in \Psi$ and $Y = |C|\downarrow$. Hence $|B|\downarrow \subseteq |C|\downarrow$. Since $|B|\downarrow \in \mathcal{O}^*_{a^*}$, there is a $D \in \Psi$ such that $|B|\downarrow = |D|\downarrow$ and $|D| \in \mathcal{O}_{a^*}$. By Lemma 4, $|B| = |D|$; hence $|B| \in \mathcal{O}_{a^*}$. Since $|B|\downarrow \subseteq |C|\downarrow$, $|B| \subseteq |C|$, also by Lemma 4. We show that $|C| \in \mathcal{O}_{a^*}$. For this, given that $|B| \subseteq |C|$ and $|B| \in \mathcal{O}_{a^*}$, it will suffice that $W - |B| \notin \mathcal{O}_{a^*}$, since M satisfies condition $(c)^m$, by hypothesis. Suppose, for *reductio,* that $W - |B| \in \mathcal{O}_{a^*}$. $W - |B| = |\neg B|$, by Proposition 1.iii. Also $\neg B \in \Psi$, since Ψ is closed under truth-functions. Consider $W^* - |B|\downarrow$. This $= |\neg B|\downarrow$. For suppose a $b \in W^* - |B|\downarrow$. Thus, $b \in W^*$ but $b \notin |B|\downarrow$, i.e., $b \notin W^* \cap |B|$. Thus $b \notin |B|$, and so $b \in |\neg B|$, by Proposition 1.iii. So, $b \in W^* \cap |\neg B|$, which is to say $b \in |\neg B|\downarrow$, as required for $W^* - |B|\downarrow \subseteq |\neg B|\downarrow$. For the converse, suppose $b \in |\neg B|\downarrow$, so that $b \in W^*$ and $b \in |\neg B|$. Hence $b \notin |B|$. Hence, $b \notin W^* \cap |B|$, i.e., $b \notin |B|\downarrow$. So, $b \in W^* - |B|\downarrow$, as required for $|\neg B| \subseteq W^* - |B|\downarrow$. Since $W^* - |B|\downarrow = |\neg B|\downarrow$, and since $\neg B \in \Psi$, and since $W - |B| \in \mathcal{O}_{a^*}$, it follows that $W^* - |B|\downarrow \in \mathcal{O}^*_{a^*}$. Hence $W^* - X \in \mathcal{O}^*_{a^*}$, in contradiction to the opening supposition. As noted, that suffices for $|C| \in \mathcal{O}_{a^*}$, which suffices for $|C|\downarrow \in \mathcal{O}^*_{a^*}$, i.e., $Y \in \mathcal{O}^*_{a^*}$, as required for condition (c).

For $(b)'$, when M satisfies $(b)'^m$, the argument is similar. Consider $X, Y \subseteq W^*$ and $a^* \in W^*$, and suppose $X \in \mathcal{O}^*_{a^*}$ and $Y \in \mathcal{O}^*_{a^*}$ and $-(X \cap Y) \notin \mathcal{O}^*_{a^*}$, where, as with (c), $-$ represents complementation with respect to W^*. Thus there are B and C such that $B \in \Psi$ and $X = |B|\downarrow$ and $|B| \in \mathcal{O}_{a^*}$ and $C \in \Psi$ and $Y = |C|\downarrow$ and $|C| \in \mathcal{O}_{a^*}$. $B \wedge C \in \Psi$ since Ψ is closed under truth-functions. As with (b), $X \cap Y = |B|\downarrow \cap |C|\downarrow = (|B| \cap W^*) \cap (|C| \cap W^*) = (|B| \cap |C|) \cap W^* = (|B| \cap |C|)\downarrow = |B \wedge C|\downarrow$. We show that $|B \wedge C| \in \mathcal{O}_{a^*}$, which will suffice then for $X \cap Y \in \mathcal{O}^*_{a^*}$. Since $-(X \cap Y) \notin \mathcal{O}^*_{a^*}$, for every $D \in \Psi$ if $-(X \cap Y) = |D|\downarrow$ then $|D| \notin \mathcal{O}_{a^*}$. We show that $-(X \cap Y) = |\neg(B \wedge C)|\downarrow$. (i) Suppose $x \in -(X \cap Y)$, i.e., $x \in W^* - (X \cap Y)$. So, $x \in W^*$ and $x \notin X \cap Y$. Thus, $x \notin |B|\downarrow \cap |C|\downarrow$, hence, $x \notin |B|\downarrow$ or $x \notin |C|\downarrow$. Consider the first; the second is similar. If $x \in W^*$ but $x \notin |B|\downarrow$, then $x \notin |B|$. Thus $M, x \not\models B$, so $M, x \not\models B \wedge C$. But in that case $M, x \models \neg(B \wedge C)$ and so $x \in |\neg(B \wedge C)|$. And since $x \in W^*$, $x \in |\neg(B \wedge C)|\downarrow$. Thus $-(X \cap Y) \subseteq |\neg(B \wedge C)|\downarrow$. (ii) Suppose $x \in |\neg(B \wedge C)|\downarrow$. Hence $x \in W^*$ and $x \in |\neg(B \wedge C)|$, i.e., $M, x \models \neg(B \wedge C)$. Then $M, x \not\models B \wedge C$, so $M, x \not\models B$ or $M, x \not\models C$, i.e., $x \notin |B|$ or $x \notin |C|$. Consider the first; the second is similar. If $x \notin |B|$, then $x \notin |B|\downarrow$, so $x \notin |B|\downarrow \cap |C|\downarrow$, and $x \notin X \cap Y$. But since $x \in W^*$, $x \in -(X \cap Y)$. Thus $|\neg(B \wedge C) \subseteq -(X \cap Y)$. Putting (i) and (ii) together, $-(X \cap Y) = |\neg(B \wedge C)|\downarrow$. Therefore, since $-(X \cap Y) \notin \mathcal{O}^*_{a^*}$, $|\neg(B \wedge C)| \notin \mathcal{O}_{a^*}$. By Proposition 1, $|\neg(B \wedge C)| = -(|B| \cap |C|)$, where here $-$ represents complementation with respect to W, i.e., $|\neg(B \wedge C)| = W - (|B| \cap |C|)$. Since obviously both $|B|, |C| \in \mathcal{E}_M$, and since $|B| \in \mathcal{O}_{a^*}$ and $|C| \in \mathcal{O}_{a^*}$ and $-(|B| \cap |C|) \notin \mathcal{O}_{a^*}$, $|B| \cap |C| \in \mathcal{O}_{a^*}$ because M meets model condition $(b)'^m$. Hence $|B \wedge C| \in \mathcal{O}_{a^*}$, by Proposition 1.i. Therefore, there is a D, namely $B \wedge C$, such that $D \in \Psi$ and $X \cap Y = |D|\downarrow$ and $|D| \in \mathcal{O}_{a^*}$. That suffices for $X \cap Y \in \mathcal{O}^*_{a^*}$, as required.

For (d), suppose for *reductio* that $\emptyset \in \mathcal{O}^*_{a^*}$. Then there is a $B \in \Psi$ such that $\emptyset = |B|\downarrow$ and $|B| \in \mathcal{O}_{a^*}$. Plainly, $\emptyset = \emptyset\downarrow$. Hence, $\emptyset\downarrow = |B|\downarrow$. So $\emptyset = |B|$,

by Lemma 4. But then $\emptyset \in \mathcal{O}_{a^*}$, contrary to M's satisfying condition $(d)^m$. Therefore, $\emptyset \notin \mathcal{O}^*_{a^*}$, as required. □

Next, we show that M and M^* are equivalent modulo ψ; or, more precisely, first:

Lemma 6. *For all $A \in \psi$ and all $a^* \in W^*$, $M, a^* \models A$ iff $M^*, a^* \models A$.*

Proof. By induction on A. Consider $a^* \in W^*$. Suppose A is atomic, i.e., $A = p$. Since $a^* \in W^*$, $a^* \in v(p)$ iff $a^* \in v(p) \downarrow$. Hence, immediately, $M, a^* \models p$ iff $M^*, a^* \models p$. Suppose the lemma holds for $B, C \in \psi$. If $A = B \wedge C$ and $A \in \psi$, then $B \in \psi$ and $C \in \psi$, since ψ is closed under subformulas. Hence, $M, a^* \models B \wedge C$ iff $M, a^* \models B$ and $M, a^* \models C$ iff, by the inductive hypothesis, $M^*, a^* \models B$ and $M^*, a^* \models C$ iff $M^*, a^* \models B \wedge C$. The cases where $A = B \vee C$ and $A = \neg B$ are similar.

Before considering the case of $A = OB$, it is helpful to note, that under the inductive hypothesis,

Proposition 7. *For all $B \in \psi$ (less than A), $|B|_{M^*} = |B|_M \downarrow$*

For consider any $B \in \psi$ (to which the inductive hypothesis applies), (i) Suppose $x \in |B|_{M^*}$, i.e., $M^*, x \models B$. So, by the inductive hypothesis $M, x \models B$ and $x \in |B|_M$. But since $|B|_{M^*} \subseteq W^*$, $x \in W^*$, so $x \in |B|_M \downarrow$. Hence $|B|_{M^*} \subseteq |B|_M \downarrow$. Likewise, (ii), suppose $x \in |B|_M \downarrow$ so that $x \in |B|_M$, i.e., $M, x \models B$, and $x \in W^*$. Hence the inductive hypothesis applies and $M^*, x \models B$, i.e., $x \in |B|_{M^*}$, and so $|B|_M \downarrow \subseteq |B|_{M^*}$. Both together yield $|B|_{M^*} = |B|_M \downarrow$, as desired.

Returning to the case of $A = OB$, if $A \in \psi$ then $B \in \psi$ since ψ is closed under subformulas. Hence $B \in \Psi$. (i) Suppose $M, a^* \models OB$. Then $|B|_M \in \mathcal{O}_{a^*}$. By the preceding Proposition, $|B|_{M^*} = |B|_M \downarrow$. Hence there is a formula D, namely B, such that $D \in \Psi$ and $|B|_{M^*} = |D|_M \downarrow$ and $|D|_M \in \mathcal{O}_{a^*}$. That suffices for $|B|_{M^*} \in \mathcal{O}^*_{a^*}$, which suffices for $M^*, a^* \models OB$. (ii) Suppose $M^*, a^* \models OB$, so that $|B|_{M^*} \in \mathcal{O}^*_{a^*}$, and thus by the preceding Proposition $|B|_M \downarrow \in \mathcal{O}^*_{a^*}$. Therefore there is a $C \in \Psi$ such that $|B|_M \downarrow = |C|_M \downarrow$ and $|C|_M \in \mathcal{O}_{a^*}$. $|B|_M = |C|_M$, by Lemma 4. Since $|C|_M \in \mathcal{O}_{a^*}$, $|B|_M \in \mathcal{O}_{a^*}$, which suffices for $M, a^* \models OB$, as required. □

The equivalence of M and M^* (modulo ψ) follows immediately, with Proposition 6.iii.

Corollary 2. *For all $A \in \psi$, $M \models A$ iff $M^* \models A$.*

From these lemmas we can now quickly establish frame-completeness for **DPM.**

Theorem 3. *(i)* **DPM.1** *is complete with respect to the class of frames, $\mathcal{F}.1$, that satisfy conditions (a), (b) and (c) as originally stated; (ii)* **DPM.2** *is complete with respect to the class of frames, $\mathcal{F}.2$, that satisfy conditions (a), (b)′, (c) and (d).*

Proof. Consider both versions simultaneously. Take any formula A such that $\nvdash A$. By the model-completeness theorem, Theorem 2, there is a model $M = \langle F, v\rangle$ meeting the respective model conditions such that A does not hold on M, i.e., with $F = \langle W, \mathcal{O}\rangle$, there is an $a \in W$ such that $M, a \not\models A$. Let ψ be the set of subformulas of A] plainly $A \in \psi$. Let $M^* = \langle F^*, v^*\rangle$, with $F^* = \langle W^*, \mathcal{O}^*\rangle$, be the model derived from the filtration of M through ψ as described above leading to Lemmas 5 and 6. There is an $a^* \in W^*$ such that $a \equiv_\psi a^*$, by Proposition 6. By Lemma 6, $M^*, a^* \not\models A$. Moreover, by Lemma 5, F^* meets the requisite frame conditions, and so $F^* \in \mathcal{F}$. Therefore, there is a model on a frame in $\mathcal{F}$ on which A does not hold, and so A *is* not valid with respect to the class of frames, $\mathcal{F}$, that meet the requisite frame conditions. Or, by contraposition and generalization, if a formula A is valid with respect to that class, it must be provable in **DPM.** □

This completes the principle result that was to be established here, that **DPM** is characterized not only by the class of models that meet conditions $(\mathrm{a})^m$, $(\mathrm{b})^m$, $(\mathrm{c})^m$, or $(\mathrm{a})^m$, $(\mathrm{b})'^m$, $(\mathrm{c})^m$, $(\mathrm{d})^m$, as appropriate, Theorem 2, but by the class of frames that meet (a), (b), (c), or (a), $(\mathrm{b})'$, (c), (d), Theorems 1 and 3. Furthermore, we notice that the models M^* that falsify non-theorems of **DPM** are based on frames F^* that are *finite.* Hence, with the corollary to Theorem 1,

Corollary 3. *(i)* **DPM.1** *is sound and complete with respect to the class of all finite frames that meet conditions (a), (b), (c); (ii)* **DPM.2** *is sound and complete with respect to the class of all finite frames that meet conditions (a), (b)′, (c), (d).*

It follows that both versions of **DPM** have the finite model property, and because of their finite axiomatizability, it follows too that **DPM** is decidable.

Corollary 4. DPM *has the finite model property.*

Corollary 5. DPM *is decidable.*

Let us now briefly consider the logics **CDPM** for conditional obligation that correspond most directly to **DPM.** The language of these systems, $\mathcal{L}_c$, has all that is necessary for PC and also the single dyadic connective $O(-/-)$ such that $O(B/A)$ is well-formed whenever A and B are. $P(B/A)$ abbreviates $\neg O(\neg B/A)$. **CDPM.1** is the least set of formulas containing, in addition to (PC), all instances of

CN) $\quad \vdash O(\top/\top)$
CAND) $\quad \vdash (O(B/A) \wedge O(C/A)) \rightarrow O(B \wedge C/A)$

and closed under the rules,

MP) If $\vdash A$ and $\vdash A \rightarrow B$, then $\vdash B$
RCE) If $\vdash A \leftrightarrow B$ then $\vdash O(C/A) \leftrightarrow O(C/B)$
CRE) If $\vdash B \leftrightarrow C$ then $\vdash O(B/A) \leftrightarrow O(C/A)$
CRPM) If $\vdash B \rightarrow C$ then $\vdash P(B/A) \rightarrow (O(B/A) \rightarrow O(C/A))$

For **CDPM.2,** the counterpart to **DPM.2,** replace (CAND) with

CPAND) $\vdash (O(A/C) \wedge O(B/C) \wedge P((A \wedge B)/C)) \rightarrow O((A \wedge B)/C)$
CP) $\vdash \neg O(\bot/A)$

For the semantics, modify the neighborhood frames of **DPM** to treat obligation now not exactly as a property of propositions, but as a relation between them, so that $O(B/A)$ represents that the proposition expressed by B is *obligatory under* the condition expressed by A. More formally, let a dyadic neighborhood frame $F = \langle W, \mathcal{O} \rangle$, with W as before, but now $\mathcal{O}$ assigns each $a \in W$ a set of ordered pairs of propositions $\langle X, Y \rangle$, i.e., $\mathcal{O}_a \subseteq \wp W \times \wp W$. As before, a model M on such an F is a pair $\langle F, v \rangle$ with $v(p) \subseteq W$ for each atomic formula p. $\models$ is defined as usual, but now with

TO(−/−)) $M, a \models O(B/A)$ iff $\langle |A|_M, |B|_M \rangle \in \mathcal{O}_a$

where as before $|A|_M = \{a : M, a \models A\}$.

CDPM.1 is sound and complete with respect to the class of frames that meet the following conditions, which are analogous to those for the frames for **DPM.1,** for all $X, Y, Z \subseteq W$, and all $a \in W$,

ca) $\langle W, W \rangle \in \mathcal{O}_a$
cb) If $\langle X, Y \rangle \in \mathcal{O}_a$ and $\langle X, Z \rangle \in \mathcal{O}_a$, then $\langle X, Y \cap Z \rangle \in \mathcal{O}_a$.
cc) If $Y \subseteq Z$ and $\langle X, Y \rangle \in \mathcal{O}_a$ and $\langle X, -Y \rangle \notin \mathcal{O}_a$ then $\langle X, Z \rangle \in \mathcal{O}_a$

while **CDPM.2** is sound and complete with respect to the class of frames that meet conditions (ca), (cc), and also

cb)′ If $\langle X, Y \rangle \in \mathcal{O}_a$ and $\langle X, Z \rangle \in \mathcal{O}_a$ and $\langle X, -(Y \cap Z) \rangle \notin \mathcal{O}_a$ then $\langle X, Y \cap Z \rangle \in \mathcal{O}_a$
cd) $\langle X, \emptyset \rangle \notin \mathcal{O}_a$

Theorem 4. *(i)* **CDPM.1** *is sound and complete with respect to the class of frames that meet conditions (ca), (cb), and (cc); (ii)* **CDPM.2** *is sound and complete with respect to the class offrames that meet (ca), (cb)′, (cc) and (cd).*

Proof Sketch. Soundness, as usual, is merely a matter of verifying that the axioms are valid and the rules preserve validity; this can be left to the reader. The same arguments as for Theorem 1 apply. The proof of completeness follows that of Theorem 3. First, define the canonical frame $F^c = \langle W^c, \mathcal{O}^c \rangle$ as usual, with W^c the set of maximal consistent extensions of **CDPM** (either version as appropriate), and with $\mathcal{O}^c$ assigning each $a \in W^c$ the set of pairs

$$\mathcal{O}^c_a = \{\langle X, Y \rangle : \exists A \exists B (X = [A] \text{ and } Y = [B] \text{ and } O(B/A) \in a)\}$$

$M^c = \langle F^c, v^c \rangle$ with $v^c(p) = \{a \in W^c : p \in a\}$. Lemma 1, that $M^c, a \models A$ iff $A \in a$ is easily extended to the new formulas $O(B/C)$. Thus, suppose $O(B/C) \in a$, then by the inductive hypothesis $|B|_{M^c} = [B]$ and $|C|_{M^c} = [C]$. So $\langle |C|_{M^c}, |B|_{M^c} \rangle \in \mathcal{O}^c_a$, and $M^c, a \models O(B/C)$. If $M^c, a \models O(B/C)$ then $\langle |C|_{M^c}, |B|_{M^c} \rangle \in \mathcal{O}^c_a$. So there are D and E such that $|C|_{M^c} = [D]$ and $|B|_{M^c} = E$ and $O(E/D) \in a$. Since $|C|_{M^c} = [C]$ and $|B|_{M^c} = [B]$, by the inductive

hypothesis, $[C] = [D]$ and $[B] = [E]$. So $\vdash C \leftrightarrow D$ and $\vdash B \leftrightarrow E$ and thus $O(B/C) \in a$, by (RCE) and (CRE), which covers this case of the induction.

Second, show that **CDPM** is model-complete with respect to the class of models that meet conditions corresponding to (ca), (cb), (cc) ((ca), $(\mathrm{cb})'$, (cc), (cd)) when the *X, Y, Z* are restricted to expressible propositions, those in $\mathcal{E}_M$. This requires showing that M^c meets these model conditions. The argument follows that under Theorem 2.

Third, given a model M that falsifies a non-theorem, derive the alternative model M^* as described for Theorem 3 with $F^* = \langle W^*, \mathcal{O}^* \rangle$ formed from the filtration through M as there described, with $\mathcal{O}^*$ assigning each $a^* \in W^*$ the set of pairs $\langle X^*, Y^* \rangle$ such that

$$\langle X^*, Y^* \rangle \in \mathcal{O}^*_{a^*} \text{ iff } \exists B \exists C (B \in \Psi \text{ and } C \in \Psi \text{ and } X^* = |B| \downarrow \text{ and } Y^* = |C| \downarrow \text{ and } \langle |B|, |C| \rangle \in \mathcal{O}_a)$$

Following the argument for Theorem 3, one can then show that F^* satisfies the conditions (ca), (cb), (cc) ((ca), $(\mathrm{cb})'$, (cc), (cd)) when M satisfies the corresponding model conditions, and also that M and M^* are equivalent modulo ψ, and hence that M^* falsifies *A*. That suffices to establish completeness. □

With completeness, by this argument, also come the corollaries, that both versions of **CDPM** have the finite model property, and so are decidable.

References

1. H. Bezzazi, D. Makinson and R Pino Pérez, "Beyond Rational Monotonicity: Some Strong non-Horn Rules for Nonmonotonic Inference Relations", *Journal of Logic and Computation,* 7 (1997), 605–631.
2. B. Chellas, *Modal Logic, an Introduction,* (Cambridge University Press, Cambridge), 1980.
3. N. C. da Costa, "New Systems of Predicate Deontic Logic", *The Journal of Non-Classical Logic,* 5 (1988), 75–80.
4. N. C. da Costa and W. A. Carnielli, "On Paraconsistent Deontic Logic", *Philosophia,* 16 (1986), 293–305.
5. J. Delgrande, "Weak Conditional Logics of Normality", unpublished manuscript.
6. J. W. Forrester, "Conflicts of Obligation", *American Philosophical Quarterly,* 32 (1995), 31–44.
7. L. Goble, "The Andersonian Reduction and Relevant Deontic Logic" in *New Studies in Exact Philosophy: Logic, Mathematics and Science,* B. Brown and J. Woods, eds., (Hermes Scientific Publishers), 2001, 213–246.
8. L. Goble "Deontic Logic with Relevance", in *Norms, Logics and Information Systems: New Studies on Deontic Logic and Computer Science,* P. McNamara and H. Prakken, eds., (IOS Press, Amsterdam), 1999, 331–345.
9. L. Goble, "A Logic of *Good, Should,* and *Would*— Part I", *Journal of Philosophical Logic,* 19 (1990), 169–199.
10. L. Goble, "A Logic of *Good, Should,* and *Would*— Part II", *Journal of Philosophical Logic,* 19 (1990), 253–276.
11. L. Goble, "The Logic of Obligation, 'Better' and 'Worse' ", *Philosophical Studies,* 70 (1993), 133–163.

12. L. Goble, "Multiplex Semantics for Deontic Logic" *Nordic Journal of Philosophical Logic,* 5 (2000), 113-134.
13. L. Goble, "Murder Most Gentle: The Paradox Deepens", *Philosophical Studies,* 64 (1991), 217–227.
14. L. Goble, "Preference Semantics for Deontic Logic – Part I: Simple Models", *Logique et Analyse, forthcoming.*
15. L. Goble, "Preference Semantics for Deontic Logic – Part II: Multiplex Models", *Logique et Analyse, forthcoming.*
16. S. O. Hansson, "Deontic Logic without Misleading Alethic Analogies", *Logique et Analyse,* 31 (1988), 337–370.
17. S. O. Hansson, *The Structure of Values and Norms,* (Cambridge University Press, Cambridge), 2001.
18. R. Hilpinen, "Deontic Logic" in *The Blackwell Guide to Philosophical Logic,* L. Goble, ed. (Blackwell, Oxford and Malden, MA), 2001, 159–182.
19. J. F. Horty, "Moral Dilemmas and Nonmonotonic Logic", *Journal of Philosophical Logic,* 23 (1994), 35–65.
20. J. F. Horty, "Nonmonotonic Foundations for Deontic Logic", in *Defeasible Deontic Logic,* D. Nute, ed., (Kluwer, Dordrecht) 1997, 17–44.
21. J. F. Horty, "Reasoning with Moral Conflicts", *Noûs,* 37 (2003), 557–605.
22. G. E. Hughes and M. J. Cresswell, *A New Introduction to Modal Logic,* (Routledge, London and New York), 1996.
23. F. Jackson, "On the Semantics and Logic of Obligation", *Mind,* 94 (1985), 177–195.
24. K. Lambert, "Free Logics", in *The Blackwell Guide to Philosophical Logic,* L. Goble, ed. (Blackwell, Oxford and Maiden, MA), 2001, 258–279.
25. D. Lewis, *Counterfactuals,* (Harvard University Press, Cambridge, MA), 1973.
26. D. Lewis, "Semantic Analysis for Dyadic Deontic Logic" In *Logical Theory and Semantic Analysis,* S. Stenlund, ed., (D. Reidel, Dordrecht), 1974, 1–14.
27. A. Loparic and L. Puga, "Two Systems of Deontic Logic", *Bulletin of the Section of Logic,* 15 (1986), 137–144.
28. C. McGinnis, "Semi-Paraconsitent Deontic Logic", unpublished manuscript.
29. C. McGinnis, "Some Quandry-Tolerant Deontic Logics", unpublished manuscript.
30. P. McNamara, "Agential Obligation as Non-Agential Personal Obligation Plus Agency", *Journal of Applied Logic,* in press.
31. D. Nute and X. Yu, "Introduction" to *Defeasible Deontic Logic,* D. Nute, ed. (Kluwer, Dordrecht), 1997, 1–16.
32. R. Routley and V. Plumwood, "Moral Dilemmas and the Logic of Deontic Notions", in *Paraconsistent Logic: Essays on the Inconsistent,* G. Priest, R. Routley, and J. Norman, eds., (Philosophia Verlag, Munich, Hamden, Vienna), 1989, 653–690.
33. P. Schotch and R. Jennings, "Non-Kripkean Deontic Logic" In *New Studies in Deontic Logic,* R. Hilpinen, ed., (D. Reidel; Dordrecht), 1981, 149–162.
34. K. Segerberg, *An Essay in Classical Modal Logic,* 3 vols., (University of Uppsala, Uppsala), 1971.
35. L. van der Torre and Y.-H. Tan, "Two-Phase Deontic Logic", *Logique et Analyse,* 43 (2000), 411–456.
36. B. van Fraassen, "The Logic of Conditional Obligation" *Journal of Philosophical Logic,* 1 (1972), 417–438.
37. B. van Fraassen, "Values and the Heart's Command", *Journal of Philosophy,* 70 (1973), 5–19.

Defeasible Logic: Agency, Intention and Obligation

Guido Governatori[1] and Antonino Rotolo[2]

[1] School of ITEE, The University of Queensland, Australia
[2] CIRSFID, University of Bologna, Italy

Abstract. We propose a computationally oriented non-monotonic multi-modal logic arising from the combination of agency, intention and obligation. We argue about the defeasible nature of these notions and then we show how to represent and reason with them in the setting of defeasible logic.

1 Introduction

This paper combines two perspectives: (a) a cognitive account of agents that specifies motivational attitudes; (b) modelling societies of agents by means of normative concepts [4]. For the first approach, our background is the belief-desire-intention (BDI) architecture, where mental attitudes are taken as primitives to give rise to a set of Intentional Agent Systems [23,2]. This view has been proved to be interesting especially when the behaviour of agents is the outcome of a rational balance among their (possibly conflicting) mental states [3,24]. The normative aspect is based on some intuitions about agents and their societies, in which it is assumed that normative concepts play a decisive role, allowing for the co-ordination of autonomous agents [22,10,12].

Our approach has in general several points of contact with the BOID architecture [4, 5,8,6], where a number of strategies are provided for solving conflicts among informational and motivational attitudes. BOID provides logical criteria (i) to retract agent's attitudes with the changing environment, and so (ii) to settle conflicts by stating different general policies corresponding to the agent type considered. A realistic agent thus corresponds to a conflict-resolution type in which beliefs override all other factors, while other agent types, such as simple-minded, selfish or social ones adopt different orders of overruling. As in the BOID architecture, our system is rule-based. In particular, it is developed in the setting of Defeasible Logic. All components are represented as defeasible conditionals. A rule such as $p \Rightarrow_K q$ means that, given *p,* this implies defeasibly agent's belief that *q.* Our claim is to develop a constructive account of BDI multi-modal logics where the rules are meant to devise suitable logical conditions for introducing modalities. If so, rules may also contain modalised literals, as for example in $Ip \Rightarrow_K q$, where *I* is a BDI operator of intention. In the same spirit, possible conversions of a modality into another can be accepted, as when the applicability of $Ip \Rightarrow_K q$ may permit to obtain *Iq.* Based on this intuitions, our focus will be on Bratman's [3] concept of *policy-based* intention [11]. The relation between mental attitudes and non-monotonicity should not sound surprising. Recent works by Thomason [27] and on BOID confirm this trend. Such a connection, with regard to epistemic logics, has already received much attention in the AI community [19]. However, the notion of defeasibility may play a new role within a constructive theory of (modal) operators. As we said, our aim is to show how to introduce modalities in a (computationally oriented) non-monotonic formalism. In

A. Lomuscio and D. Nute (Eds.): DEON 2004, LNAI 3065, pp. 114–128, 2004.

this way, the notion of defeasible derivability is crucial since rules for mental states and conditions for derivation involving them allow to introduce modal operators. This approach is motivated by the inherent computational complexity of multimodal logics [13] and, often, the notion of modality adopted for agents systems is by its own nature non-monotonic and so does not lend itself to necessitation [11]. The use of non-monotonic logics in intention reasoning allows the agent to reason with partial knowledge without having a complete knowledge of the environment. This also helps the agent in avoiding a complete knowledge of the consequences. We outline a proof theory whereby one can reason about ways of maintaining intention consistency in BDI like agent systems. The new approach facilitates the designer of an agent system like BDI in describing rules for constructing intentions from goals and goals from knowledge.

BOID system incorporates also obligations. This is crucial in characterising the interplay between internal and external factors. Such intuition is also adopted here and is framed as well within a non-monotonic setting. Even for this component, the logical aim is to devise suitable conditions for introducing modalities. Two questions may be decisive in this regard. First, it would be important to recast the logical nature of obligations and to investigate how defeasible logic, as described in the following sections, might capture the well-known defeasible character of deontic reasoning. A full analysis of the above issue is outside the scope of the paper. However, it is at least worth mentioning that our framework avoids a difficulty that is recognised in the deontic literature [7]. The source of this difficulty is the closure, classically accepted in Standard Deontic Logic, of the obligation operator under logical consequence. We simply point out that these difficulties are avoided by developing a suitable notion of logical derivation of obligations. In general, with the adoption of this strategy we preserve at least some basic properties of obligations such as the closure under logical equivalence and consistency. An important issue concerns the relation between obligations and mental states. As it is pointed out [5], a number of possible approaches are available. Here we focus shortly on some minimal principles that emerge from the agent specification approach considered in [6]. In particular, as argued there, we may adopt, for example, the schema $Op \rightarrow Ip$, or analogous versions for the other mental attitudes. This axiom is the strong version of intentional *norm regimentation* as it does not simply prescribe the consistency between obligations and intentions but states the inclusion of the former in the latter ones. This of course means that what is not intended is also not obligatory. Other principles, such as $Op \rightarrow \neg I\neg p$, correspond to weak forms of norm regimentation with regard to agent's mental states. In this sense, they also express hard constraints on agent systems. A different principle that regulate the interaction between obligations and desires may be $(Op \wedge \mathrm{GOAL}\neg p) \rightarrow \neg I\neg p$. This avoids that the output of a conflict between an obligation and a desire is that of adopting a plan for obtaining what is desired. These principles can be easily encoded in our framework.

Last but not least, our framework is enriched by the notion of modal agency [9]. This aspect differentiates this system if compared, for example, to BOID architecture. The same logical strategy—a rule-based approach to introduce modalities—is also applied to this case. In particular, we will devise a set of rules to encode the action transitions occurring, under certain circumstances, as the results of actions.

We will focus on the idea of personal and direct action to realise a state of affairs. This concept is usually formalised by the well-known modal operator E, such that a formula like $E_i p$ means that the agent i brings it about that p. Different axiomatisations have been provided for it but almost all include $E_i p \rightarrow p$ (T, i.e., successfulness), $\neg E_i \top$ **(No)**, $(E_i p \wedge E_i q) \rightarrow E_i(p \wedge q)$ **(C)**, and are closed under logical equivalence [25,9]. This analysis, however, is here integrated by focusing on the intentional character of actions. This is done for two reasons. First, in the light of the logical framework we have defined so far it is interesting to devise criteria for handling the specific interaction between actions, intentions and the other mental states. Second, the aim is to make more precise the logical meaning of the notion of direct action. In fact, as found in the literature [26], it is not possible to capture with E the difference between the modal qualifications "sees to it" and "brings it about". Both are usually represented by this modal operator, despite the fact that the former expression exhibits a clear intentional character, whereas the latter may refer as well to unintentional actions [14]. Thus we introduce the operator Z to express intentional actions. It is characterised by all basic properties of E plus the schema $Zp \rightarrow Ip$, which cannot be in general valid for E.

The interest of adding agency to a framework that includes cognitive states and obligations is evident. First, the simple combination of agency and deontic operators makes possible a more accurate representation of obligations directed to agents' behaviour, such as in the case of OZp. In addition, it allows to express the creation of obligations, as in ZOp. As regards handling conflicts between rules, new possible types of agents can be defined, according to the order of overruling we want to adopt. In this perspective, forms of regimentation may be introduced especially for the operator Z. Finally, it is possible to embed in the system a number of interesting properties, such as $ZOp \rightarrow Ip$, which completes what is stated by a reasonable and analogous schema without the operator of agency, namely, $IOp \rightarrow Ip$.

Finally, a few notes on the meaning of rules for obligation, which emerge from focusing on their interplay with the other components we have described so far. If rules define the conditions for the introduction of modal operators, when we deal with obligations defeated by other components we may in fact adopt two different views. Suppose we have two rules like $r_1 : p \Rightarrow_Z q$ (a rule for action) and $r_2 : s \Rightarrow_O \neg q$ (a rule for obligation). Both are applicable and r_1 defeats r_2. If so we cannot derive $\neg q$ via r_2 and so $O\neg q$. In a first interpretation (applicability-based obligation), that a rule for action (but the same applies to other components such as beliefs) collides with a rule for obligation means that a normative violation has occurred [4]. But if r_1 prevails, in our setting we cannot argue in favour of the occurrence of $O\neg q$. On the other hand, a violation of an obligation does not imply the cancellation of such an obligation [28]. The obligation is still in force. This means that the existence of the actual obligation $O\neg q$ depends on the applicability of r_2, independently of the effective derivation of its consequent. In a second interpretation (pure-derivability-based obligation) the existence of actual obligations depends on the effective derivation of the consequent of a rule. In this case we can argue as follows. On the one hand, the non-derivation of $O\neg q$ means that, as soon as a violation occurs, r_2 is nothing but a special kind of *prima facie* obligation: when violated, it does not make sense to deduce its consequent as a real obligation. On the other, and more radically, since the obligations that count in the system are those which

are derivable, we may say that, in the event the action of the agent blocks the inference of $O\neg q$, the agent is a sort of legislator within the system; similar considerations apply to when intentions override obligations.

2 Basic Defeasible Logic of Agency, Intention and Obligation

Usually modal logics are extensions of classical propositional logic with some intensional operators. Thus any classical (normal) modal logic should account for two components: (1) the underlying logical structure of the propositional base and (2) the logic behaviour of the modal operators. Alas, as is well-known, classical propositional logic is not well suited to deal with real life scenarios. The main reason is that the descriptions of real-life cases are, very often, partial and somewhat unreliable. In such circumstances classical propositional logic might produce counterintuitive results insofar as it requires complete, consistent and reliable information. Hence any modal logic based on classical propositional logic is doomed to suffer from the same problems. On the other hand the logic should specify how modalities can be introduced and manipulated. Common rules for modalities are necessitation and RM. Consider the necessitation rule of normal modal logic which dictates the condition that an agent knows all the valid formulas and thereby all the tautologies. Such a formalisation might suit for the knowledge an agent has but definitely not for the intention part and, consequently, not for a logic of intentional agency. Furthermore, many authors have expressed concerns about the meaningfulness of $O\top$. Moreover, an agent need not be intending all the consequences of a particular action it does. It might be the case that it is not confident of them being successful. Thus the two rules are not appropriate for a logic of deontic agency. A logic of deontic agency should take care of the underlying principles governing the intention and the action of an agent. It should have a notion of the direct and indirect knowledge of the agent, where the former relates to facts as literals whereas the latter to that of the agent's theory of the world in the form of rules. Similarly the logic should also be able to account for general intentions as well as the policy-based (derived ones) intentions of the agent. Finally it should offer facilities to describe obligations and the relationships between the various modalities.

These are in short the main guidelines we will follow in this and the subsequent sections to develop a suitable framework to deal with agency, intention and obligation components. As we have argued so far, reasoning about intentions and other mental attitudes has a defeasible nature, and defeasibility is one of the proper characteristic of normative reasoning. Thus any system that aims at the integration of intentions and obligations, for example a multi-agent system, should cater for defeasibility. The two phenomena (mental attitudes and deontic notions) are both subject to defeasibility, but they might obey different and sometimes incompatible intuitions; thus we need a non-monotonic formalism that is able to deal with them in a flexible, efficient and modular way and should offers itself to a seamless integration of the relevant modal operators. Moreover we need an efficient and easily implementable system to capture the required defeasible instances.

Defeasible logic, as developed by Nute [20] with a particular concern about computational efficiency and developed over the years by [17,1], is our choice. The reason

being ease of implementation [18], flexibility [1] (it has a constructively defined and easy to use proof theory which allows us to capture a number of different intuitions of non-monotonicity) and it is efficient: it is possible to compute the complete set of consequences of a given theory in linear time [16].

A defeasible theory contains five different kinds of knowledge: facts, strict rules, defeasible rules, defeaters, and a superiority relation. In this section we consider only essentially prepositional rules. Rules containing free variables are interpreted as the set of their variable-free instances.

Facts are indisputable statements, for example, "John is a minor". In the logic, this might be expressed as *minor*(*John*).

Strict rules are rules in the classical sense: whenever the premises are indisputable (e.g., facts) then so is the conclusion. An example of a strict rule is "every minor is a person". Written formally: $minor(X) \rightarrow person(X)$.

Defeasible rules are rules that can be defeated by contrary evidence. An example of such a rule is "every person has the capacity to perform legal acts to the extent that the law does not provide otherwise"; written formally: $person(X) \Rightarrow hasLegalCapacity(X)$. The idea is that if we know that someone is a person, then we may conclude that he/she has legal capacity *unless there is other evidence suggesting that h/she may not have.*

Defeaters are a special kind of rules. They are used to prevent conclusions not to support them. For example: $WeakEvidence \rightsquigarrow \text{-}guilty$ This rule states that if pieces of evidence are assessed as weak, then they can prevent the derivation of a "guilty" verdict; on the other hand they cannot be used to support a "not guilty" conclusion.

The *superiority relation* among rules is used to define priorities among rules, that is, where one rule may override the conclusion of another rule. For example, given the defeasible rules

$$r: \quad person(X) \Rightarrow hasLegalCapacity(X)$$
$$r': minor(X) \Rightarrow \neg hasLegalCapacity(X)$$

which contradict one another, no conclusive decision can be made about whether a minor has legal capacity. But if we introduce a superiority relation $>$ with $r' > r$, then we can indeed conclude that the minor does not have legal capacity.

A *rule r* consists of its *antecedent* (or *body*) $A(r)$ ($A(r)$ may be omitted if it is the empty set) which is a finite set of literals, an arrow, and its *consequent* (or *head*) $C(r)$ which is a literal. Given a set R of rules, we denote the set of all strict rules in R by R_s, the set of strict and defeasible rules in R by R_{sd}, the set of defeasible rules in R by R_d, and the set of defeaters in R by R_{dft}. $R[q]$ denotes the set of rules in R with consequent q. If q is a literal, $\sim q$ denotes the complementary literal (if q is a positive literal p then $\sim q$ is $\neg p$; and if q is $\neg p$, then $\sim q$ is p).

A *defeasible theory* D is a structure $(F, R^K, R^I, R^Z, R^O, >)$ where F is a finite set of facts; R^K, R^I, R^Z and R^O are, respectively, finite set of rules (strict, defeasible rules and defeaters) for knowledge, intentions, agency, and obligations; and $>$, the superiority relation, is a binary relation over the set of rules (i.e., $> \subseteq (R^K \cup R^I \cup R^Z \cup R^O)^2$).

Intuitively, given an agent, F consists of the information the agent has about the world, its immediate intentions, its actions and the absolute obligations; R^K corresponds to the agent's theory of the world, while R^Z, R^I and R^O encode its actions, policy, and normative system; $>$ captures the strategy of the agent (or its preferences). The policy

part of a defeasible theory captures both intentions and goals. The main difference is the way the agent perceives them: goals are possible outcomes of a given context while intentions are the actual goals the agent tries to achieve in the actual situation. In other words goals are the choices an agent has and intentions are the chosen goals; in case of conflicting goals (policies) the agent has to evaluate the pros and cons and then decide according to its aims (preferences), which are encoded by the superiority relation.

A *conclusion* of D is a tagged literal and can have one of the following four forms:

$+\Delta q$ meaning that q is definitely provable in D (i.e., using only facts and strict rules).
$-\Delta q$ meaning that we have proved that q is not definitely provable in D.
$+\partial q$ meaning that q is defeasibly provable in D.
$-\partial q$ meaning that we have proved that q is not defeasibly provable in D.

Over the years a number of formulations of the proof theory of defeasible logic have been proposed (sometimes for variants of defeasible logic); here we will adopt the meta-program formalisation of [17].

The meta-program M assumes that the predicates, `fact(Head)`, `superior(Rule1, Rule2)`, `strict(Name,Operator,Head,Body)`, `defeasible(Name,Operator, Head,Body)`, and `defeater(Name,Operator,Head,Body)`, which are used to represent a defeasible theory, are defined. The interpretation of the basic predicates of the meta-program is as follows:

$$\texttt{fact(p)} \textit{ iff } p \in F$$
$$\texttt{strict(r, m, p, [a}_1,\ldots,\texttt{a}_n\texttt{])} \textit{ iff } r: a_1,\ldots,a_n \rightarrow_m p \in R_s[p]$$
$$\texttt{defeasible(r, m, p, [a}_1,\ldots,\texttt{a}_n\texttt{])} \textit{ iff } r: a_1,\ldots,a_n \Rightarrow_m p \in R_d[p]$$
$$\texttt{defeater(r, m, p, [a}_1,\ldots,\texttt{a}_n\texttt{])} \textit{ iff } r: a_1,\ldots,a_n \rightsquigarrow_m p \in R_{dft}[p]$$
$$\texttt{superior(r, s)} \textit{ iff } r > s$$

According to the above predicates we introduce the definition of a rule.

```
rule(R, X, P, [A1,...,An]):- strict(R, X, P, [A1,...,An]).
rule(R, X, P, [A1,...,An]):- defeasible(R, X, P, [A1,...,An]).
rule(R, X, P, [A1,...,An]):- defeater(R, X, P, [A1,...,An]).
```

We are now ready for the clause defining the meta-program describing the proof-theory of defeasible logic[1]. If we disregard the modal operator it is immediate to see that the following meta-program has the same structure as the meta-programs given for propositional defeasible logic in [1,17]. Essentially we have four (independent) copies of the same meta-program, one for each modality.

[1] We have permitted ourselves some syntactic flexibility in presenting the meta-program. However, there is no technical difficulty in using conventional logic programming syntax to represent this program. As usual with logic programming capital letters stand for variables, however we reserve *K,O, I* and *Z* for modalities, and we will use `X, Y, W` for variables ranging over modal operators.

```
strictly(P, K):- fact(P).
strictly(P, X):- fact(XP).
strictly(P, X):- strict(R, X, P, [A1,...,An,Y1B1,...,YmBm]),
   strictly(A1, K), ..., strictly(An, K),
   strictly(B1, Y1), ..., strictly(Bm, Ym).
```

The first two clauses establish that a conclusion in strictly provable if it is one of the facts, while the third corresponds to modus ponens for strict rales and strictly derivable literals. Notice that the first clause is relative to rule for knowledge; as we have argued before the rules in R^K are used to encode the description of the environment (and there is no modal operator K!). Thus unmodalized literals can be thought of as prefixed by a virtual K modal operator.

```
defeasibly(P, X):- strictly(P, X).
defeasibly(P, X):- consistent(P, X),
   supported(R, X, P),
   not defeated(P, X, S).

consistent(P, X):- not strictly(~P, X).

defeated(P, X, S):- applicable(S, X, ~P),
   not overruled(~P, X, T, S).

overruled(P, X, T, S):- supported(T, X, P),
   superior(T, S).

applicable(R, X, P):- rule(R, X, P, [A1,...,An,Y1B1,...,YmBm]),
   defeasibly(A1, K), ..., defeasibly(An, K),
   defeasibly(B1, Y1), ..., defeasibly(Bm, Ym).

supported(R, X, P):- rule(R, X, P, [A1,...,An,Y1B1,...,YmBm]),
   defeasibly(A1, K), ..., defeasibly(An, K),
   defeasibly(B1, Y1), ..., defeasibly(Bm, Ym).
   not defeater(R, X, P, [A1,...,An,Y1B1,...,YmBm]).
```

The first clause allows the transformation of a strict conclusion in a defeasible conclusion. A defeasible derivation of a literal p consists of three phases. In the first phase we establish that the opposite literal is not strictly provable and then have to provide an applicable supportive rule for p (i.e., using the predicate `supported(r,p)`, where r is a supportive rule for p), then in the second phase we build all possible counterarguments against p (i.e., `defeated(p, s)` meaning that the literal p is defeated by rale s) and we have to verify that the conclusion is not defeated by the attacking arguments, so we try to rebut the counterarguments (i.e., `overruled(~p, t, s)`) by stronger arguments for the intended conclusion.

The relationship between proof tags on one hand and the predicates `strictly` and `defeasibly` on the other is as follows:

$D \vdash +\Delta_X p$ *iff* $M \vdash$ `strictly(p,X)` $\quad D \vdash -\Delta_X p$ *iff* $M \vdash$ `not strictly(p,X)`

$D \vdash +\partial_X p$ *iff* $M \vdash$ `defeasibly(p,X)` $\quad D \vdash -\partial_X p$ *iff* $M \vdash$ `not defeasibly(p,X)`

Let us consider a theory where $F = \{Ia,b,Od,e\}$ and $R = \{Ia,b \Rightarrow_Z c;\ e,Zc \Rightarrow_O f\}$. Here we can prove $+\partial_I a$, $+\partial_K b$, $+\partial_K e$ and $+\partial_O d$ since they are facts. Then the first rule is applicable and we can derive $+\partial_Z c$, and now the second rule is applicable and we obtain $+\partial_O f$. If we replace the first rule with $Ia,b \Rightarrow_K c$ we conclude $+\partial_K c$ instead of $+\partial_K c$ and now the second rule is no longer applicable. We illustrate the theory with the help of a concrete example. A drunk surgeon intends to operate a patient. The surgeon is aware that operating under the influence of alcohol will result in a failure. Moreover the legal system under which the surgeon operates prescribes that people causing permanent damages as a result of negligence are responsible. Thus the two rules can be rewritten, respectively as

$$I(operate),\ drunk \Rightarrow_Z fail$$
$$permanentDamages,\ Z(fail) \Rightarrow_O responsible$$

The conclusion is that the surgeon is responsible, because the damages are the result of an intentional negligence. What about when the surgeon, not on duty and being the only person able to complete the required medical procedure, is drunk and the patient will die without the operation? The surgeon knows that the patient will suffer permanent damages as a result of the operation, but he operates anyway. In this case we have to change the first rule in $I(operate),\ drunk \Rightarrow_K fail$. Here we derive $+\partial_K fail$ instead of $+\partial_Z fail$, and thus we block the application of the second rule. Hence we cannot conclude that the surgeon is responsible.

3 Interaction among Agency, Intention and Obligation

The program given in the previous section does not account for the properties of the modal operators and their mutual relationships. For these we have to introduce more clauses in the meta-program.

```
strictly(P, K):- strictly(P, Z).
defeasibly(P, K):- defeasibly(P, Z).
```

These two clauses enable us to convert a conclusion in *Z* in a conclusion in *K*, and thus they mimic the successfulness of the modal operator *Z*.

Let us see now the relationship between the different kinds of rule we have introduced so far. Table 1 shows all possible cases and, for each kind of rule, indicates all *potential* attacks on it. Since we have defined four kinds of rules, we have to analyse twelve combinations, which are gathered in the table in six columns. Each column corresponds to a type of potential attack, such that the second rule placed in each box is nothing but the potential attack on the first one. If the potential attack fails, since the superiority does not play here any role, this means that the case at stake does not correspond to a real attack: The type of rule that wins does so in any case and independently of inspecting the strength of the rules involved (i.e., without considering superiority relation).

To represent the possible attacks we have to strengthen the definitions of the predicate `consistent`.

```
consistent(P, X):- not strictly(~P, K),
   not strictly(~P, Y1), ..., not strictly(~P, Yn).
```

Table 1. Basic Attacks

$\Rightarrow_K p$ $\Rightarrow_O \sim p$	$\Rightarrow_K p$ $\Rightarrow_I \sim p$	$\Rightarrow_K p$ $\Rightarrow_Z \sim p$	$\Rightarrow_I p$ $\Rightarrow_Z \sim p$	$\Rightarrow_O p$ $\Rightarrow_I \sim p$	$\Rightarrow_O p$ $\Rightarrow_Z \sim p$
$+\partial_K p$	$+\partial_K p$	$-\partial_K p$	$-\partial_I p$	type of agent	type of agent
$\Rightarrow_O p$ $\Rightarrow_K \sim p$	$\Rightarrow_I p$ $\Rightarrow_K \sim p$	$\Rightarrow_Z p$ $\Rightarrow_K \sim p$	$\Rightarrow_Z p$ $\Rightarrow_I \sim p$	$\Rightarrow_I p$ $\Rightarrow_O \sim p$	$\Rightarrow_Z p$ $\Rightarrow_O \sim p$
$-\partial_O p$	$-\partial_I p$	$-\partial_Z p$	$-\partial_Z p$	type of agent	type of agent

where $\mathtt{Y}_1, \ldots, \mathtt{Y}_n$ are the modalities that attack the modality X, according to Table 1. At the same time, we have to allow more types of rule in the attack phase.

```
applicable(R, X, P):- rule(R, Y, P, [A1,...,An,W1B1,...,WmBm]) ,
   defeasibly(A1, K), ..., defeasibly(An, K),
   defeasibly(B1, W1), ..., defeasibly(Bm, Wm).
```

This clause is required for all Y that attack X in Table 1. Moreover, if $Y = Z$ we have to include, due to the successfulness of the operator, the additional clause

```
applicable(R, X, P):- rule(R, Z ~XP, [A1,...,An,Y1B1,...,YmBm]) ,
   defeasibly(A1, K), ..., defeasibly(An, K),
   defeasibly(B1, Y1), ..., defeasibly(Bm, Ym).
```

Table 1 (and, as we shall see, Tables 2 and 3) provides some basic criteria for classifying cognitive agents [8, 4]. The general assumption of Table 1 is to deal with realistic agents. In other words, we set criteria for solving conflicts in which beliefs in general override the other components. In fact, our approach considers epistemic rules as agent's basic principles of rationality about the world. The only exception to this view is that rules for action may attack rules for belief, since the former ones capture the mechanism that governs the factual results of (intentional) actions. We can speak in this case of quasi-realistic agents since, given a certain belief, a contrary evidence based on rules for action may prove that such a belief is false. Given this background, Tables 2 and 3 will consider other agent's types, such as selfish and social, plus further specifications deriving from more articulated criteria for solving conflicts. As we shall see, the double reading assigned to the rules for obligation will allow us to provide an alternative interpretation of some already established criteria for handling conflicts between deontic factors, on one hand, and mental as well as action components, on the other.

Let us focus on some examples for each type of potential attack described in Table 1. Suppose we have (first column from the left) r_1 : *forest*, *dry*, *spark* $\Rightarrow_K$ *fire* and r_2 : *forest* $\Rightarrow_O \neg$*fire*. It is clear that rule r_2 does not determine a real attack on r_1. Since we assume the agent is realistic, rule r_1 is nothing but a principle of rationality of the agent: It says that a fire is (defeasibly) the consequence of a spark in a dry forest. Rules like r_1 must prevail with regard to deontic rules, such as r_2 that prohibits to light a fire in a forest. When r_1 is attacked by r_2, the output that follows from r_1 is not affected by this attack and the fire should be obtained since this fact is independent from any rule that forbids to light fires in the forest. Vice versa, the derivation of the obligation not to light a fire is blocked since such an obligation is meaningless when the conditions for r_1 occur: Of course, r_2 does not apply when *fire* is obtained according to agent's rationality.

Similar remarks apply to the case that involves rules for knowledge and intention (second column). Let us consider the rule r_3 : ***cautious*** $\Rightarrow_I$ ***¬fire***. Even here it is reasonable to argue in favour of r_1. Although agent's being cautious means to intend not to light a fire, this intention does not necessarily override r_1, namely *the fact,* according to agent's knowledge, that a spark normally causes a fire in a dry forest. This means that, when r_3 attacks r_1, the consequent of the latter must be obtained, while the reverse attack should prevent to get $I\neg fire$ since such an intention is meaningless when the agent assumes rationally that the fire must spread through the forest. Different arguments may be put forward when a rule for action, r_4 : ***protect_spark*** $\Rightarrow_Z$ ***¬fire***, is considered in combination with r_1. Rule r_4 states *the fact that fire* obtains and may be viewed as a (factual) contrary evidence with regard to r_1. In general, rules like $p \Rightarrow_Z q$ say that a specific action preformed by agent, under certain circumstances, defeasibly determines through such action the occurrence of *q,* and so that *Zq.* The applicability of these rules may thus be a factual and contrary evidence with respect to *k*-rules that would allow to infer $\neg q$. For similar (but opposite) reasons, the reverse attack (r_1 on r_4) should block the derivation of *¬fire.*

Since we assume the rationality of the agent with regard to its knowledge about the world, we have set that rules for knowledge be greater in strength with regard to rules for obligation and intention. Actions may override knowledge while mutual attacks involving intentions and actions determine real attacks for the trivial reason that actions are intentional in character. It is obvious that, when we have a rule such as r_5 : ***incautious*** $\Rightarrow_I$ ***fire***, the attack of r_5 on r_4 prevents from obtaining *¬fire* while the reverse attack blocks the derivation of *fire:* Actions defined by rules for *Z* are intentional.

On the other hand, as we have indicated in Table 1, the interplay between obligations, intentions and actions cannot be settled so easily. In the light of well-known distinctions among different kinds of agent, Table 2 summarises all combinations related to the cases indicated in Table 1, first and second columns from the right.

Table 2. Type of Agent: Basic Attacks

$\Rightarrow_O p / \Rightarrow_I \sim p$			$\Rightarrow_O p / \Rightarrow_Z \sim p$		
$+\partial_O p$	$+\partial_I \sim p$	Independent	$+\partial_O p$	$+\partial_Z \sim p$	Strongly independent
$+\partial_O p$	$-\partial_I \sim p$	Strongly social	$+\partial_O p$	$-\partial_Z \sim p$	Social
$-\partial_O p$	$+\partial_I \sim p$	Selfish	$-\partial_O p$	$+\partial_Z \sim p$	Strongly selfish
$-\partial_O p$	$-\partial_I \sim p$	Strongly pragmatic	$-\partial_O p$	$-\partial_Z \sim p$	Pragmatic

Let us provide some brief comments. Independent and strongly independent agents are free respectively to adopt intentions and to perform intentional actions in conflict with obligations. In particular, within a pure-derivability-based interpretation of obligations (see Section 1), strongly independent agents may correspond to true cases of normative violation, since the actual obligation is derived in presence of a contrary action. As expected, for social and strongly social agents obligations override rules for action and for intention. In addition to the standard view [4], the overruling of intentions or actions with regard to obligations may configure a case of agent legislator, when, within a pure-derivability-based interpretation, only derived obligations count as such in the system. Pragmatic and strongly pragmatic are cases where no derivation is

possible and so the agent's behaviour is open to any other course of action other than those specified in the rules considered. To illustrate the potential conflicts between obligations and intentional acts we examine the well-known prisoner dilemma. Two people are arrested for a major crime, however the police does not have enough evidence to incriminate them, but they can be charged with and convicted for a minor crime. However if one of them confesses the crime she will be sentenced to one year and the other to twenty-five years. If both confess they will be imprisoned for ten years each. Finally if none of them confesses then they have to serve for three years each. The two criminals are part of a criminal organisation renowned for its code of honour that prescribes to not betray your fellows. The best individual outcome is to confess the crime, while the best outcome according to the organisation code is not confessing it. Hence this situation can be represented by the following theory:

$$\Rightarrow_Z \mathit{confess} \qquad \Rightarrow_O \neg \mathit{confess}$$

A "selfish criminal" will confess $(+\partial_Z \mathit{confess}, -\partial_O \neg \mathit{confess})$, giving thus priority to his welfare, while a "social criminal" will stick with the code of honour and will not confess the crime $(+\partial_O \neg \mathit{confess}, -\partial_Z \mathit{confess})$.

Table 2 does not cover all possible types of agent. In fact, the focus is there on possible attacks that involve only two rules. Table 3 completes the scenario and provides all possible combinations when we deal with three rules. It is worth noting that we consider only the case with $\Rightarrow_O p$, $\Rightarrow_I \sim p$ and $\Rightarrow_Z \sim p$: The case with $\Rightarrow_I \sim p$ and $\Rightarrow_Z p$ is meaningless since rules for Z govern only intentional actions. For similar reasons, some combinations in Table 3 are excluded (as highlighted by adding three question marks). Some comments on Table 3. Strongly independent agents are basically as in Table 2 because Z implies I. The types hypersocial and hyperpragmatic do not add conceptually anything with respect to their corresponding and weaker versions of Table 2. The new cases are the selfish saint, sinner and social sinner types. The first is given when the content of agent's intention is in conflict with an obligation, but no intentional action to realise such a content is performed. The sinner performs this action and, in parallel, the obligation is defeated. The social sinner has this intention, the derivation of the obligation is blocked but no violating action is performed. Once again, notice that sinner and social sinner may viewed, within a pure-derivability-based interpretation, as peculiar cases of legislator.

Table 3. Type of Agent: Other Attacks

$\Rightarrow_O p / \Rightarrow_Z \sim p / \Rightarrow_I \sim p$			
$+\partial_O p$	$+\partial_Z \sim p$	$+\partial_I \sim p$	Strongly independent
$+\partial_O p$	$+\partial_Z \sim p$	$-\partial_I \sim p$	???
$+\partial_O p$	$-\partial_Z \sim p$	$+\partial_I \sim p$	Selfish saint
$+\partial_O p$	$-\partial_Z \sim p$	$-\partial_I \sim p$	Hypersocial
$-\partial_O p$	$+\partial_Z \sim p$	$+\partial_I \sim p$	Sinner
$-\partial_O p$	$+\partial_Z \sim p$	$-\partial_I \sim p$	???
$-\partial_O p$	$-\partial_Z \sim p$	$+\partial_I \sim p$	Social sinner
$-\partial_O p$	$-\partial_Z \sim p$	$-\partial_I \sim p$	Hyperpragmatic

Another interesting feature that could be explained using our formalism is that of *rule conversion.* For instance, suppose that a rule of a specific type is given and also suppose that all the literals in the antecedent of the rule are provable in one and the same modality. If so, it is possible to argue that the conclusion of the rule inherits the modality of the antecedent. To give an example let $p, q \Rightarrow_K r$ denote that an agent knows r given p and q (or r is a consequence of p and q). Now suppose *I(p)* and *I(q)* are given. Can we conclude *I(r)*? Here we should be careful about the interpretation of the rules as $p \rightarrow_K q$ (q is a consequence of p), $p \Rightarrow_O q$ (given p, q is obligatory), $p \Rightarrow_I q$ (given p the agent has the intention q), and $p \Rightarrow_Z q$ (given p the agent sees to it that q).

The adoption of conversions should not sound strange. In many formalisms it is possible to convert from one type of conclusion into a different one. Take for example the right weakening rule of non-monotonic consequence relations, where $B \vdash C$ and $A \mid\sim B$ imply $A \mid\sim C$ (see, e.g., [15]). In other words, it allows the combination of non-monotonic consequence with classical consequences. While not every combination of obligations and mental attitudes or action concepts will produce meaningful results for the conversion, some of them can prove useful in the present context. For example if we want to convert rules for knowledge/belief into rules for obligations we have to determine conditions under which a rule for knowledge can be used to directly derive an obligation. The condition we have after is that all the antecedents on the rule can be shown to be obligatory. In general, when we admit conversion of rules, the situation is such that when given environmental conditions are satisfied a rule for X is transformed in a rule for Y; accordingly we have to use the "transformed" rule both in the support and attack phases. The conditions under which a rule can be converted are that all impersonal literals are (defeasibly) provable in K and all personal literals are (defeasibly) provable in the modalities required by the conversion (see Table 4). Formally we have

```
supported(R, X, P):- rule(R, Y, P, [A1,...,An]),
   environment(A1, W), ..., environment(An, W),
   not defeater(R, X, P, [A1,...,An]).

applicable(R, X, P):- rule(R, Y, P, [A1,...,An]),
   environment(A1, W), ..., environment(An, W),
```

where

```
environment(P, X):- personal(P), defeasibly(P, X).
environment(P, X):- not personal(P), defeasibly(P, K).
```

The relationships among the modalities X, Y and W are described in Table 4[2]. Notice that not all cases in the Table 4 can be accepted for all types of agents: the first column

[2] Table 4 should be read as follows. The first and second columns indicate the modal qualifications of the antecedents of a rule. The third column specifies the type of rule while the fourth provides the possible modal qualification we may obtain in the light of the antecedents and the rule type. Fifth column says whether the corresponding conversion holds in all cases or characterises only some particular agent types. For example (fourth row from the top), if $Zp, Iq \Rightarrow_K r$ is applicable we may obtain $+\partial_I r$. On the other hand, the derivation of $+\partial_I r$ from $Ip, Iq \Rightarrow_O r$ is possible only if we assume a kind of norm regimentation, with which we impose that all agents intend what is prescribed by deontic rules.

Table 4. Conversions

X	Y	$\Rightarrow$	W	
O	O	K	O	For all agents
I	I	K	I	For all agents
Z	Z	K	Z	For all agents
Z	I	K	I	For all agents
I	I	O	I	quasi-intentional-regimentation
Z	Z	O	I	quasi-intentional-regimentation
Z	Z	O	Z	quasi-behavioural-regimentation
Z	I	O	I	quasi-intentional-regimentation
I	I	Z	I	For all agents
O	O	Z	O	For all agents
I	O	Z	I	quasi-socio-intentional-regimentation

from the right indicates new types of agent corresponding to each rule conversion. This is particularly evident when obligations, actions and intentions are considered. Other combinations than those here defined are possible but they are problematic. As we can see, some of the conversions above logically characterise new types of agents. Let us focus on them. All these types correspond to weak versions of strong norm regimentation. Strong regimentation, as maintained in [6], corresponds to adopting schemata like $Op \rightarrow Ip$. The just mentioned conversions configure weak forms of regimentation. For instance, consider the conversion described in the fifth row from the top. Roughly speaking, if we want to give an intuitive reading we could conceive it as follows: Ip and $O(p \rightarrow q)$ entail Iq. Let us see with a concrete example the meaning of some conversions. The Yale Shooting Problem can be described as follows[3]

$$liveAmmo, load, shoot \Rightarrow_K kill$$

This rule encodes the knowledge of an agent that knows that loading the gun with live ammunitions, and then shooting will kill her friend. This example clearly shows that the qualification of the conclusions depends on the modalities relative to the individual acts "load" and "shoot". If the agent intends to load and shoot the gun $(I(load), I(shoot))$, then, since she knows that the consequence of these actions is the death of her friend, she intends to kill him $(+\partial_I kill)$. Similarly if she intentionally loads the gun and intends to shoot $(Z(load), I(shoot))$. To intentionally killing him she has to load and to shoot the gun intentionally $(Z(load), Z(shoot))$. If she intentionally loads the gun $(Z(load))$ and accidentally shoots it $(shoot)$, she kills the friend $(+\partial_K kill)$ but this is not an intentional act $(-\partial_Z kill)$, since not all the actions leading to this dramatic conclusion are intentional. Finally in the case she has the intention to load the gun $(+\partial_I load)$ and then for some reason shoot it $(shoot)$, then the friend is still alive $(-\partial_K kill)$.

So far we have only examined cases where we pass from a single modality to a different modality. However axioms $ZOp \rightarrow Ip$ and $IOp \rightarrow Ip$ provide modal reductions. These principles can be described in the meta-program by the following clause

[3] Here we will ignore all temporal aspects and we will assume that the sequence of actions is done in the correct order.

```
defeasibly(P, I):- defeasibly(OP, Z).
defeasibly(P, I):- defeasibly(OP, I).
```

Moreover we have to add clauses for applicable and supported where we consider rules for Z and I with conclusion Op when rules for I with conclusion p/~p are admissible.

4 Related and Future Work

Let us sketch just some short conclusions, also for future research.

Nute [21] proposed a Deontic Defeasible Logic which, in some respect, is similar to the framework presented here. Beside some minor differences in the way rules are handled at the propositional level, the main difference is that he uses only one type of rule. Traditionally, in proof-theory, rules to introduce operators give the meaning of them. Thus using one and the same type of rule both for obligation and factual conclusion does not show the real meaning of the operators involved. Moreover it is not clear to us whether and how complex conversions and reductions can be dealt with in a system with only a single type of rules.

As we said, another reference of this paper is to BOID. Its calculation scheme is similar to the one proposed here. For example, as in BOID it is possible to state general orders of overruling but also local preferences involving single rules. This last job is made here by means of the superiority relation. However, our system, which also deals with agency, is designed to take care of modalised literals and modal conversions. This is due to the logical task assigned to the rules. For this reason, but in a different perspective, our logical view may be also useful to study the notion of negative permission. In fact, conditions for $\partial_{O}p$ may also determine the implicit introduction of a modal operator of permission in terms of non-derivability of an obligation.

As regards the complexity of the system, [16] has proved, for the propositional case, that the set of tagged literals can be derived from the theory in linear time in the number of rules in it. It is not hard to extend this result to the modal case. The distinction of different kinds of rules does not affect the complexity of the theory. The case for K is the same adopted in standard Defeasible Logic while, for the other components, we convert relevant rules into the appropriate"extended" modal literals. At this point, the inference mechanism is the same as the standard one.

Due to space limitations it was not possible to show how to model other notions of agency—such as capability (both practical [9] and deontic [12]), attempt, and so on—that have received some attention in the literature in the past few years. Here it suffices to say that those notions can be easily represented (modularly) by adopting a strategy similar to that used in [11] to derive goals from intentions in a BDI defeasible logic.

References

1. Antoniou, G., D. Billington, G. Governatori, and M. J. Maher. A flexible framework for defeasible logics. In *AAAI-2000.* AAAI/MIT Press, 2000.
2. Bratman, M., D. Israel, and M. Pollack. Plans and resource-bounded practical reasoning. *Computational Intelligence,* 4, 1988.

3. Bratman, M. E. *Intentions, Plans and Practical Reason.* Harvard University Press, 1987.
4. Broersen, J., M. Dastani, J. Hulstijn, Z. Huang, and L. van der Torre. The BOID architecture. *In Agents-01.* 2001.
5. Broersen, J., M. Dastani, and L. van der Torre. Resolving Conflicts between Beliefs, Obligations, Intentions, and Desires. In *ECSQARU 2001,* Benferhat, S. and P. Besnard, eds. Springer, 2001.
6. Broersen, J., M. Dastani, and L. van der Torre. $\mathbf{BDIO}_{CTL}$: Obligations and the specification of agent behavior. In *IJCAI-03.* 2003.
7. Carmo, J. and A. J. Jones. Deontic logic and contrary-to-duties. In *Handbook of Philosophical Logic (2nd edition), vol. 8,* Gabbay, D. and F. Guentner, eds. Kluwer, 2000.
8. Dastani, M. and L. van der Torre. A classification of cognitive agents. In *Cogsci'02.* 2002.
9. Elgesem, D. The modal logic of agency. *Nordic Journal of Philosophical Logic,* 2, 1997.
10. Gelati, J., G. Governatori, A. Rotolo, and G. Sartor. Declarative power, representation, and mandate: A formal analysis. In *JURIX02,* Bench-Capon, T., A. Deskalopulu, and R. Winkels, eds. IOS Press, 2002.
11. Governatori, G. and V. Padmanabhan. A defeasible logic of policy-based intention. In *AI 2003: Advances in Artificial Intelligence,* Gedeon, T. D. and L. C. C. Fung, eds., vol. 2903 *of LNAI.* Springer, 2003.
12. Governatori, G. and A. Rotolo. A defeasible logic of institutional agency. In *NRAC'03,* Brewka, G. and P. Peppas, eds. 2003.
13. Halpern, J. Y and Y. Moses. A guide to completeness and complexity for modal logic of knowledge and belief. *Artificial Intelligence,* 54, 1992.
14. Hilpinen, R. On action and agency. In *Logic, Action and Cognition: Essays in Philosophical Logic,* Ejerhed, E. and S. Lindström, eds. Kluwer, 1997.
15. Kraus, S., D. Lehmann, and M. Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence,* 44, 1990.
16. Maher, M. Propositional defeasible logic has linear complexity. *Theory and Practice of Logic Programming,* (6), 2001.
17. Maher, M. J. and G. Governatori. A semantic decomposition of defeasible logic. In *AAAI-99.* AAAI Press, 1999.
18. Maher, M. J., A. Rock, G. Antoniou, D. Billignton, and T. Miller. Efficient defeasible reasoning systems. *International Journal of Artificial Intelligence Tools,* (4), 2001.
19. Meyer, J.-J. C. and W. van der Hoek. *Epistemic Logic for AI and Computer Science.* Cambridge University Press, 1995.
20. Nute, D. Defeasible logic. In *Handbook of Logic in Artificial Intelligence and Logic Programming,* vol. 3. Oxford University Press, 1987.
21. Nute, D. Norms, priorities, and defeasible logic. In *Norms, Logics and Information Systems,* McNamara, P. and H. Prakken, eds. IOS Press, 1998.
22. Pitt, J. (ed.). *Open Agent Societies.* Wiley, 2004.
23. Rao, A. and M. Georgeff. Modelling rational agents within a BDI-architecture. In *KR'91,* Fikes, A. J. and R. E. Sandewall, eds. Morgan Kaufmann, 1991.
24. Rao, A. and M. Georgeff. BDI agents: From theory to practice. In *ICMAS'95.* 1995.
25. Santos, F. and J. Carmo. Indirect action: Influence and responsibility. In *Deontic Logic, Agency and Normative Systems,* Brown, M. and J. Carmo, eds. Springer, 1996.
26. Sergot, M. and F. Richards. On the representation of action and agency in the theory of normative positions. *Fundamenta Informaticae,* 48, 2001.
27. Thomason, R. H. Desires and defaults: A framework for planning with inferred goals. In *KR2000,* Cohn, A. G., F. Giunchiglia, and B. Selman, eds. Morgan Kaufmann, 2000.
28. van der Torre, L. and Y. Tan. The many faces of defeasibility. In *Defeasible Deontic Logic,* Nute, D., ed. Kluwer, 1997.

Collective Obligations and Agents: Who Gets the Blame?

Davide Grossi[1], Frank Dignum[1],
Lambèr M.M. Royakkers[2], and John-Jules Ch. Meyer[1]

[1] Utrecht University,
The Netherlands
{davide,dignum,jj}@cs.uu.nl
[2] Eindhoven University of Technology,
The Netherlands
L.M.M.Royakkers@tm.tue.nl

Abstract. This work addresses the issue of obligations directed to groups of agents. Our main concern consists in providing a formal analysis of the structure connecting collective obligations to individual ones: which individual agent in a group should be held responsible if an obligation directed to the whole group is not fulfilled? To this aim, concepts from planning literature (like plan and task allocation) are first used in order to conceptualize collective agency, and then formalized by means of a dynamic deontic logic framework. Within this setting, a formal account of the notion of coordination, intended as management of interdependencies among agents' activities, is also provided.

1 Introduction

In multi-agent systems, the cooperation between agents is an important issue. For that purpose several theories have been developed about joint goals, plans and intentions. All these joint attitudes try to capture some of the team elements of agents that work together. In this paper we follow up on that work and will look at collective obligations. What are the consequences for an agent if the group in which an agent performs its tasks gets an obligation to fulfill a certain goal? E.g. a program committee of DEON'04 may have the collective obligation to review all submitted papers before a certain time. We are interested to explore how this collective obligation translates into individual obligations for the program committee members, e.g. review two or three papers, and the extra obligations for the program chair to divide the papers and monitor the process and make final decisions.

In [12], the collective obligation is formalized in a framework of deontic logic, which gives the opportunity to express which agent (or which group of agents) has the responsibility to bring about a certain situation (to express group liability, e.g. liability for a trading partnership) and to express the relation between the agents of a group. However, in the theory developed in [12] we cannot express the individual responsibility that follows from the task to achieve the fulfillment of the collective obligation. A consequence is that we cannot indicate which individual is responsible for a violation of a collective obligation.

A. Lomuscio and D. Nute (Eds.): DEON 2004, LNAI 3065, pp. 129–145, 2004.

We believe that the main differences between individual and collective obligations can be explained through the distribution of *responsibility* and *knowledge*.

If an individual has an obligation he has to perform all tasks for the fulfillment of the obligation (including planning the tasks) himself. Therefore whenever the obligation is not fulfilled the cause is that the individual did not perform a task that should be done to fulfill the obligation. It follows that, whatever the actual task was that was not performed, the individual is responsible. This differs fundamentally from the situation where a group has to fulfill the obligation. In this case the tasks are distributed over the group. Each individual can independently (autonomously) decide to perform his task or not. The responsibility therefore also is distributed over the members of the group. If the group does not fulfill the obligation the individual that did not perform his task will be held responsible! So, it becomes important to check exactly which are the obligations for each member of the group that follow from the collective obligation of that group.

The second aspect that is very different between collective and individual obligation is the distribution of knowledge. An individual can decide for himself whether to fulfill an obligation or not, based on its utility, beliefs, goals etc. The individual also knows the complete plan, whether he is capable to perform the actions in the plan, when he has performed actions or decided not to perform them. All this information is readily available and can be reasoned about. However, when a collective obligation is divided over the individuals of that collective, they might not know the whole plan, typically do not have information about actions that are performed, etc.

We make the important assumption that an individual can only be responsible for violating an obligation if he knows (or could have known) that he has the obligation. If a group has a collective obligation, but a member of the group does not know what are the consequences of that obligation for himself, he cannot be held responsible for violating that part of the obligation (unless he knows he has the obligation to find out what he has to do, but never bothered to do it). This assumption has many consequences for the desired dissemination of knowledge through the group about the plan, plan allocation and current execution of the plan, etc. In the ideal case all members of the group have complete knowledge about all these aspects, such that no member can avoid his responsibility based on a lack of knowledge. It leads to the introduction of explicit coordination actions in our model to incorporate this aspect of the collective obligation. We will discuss these in detail in Section 2.

Coordination actions are actually only one type of meta actions that should be considered. Besides the plan to achieve the content of the obligation the group should create that plan, allocate agents to parts of the plan, create a plan for what to do when the original plan fails, etc. These meta actions should also be coordinated again creating in the end an infinite regression of meta actions. In this paper we will not take all these layers into account, but will limit us to the coordination actions that are necessary during the execution of a plan to fulfill the obligation.

In the next section we will discuss the notions of collective agency, plans, task allocations and coordination of tasks. This will indicate which concepts

are needed in the formal framework to describe collective obligations and their consequences. The formal framework is described in Section 3. In Section 4, we show how the formal framework can be used to model an example and how some consequences of collective obligations can be derived depending on characteristics of the obligated task, structure of the group and their knowledge. In Section 5 we draw some conclusions and give directions for future research.

2 Conceptualizing Collective Agency: Plans, Allocation, Coordination

In this section we discuss two basic issues that underlie the analysis of the consequences of a collective obligation for the individuals of the group. First we discuss the notion of a (distributed) plan to achieve the collective obligation. After that we discuss some of the coordination issues around these distributed plans. During the discussion we will make use of the notation of the formal framework that is fully described in Section 3. The intuitions and properties described here do not depend on this formalism though!

2.1 Concepts of Plan and Task Allocation

To fulfill a collective obligation $O(X : \gamma)$, the group X has to perform a complex action γ [1]. The concrete manner to deal with such a complex action is planning. The group has to decompose the complex action into a number of individual sub-actions. For example, if the program committee is obliged to notify the authors of the submitted papers of acceptance before a certain deadline, this obligation can only be fulfilled if the work that has to be done, is shared out in several tasks over the members of the program committee. Therefore, the group needs a plan: a concrete manner to achieve the task of the group.

We can define a plan to perform the complex action γ as a decomposition of the complex action γ by a sequence of (possible simultaneous) individual actions:

$$Plan(\gamma) = \langle \alpha_1 \bullet \alpha_2 \bullet \ldots \bullet \alpha_n \rangle \text{ such that } \gamma = \alpha_1 \bullet \alpha_2 \bullet \ldots \bullet \alpha_n,$$

where $\bullet$ stands for the simultaneous operator '&:' or the sequential operator ';'. The action $\alpha_1 \& \alpha_2$ stands for the simultaneous performance of α_1 and α_2, and action $\alpha_1; \alpha_2$ stands for the sequential composition of α_1 and α_2.

We need the simultaneous operator, since some actions have to be performed at the same time. The sequential operator is needed because some actions might depend on other ones: a certain action can only be performed if an other action is done. So, the plan must at least determine the *order* of sub-actions. For example, the notification of acceptance of a certain paper can only be done if it is reviewed by the delegated members of the program committee. The responsibility of the performance of an action α by an agent depends not only on the individual who is committed to perform the action α, but also on agents who have to perform actions which are necessary to perform action α.

[1] Such a complex action γ may be seen as equivalent to an action of the type $achieve(\tau)$, where τ is a state of affairs.

The complex action γ does not contain a non-deterministic choice, because we only use totally ordered plans [15]. However, there can be several totally ordered plans to execute a certain action γ, and it therefore makes sense to talk about a choice of plans. In this paper we restrict ourselves to *acceptable* plans, which are plans containing individual actions that can be performed by an agent in the group. In other words, for every individual action α there is an agent of the group which is *capable* of performing α. Note that there are better and worse plans. E.g. a plan that consists of individual actions that only can be performed by one agent seems worse than a plan that consists of individual actions that can be performed by several agents. Since there are several possible ways to divide the task, the group has to decide which plan should be followed.

Besides task division (the decomposition of a complex actions γ into individual sub-actions $\alpha_1, \ldots, \alpha_n$), task allocation is needed, which indicates which (capable) agent of the group has to perform which sub-action of the complex action. We use the following definition for (partial) task allocation:

Definition 1. (Partial task allocation) *A partial task allocation for a task γ within a group of agents X is defined as follows:*

$$< I_1 : \alpha_1 \bullet I_2 : \alpha_2 \bullet \ldots \bullet I_n : \alpha_n > \text{ such that}$$

$$\gamma = \alpha_1 \bullet \alpha_2 \bullet \ldots \bullet \alpha_n, \text{ and } I_j \in \mathcal{P}^+(X) \text{ for } j = 1, 2, \ldots, n.$$

We refer to the set of all partial task allocations of γ within X as $PPlan(X : \gamma)$

A partial task allocation is thus a composition of constructs of the type "group I performs the (possibly complex) action α" $(I : \alpha)$. These constructs are called *(collective) events* ([12]).

Definition 2. (Complete task allocation) *CP is a complete task allocation for a task γ to a group X, iff $CP \in PPlan(X : \gamma)$ and for all I_j occurring in CP holds that $I_j = \{a_j\}$* [2].

E.g. let $CP \in PPlan(\{a_1, a_2, a_3\}, \gamma)$ and $CP = \langle a_1 : \alpha_1 \ ; \ a_2 : \alpha_2 \ \& \ a_3 : \alpha_3 \rangle$. In CP agent a_2 is responsible for the performance of action α_2, but according to the plan he can only perform his action after agent a_1 has done action α_1. The same holds for agent a_3.

The collective obligation $O(X : \gamma)$ can be translated by the group X in internal obligations of the group given a plan. E.g. given that we have $O(\{a_1, a_2, a_3\}, \gamma)$ and plan CP above. Then we would at least have that:

$$O(a_1 : \alpha_1) \wedge [a_1 : \alpha_1](O(a_2 : \alpha_2) \wedge O(a_3 : \alpha_3))$$

Which states that first a_1 is obliged to perform α_1 and if a_1 has performed α_1 ($[a_1 : \alpha_1]$) then a_2 and a_3 are obliged to perform α_2 and α_3 respectively. In general, we have that given a complex collective obligation, a disjunction of all internal obligations, which are determined by the possible acceptable plans,

[2] In what follows, in order to keep formal expressions more readable, we will often refer to singletons omitting the {.} standard notation.

can be derived. The fulfillment of one of these internal obligations leads to the fulfillment of the collective obligation. Violation of the internal obligation leads, typically, to a decision to make another plan establishing an alternative internal obligation, or to the violation of the collective obligation itself.

The above indicates how a plan can be used to distribute a collective obligation over the members of a group. However, it also shows that the obligations of at least some members (in the example above a_2 and a_3) depend on the performance of an action of another member of the group (in the example above a_1). This observation leads us to the issue of coordination between the actions in a plan.

2.2 The Problem of Coordination

The view on coordination, which is assumed here, can be perfectly summarized by the following quotation from [5]:

> "coordination is the process of managing interdependencies between activities".

Plans and task allocations, as they have been defined in Section 2.1, are essentially sequences of, respectively, simpler actions and events. This sequential structure is precisely what should be managed by means of coordination. Let us reconsider our example about the group organizing the DEON'04 workshop: for each submitted paper a reviewing process takes place, which ends, in some cases, with a notification of acceptance. The first question coming naturally about is: when should the agent responsible for notifying acceptances start his activity? I.e. how does he *know* that the condition for his obligation to send the notifications is true? In the introduction we mentioned that we only consider an agent responsible for violating an obligation when he knows he has the obligation. We assume that if an agent knows the plan to fulfill the obligation and knows that the actions he depends on are performed he will also know that he has an obligation to perform his own part. Using the example and the notation used before plus B_i to denote the beliefs of agent a_i and $DONE(a : \alpha)$ to indicate that agent a has performed α we get:

$$B_2([a_1 : \alpha_1](O(a_2 : \alpha_2)) \wedge B_2(DONE(a_1 : \alpha_1)) \rightarrow B_2(O(a_2 : \alpha_2)) \qquad (1)$$

This suggests that a_2 should somehow come to believe that a_1 performed its task[3]. We do assume that an agent believes it performed a task whenever it did so (internal monitoring or conscious behavior) but this leads only to:

$$[a_i : \alpha](DONE(a_i : \alpha) \wedge B_i DONE(a_i : \alpha))$$

[3] Note, in passing, that beliefs update issues might come about in relation with formula 1. For example: how does the set of beliefs of an agent evolve? We leave these issues aside focusing exclusively on some formal aspects concerning the interaction between beliefs and obligations. This makes sense especially because beliefs will be considered essentially as the effect of coordination activities.

and not to:

$$[a_i : \alpha](DONE(a_i : \alpha) \wedge B_j DONE(a_i : \alpha))$$

for any $j \neq i$.

The latter can only be achieved through the introduction of explicit coordination actions. This constraint determines the necessity for agents to be informed *at least* about what the previous agent did. And such a necessity impinges on the task decomposition itself forcing it to take into consideration how to provide agents with the necessary information to act according to the established plan. A task allocation $a_1 : \alpha \ ; \ a_2 : \beta$ for the complex action γ for the group X, once taken into account this type of coordination problem concerning informational issues, becomes something of this kind:

$$a_1 : \alpha \ ; \ a_i : \text{coordinate}(a_1, \alpha, Y) \ ; \ a_2 : \beta$$

where $a_i \in X$ and $Y \subseteq X$ with $Y \neq \emptyset$. The *coordinate* action is here understood as an action after each execution of which a belief holds in each agent belonging to Y which concerns an activity executed by the previous agent. Notice that Y should always be such that it contains at least the agent responsible for the task that follows within the task allocation sequence: so if a_{j-1} is the preceding agent, Y should be such that $a_j \in Y$. Notice moreover how the coordination action assumes different intuitive meanings depending on the agent appointed to its performance. In the example above we have that possibly $a_i = a_1$ or $a_i = a_2$. In the first case the *coordinate* action becomes a kind of *informative,* the agent itself being the one providing information about its performance:

$$[a_{j-1} : \alpha][a_{j-1} : \text{coordinate}(a_{j-1}, \alpha, Y)] \bigwedge_{a_k \in Y} B_k DONE(a_{j-1} : \alpha)$$

and $\exists a_k \in Y$ s.t. $a_k = a_j$. In the second case the *coordinate* action turns out to correspond to a sort of *checking,* since it is the agent himself which acquires the necessary information before acting:

$$[a_{j-1} : \alpha][a_j : \text{coordinate}(a_{j-1}, \alpha, Y)] \bigwedge_{a_k \in Y} B_k DONE(a_{j-1} : \alpha)$$

and $\exists a_k \in Y$ s.t. $a_k = a_j$. Notice, in the end, that even a "third party" might be responsible for this coordination task.

Due to the introspection principle for agents concerning their actions the coordination action between two actions of the same agent becomes superfluous and takes the form of a so-called *skip* action:

$$a_i : \alpha \ ; \ a_i : \text{coordinate}(a_i, \alpha, \{a_i\}) \equiv a_i : \alpha \ ; \ a_i : skip.$$

The property expressed in formula (1) relates the beliefs of an agent about the plan and the directly preceding actions to his own obligations. One might also be interested in the fact that an agent that is obliged to perform an action on which your actions depends does not perform his action. In the above example, in which $Y = \{a_2, a_3\}$, we could have:

$$O(a_1 : \alpha_1) \wedge [\overline{a_1 : \alpha_1}][a_j : \text{coordinate}(a_1, \overline{\alpha_1}, \{a_2, a_3\})]$$

$$(B_2(\neg DONE(a_1 : \alpha_1)) \wedge B_3(\neg DONE(a_1 : \alpha_1)).$$

where $\overline{a_1 : \alpha_1}$ denotes the negation of an event. Adding this feature enables the coordination task to take care that both in the case when an action is performed as scheduled as well as when an action is not performed as scheduled the dependent agents get to know about it. Further coordination mechanisms are usually devised which remedy the violation of an obligation by one of the agents ([16]). We will not get into this aspects in this paper yet.

In the next section we will describe the formal framework in which the discussions above can be expressed formally and which provides the right validities to indeed prove the properties that we would like to have in order to distribute the collective obligations over the individuals of a group.

3 A Formal Framework

Our aim is to further pursue a line of analysis, which has been already proposed in [13,12], trying to provide a formal framework in which to account for the problems of collective obligation and collective agency in terms of the concepts introduced in the previous section. Some attempts to formally address these problems have been already proposed ([1,3,4]), but none of the frameworks presented so far possesses the necessary expressive power for dealing with concepts such as plan and task allocation. One of the main reasons is that they are basically grounded on *stit* or *bringing-it-about* action logics, which cannot cope with any specification of the internal structure of plans.

More technically, this framework should handle event expressions, that is expressions about the performance of some action by some agents, and their composition, doxastic expressions, deontic expressions concerning event expressions, predicates on events. The framework is obtained expanding the proposal contained in [12] in several directions: first of all adding transactions and negations on transactions, following the line described in [6]; then adding doxastic logic; and finally adding a type of dynamic assertions, by means of the *DONE* operator, in order to express, in analogy with [7,8], performances actually (and not just possibly) taking place in a backward direction.

3.1 Language

The alphabet consists of a set P of propositional symbols (p), the propositional constant V, the operator *DONE,* a set of agent identifiers I (groups of agents identifiers are denoted by *X, Y,...*), a set A of atomic action symbols typically denoted by $\underline{a}$ (this set at least includes the coordination actions $coordinate(a_i, \alpha_i, Y)$ with $a_i \in I$), the doxastic operator B, and the dynamic operators [ ] and $\langle\ \rangle$. The language $\mathcal{L}$ is based on three types of syntactic constructs that we are now going to define.

The set *Act* of action expressions (α) is defined through the following BNF:

$$\alpha ::= \underline{a} \mid skip \mid \overline{\alpha} \mid \alpha_1 + \alpha_2 \mid \alpha_1 \& \alpha_2 \mid \alpha_1; \alpha_2.$$

where *skip* represents a "doing nothing"action, $\bar{}$ stands for the negation operator, + stands for the indeterministic choice operator, & for the parallel performance operator and ; for the sequencing operator. The set *Evt* of event expressions (ξ) is defined through the following BNF:

$$\xi ::= X : \alpha \mid \overline{X : \alpha} \mid \xi_1 + \xi_2 \mid \xi_1 \& \xi_2 \mid \xi_1 ; \xi_2.$$

Notice that the same notation for actions and event operators (negation, +, &,;) is used. It is nevertheless obvious that they belong to different categories of operators. We chose, however, for keeping notation not too rich.

The set *Ass* of assertions (ϕ) is defined through the following BNF:

$$\phi ::= p \mid V \mid DONE(\xi) \mid \neg\phi \mid \phi_1 \vee \phi_2 \mid \phi_1 \wedge \phi_2 \mid \phi_1 \rightarrow \phi_2 \mid [\xi]\phi \mid B_i\phi.$$

3.2 Models

In order to give a semantics to the language introduced above we start defining the notion of model for $\mathcal{L}$.

Definition 3. (Models) *A model M is defined as follows:*

$$M = \langle \mathcal{P}^+(\mathbb{I}), \mathbb{A} \cup \text{skip}, \mathbb{W}, [\![\ ]\!]_R, \{\mathbb{R}_i\}_{i\in\mathbb{I}}, \prec, \pi \rangle$$

where:

- $\mathcal{P}^+(\mathbb{I})$ *is the non-empty powerset of the finite set of actors* $\mathbb{I}$*, that means the possible groups of actors. We assume* $\mathbb{I} = I$.
- $\mathbb{A} \cup \text{skip}$ *is the set of actions,*
- $\mathbb{W}$ *is the set of possible states.*
- $[\![\ ]\!]_R$ *is a function* f *s.t.* $f : Evt \times \mathbb{W} \longrightarrow \mathcal{P}(\mathbb{W})$*, to each event expression-world couple it associates the set of states to which the performance of that event in that world leads. It consists of a composition of the two functions* $[\![\ ]\!]$ *and R which will be introduced in Section 3.3.*
- $\{\mathbb{R}_i\}_{i\in\mathbb{I}}$ *is a family of serial symmetric and transitive accessibility relations which are indexed by actors indicating the believable worlds of agent* a_i [4].
- $\prec$ is *a partial ordering on* $\mathbb{W}$ *denoting the order in which worlds are reached through actual performances of events. This ordering is constrained as follows: if* $w_1 \prec w_2$ *and* $\exists w_3$ *s.t.* $w_3\mathbb{R}_i w_1$ *or* $w_3\mathbb{R}_i w_2$ *then* $w_3\mathbb{R}_i w_2$ *and* $w_3\mathbb{R}_i w_1$ [5]*. From an intuitive point of view, this condition guarantees the whole path of actual performances through* $\mathbb{W}$ *to be doxastically accessible,*
- π *is a usual truth function* f *s.t.* $f : Ass \times W \longrightarrow \{1, 0\}$.

Like in [9,6,7] our semantics consists of two parts: first event expressions are interpreted as set theoretic constructs on $\mathbb{A}$ where events get a so-called *open* interpretation; successively event expressions are interpreted as state-transition functions determining the accessibility relation $[\![\ ]\!]_R$ on $\mathbb{W}$.

[4] The doxastic part of the framework will be of interest in particular for the modelling of an example in Section 4.

[5] This condition is important for proving validity (10) in Proposition 1.

3.3 Synchronicity Sets, Steps, Synchronicity Traces, and Worlds

The interpretation of events is based on the basic notion of *Synchronicity set* (s-set).

Definition 4. (s-set) *The set $\mathcal{S}$ of s-sets is defined as follows:* $\mathcal{S} = \mathcal{P}^+(\mathbb{I}) \times \{\text{skip}\} \cup \mathcal{P}^+(\mathbb{I}) \times \mathcal{P}^+(\mathbb{A})$.

Synchronicity sets, that is elements of $\mathcal{S}$, are denoted by $S_1, S_2, \ldots$. Informally, a s-set is nothing but a set of parallel executions of events by a group of agents, and formalizes the aforementioned open interpretation view on events. Based on the notion of s-set we define the notion of *step*[6].

Definition 5. (Step) *The set Step of steps is defined as follows:*

$$Step = \{\times_{X \in \mathcal{P}^+(\mathbb{I})} \mathcal{S}_X \mid \forall X, Y \in \mathcal{P}^+(\mathbb{I}) : Y \subseteq X \Rightarrow act(S_Y) \subseteq act(S_X) \;\&$$
$$\forall X, Y \in \mathcal{P}^+(\mathbb{I}) : act(S_Y) = \text{skip} \Rightarrow act(S_{X \cup Y}) = act(S_X)\}$$

where act is a function that extracts the action component from a given s-set $(act(X : \{a_1, a_2\}) = \{a_1, a_2\})$.

Steps represent a sort of snapshot of the activity of each subgroup of $\mathbb{I}$ at a certain moment, depicting how all agents move one "step" ahead. Steps are therefore sets of s-sets of cardinality $2^n - 1$ where n is the number of agents in $\mathbb{I}$. They are constrained in such a way that whatever action is performed by a subgroup is also performed by a supergroup, and subgroups remaining inactive are treated as performing a skip action. Steps, that is elements of *Step,* are denoted by $s_1, s_2, \ldots$.

In order to provide a semantics for sequential expressions the concept of *Synchronicity trace* (s-trace) is needed. Notice that this concept uses steps instead of s-sets like in [9,6].

Definition 6. (s-trace) *The set $\mathcal{T}$ of s-traces is defined as follows:*

$$\mathcal{T} = \{\langle s_1, \ldots, s_n, \ldots\rangle \mid s_1, \ldots, s_n, \ldots \in \mathcal{S}\}.$$

The length of an s-trace t is denoted by $dur(t)$. We assume $dur(t)$ to be finite.

An event will be interpreted as a set of s-traces. The range for our interpretation of events is a set $\mathcal{E}$ such that $\mathcal{E} = \mathcal{P}(\mathcal{T})$. Elements of $\mathcal{E}$ (sets of s-traces) are denoted as $T_1, T_2, \ldots$ The length *dur(T)* of a set T is defined as $max\{dur(t) | t \in T\}$.

We can now introduce the operations that constitute the semantic counterpart of our syntactic operators.

Definition 7. (Operations on events) *Let $T_1, T_2 \in \mathcal{T}$:*

$$T_1 \circ T_2 = \{t_1 \circ t_2 \mid t_1 \in T_1, t_2 \in T_2\}$$

[6] Notice that in [7] s-sets axe called steps, and no notion of step as it will be defined in this work occurs there.

$$T_1 \Cap T_2 = \bigcup\{t_1 \Cap t_2 \mid t_1 \in T_1, t_2 \in T_2\}$$

$$T_1 \Cup T_2 = T_1 \cup T_2 - (\bigcup\{t_1 \Cap t_2 \mid t_1 \in T_1, t_2 \in T_2 \ \ and \ \ t_1 \neq t_2\})$$

$$\tilde{T} = \begin{cases} if \ \ T \neq \emptyset, \tilde{T} = \Cap\{\tilde{s} \mid s \in T\} \\ if \ \ T = \emptyset, \tilde{T} = Step \end{cases}$$

Where:

- $t_1 \circ t_2$ *is defined as follows: if* $t_1 = \langle s_1, ..., s_n \rangle$ *and* $t_2 = \langle s'_1, ..., s'_m \rangle$ *then,* $t_1 \circ t_2 = \langle s_1, ..., s_n, s'_1, ..., s'_m \rangle$.
- $t_1 \Cap t_2$ *is defined as follows:* $t_1 \Cap t_2 = \begin{cases} t_1 & if \ \ t_2 \in start(t_1) \\ t_2 & if \ \ t_1 \in start(t_2) \\ \emptyset & otherwise \end{cases}$
  where start is a function which associates to a given s-trace all its starting possible s-traces: $start(t) = \{t' \mid t' = t \ or \ \exists t'' \neq \emptyset \ s.t. \ t' \circ t'' = t\}$.
- $\tilde{t}$ *is defined as follows:* $\tilde{t} = \bigcup_{1 \leq n \leq dur(t)} \langle s_1, ..., \tilde{s_n} \rangle$, *where* $\tilde{s} = Step - \{s\}$ [7].

Intuitively, we want $\Cup$ to yield the property: $a \equiv a + a;b$ for event expressions. In order to establish this property we cannot just use a union of the sets of s-traces representing a and $a;b$ but have to do some "cleaning up" by subtracting superfluous parts.

The semantics of events are obtained by means of a function $[\![\ ]\!] : Evt \longrightarrow \mathcal{E}$ such that:

Definition 8. (Semantics of events)

$$\begin{aligned} [\![X : \underline{a}\ ]\!] &= \{s \mid S_X \in s, a \in act(S_X)\} \\ [\![\xi_1;\xi_2\ ]\!] &= [\![\xi_1\ ]\!] \circ [\![\xi_2\ ]\!] \\ [\![\xi_1 + \xi_2\ ]\!] &= [\![\xi_1\ ]\!] \Cup [\![\xi_2\ ]\!] \\ [\![\xi_1 \& \xi_2\ ]\!] &= [\![\xi_1\ ]\!] \Cap [\![\xi_2\ ]\!] \\ [\![\overline{\xi}\ ]\!] &= [\![\tilde{\xi}\ ]\!] \\ [\![skip\ ]\!] &= \{\text{skip}\}. \end{aligned}$$

The basic clause stipulates that the meaning of an atomic event consists of the set of steps where that action at least is performed by that specific group of agents.

On the basis of this evaluation for events, an evaluation of groups performing complex actions is obtained:

Definition 9. (Semantics of collective actions)

$$\begin{aligned} [\![X : \alpha_1;\alpha_2\ ]\!] &= [\![X : \alpha_1\ ]\!] \circ [\![X : \alpha_2\ ]\!] \\ [\![X : \alpha_1 + \alpha_2\ ]\!] &= [\![X : \alpha_1\ ]\!] \Cup [\![X : \alpha_2\ ]\!] \\ [\![X : \alpha_1 \& \alpha_2\ ]\!] &= [\![X : \alpha_1\ ]\!] \Cap [\![X : \alpha_2\ ]\!] \\ [\![X : \overline{\alpha'}\ ]\!] &= [\![\overline{X : \alpha'}\ ]\!]. \end{aligned}$$

[7] Negation of sequences constitutes a delicate matter. For a deeper discussion of this issue we refer to [6].

To connect this interpretation of events to a possible world semantics a function $R : \mathcal{E} \times \mathbb{W} \longrightarrow \mathbb{W}$ is defined, which couples events with state-transitions.

Definition 10. (Function R)
$R(T, w_1) = \{w_2 \mid \exists t \in T \ \ s.t. \ \ w_2 = R(t, w_1)\}$ *where* R *on transitions is inductively defined as follows:*

$$R(s_1, w_1) = r(s_1, w_1)$$
$$R(t_1 \circ t_2, w_1) = R(t_2, R(t_1, w_1)).$$

and $r : \mathcal{S} \times \mathbb{W} \longrightarrow \mathbb{W}$*, that is a function that, given a state, returns the following state reachable through a given synchronicity set, and such that* $r(\{\text{skip}\}, w) = w$.

3.4 Evaluating Formulas

The meaning of formulas ϕ in a world w, given the structure M, is defined as usual. For space reasons we report here only clauses for dynamic operators, and the *DONE* unary operator.

Definition 11. (Semantics of assertions) *In the following let* $dur([\![\xi_1]\!]) = 1$,

$M, w_1 \models [\xi]\phi$ *iff* $\forall w_2 \in [\![\xi]\!]_R(w_1), M, w_2 \models \phi$
$M, w_1 \models \langle\xi\rangle\phi$ *iff* $\exists w_2 \in [\![\xi]\!]_R(w_1), M, w_2 \models \phi$
$M, w_1 \models DONE(\xi_1)$ *iff* $\forall w_2 \in \mathbb{W}, \ w_2 \prec w_1 \Rightarrow w_1 \in [\![\xi_1]\!]_R w_2$;
$M, w_1 \models DONE(\xi; \xi_1)$ *iff* $\forall w_2 \in \mathbb{W}, \ w_2 \prec w_1 \Rightarrow M, w_1 \models DONE(\xi_1)$ *and* $M, w_2 \models DONE(\xi)$.

Informally, a sentence $[\xi]\phi$ ($\langle\xi\rangle\phi$) is true in w iff ϕ is true in every world (respectively in at least one world) accessible through a performance of ξ. As to the semantics of $DONE(\xi)$, the two clauses should be read as a basis and an induction step: intuitively, a sentence $DONE(\xi)$ is evaluated as true in a world w_1 iff that world can be reached via a sequence of events of length one from all the worlds w_2 which are connected with w_1 along the $\prec$ ordering.

We do not provide a separate semantics for the coordination actions. Instead we give the following constraint on the possible models which indicates that the effect of a coordination action is a certain type of belief in the recipients of that action.

Definition 12. (Coordination action) *Let* $a_j \in \mathbb{I}$ *with* $1 \leq j \leq n$ *and* $n = ||\mathbb{I}||$*, then:*

$$M, w_1 \models DONE(\{a_1\} : \alpha_1) \ \Rightarrow \ \forall w_2 \in [\![\{a_i\} : coordinate(a_1, \alpha_1, Y)]\!]_R(w_1) : M, w_2 \models \bigwedge_{a_j \in Y} B_j(DONE(a_1 : \alpha_1))$$

We deem worth stressing that the constraint above determines only a kind of minimal characterization of a notion of *coordination,* which is based on the informal discussion of Section 2.2.

The deontic notions, which range over events, are defined according to the following reduction principles:

Definition 13. (Deontic notions)

$$F(\xi) \equiv [\xi]V;$$
$$O(\xi) \equiv [\bar{\xi}]V;$$
$$P(\xi) \equiv \neg[\xi]V.$$

For an extensive account of how these notions are related we refer to [12]. In what follows our attention is exclusively focused on expressions concerning obligations.

3.5 Some Relevant Validities

We will now focus on some validities that are relevant in order to model the collective obligations appropriately. The validities listed below are of three kinds: validities (2)-(9) show how obligations propagate within the group, that is how the sub/supergroup relation is connected with the enactment of a group obligation and in particular how the obligation on a plan distributes over the obligations on the single components of that plan; second, validity (10) has to do with the beliefs of each individual agent about the task allocation and related obligations addressed to him; third, validities (11)-(13) are of a more general kind, and express some very basic features of the framework.

Proposition 1. (Validities) *Let* $X, Y \in \mathcal{P}^+(\mathbb{I})$ *and* $\alpha_1, \alpha_2, \gamma_1, \gamma_2 \in Act$ *and* γ_1, γ_2 *do not contain any occurrence of the action negation symbol:*

$$\models O(X \cup Y : \overline{\gamma_1}) \rightarrow O(X : \overline{\gamma}) \tag{2}$$
$$\models O(X : \gamma) \rightarrow (X \cup Y : \gamma) \tag{3}$$
$$\models O(X : \overline{\gamma_1} + Y : \overline{\gamma_1}) \rightarrow O(X \cap Y : \overline{\gamma_1}), \quad \textit{with} \quad X \cap Y \neq \emptyset \tag{4}$$
$$\models O(X : \gamma_1 + Y : \gamma_1) \rightarrow (X \cup Y : \gamma) \tag{5}$$
$$\models O(X : \overline{\gamma_1}\ \&\ Y : \overline{\gamma_2}) \rightarrow (X \cap Y : \overline{\gamma_1}\&\overline{\gamma_2}), \quad \textit{with} \quad X \cap Y \neq \emptyset \tag{6}$$
$$\models O(X : \gamma_1\ \&\ Y : \gamma_2) \rightarrow (X \cup Y : \gamma_1\&\gamma_2) \tag{7}$$
$$\models O(X : \gamma_1\ ;\ Y : \gamma_2) \rightarrow O(X \cup Y : \gamma_1; \gamma_2) \tag{8}$$
$$\models O(X : \alpha_1\ \&\ Y : \alpha_2) \leftrightarrow O(X : \alpha_1) \wedge O(Y : \alpha_2) \tag{9}$$
$$\models B_i(O(\{j\} : \alpha_1\ ;\ \{i\} : \alpha_2)) \rightarrow B_i(DONE(\{j\} : \alpha_1) \rightarrow O(\{i\} : \alpha_2)) \tag{10}$$
$$\models O(X : \alpha_1\ ;\ Y : \alpha_2) \leftrightarrow O(X : \alpha_1) \wedge [X : \alpha_1]O(Y : \alpha_2) \tag{11}$$
$$\models O(X : \alpha_1) \vee O(Y : \alpha_2) \rightarrow O(X : \alpha_1 + Y : \alpha_2) \tag{12}$$
$$\models [X : \alpha_1]DONE(X : \alpha_1). \tag{13}$$

Proofs are omitted but can be easily obtained from the semantics presented. Nevertheless, some of these validities deserve some remarks. We start from (10), which shows that, given a (possibly limited) knowledge of the strategy of the group, that is, of a task allocation sequence, and given the knowledge of what just happened, an individual agent is aware of the obligations stemming from that task allocation and addressed to him. This validity is in some sense central for both coordination and knowledge problems raised in the Introduction and in Section 2.

Other interesting validities are (3) and (8), showing how each action of a subgroup is also an action of a supergroup (this manifests exactly the constraints contained in Definition 5). To get from the action of a supergroup to the ones of the subgroups is a more complicated matter. Highly desirable in this sense would be the following property:

$$\models O(X:\gamma_1;\gamma_2) \rightarrow O((X_{11}:\gamma_1);(X_{12}:\gamma_2)+\ldots+(X_{n1}:\gamma_1);(X_{n2}:\gamma_2)) \quad (14)$$
where: $X_{i1}, X_{i2} \subseteq X$, and $1 \leq i \leq n$
with $n = ||\mathcal{P}^+(X) \times \mathcal{P}^+(X)||$

This formula says that the obligation to execute a (possibly partial) plan, addressed to a group, implies the obligation to choose to perform at least one among all the task allocations, which are possible given that group of agents and that plan. Such a formula is valid under precise conditions. Note first of all that, given the properties of the $\Cup$ operation (Definition 7), it does not hold in general that if $T_1 \subseteq T_2$ then $T_2 = T_1 \Cup T_2$. Let us now consider a $\Subset$ relation on sets of s-traces defined as follows: $T_1 \Subset T_2$ iff 1. $T_1 \subseteq T_2$ and 2. $\forall t_3 \in T_2/T_1$, $\nexists t_1 \in T_1$ s.t. $t_1 \in start(t_3)$ [8], that is to say that T_1 is not *sequentially extended by* T_2. Intuitively, $T_1 \Subset T_2$ iff T_1 and T_2 are in a subset relation and the elements belonging to T_2 but not to T_1 are not sequences obtained concatenating new s-traces to elements of T_1. Considering the $\Subset$ relation, the following can be proved:

Proposition 2. (Choice on task allocations)
If $\forall i$ *s.t.* $1 \leq i \leq n$ *and* $X_{i1}, X_{i2} \subseteq X$, $[[X_{i1} : \gamma_1]] \circ [[X_{i2} : \gamma_2]]$ is not sequentially extended by $[\![X : \gamma_1]\!] \circ [\![X : \gamma_2]\!]$, *with* $n = ||\mathcal{P}^+(X) \times \mathcal{P}^+(X)||$, *then:*

$$\models O(X:\gamma_1;\gamma_2) \rightarrow O((X_{11}:\gamma_1);(X_{12}:\gamma_2)+\ldots+(X_{n1}:\gamma_1);(X_{n2}:\gamma_2)).$$

Proof. Formula (14) follows from two facts: 1) on the ground of Definitions 5 and 7, for each element X_{i1}, X_{i2} we have that $[X_{i1} : \gamma_1] \circ [X_{i2} : \gamma_2] \subseteq [X : \gamma_1] \circ [X : \gamma_2]$; 2) from Definition 7, holding that $\forall T_1, T_2$ s.t. $T_1 \Subset T_2, T_2 = T_1 \Cup T_2$ we obtain that: $[X : \gamma_1] \circ [X : \gamma_2] = [X_{11} : \gamma_1] \circ [X_{12} : \gamma_2] \Cup \ldots \Cup [X_{n1} : \gamma_1] \circ [X_{n2} : \gamma_2]$, since $\exists j$ s.t. $1 \leq j \leq n$ and $X_{j1} = X_{j2} = X$. ■

Despite its complex formulation this property states indeed something remarkably intuitive: (14) holds whenever within a given group each action performed by a subgroup is always "complete" from a sequential point of view in the sense that no supergroup performs any action which contains an action performed by the subgroup as a starting action (sequence).

4 Collective Obligations and Coordination as Action: Modeling an Example

In this section we will now show the use of the formalism developed in the previous section to model a collective obligation and show how we can use the

[8] See Definition 7.

validities to derive some individual obligations. We need the properties of the coordinating actions to also make all agents aware of their obligations at the right moment.

Let us consider again the example we have several times touched upon. We have the program committee *PC* of the DEON'04 workshop, with a chairman c, and the other members $a_1, \ldots, a_n$, such that $PC = \{c, a_1, \ldots, a_n\}$. The collective obligation of the program committee is to notify the authors of the acceptance of their papers: $O(PC : \text{notify})$. It seems reasonable to consider the constraint contained in Proposition 2, which enables formula (14), to be met by the example at issue (each agent perform a sequentially "complete" action), and therefore to be able to infer from $O(PC : \text{notify})$ an obligation on the choice among all the possible task allocations. Let us then suppose that the program committee has selected, out of that choice, the following plan for the notification of acceptance: the chairman collects the submitted papers and divides the papers among the other PC members; the PC members review the papers they have received from the chairman and send their results to the chairman; the chairman makes the final decision which papers are selected for the workshop and informs the authors about the decision. Including the coordination aspect this corresponds with the following task allocation *CP,* which belongs to the set $PPlan(PC : \text{notify})$ [9]:

$$
\begin{aligned}
CP = \langle\ & \{c\} : \text{collect}\ ;\ \{c\} : \text{coordinate}(c, \text{collect}, \{c\})\ ;\ \{c\} : \text{divide}\ ; \\
& \{c\} : \text{coordinate}(c, \text{divide}, \{a_1, ..., a_n\})\ ; \\
& (\{a_1\} : \text{review}_1; \text{coordinate}(a_1, \text{review}_1, \{c\})) \\
& \&\ \ldots\ \&\ (\{a_n\} : \text{review}_n; \text{coordinate}(a_n, \text{review}_n, \{c\}))\ ; \\
& \{c\} : \text{decide}\ ;\ \{c\} : \text{coordinate}(c, \text{decide}, \{c\})\ ;\ \{c\} : \text{inform\_ authors}\rangle
\end{aligned}
$$

where *skip* can be substituted for $\text{coordinate}(c, \alpha, \{c\})$. Given the collective obligation $O(G : \text{write})$ and the more concrete obligation on the task allocation sequence *CP* many consequences, which we here omit for space reasons, can be drawn through validities (2)-(13).

We can now also come back to the question who is taking the initiative for the coordination action. If it is the agent that also performed the action (as in the example above) the coordination becomes an informative action. If the initiative lays with the other party it becomes a checking action. The difference between these two kinds of coordination actions is very important for the issue of responsibility. In the example, the chairman has to inform the other members that he has divided the papers among them. If a member of the PC has not received any paper to review and got no information of the chairman, the chairman is responsible for the violation of the obligation that some papers are not reviewed. If the coordination action, however, is defined as a checking action: A : coordinate(c, divide, A), then all the members of the PC have to check whether the chairman has divided the papers, and are therefore also responsible if e.g. the chairman has forgotten to send a member the papers to be reviewed. So, the question who is responsible for a certain action depends on the coordination actions. In a structured group, for example in a formal organization,

[9] See Definitions 1 and 2.

the coordination can be defined explicitly and therefore also the responsibilities. This is especially necessary, if not everyone in the group has knowledge about the complete plan. In a group which is not so structured it is imaginable that all the members of the group know the plan, and that the coordination consists of informing and checking.

In this example we could assume that every member of the PC has knowledge of the complete plan, which implicitly has the consequence that a member who has not received papers from the chairman, has the obligation to check whether the chairman has done his task. In our formalization we could also indicate that only, e.g., two members a_1 and a_2 have to check whether everyone has received the papers by replacing of A : coordinate(c, divide, A) by $\{a_1, a_2\}$: coordinate(c, divide, A). Of course, it is now up to a_1 and a_2 to make sure that all the PC members know that they have received the papers and have the obligation to review them!

Given the collective obligation $O(PC$: notify) and the more concrete obligation on the task allocation sequence CP it is now possible to check all kinds of properties of the plan, a certain knowledge about the plan and a coordination scheme with the plan. It becomes possible to check whether all agents know that they have an obligation whenever they have one (true in the example above). Whether all agents know whenever another agent violated his obligation (not true in the example). Whether the coordination is the most efficient possible, given a certain knowledge of the agents (and maybe observable actions), etc. Due to space limitations we omit these discussions in the present paper.

5 Discussion and Conclusions

Before summarizing the contributions of this work and showing some possible lines of development, it is worth adding a few remarks about the role the notions of plan and allocation come to play in our approach. We stated that a plan is a way of performing a complex action, and that a task allocation determines who in a group carries out which part of the plan. The problem of how a plan and an allocation are chosen, that is, how they can be selected among all the possible ones, has been disregarded. We chose to do that basically for two reasons.

- Firstly, because our central concern is to understand what are the formal properties of obligations distribution within groups, and not how this distribution can be generated.
- Secondly, because the ways plans and allocations might be selected (that is, *how* they are established) are several and can obey different requirements: it might be required that plans are such that each agent is capable of performing the task he is appointed to (and this is indeed a requirement we assumed); or it might be required that each task assignment is "morally acceptable" for each agent and should not conflict with the general deontic principles each agent might be endowed with; besides, plans might be required to be of a "best choice" kind and provide optimal solutions, etc. Several are also the effective sources of these plans and allocations (that is, *who* establishes them): they might be hard-coded by a "designer" in the

group, or emerge from specific power relations holding within the group itself (see [2]), or from individual commitments of the agents (like in [3,4]), or be the result of delegation mechanisms (see [2]), etc.

Following these remarks, some further words can be spent, in particular with respect to the concept of *delegation.* Notice that the example of the program committee of DEON'04 can be easily analyzed also in terms of a notion of delegation: the program committee chair, being obliged to review all submitted papers, delegates reviews to the committee members. In this reading of the example no notion of collective agency and obligation is involved. However, such an approach seems to be less general, it presupposing the existence of a starting individual obligation (which is not always available), and of a delegation mechanism. Examples of obligations addressed exclusively to groups can be easily devised and are indeed discussed in the literature (e.g. a mother ordering her two sons to set the table [14,3]). In those cases no analysis in terms of individual obligations and delegation is feasible, while the approach proposed here remains applicable. The theoretical point that should be stressed is that the effect of a delegation process consists exactly in the establishment of a precise task allocation for the group: when the program committee chair delegates reviews to the members of the program committee, he establishes a precise task allocation, and therefore a determined distribution of obligations within the program committee itself. For these reasons, the formal analysis of obligations distribution, given a task allocation, seemed to us to constitute a more primitive issue.

Such a logical theory of collective obligation has to face many complex and interesting questions:

> From the point of view of methodological individualism, the action of a collective is in some sense composed of or determined by individual actions performed by the members of the collective. The general study of the nature of that composition, of the dependence of collective actions on individual ones, could be said to constitute the theory of collective action in a narrower sense. ([11])

A first attempt is made to "constitute the theory of collective action in a narrower sense" by the introduction of a plan (task allocation) and coordination. We discussed collective agency on the basis of some inputs from planning literature in AI in order to provide some definitions of concepts of relevance for our analysis, especially coordination. The notion of coordination given a plan is very useful to determine which agent has to perform and which agent is responsible for a certain sub-action necessary for the fulfillment of the collective action.

We provided a formal framework in which relevant notions for explaining collective agency can be formalized, such as task allocation, collective obligation, and coordination. This framework can help in representing coordination issues to indicate the individual responsibilities, though of a quite simple kind. We believe it provides a valuable basis for further research in collective obligations. First, it is our aim to embed in our framework a more comprehensive theory of responsibility. Secondly, we would like to consider more types of meta-actions. And finally we would like to check the influence of group structures on collective obligations.

References

1. J. Carmo and O. Pacheco. *Deontic and Action Logics for Collective Agency and Roles.* In R. Demolombe and R. Hilpinen, editors, Proc. Fifth International Workshop on Deontic Logic in Computer Science (DEON'00), pages 93-124. ONERA-DGA, 2000.
2. C. Castelfranchi *Modelling Social Action for AI Agents.* In Artificial Intelligence, Volume 103, pp. 157–182, 1998.
3. L. Cholvy, Ch. Garion *Collective obligations, commitments and individual obligations: a preliminary study* In Proceedings of 6th international Workshop on Deontic Logic in Computer Science (DEON'02), London, May 2002.
4. L. Cholvy, Ch. Garion *Distribution of Goals Addressed to a Group of Agents.* In Proc. of 2nd Int. Joint. Conf. on Autonomous Agents and Multi-Agents Systems (AAMAS 2003), Melbourne, July 2003.
5. K. S. Decker and V. R. Lesser. *Designing a Family of Coordination Algorithms.* Technical Report No. 94-14, Department of Computer Science, University of Massachussets, Amherst, MA01003, 1995.
6. F. Dignum J.-J.Ch. Meyer. *Negations of Transactions and Their Use in the Specification of Dynamic and Deontic Integrity Constraints.* In M. Kwiatkowska, M.W. Shields, and R.M. Thomas, editors, Semantics for Concurrency, Leicester 1990, pages 61–80, Springer-Verlag, Berlin, 1990.
7. F. Dignum, J.-J.Ch. Meyer, R. Wieringa and R. Kuiper. *A Modal Approach to Intentions, Commitments and Obligations: Intention plus Commitment Yields Obligation.* In M.A. Brown and J. Carmo, editors, Deontic Logic, Agency and Normative Systems (Workshops in Computing), pages 80–97. Springer-Verlag, 1996.
8. F. Dignum and B. van Linder. *Modelling Social Agents: Communication as Action.* In M. Wooldridge J. Muller and N. Jennings, editors, Intelligent Agents III (LNAI-1193), pages 205–218. Springer-Verlag, 1997.
9. J.-J. Ch. Meyer. *A Different Approach to Deontic Logic: Deontic Logic Viewed as a Variant of Dynamic Logic.* In Notre Dame Journal of Formal Logic, Volume 29, pages 106–136, 1988.
10. J.-J. Ch. Meyer and W. van der Hoek. *Epistemic Logic for AI and Computer Science.* CUP, 1995.
11. I. Pörn. *The Logic of Power.* Basil Blackwell, Oxford, 1970.
12. L. Royakkers. *Extending Deontic Logic for the Formalization of Legal Rules.* Kluwer Academic Publishers, Dordrecht 1998.
13. L. Royakkers and F. Dignum. *Collective Obligation and Commitment.* In Proceedings of 5th Int. conference on Law in the Information Society, Florence, December, 1998.
14. L. Royakkers and F. Dignum. *No Organization without Obligations: How to Formalize collective obligation?.* In M. Ibrahim, J. Kung and N. Revell, editors, Proceedings of 11th International Conference on Databases and Expert Systems Applications (LNCS-1873), pages 302–311. Springer-Verlag, 2000.
15. S. Russell and P. Norvig. *Artificial Intelligence. A Modern Approach.* Prentice Hall International, 1995.
16. M. Tambe and W. Zhang. *Towards Flexible Teamwork in Persistent Teams: Extended Report.* Journal of Autonomous Agents and Multi-Agent Systems, 3(2):159–183, 2000.

Conflicting Imperatives and Dyadic Deontic Logic

Jörg Hansen

Institut für Philosophie
Universität Leipzig
Beethovenstraße 15, D-04107 Leipzig
jhansen@uni-leipzig.de

Abstract. Often a set of imperatives or norms seems satisfiable from the outset, but conflicts arise when ways to fulfill all are ruled out by unfortunate circumstances. Semantic methods to handle normative conflicts were devised by B. van Fraassen and J. F. Horty, but these are not sensitive to circumstances. The present paper extends these resolution mechanisms to circumstantial inputs, defines according dyadic deontic operators, and provides a sound and (weakly) complete axiomatic system for such a deontic semantics.

1 The Question of Normative Conflicts

Are there moral conflicts? The orthodox belief in the 1950's was that such conflicts only exist at first glance, the seemingly conflicting obligations arising from the application of merely incomplete principles. Instead, what is actually obligatory must be determined by careful moral deliberation that involves considering and weighing all relevant facts and reasons, and cannot produce conflicting outcomes[1]. Among the first that came to reject this view were E. J. Lemmon [18] and B. Williams [32]: Lemmon observed that in cases of true moral dilemma, one does not know the very facts needed to determine which obligation might outweigh the other. Williams argued *in reductio* that if, in case of conflicting oughts, there is just one thing one 'actually' ought to do, then feelings of regret about having not acted as one should have are out of place and one should not mind getting into similar situations again. To avoid having to accept the derivation of the ought of a contradiction from two oughts with contradictory contents, Williams argued that deontic logic should give up the agglomeration principle

(C) $\quad OA \wedge OB \to O(A \wedge B)$

Lemmon had no such qualms: he advocated dropping the Kantian Principle 'ought implies can'

(KP) $\quad OA \to \Diamond A$

and concluded:

[1] W. D. Ross' [22] *prima facie* obligations and R. M. Hare's [10] 'rules of thumb' must be mentioned, though I cannot do these authors justice here.

A. Lomuscio and D. Nute (Eds.): DEON 2004, LNAI 3065, pp. 146–164, 2004.

> "I should like to see a proper discussion of the arguments that go to resolve moral dilemmas, because I do not believe that this is an area of total irrationality, though I do not believe that a traditional logical approach (the logic of imperatives, deontic logic, and whatnot) will do either."

Regarding commands and legal norms, G.H. von Wright ([29] ch.7), like H. Kelsen ([16] p.211) at the time, excluded the coexistence of conflicting norms from the same source: The giving of two conflicting norms is the expression of an irrational will; it is a performative self-contradiction and as such a pure fact that fails to create a norm. E. Stenius [25] and later C.E. Alchourrón and E. Bulygin [3] rejected this view: A system of norms that is impossible to obey might be unreasonable and its norm-giver blameworthy, but its existence does not constitute a logical contradiction – conflicts are ubiquitous in systems of positive law and logic cannot deny this fact. In his later theory, von Wright [31] concedes that factual normative orders may or may not be contradiction-free, and reformulates deontic principles as meta-norms for consistent norm-giving. Kelsen [17] later came to view logic as inapplicable to law.

2 Van Fraassen's Proposal and Horty's Variation

2.1 Van Fraassen's Operator O^F

Not taking sides, *pro* or *contra* the existence of genuine normative conflicts, but arguing that the view in favor seems at least tenable, B. van Fraassen [28] took up the burden of finding plausible logical semantics that could accommodate conflicting obligations. The intended semantics should accept the possible truth of two deontic sentences OA, $O\neg A$ without committing the norm-subject to the absurd by making $O(A \wedge \neg A)$ true, for van Fraassen wanted to keep the Kantian Principle. Given the existence of certain imperatives in force, i.e. imperatives that are left as valid, relevant, not overridden etc. by some unspecified deliberation process, van Fraassen's idea was to make these imperatives part of the logical model, and to describe something as obligatory if it serves to satisfy some, not necessarily all, imperatives. Formally, let I be the set of imperatives in force, $\mathbf{B}$ be the set of possible states of affairs, and $i^+ \subseteq \mathbf{B}$ be the possible states of affairs where the imperative $i \in \mathrm{I}$ is considered fulfilled. Let $\|A\| \subseteq \mathbf{B}$ be the set of possible states of affairs where the indicative sentence A is considered true. Finally, let $score(v)$ be the set of all imperatives that are fulfilled in the state of affairs v: $score(v) = \{i \in \mathrm{I} \mid \mathrm{v} \in \mathrm{i}^+\}$. Van Fraassen then defines[2]:

[Df-F] $O^F A$ is true iff $\exists v \in \|A\| : \forall v' \in \|\neg A\| : score(v) \not\subseteq score(v')$

So A is obligatory if and only if (iff) the falsity of A would commit us to a lower 'score' than one which could be achieved when A is true. In other words, A is obligatory if the truth of A is necessary for achieving a maximal score.

[2] The definition given is van Fraassen's final proposal in [28] p. 18.

By slightly changing the viewpoint, van Fraassen's proposal might also be described in the following way: Let I be a set not of imperatives, but of indicative sentences in the language $\mathrm{L_{BL}}$ of some basic logic *BL*. The motivation is that I contains one sentence A for each imperative i in force that is true in exactly those states of affairs in which the imperative is fulfilled, i.e. $\|A\| = i^+$. *BL* is assumed to be compact and the turnstile in $\Gamma \vdash_{BL} A$ means a classical consequence relation that characterizes *BL*, $\Gamma \subseteq \mathrm{L_{BL}}$, $A \in \mathrm{L_{BL}}$. Let the *remainder set* $\Gamma \perp A$ be the set of all maximal subsets that do not derive A, i.e. of all $\Gamma' \subseteq \Gamma$ such that (i) $\Gamma' \nvdash_{BL} A$, and (ii) there is no Γ'' such that $\Gamma' \subset \Gamma'' \subseteq \Gamma$ and $\Gamma'' \nvdash A$. Then Df-F is equivalent to Df-F*(k means an arbitrary contradiction)[3]:

[Df-F*] $O^F A$ is true iff $\exists I' \in I \perp k : I' \vdash_{BL} A$

Acccordingly, A is obligatory iff it is derivable from a maximally consistent subset of the imperative-corresponding sentences.

To see how van Fraassen's semantics work, first let $I = \{A, B\}$, where A, B are supposedly contingent and independent. There are no conflicts, I is consistent and $O^F A$, $O^F B$ and $O^F(A \wedge B)$ are all true since I derives $A, B, A \wedge B$. Thus the semantics permit agglomeration of contents when the underlying imperatives do not conflict. For the conflict case, change I into $\{A \wedge C, B \wedge \neg C\}$; $O^F A$ and $O^F B$ are true since A and B derive from the maximally consistent sets $\{A \wedge C\}$ and $\{B \wedge \neg C\}$, but $O^F(C \wedge \neg C)$ is false since no consistent subset derives $C \wedge \neg C$. The same is true for $O^F(A \wedge B)$ though $\{A \wedge B\}$ is consistent: the truth of $A \wedge B$ is not necessary for maximal norm satisfaction.

An axiomatic system *DF* that is (weakly) complete with regard to van Fraassen's semantics is defined by the following axiom-schemes, in addition to *BL*-instances and *modus ponens* (cf. [8], t means an arbitrary tautology, and the index 'F' here and below indicates that the deontic operators occurring in the axiom scheme are thus indexed):

(M^F) $O^F(A \wedge B) \rightarrow (O^F A \wedge O^F B)$
(D^F) $\neg O^F k$
(N^F) $O^F t$
(Ext^F) If $\vdash_{BL} A \leftrightarrow B$ then $\vdash_{DF} O^F A \leftrightarrow O^F B$

To van Fraassen's own puzzlement, the cases where agglomeration remains permissible seem not axiomatizable: object language does not reveal whether particular A, B of some $O^F A, O^F B$ are derived from the demands of imperatives that do not collide and so $O^F(A \wedge B)$ should be supported[4].

[3] Cf. Horty's [13] Theorem 2.

[4] Agglomeration thus requires a consistency check of the underlying imperatives' contents. As a solution, [26] and [27] have proposed a two-phase deontic logic, where 'consistent aggregation' must take place before weakening of norm-contents. For consistency checks before agglomeration also cf. [14] and my [8] which provides a bimodal axiomatization.

2.2 The 'Sceptical' Operator O^S

The invalidation of the agglomeration principle by van Fraassen's semantics did not make them popular (cf. [5] p. 298). Moreover, let a P^F-operator for permission be defined in the usual way, i.e. $P^F A$ is true iff $\neg O^F \neg A$ is true, and modify *DF* to additionally contain

$(\mathrm{Def}^F) \quad P^F A \leftrightarrow \neg O^F \neg A$

Consider again $I = \{A \wedge C, B \wedge \neg C\}$: $O^F(A \wedge C)$ is true, and so is $O^F \neg C$. Applying (M^F) and (Ext^F), $O^F \neg (A \wedge C)$ must be true, and by the above definition $P^F(A \wedge C)$ is false. So not even what is obligatory is always permitted, which was considered unintelligible [15].

In reaction to the dismissal of the agglomeration principle, Donagan [5] and Brink [4] have claimed that even if there could be a normative demand for *A* and a conflicting demand for *B*, with $\vdash_{BL} A \rightarrow \neg B$, it need not follow that the norm-subject have an obligation to realize *A and* an obligation to realize *B*. Rather, there should just be a disjunctive obligation to realize *A or B*. Given competing normative standards of equal weight, the strategy of this reasoning is not to trust a single standard, but to consider obligatory only what *all* standards demand. Let *I* be as before. Varying van Fraassen's truth definition, Horty [13] has formalized this 'skeptical' ought as follows[5]:

[Df-S] $O^S A$ is true iff $\forall I' \in I \perp k : I' \vdash_{BL} A$

So $O^S A$ is true iff *A* is derivable from all maximally consistent subsets of *I*.

Let again $I = \{A \wedge C, B \wedge \neg C\}$. $O^S A$ and $O^S B$ are false and $O^S(A \vee B)$ is true: just $A \vee B$, but neither *A* nor *B* are derived by both of the two consistent subsets $\{A \wedge C\}$ and $\{B \wedge \neg C\}$. $P^S(A \wedge C)$ is also true: *A, C* were assumed to be contingent and independent, so the maximally consistent subset $\{A \wedge C\} \subseteq I$ does not derive $\neg(A \wedge C)$. Ergo what is O^F-obligatory is at least P^S-permitted.

A complete axiomatic system *DS* is defined by the axiom-schemes (Def^S), (M^S), (C^S), (D^S), (N^S), and (Ext^S), together with *BL*-instances and *modus ponens.* Since the truth definitions for $O^F A, P^F A$ and $O^S A, P^S A$ merely depend on the set *I* and *BL*, mixed expressions such as $O^F A \wedge \neg O^S A$ are meaningful and may be accepted as well-formed. Then

$(\mathrm{C}^{FS}) \quad O^F A \wedge O^S B \rightarrow O^F(A \wedge B)$

is valid, and the mixed system *DFS* – containing the axiom schemes for *DF*, *DS*, the axiom scheme (C^{FS}), all instances of *BL*-theorems and *modus ponens* – is sound and (weakly) complete (cf. [8]).

---

[5] More in parallel to van Fraassen's original definition, one may equivalently define

[Df-S*] $O^S A$ is true iff $\forall v \in \|\neg A\| : \exists v' \in \|A\| : score(v) \subset score(v')$

The proof is easy and left as entertainment for the reader.

3 Predicaments and Dyadic Deontic Logic

Arguing for the possibility of moral conflicts, R. Barcan Marcus [20] gave the following example:

> "Under the single principle of promise keeping, I might make two promises in all good faith and reason that they will not conflict, but then they do, as a result of circumstances that were unpredictable and beyond my control."

Note that there is no conflict at the outset: Any dilemma could have been averted by not promising anything. Moreover, there might have been some point in time at which keeping both promises was possible: Having 500 $ with me and another 1000 $ in the office, on Saturday I promise Sally and Jane 500 $ each with every intention of paying them on Monday, only to find out that the office had been burglarized over the weekend. Donagan [5] argues that this is not a genuine conflict, because three resolving principles apply: (i) one must not make promises one cannot or must not keep, (ii) all promises are made with the implicit condition that they are void if they cannot or must not be kept, (iii) one must not make promises when one does not believe that the other party has fully understood (ii). Now suppose whatever happens at the office, neither Sally nor Jane are going to let me off the hook, and I could have known this. According to (iii), I was wrong to make the promise, so am I entitled to break it? – We have here what G.H. von Wright terms a 'predicament': a situation from which there is no permitted way out, but to which there also is no permitted inlet (cf. [30] p. 78). The normative order is consistent, it is only through one's own fault that one finds oneself in a predicament[6]. Von Wright then asks:

> "The man in a predicament will, of necessity, react in some way or other, either do something or remain passive. Even though every reaction of his will be a sin, is it not reasonable to think that there is yet something he ought to do rather than anything else? To deny this would be to admit that it makes, deontically, no difference what he does. But is this reasonable? (...) If all our choices are between forbidden things, our duty is to choose the least bad thing."

Sub-ideal demands are usually represented by a dyadic deontic sentence *O(A/C),* meaning that in case *C* is true it ought to be that *A*. By accepting all instances

[6] That predicaments *only* arise from an agent's own faults, and not through misfortune or the wrongdoings of others, is a view von Wright and Donagan ascribe to Thomas Aquinas, but this does not seem quite correct: In the discussion of oaths *(Summa Theologica* II.II Qu. 89 art. 7 ad 2), Thomas considers the objection that it would sometimes be contrary to virtue, or an obstacle to it, if one were to fulfill what one has sworn to do – so oaths need not always be binding. In answering, Thomas distinguishes oaths that are unlawful from the outset, where a man sinned in swearing, and oaths that could be lawful at the outset but lead to an evil result through some new and unforeseen emergency: fulfillment of such oaths is unlawful.

of $O(A/t) \rightarrow P(A/t)$ as a logical truth in [30], von Wright dismisses an inconsistent normative system as 'conceptual absurdity': if A is obligatory on tautological conditions (i.e. unconditionally obligatory), then there cannot be a likewise unconditional obligation to the contrary. But as far as I can see, von Wright does not similarly advocate excluding predicaments similarly on grounds of logic alone: $O(A/C) \rightarrow P(A/C)$ is not a theorem, so it remains possible that in circumstances C it ought to be that A and simultaneously it ought to be that $\neg A$[7]. Dyadic operators are needed to express this different deontic-logical treatment of conflicts and predicaments, for otherwise it would be difficult to tell whether $O(C \rightarrow A)$ and $O(C \rightarrow \neg A)$ are oughts that are conditional on C and so their contents may not be agglomerated, or oughts with material implications for contents that permit agglomeration.

Turning object language oughts into a special sort of conditionals does not mean that there must be a change in the formalization of background imperatives as well: Consider the set $I = \{(C \rightarrow A), (C \rightarrow \neg A)\}$, corresponding to background imperatives in the usual way. There is just one maximally consistent subset, which derives $\neg C$ so $O^F\neg C$ and $O^S\neg C$ are both true. But there is no single standard available once C becomes true: the imperatives have not all been fulfilled (otherwise one would not be in condition C), and any maximal set of imperatives that is consistent with the given circumstances cannot contain all. So the proposal is to call A obligatory in case C iff, given the truth of C, A is necessary to fulfill as many norms as is still possible. Formally:

[Df-DF] $O^F(A/C)$ iff $\exists I' \in I \perp \neg C : I' \cup \{C\} \vdash_{BL} A$

According to this definition, $O^F(A/C)$ is true iff there is some set, among the maximal subsets of I consistent with C, that together with C derives A. This is obviously a conservative extension of the definition given for the unconditional case, so we may define $O^F A =_{def} O^F(A/t)$.

If a cautious, disjunctive approach were appropriate for cases of conflict, then it would be hard to see why predicaments should be treated differently: That conflicts must be accounted for at the outset, but analogues of Buridan's ass cannot occur on the level of predicaments brought about by fate or unpredictable human nature, would hardly be plausible. Distrusting any single standard, such an approach would accept, given the circumstances C, only what is necessary by any standard that could still be met - no worrying about spilled milk. Formally:

[Df-DS] $O^S(A/C)$ iff $\forall I' \in I \perp \neg C : I' \cup \{C\} \vdash_{BL} A$

According to this definition, $O^S(A/C)$is true iff all the maximal subsets of I consistent with C derive A, given the truth of C. This is again just a conservative extension of the unconditional case, so one may define $O^S A =_{def} O^S(A/t)$.

In the remaining section, I give an axiomatic dyadic deontic system *DDFS*, and prove that this is sound and (only) weakly complete with respect to the above semantics.

[7] Cf. [30] pp. 36, 81, 89.

4 The Dyadic Deontic Logic *DDFS*

Let the *basic logic* be prepositional logic: The alphabet has proposition letters $Prop = \{p_1, p_2, ...\}$, truth-functional operators '$\neg$', '$\wedge$', '$\vee$', '$\rightarrow$', '$\leftrightarrow$' and brackets '(',')'. The set of sentences is defined as usual. $\bigwedge$, $\bigvee$ in front of a set of sentences means their conjunction and disjunction, and e.g. $\bigwedge_{i=1}^{n} A_i$ further abbreviates $\bigwedge\{A_i, ..., A_n\}$. In the semantics, valuation functions $v : Prop \rightarrow \{1, 0\}$ define the truth of sentences $A \in \mathrm{L_{PL}}$ as usual (written $v \vDash A$), $\mathbf{B}$ is the set of all such valuations, and $\|A\|$ means $\{v \in \mathbf{B} \mid v \vDash A\}$. *PL* is a sound and complete axiomatic system, and $\vdash_{PL} A$ means that *A is* provable in *PL.*

The alphabet of the *language* $\mathrm{L_{DDFS}}$ additionally has the operators 'O^F', 'P^F', 'O^S', 'P^S', and the auxiliary '/'. $\mathrm{L_{DDFS}}$ is then the smallest set such that

a) for all $A, C \in \mathrm{L_{PL}}$, $O^F(A/C)$, $P^F(A/C)$, $O^S(A/C)$, $P^S(A/C) \in \mathrm{L_{DDFS}}$,
b) if $A, B \in \mathrm{L_{DDFS}}$, so are $\neg A, (A \wedge B)$, $A \vee B)$, $(A \rightarrow B)$, $(A \leftrightarrow B)$.

Outer brackets will mostly be omitted. For simplification we do not have mixed expressions and nested deontic operators as in $p_1 \wedge O^S(p_2/p_1)$, $P^S(O^F(p_2/p_2)/p_1)$.

For *DDFS-semantics,* the truth of DDFS-sentences is defined with respect to a set $I \subseteq \mathrm{L_{PL}}$ by the following clauses (Boolean operators being as usual):

$I \vDash O^F(A/C)$ iff $\exists I' \in I\bot\neg C :\ I' \cup \{C\} \vdash_{PL} A$
$I \vDash P^F(A/C)$ iff $\forall I' \in I\bot\neg C :\ I' \cup \{C\} \nvdash_{PL} \neg A$
$I \vDash O^S(A/C)$ iff $\forall I' \in I\bot\neg C :\ I' \cup \{C\} \vdash_{PL} A$
$I \vDash P^S(A/C)$ iff $\exists I' \in I\bot\neg C :\ I' \cup \{C\} \nvdash_{PL} \neg A$

If $I \vDash A$, *A is* called *DDFS-satisfiable,* and called *DDFS-valid* if $I \vDash A$ for all $I \subseteq \mathrm{L_{PL}}$ (we write $\vDash_{DDFS} A$).

Consider the following axiom-schemes (* is the uniform index *F* or *S*):

(DDef) $O^*(A/C) \leftrightarrow \neg P^*(\neg A/C)$
(DM) $O^*(A \wedge B/C) \rightarrow (O^*(A/C) \wedge O^*(B/C))$
(DC^S) $O^S(A/C) \wedge O^S(B/C) \rightarrow O^S(A \wedge B/C)$
(DC^{FS}) $O^F(A/C) \wedge O^S(B/C) \rightarrow O^F(A \wedge B/C)$
(CExt) If $\vdash_{PL} C \rightarrow (A \leftrightarrow B)$ then $\vdash_{DDFS} O^*(A/C) \leftrightarrow O^*(B/C)$
(ExtC) If $\vdash_{PL} C \leftrightarrow D$ then $\vdash_{DDFS} O^*(A/C) \leftrightarrow O^*(A/D)$
(DN^S) $O^S(t/C)$
(DN^F) If $\nvdash_{PL} \neg C$ then $\vdash_{DDFS} O^F(t/C)$
(DD^S) If $\nvdash_{PL} \neg C$ then $\vdash_{DDFS} P^S(t/C)$
(DD^F) $P^F(t/C)$
(Up) $O^*(A/C \wedge D) \rightarrow O^*(D \rightarrow A/C)$
(Down1) $O^F(A/C \vee D) \wedge \neg O^F(A \wedge \neg C/C \vee D) \rightarrow O^F(A/C)$
(Down2) $O^S(A/C \vee D) \wedge \neg O^F(A \wedge \neg C/C \vee D) \rightarrow O^S(A/C)$
(Down3) $O^S(A/C \vee D) \wedge \neg O^S(A \wedge \neg C/C \vee D) \rightarrow O^F(A/C)$

(DDef), (DM), (DC) and (DC^{FS}) are the dyadic analogues of the monadic axiom schemes given above. (CExt) is a contextual 'extensionality' rule for consequents, and (ExtC) an extensionality rule for antecedents. A system that contains (CExt), (DDef) and $O^*(t/C)$, $P^*(t/C)$ is inconsistent if $C = k$ is allowed,

so (DN^F) and (DD^S) are accordingly restricted. (Up) transfers obligations conditionally from stronger to weaker circumstances, and (Down1-3) allow for corresponding transfers of obligations from weaker to stronger circumstances if these can be excluded to be 'contrary-to-this-duty'.

The *axiomatic system DDFS* is then the set such that (i) all $\mathrm{L}_{\mathrm{DDFS}}$-instances of *PL*-tautologies are in *DDFS,* (ii) all L_{PL}-instances in the above axiom schemes are in *DDFS,* and (iii) *DDFS* is closed under *modus ponens.* If $A \in DDFS$ we write $\vdash_{DDFS} A$ and call *A provable* in *DDFS.* $\Gamma \subseteq \mathrm{L}_{\mathrm{DDFS}}$ is *DDFS-inconsistent* iff there are $A_1,...,A_n$ in Γ, $n \geq 1$, with $\vdash_{DDFS} (A_1 \wedge ... \wedge A_n) \rightarrow k$, otherwise Γ is *DDFS-consistent.* $A \in \mathrm{L}_{\mathrm{DDFS}}$ is *DDFS-derivable* from $\Gamma \subseteq \mathrm{L}_{\mathrm{DDFS}}$ (written $\Gamma \vdash_{DDFS} A$) iff $\Gamma \cup \{\neg A\}$ is *DDFS*-inconsistent.

Theorem 1. *The following are DDFS-provable (* is F or S as indicated):*

		F	S	*Mixed*
(FH)	$O^*(A/C \vee D) \wedge P^*(C/D) \rightarrow O^*(A/C)$	+	−	*SFS, SSF*
(REF)	$O^*(A/A)$	−	+	
(OR)	$O^*(A/C) \wedge O^*(A/D) \rightarrow O^*(A/C \vee D)$	−	+	*SFF, FSF*
(DR)	$O^*(A/C \vee D) \rightarrow O^*(A/C) \vee O^*(A/D)$	+	−	*SFS, SSF*
(RM)	$O^*(A/C) \wedge P^*(D/C) \rightarrow O^*(A/C \wedge D)$	+	−	*SFS, SSF*
(CM)	$O^*(A/C) \wedge O^*(D/C) \rightarrow O^*(A/C \wedge D)$	−	+	*FSF, SFF*
(CUT)	$O^*(A/C \wedge D) \wedge O^*(D/C) \rightarrow O^*(A/C)$	−	+	*FSF, SFF*

Proof. (FH) is the 'down-theorem' proposed in [6], it derives (Down1-3) given agglomeration, and the other theorems are well-known from the study of nonmonotonic logics. All proofs are straightforward and I just give those for (FH) in version *FFF* and (RM) in version *SFS;* both will be employed below.

(FH): (Down1) is $O^F(A/C \vee D) \wedge \neg O^F(A \wedge \neg C/C \vee D) \rightarrow O^F(A/C)$. By contraposition $O^F(A/C \vee D) \wedge \neg O^F(A/C) \rightarrow O^F(A \wedge \neg C/C \vee D)$. (DM) derives $O^F(A \wedge \neg C/C \vee D) \rightarrow O^F(\neg C/C \vee D)$, and ($\mathrm{DD}^F$) derives $P^F(C \vee D/C \vee D)$, which using (DDef), (CExt), (ExtC) is equivalent to $\neg O^F(\neg C \wedge \neg D/D \vee C)$. $O^F(\neg C/C \vee D) \wedge \neg O^F(\neg C \wedge \neg D/D \vee C) \rightarrow O^F(\neg C/D)$ is an instance of (Down1), so $O^F(A/C \vee D) \wedge \neg O^F(A/C) \rightarrow O^F(\neg C/D)$ is derived with *modus ponens* and *PL,* which is (FH) in contraposition.

(RM): $O^S(A/(C \wedge D) \vee C) \wedge \neg O^F(A \wedge \neg(C \wedge D)/(C \wedge D) \vee C) \rightarrow O^S(A/C \wedge D)$ is an instance of (Down2). By use of (ExtC), $O^S(A/(C \wedge D) \vee C)$ is equivalent to $O^S(A/C)$. By use of (DDef) and (ExtC) $\neg O^F(A \wedge \neg(C \wedge D)/(C \wedge D) \vee C)$ is equivalent to $P^F(A \rightarrow (C \wedge D)/C)$ that derives from $P^F(C \wedge D)/C)$ with (CExt) and (DM), which derives from $P^F(D/C)$ with (CExt). Then (RM) is obtained by equivalent substitution and strengthening of the antecedent.

Theorem 2. DDFS *is sound.*

Proof. The validity of (DDef), (DM), (DC^S), (DC^F), (CExt), and (ExtC) is immediate. (DN^S), (DD^F) are valid since any subset of L_{PL} derives t, and any maximally consistent subset is consistent. If $\nvdash_{PL} \neg C$ then at least $\varnothing$ is in $I \perp \neg C$, so $I \perp \neg C \neq \varnothing$ and (DN^F), (DD^S) are likewise true. Consider (Up),(Down1-3):

(Up) Assume $O^F(A/C \wedge D)$, so there is an $I' \in I\perp\neg(C \wedge D)$ such that $I' \cup \{C \wedge D\} \vdash_{PL} A$ and $I' \cup \{C\} \vdash_{PL} D \rightarrow A$. Since $I' \nvdash_{PL} \neg(C \wedge D)$, also $I' \nvdash_{PL} \neg C$, so by maximality there is an $I'' \in I\perp\neg C$ such that $I' \subseteq I''$, so there is an $I'' \in I\perp\neg C : I'' \cup \{C\} \vdash_{PL} D \rightarrow A$, so $O^F(D \rightarrow A/C)$. Assume $O^S(A/C \wedge D)$. So for all $I' \in I\perp\neg(C \wedge D) : I' \cup \{C \wedge D\} \vdash_{PL} A$. Suppose there is an $I'' \in I\perp\neg C : I'' \cup \{C\} \nvdash_{PL} D \rightarrow A$. So also $I'' \cup \{C\} \nvdash_{PL} \neg D$ and $I'' \nvdash_{PL} \neg(C \wedge D)$. So by maximality there is an $I' \in I\perp\neg(C \wedge D) : I'' \subseteq I'$. Since $I' \nvdash_{PL} \neg(C \wedge D)$, $I' \nvdash_{PL} \neg C$, there is an $I'' \in I\perp\neg C : I' \subseteq I''$, so by maximality of each $I'' \in I\perp\neg C$, $I' = I''$. So there is an $I' \in I\perp\neg(C \wedge D) : I' \cup \{C \wedge D\} \nvdash_{PL} A$, which violates the assumption. So for all $I'' \in I\perp\neg C : I'' \cup \{C\} \vdash_{PL} D \rightarrow A$, and $O^S(D \rightarrow A/C)$.

(Down1) Assume $O^F(A/C \vee D)$, so $\exists I' \in I\perp\neg(C \vee D) : I' \cup \{C \vee D\} \vdash_{PL} A$, and $\neg O^F(A \wedge \neg C/C \vee D)$, so $\forall I'' \in I\perp\neg(C \vee D) : I'' \cup \{C \vee D\} \nvdash_{PL} A \wedge \neg C$. So $I' \nvdash_{PL} \neg C$. So by maximality $\exists I''' \in I\perp\neg C : I' \subseteq I'''$, so $I''' \cup \{C \vee D\} \vdash_{PL} A$, so $I''' \cup \{C\} \vdash_{PL} A$, so $O^F(A/C)$ is true.

(Down2) Assume $O^S(A/C \vee D)$, so $\forall I' \in I\perp\neg(C \vee D) : I' \cup \{C \vee D\} \vdash_{PL} A$, and $\neg O^F(A \wedge \neg C/C \vee D)$, so $\forall I' \in I\perp\neg(C \vee D) : I' \cup \{C \vee D\} \nvdash_{PL} A \wedge \neg C$. Suppose $I'' \in I\perp\neg C$, so $I'' \nvdash_{PL} \neg C$ and $I'' \nvdash_{PL} \neg(C \vee D)$, so by maximality $\exists I''' \in I\perp\neg(C \vee D) : I'' \subseteq I'''$. By the first assumption $I''' \cup \{C \vee D\} \vdash_{PL} A$. If $I'' \cup \{C \vee D\} \nvdash_{PL} A$ then $\exists \{i_1, ..., i_n\} \subseteq I''' : \{i_1, ..., i_n\} \nsubseteq I''$ by compactness of *PL*. By maximality $I'' \cup \{i_1, ..., i_n\} \vdash_{PL} \neg C$, but $I'' \cup \{i_1, ..., i_n\} \subseteq I'''$, so $I''' \cup \{C \vee D\} \vdash_{PL} A \wedge \neg C$, which violates the second assumption. So $I'' \cup \{C \vee D\} \vdash_{PL} A$, $I'' \cup \{C\} \vdash_{PL} A$, and $\forall I'' \in I\perp\neg C : I'' \cup \{C\} \vdash_{PL} A$ since I'' was arbitrary, so $O^S(A/C)$ is true.

(Down3) Assume $O^S(A/C \vee D)$, so $\forall I' \in I\perp\neg(C \vee D) : I' \cup \{C \vee D\} \vdash_{PL} A$, and $\neg O^S(A \wedge \neg C/C \vee D)$, so $\exists I'' \in I\perp\neg(C \vee D) : I'' \cup \{C \vee D\} \nvdash_{PL} A \wedge \neg C$. So $I'' \nvdash_{PL} \neg C$. So by maximality $\exists I''' \in I\perp\neg C : I'' \subseteq I'''$, so $I''' \cup \{C \vee D\} \vdash_{PL} A$, so $I''' \cup \{C\} \vdash_{PL} A$, so $O^F(A/C)$ is true.

Theorem 3. *DDFS-semantics are not compact.*

Proof. In [8] I provide a counterexample to the compactness of semantics that just employ the monadic deontic operator O^F. Since $O^F A$ can be defined as $O^F(A/t)$, this also refutes compactness of *DDFS* and of the subsystem that contains just the dyadic operators O^F and P^F. The following counterexample is expressed in terms of the dyadic operators O^S and P^S only, which also refutes compactness of the subsystem that contains just these operators: Let

$$\begin{aligned}\Gamma = \quad & \{O^S(p_2/t)\} \\ & \cup \{P^S(\neg p_2/p_1)\} && \cup \textstyle\bigcup_{i=3}^{\infty}\{O^S(p_i/p_1)\} \\ & \cup \{P^S(\neg p_2/\neg p_1)\} && \cup \textstyle\bigcup_{i=3}^{\infty}\{O^S(p_i/\neg p_1)\} \\ & \cup \{P^S(\neg p_2/p_1 \leftrightarrow p_2)\} && \cup \textstyle\bigcup_{i=3}^{\infty}\{O^S(p_i/p_1 \leftrightarrow p_2)\} \\ & \cup \{P^S(\neg p_2/p_1 \leftrightarrow \neg p_2)\} && \cup \textstyle\bigcup_{i=3}^{\infty}\{O^S(p_i/p_1 \leftrightarrow \neg p_2)\}\end{aligned}$$

Γ is finitely *DDFS*-satisfiable: Let n be the greatest index of any proposition letter occurring in some finite $\Gamma_f \subseteq \Gamma$. Then $I_f = \{ p_{n+1} \wedge (p_1 \rightarrow \neg p_2)\ ,\ \neg p_{n+1} \wedge (\neg p_1 \rightarrow \neg p_2)\ ,\ p_2, p_3, ..., p_n\}$ satisfies Γ_f.

For easy verification, I list the relevant sets of maximal subsets:

$$I_f \perp k = \left\{ \begin{array}{l} \{p_{n+1} \wedge (p_1 \rightarrow \neg p_2), p_2, p_3, ..., p_n)\}, \\ \{\neg p_{n+1} \wedge (\neg p_1 \rightarrow \neg p_2), p_2, p_3, ..., p_n)\} \end{array} \right\}$$

$$\begin{array}{r} I_f \perp \neg p_1 \\ I_f - \perp \neg (p_1 \leftrightarrow p_2) \end{array} = \left\{ \begin{array}{l} \{p_{n+1} \wedge (p_1 \rightarrow \neg p_2), p_3, ..., p_n)\}, \\ \{\neg p_{n+1} \wedge (\neg p_1 \rightarrow \neg p_2), p_2, p_3, ..., p_n)\} \end{array} \right\}$$

$$\begin{array}{r} I_f \perp p_1 \\ I_f \perp \neg (p_1 \leftrightarrow \neg p_2) \end{array} = \left\{ \begin{array}{l} \{\ p_{n+1} \wedge (p_1 \rightarrow \neg p_2),\ p_2, p_3, ..., p_n)\ \}, \\ \{\ \neg p_{n+1} \wedge (\neg p_1 \rightarrow \neg p_2),\ p_3, ..., p_n)\} \end{array} \right\}$$

However, Γ is not DDFS-satisfiable: Suppose $I \subseteq \mathrm{L_{PL}}$ satisfies Γ, and let $A \in \{p_1, \neg p_1, p_1 \leftrightarrow p_2, p_1 \leftrightarrow \neg p_2\}$. Observe that

(i) There are $I_1, I_2 \in I \perp k$ such that $I_1 \vdash_{PL} p_1 \wedge p_i$, $I_2 \vdash_{PL} \neg p_1 \wedge p_i$, $i \geq 2$.
Proof: From $O^S(p_2/t), P^S(\neg p_2/\neg p_1) \in \Gamma$ and the validity of (Down2) it follows that there is an $I_1 \in I \perp k : I_1 \vdash_{PL} p_1$. Likewise from $O^S(p_2/t)$, $P^S(\neg p_2/p_1) \in \Gamma$ it follows that there is an $I_2 \in I \perp k : I_2 \vdash_{PL} \neg p_1$. To satisfy $O^S(p_2/t)$ it is necessary that all $I' \in I \perp k : I' \vdash_{PL} p_2$, and from $O^S(p_i/p_1), O^S(p_i/\neg p_1) \in \Gamma$ and the validity of (Up), (DC^S) one obtains that for all $I' \in I \perp k : I' \vdash_{PL} p_i$, $i \geq 3$.

(ii) For each A, there is an $I_A \in I \perp \neg A : I_A \cup \{A\} \vdash_{PL} \neg p_2$.
Proof: Let $A \in \{p_1, p_1 \leftrightarrow p_2\}$. Then by observation (i) $I_1 \in I \perp \neg A$. Since $I_1 \vdash_{PL} p_2$, to satisfy $P^S(\neg p_2/A) \in \Gamma$ there is an $I_A \in I \perp \neg A$ such that $I_A \cup I_1 \vdash_{PL} \neg A$. So $I_A \cup \{A\} \vdash_{PL} \neg(p_1 \wedge p_2 \wedge ... \wedge p_n)$ for some n. If $n \geq 3$ then $I_A \cup \{A\} \vdash_{PL} \neg(p_1 \wedge p_2 \wedge ... \wedge p_{n-1})$, since $I_A \cup \{A\} \vdash_{PL} p_n$ is necessary for $O^S(p_n/A) \in \Gamma$. So $I_A \cup \{A\} \vdash_{PL} \neg(p_1 \wedge p_2)$, so $I_A \cup \{A\} \vdash_{PL} \neg p_2$. Likewise, the proof for $A \in \{\neg p_1, p_1 \leftrightarrow \neg p_2\}$ is obtained from $I_2 \in I \perp \neg A$.

(iii) If $A \in \{p_1, p_1 \leftrightarrow \neg p_2\}$ then $I_A \cup \{p_1, \neg p_2, p_3, p_4, ...\} \nvdash_{PL} k$. If $A \in \{\neg p_1, p_1 \leftrightarrow p_2\}$ then $I_A \cup \{\neg p_1, \neg p_2, p_3, p_4, ...\} \nvdash_{PL} k$.
Proof: Suppose $A \in \{p_1, p_1 \leftrightarrow \neg p_2\}$ and $I_A \cup \{p_1, \neg p_2, p_3, p_4, ...\} \vdash_{PL} k$. Then $I_A \cup \{A, \neg p_2, p_3, p_4, ...\} \vdash_{PL} k$. So $I_A \cup \{A\} \vdash_{PL} \neg(\neg p_2 \wedge p_3 \wedge p_4 \wedge ... \wedge p_n)$ for some n. But also $I_A \cup \{A\} \vdash_{PL} \neg p_2 \wedge p_3 \wedge p_4 \wedge ... \wedge p_n$ by observation (ii) and from the fact that I satisfies $O^S(p_i/A) \in \Gamma$, $3 \leq i \leq n$. So $I_A \vdash_{PL} \neg A$, but this contradicts $I_A \in I \perp \neg A$. The proof for $A \in \{\neg p_1, p_1 \leftrightarrow p_2\}$ and the set $I_A \cup \{\neg p_1, \neg p_2, p_3, p_4, ...\}$ is done likewise.

It follows that $I_{p_1} \cup I_{(p_1 \leftrightarrow \neg p_2)} \nvdash_{PL} k$ and $I_{\neg p_1} \cup I_{(p_1 \leftrightarrow p_2)} \nvdash_{PL} k$. This is most easily seen by appealing to *PL*-semantics: some $v \in \mathbf{B}$ satisfies $\{p_1, \neg p_2, p_3, p_4, ...\}$ and by (iii) all elements of I_{p_1} as well as all of $I_{(p_1 \leftrightarrow \neg p_2)}$, so their union is satisfiable and therefore consistent (likewise for $\{\neg p_1, \neg p_2, p_3, p_4, ...\}$ and $I_{\neg p_1} \cup I_{(p_1 \leftrightarrow p_2)}$). From (ii) it follows that

$$I_{p_1} \cup I_{(p_1 \leftrightarrow \neg p_2)} \vdash_{PL} (p_1 \rightarrow \neg p_2) \wedge ((p_1 \leftrightarrow \neg p_2) \rightarrow \neg p_2)$$
$$I_{\neg p_1} \cup I_{(p_1 \leftrightarrow p_2)} \vdash_{PL} (\neg p_1 \rightarrow \neg p_2) \wedge ((p_1 \leftrightarrow p_2) \rightarrow \neg p_2)$$

But the conclusions are tautologically equivalent to $\neg p_2$, so there are consistent subsets of I that derive $\neg p_2$, and $I \not\models O^S(p_2/t)$, although $O^S(p_2/t) \in \Gamma$.

Theorem 4. *DDFS is weakly complete.*

Proof. We must prove that if $\models_{DDFS} A$ then $\vdash_{DDFS} A$ for any $A \in \mathrm{L_{PL}}$. We assume $\nvdash_{DDFS} A$ so $\neg A$ is DDFS-consistent. We build a disjunctive normal form of $\neg A$ and eliminate all negation signs in front of deontic operators by use of (DDef). The result is a disjunction of conjunctions, where each conjunct is either $O^F(B/C)$, $O^S(B/C)$, $P^F(B/C)$, or $P^S(B/C)$. One disjunct must then be *DDFS*-consistent. Let δ be that disjunct. Let $\mathrm{L}^{\delta}_{\mathrm{PL}}$ be the *PL*-sentences that contain only proposition letters occurring in δ. Let $r(\mathrm{L}^{\delta}_{\mathrm{PL}})$ be a set of 2^{2^n} mutually non-equivalent representatives of $\mathrm{L}^{\delta}_{\mathrm{PL}}$, where n is the number of proposition letters in δ. By writing *PL*-sentences (including t and k) we now mean their unique representatives in $r(\mathrm{L}^{\delta}_{\mathrm{PL}})$. We construct a set Δ such that:

(a) Any conjunct of δ is in Δ.

(b) For all $B, C \in r(\mathrm{L}^{\delta}_{\mathrm{PL}})$:
- either $P^F(B/C)$ or $O^F(\neg B/C) \in \Delta$, and
- either $P^S(B/C)$ or $O^S(\neg B/C) \in \Delta$.

(c) Δ is DDFS-consistent.

It then suffices to find a set $I \subseteq \mathrm{L_{PL}}$ that satisfies all $B \in \mathcal{D}$. – The proof follows the completeness proof of Spohn [24] for B. Hansson's [9] preference-based dyadic deontic logic *DSDL3,* and I will remark on the parallels as they arise.

Definition 1. *For any* $C \in r(\mathrm{L}^{\delta}_{\mathrm{PL}})$, *let*

- $\mathcal{O}^S_C = \bigwedge\{A \in r(\mathrm{L}^{\delta}_{\mathrm{PL}}) \mid \mathrm{O^S(A/C)} \in \Delta\}$,
- $\mathbb{O}^F_C = min\,\{A \in r(\mathrm{L}^{\delta}_{\mathrm{PL}}) \mid \mathrm{O^F(A/C)} \in \Delta\}$.

where $min\,\Gamma = \{A \in \Gamma \mid \forall B \in \Gamma, \textit{if } \vdash_{PL} B \to A \textit{ then } \vdash_{PL} B \leftrightarrow A\}$, $\Gamma \subseteq \mathrm{L_{PL}}$.

We have $O^S(t/C) \in \Delta$ due to (DN^S) and *DDFS*-consistency of Δ, so $\mathcal{O}^S_C$ is well defined for any C. From the definitions of Δ, $\mathcal{O}^S_C$, and $\mathbb{O}^F_C$, we obtain:

(L1) $O^S(A/C) \in \Delta$ iff $\vdash_{PL} \mathcal{O}^S_C \to A$

(L2) $O^F(A/C) \in \Delta$ iff $\exists \mathcal{O}_C \in \mathbb{O}^F_C : \vdash_{PL} \mathcal{O}_C \to A$

Remark 1. Definition 1 identifies syntactically what Hansson called the *deontic basis* (Spohn [24] writes $\widetilde{C}$) in an extension $\|C\|$. Monadic deontic logic has just one basis, dyadic deontic logic usually has one basis for any C, and here there may be several bases $\mathcal{O}_C \in \mathbb{O}^F_C$ which expresses some conflict or predicament in case C. Semantically, we want to identify $\|\mathcal{O}_C\|$ with the set of 'best' states of affairs in $\|C\|$, where the particular standard can be made explicit here as the satisfaction of some maximum of imperative-corresponding sentences $I' \in I \perp \neg C$.

Definition 2. *For any* $A \in r(\mathrm{L}^{\delta}_{\mathrm{PL}})$, *let:*

$$\mathbb{C}_A = max\,\{C \in r(\mathrm{L}^{\delta}_{\mathrm{PL}}) \mid \mathrm{P^F(A/C)} \in \Delta\}$$

where $max\,\Gamma = \{A \in \Gamma \mid \forall B \in \Gamma : \textit{if } \vdash_{PL} A \to B \textit{ then } \vdash_{PL} B \leftrightarrow A\}$, $\Gamma \subseteq \mathrm{L_{PL}}$.

Remark 2. Definition 2 identifies the most general circumstances C in which A is P^F-permitted. As we shall see (L7), for any A there is just one such C (we write $\mathcal{C}_A$) that also has the useful property of owning, for any $\mathcal{O}_A \in \mathbb{O}_A$, some basis $\mathcal{O}_{\mathcal{C}_A} \in \mathbb{O}_{\mathcal{C}_A}$ such that $\vdash_{PL} \mathcal{O}_A \leftrightarrow (A \wedge \mathcal{O}_{\mathcal{C}_A})$ (cf. L9), which in turn means that just these general circumstances need to be considered in the construction of the canonical I. To the same effect, Spohn [24] identifies the most general circumstances by the use of equivalence classes $[A]^{\approx}$ defined via 'permission circles': $A \approx B$ iff B is in some $\{B_1, ..., B_n\} \subseteq r(\mathrm{L}^{\delta}_{\mathrm{PL}})$ such that $P^F(B_1/A), P^F(B_2/B_1), ..., P^F(B_n/B_{n-1}), P^F(A/B_n)$ are in Δ. As can be shown, the set of all such classes is $\{[A]^{\approx} \mid A = \mathcal{C}_B \text{ for some } B \in r(\mathrm{L}^{\delta}_{\mathrm{PL}})\}$.

We prove some observations regarding $\mathbb{C}_A$:

(L3) If $P^F(A/D) \in \Delta$ then there is a $\mathcal{C} \in \mathbb{C}_A : \vdash_{PL} D \rightarrow \mathcal{C}$.

Proof: Immediate from the definition of $\mathbb{C}_A$ and the finiteness of $r(\mathrm{L}^{\delta}_{\mathrm{PL}})$.

(L4) For all $A \in r(\mathrm{L}^{\delta}_{\mathrm{PL}})$: $\mathbb{C}_A \neq \varnothing$.

Proof: $P^F(A/A) \in \Delta$ follows from (L2) and (DD^F), (CExt), so $\mathbb{C}_A \neq \varnothing$ follows from (L3).

(L5) For all $\mathcal{C} \in \mathbb{C}_A$ for some $A \in r(\mathrm{L}^{\delta}_{\mathrm{PL}})$, we have $\mathbb{C}_{\mathcal{C}} = \{\mathcal{C}\}$.

Proof: If $\mathcal{C}' \in \mathbb{C}_{\mathcal{C}}$ then $P^F(A/\mathcal{C}), P^F(\mathcal{C}/\mathcal{C}), P^F(\mathcal{C}/\mathcal{C}') \in \Delta$, using (FH) we obtain $P^F(A/\mathcal{C} \vee \mathcal{C}'), P^F(\mathcal{C}/\mathcal{C} \vee \mathcal{C}') \in \Delta$, so $\mathcal{C} = (\mathcal{C} \vee \mathcal{C}') = \mathcal{C}'$.

(L6) For all $\mathcal{C} \in \mathbb{C}_A$ for some $A \in r(\mathrm{L}^{\delta}_{\mathrm{PL}})$, if $P^F(\mathcal{C}/D) \in \Delta$ then $\vdash_{PL} D \rightarrow \mathcal{C}$.

Proof: If $P^F(\mathcal{C}/D) \in \Delta$ then from $P^F(A/\mathcal{C}) \in \Delta$ and (FH) it follows that $P^F(A/\mathcal{C} \vee D)$, so $\mathcal{C} = (\mathcal{C} \vee D)$, $\vdash_{PL} D \rightarrow \mathcal{C}$.

(L7) For all $A \in r(\mathrm{L}^{\delta}_{\mathrm{PL}})$, there is some $\mathcal{C}_A$ such that $\mathbb{C}_A = \{\mathcal{C}_A\}$.

Proof: $\mathbb{C}_A \neq \varnothing$ (L4). Assume $\mathcal{C}, \mathcal{C}' \in \mathbb{C}_A$. $P^F(A/A) \in \Delta$, so $\vdash_{PL} A \rightarrow \mathcal{C}$ (L3). $P^F(A/\mathcal{C}'), P^F(A/\mathcal{C}) \in \Delta$ by definition, $P^F(\mathcal{C}/\mathcal{C}') \in \Delta$ by use of (DM),(CExt), so we get $P^F(A/\mathcal{C} \vee \mathcal{C}') \in \Delta$ from (FH), so $\mathcal{C} = (\mathcal{C} \vee \mathcal{C}') = \mathcal{C}'$.

(L8) For all $A \in r(\mathrm{L}^{\delta}_{\mathrm{PL}})$: $\vdash_{PL} \mathcal{O}^S_A \leftrightarrow (A \wedge \mathcal{O}^S_{\mathcal{C}_A})$.

Proof: We have $\vdash_{PL} A \rightarrow \mathcal{C}_A$ and $O^S(\mathcal{O}^S_A/A) \in \Delta$, so with (Up) we obtain $O^S(A \rightarrow \mathcal{O}^S_A/\mathcal{C}_A) \in \Delta$. So $\vdash_{PL} (A \wedge \mathcal{O}^S_{\mathcal{C}_A}) \rightarrow \mathcal{O}^S_A$ which is the right-to-left direction. For the left-to-right direction, $\vdash_{PL} \mathcal{O}^S_A \rightarrow A$ follows from (CExt), and from $O^S(\mathcal{O}^S_{\mathcal{C}_A}/\mathcal{C}_A), P^F(A/\mathcal{C}_A) \in \Delta$ we get $O^S(\mathcal{O}^S_{\mathcal{C}_A}/A) \in \Delta$ with (RM). So $\vdash_{PL} \mathcal{O}^S_A \rightarrow (A \wedge \mathcal{O}^S_{\mathcal{C}_A})$.

(L9) For all $A \in r(\mathrm{L}^{\delta}_{\mathrm{PL}})$, $\mathcal{O}_A \in \mathbb{O}^F_A$: $\exists \mathcal{O}_{\mathcal{C}_A} \in \mathbb{O}^F_{\mathcal{C}_A} : \vdash_{PL} \mathcal{O}_A \leftrightarrow (A \wedge \mathcal{O}_{\mathcal{C}_A})$.

Proof: Let $\mathcal{O}_A \in \mathbb{O}^F_A$, so $O^F(\mathcal{O}_A/A) \in \Delta$. $\vdash_{PL} A \rightarrow \mathcal{C}_A$, with (Up) we get $O^F(A \rightarrow \mathcal{O}_A/\mathcal{C}_A) \in \Delta$, and so there is some $\mathcal{O}_{\mathcal{C}_A} \in \mathbb{O}^F_{\mathcal{C}_A}$ with $\vdash_{PL} (A \wedge \mathcal{O}_{\mathcal{C}_A}) \rightarrow \mathcal{O}_A$. Assume $P^F(\neg\mathcal{O}_{\mathcal{C}_A}/A) \in \Delta$. From $O^F(\mathcal{O}_{\mathcal{C}_A}/\mathcal{C}_A) \in \Delta$ and (Down1) we get $O^F(\mathcal{O}_{\mathcal{C}_A} \wedge \neg A/\mathcal{C}_A)$ and $O^F(\neg A/\mathcal{C}_A) \in \Delta$, which contradicts that $P^F(A/\mathcal{C}_A) \in \Delta$ by the definition of $\mathbb{C}_A$. So $O^F(\mathcal{O}_{\mathcal{C}_A}/A) \in \Delta$, and $O^F(A \wedge \mathcal{O}_{\mathcal{C}_A}/A) \in \Delta$ by (CExt). Since $\vdash_{PL} (A \wedge \mathcal{O}_{\mathcal{C}_A}) \rightarrow \mathcal{O}_A$ we then have $\vdash_{PL} \mathcal{O}_A \leftrightarrow (A \wedge \mathcal{O}_{\mathcal{C}_A})$ from the minimality of $\mathcal{O}_A$.

Definition 3. *Let* $\mathbb{C} = \{\mathcal{C} \in r(\mathrm{L}^{\delta}_{\mathrm{PL}}) \mid \mathcal{C} \in \mathbb{C}_{\mathrm{A}}) \textit{ for some } A \in r(\mathrm{L}^{\delta}_{\mathrm{PL}})\}$.

Remark 3. If this were 'ordinary' dyadic deontic logic with agglomeration and so just one basis $\mathcal{O}_C$ for any C, we would be almost done: Like Spohn [24] orders his equivalence classes $[C]^{\approx}$ by a relation *before,* $\mathbb{C}$ could be ordered into $\langle \mathcal{C}_1, ..., \mathcal{C}_n \rangle$ with $\mathcal{C}_1 = t$, $\mathcal{C}_{i+1} = \mathcal{C}_i \wedge \neg\mathcal{O}_{\mathcal{C}_i}$, and $\mathcal{C}_n = k$. $S = \langle S_1, ..., S_n \rangle$ with $S_i = (\mathcal{C}_i \wedge \neg\mathcal{C}_{i+1})$, $1 \leq i < n$, is then the 'system of spheres', and $v \succeq v'$ iff $v \in S_i$, $v' \in S_j$, $i \leq j$ defines the corresponding preference relation. – Here, no sphere $\mathcal{C} \in \mathbb{C}$ is guaranteed to have a single basis. But as it turns out, $\mathbb{C}$ has the structure of a 'multiple' system of spheres that is similarly identified.

The following observations hold for any $\mathcal{C} \in \mathbb{C}$, $\mathcal{O}_{\mathcal{C}} \in \mathbb{O}^F_{\mathcal{C}}$, $D \in r(\mathrm{L}^{\delta}_{\mathrm{PL}})$:

(L10) If $\{\mathcal{C} \to \mathcal{O}_{\mathcal{C}}\} \cup \{D\} \nvdash k$ then $O^F(\mathcal{C} \to \mathcal{O}_{\mathcal{C}}/D) \in \Delta$.

Proof: Assume $\{\mathcal{C} \to \mathcal{O}_{\mathcal{C}}\} \cup \{D\} \nvdash k$. If $O^F(\neg\mathcal{C}/D) \in \Delta$ then the conclusion holds trivially. Otherwise $P^F(\mathcal{C}/D) \in \Delta$, so $\vdash_{PL} D \to \mathcal{C}$ by (L6). For r.a.a. suppose $P^F(\neg\mathcal{O}_{\mathcal{C}}/D) \in \Delta$. With $O^F(\mathcal{O}_{\mathcal{C}}/\mathcal{C}) \in \Delta$ we obtain $O^F(\mathcal{O}_{\mathcal{C}} \wedge \neg D/\mathcal{C}) \in \Delta$ by (Down1), and $\vdash_{PL} \mathcal{O}_{\mathcal{C}} \to \neg D$ by minimality of $\mathcal{O}_{\mathcal{C}}$. But then $\vdash_{PL} D \to (\mathcal{C} \wedge \neg\mathcal{O}_{\mathcal{C}})$, which refutes the assumption. So instead $O^F(\mathcal{O}_{\mathcal{C}}/D) \in \Delta$ and $O^F(\mathcal{C} \to \mathcal{O}_{\mathcal{C}}/D) \in \Delta$ by use of (CExt).

(L11) $\{t, k\} \subseteq \mathbb{C}$

Proof: $P^F(t/t) \in \Delta$ by (DD^F), and $\vdash_{PL} C \to t$ for any $P^F(t/C) \in \Delta$, so $t \in \mathbb{C}_t$, $t \in \mathbb{C}$. Concerning k, $P^F(k/k) \in \Delta$ by (DD^F), and due to (DN^F) $C = k$ for any C such that $P^F(k/C) \in \Delta$, so $k \in \mathbb{C}_k$, $k \in \mathbb{C}$.

(L12) $\mathcal{C}_{\mathcal{C}\wedge\neg\mathcal{O}_{\mathcal{C}}} = \mathcal{C} \wedge \neg\mathcal{O}_{\mathcal{C}}$.

Proof: The right-to-left direction is obvious from (DD^F), (CExt), and (L3). For the left-to-right direction, suppose $\nvdash_{PL} \mathcal{C}_{t\wedge\neg\mathcal{O}_{\mathcal{C}}} \to (\mathcal{C} \wedge \neg\mathcal{O}_{\mathcal{C}})$, so $\{\mathcal{C} \to \mathcal{O}_{\mathcal{C}}\} \cup \{\mathcal{C}_{\mathcal{C}\wedge\neg\mathcal{O}_{\mathcal{C}}}\} \nvdash_{PL} k$. We obtain $O^F(\mathcal{C} \to \mathcal{O}_{\mathcal{C}}/\mathcal{C}_{\mathcal{C}\wedge\neg\mathcal{O}_{\mathcal{C}}}) \in \Delta$ by application of (L10), but $P^F(\mathcal{C} \wedge \neg\mathcal{O}_{\mathcal{C}}/\mathcal{C}_{\mathcal{C}\wedge\neg\mathcal{O}_{\mathcal{C}}}) \in \Delta$ by the construction of $\mathbb{C}_{\mathcal{C}\wedge\neg\mathcal{O}_{\mathcal{C}}}$ so Δ is *DDFS*-inconsistent, but we assumed otherwise.

(L13) If $\vdash_{PL} \mathcal{C} \to D$, then $\mathcal{C} = D$ or $\vdash_{PL} \mathcal{C} \to (D \wedge \neg\mathcal{O}_D)$ for some $\mathcal{O}_D \in \mathbb{O}^F_D$.

Proof: Either $P^F(\mathcal{C}/D) \in \Delta$, so $\vdash_{PL} D \to \mathcal{C}$, $\mathcal{C} = D$ (L6). Or $O^F(\neg\mathcal{C}/D) \in \Delta$, so $\vdash_{PL} \mathcal{O}_D \to \neg\mathcal{C}$ for some $\mathcal{O}_D \in \mathbb{O}^F_D$, and $\vdash_{PL} \mathcal{C} \to (D \wedge \neg\mathcal{O}_D)$.

(L14) $D \neq (D \wedge \neg\mathcal{O}_D)$, and if $D \neq k$ then $D \neq (D \wedge \neg\mathcal{O}^S_D)$.

Proof: If $D = (D \wedge \neg\mathcal{O}_D)$ then $\vdash_{PL} D \to \neg\mathcal{O}_D$. But also $\vdash_{PL} \mathcal{O}_D \to D$ due to (CExt), so $\mathcal{O}_D = k$ and $O^F(k/D) \in \Delta$ by (L2). But $P^F(t/D) \in \Delta$ by (DD^F), so with (DDef) this contradicts *DDFS*-consistency of Δ. $D \neq (D \wedge \neg\mathcal{O}^S_D)$, $D \neq k$, is proved likewise by use of (DD^S).

(L15) Let $\mathbb{C}^*$ be such that (i) $t \in \mathbb{C}^*$, and (ii) for any $\mathcal{C}^* \in \mathbb{C}^*$, $\mathcal{O}_{\mathcal{C}^*} \in \mathbb{O}^F_{\mathcal{C}^*}$: $(\mathcal{C}^* \wedge \neg\mathcal{O}_{\mathcal{C}^*}) \in \mathbb{C}^*$. Then $\mathbb{C} = \mathbb{C}^*$.

Proof: $\mathbb{C}^* \subseteq \mathbb{C}$ is immediate from (L11), (L12). As for $\mathbb{C} \subseteq \mathbb{C}^*$, for each $\mathcal{C} \in \mathbb{C}$ there is some $\mathcal{C}^* \in \mathbb{C}^*$ such that (a) $\vdash_{PL} \mathcal{C} \to \mathcal{C}^*$, and (b) for no $\mathcal{O}_{\mathcal{C}^*} \in \mathbb{O}^F_{\mathcal{C}^*}$:$\vdash_{PL} \mathcal{C} \to (\mathcal{C}^* \wedge \neg\mathcal{O}_{\mathcal{C}^*})$. (a) is guaranteed by $t \in \mathbb{C}^*$, and (b) follows from (L13), (L14), and the finiteness of $r(\mathrm{L}^{\delta}_{\mathrm{PL}})$.

Definition 4 (Canonical Construction). *For any* $\mathcal{C} \in \mathbb{C}, D \in r(\mathrm{L}^{\delta}_{\mathrm{PL}})$, *let*

- $SUCC(\mathcal{C}) = \{\mathcal{C}' \in \mathbb{C} \mid \exists \mathcal{O}_{\mathcal{C}} \in \mathbb{O}^F_{\mathcal{C}} : \mathcal{C}' = (\mathcal{C} \wedge \neg \mathcal{O}_{\mathcal{C}})\}$,
- $F\text{-}CHAIN(\mathcal{C})$ *be the set of all* $\langle \mathcal{C}_1, ..., \mathcal{C}_n \rangle$ *such that* $\mathcal{C}_1 = t$, $\mathcal{C}_n = \mathcal{C}$, *and for any* i *with* $1 \leq i < n$, $\mathcal{C}_{i+1} \in SUCC(\mathcal{C}_i)$,
- $S\text{-}CHAIN(\mathcal{C}, D)$ *be the set of all* $\langle D_1, ..., D_n \rangle$ *such that* $D_1 = t$, $D_n = D$, $\langle D_1, ..., D_k \rangle \in F\text{-}CHAIN(\mathcal{C})$, $1 \leq k < n$, $D_n \neq D_{n-1}$, *and for any* i *with* $k \leq i < n$, $D_{i+1} = D_i \wedge \neg \mathcal{O}^S_D$.

For any $\mathcal{C} \in \mathbb{C}$, $\mathcal{C}' \in SUCC(\mathcal{C})$, *let*

- $\pi : \mathbb{C} \to [Prop - \mathrm{L}^{\delta}_{\mathrm{PL}}]$ *be a function that associates a unique proposition letter not occurring in* δ *with each element of* $\mathbb{C}$,
- $\phi(\mathcal{C}, \mathcal{C}') = \pi(\mathcal{C}') \wedge \bigwedge\{\neg\pi(\mathcal{C}'') \mid \mathcal{C}'' \in SUCC(\mathcal{C}), \mathcal{C}' \neq \mathcal{C}''\}$,
- $\sigma(\mathcal{C}) = \bigwedge\{\neg\pi(\mathcal{C}') \mid \mathcal{C}' \in SUCC(\mathcal{C})\}$.

For any $\mathcal{C} \in \mathbb{C}$, $\mathcal{C} \neq t$, $ch(\mathcal{C}) = \langle \mathcal{C}_1, ..., \mathcal{C}_n \rangle \in F\text{-}CHAIN(\mathcal{C})$, *let*

- $i^F[ch(\mathcal{C})] = \neg\, \mathcal{C} \wedge \bigwedge_{i=1}^{n-1} \phi(\mathcal{C}_i, \mathcal{C}_{i+1})$.

For any $\mathcal{C} \in \mathbb{C}$, $ch(\mathcal{C}, D) = \langle D_1, ..., D_n \rangle \in S\text{-}CHAIN(\mathcal{C}, D)$, $D_k = \mathcal{C}$, *let*

- $i^S[ch(\mathcal{C}, D)] = \neg D \wedge \begin{cases} \sigma(\mathcal{C}) \wedge \bigwedge_{i=1}^{k-1} \phi(\mathcal{C}_i, \mathcal{C}_{i+1}) & \text{if } \mathcal{C} \neq t, \\ \sigma(\mathcal{C}) & \text{otherwise.} \end{cases}$

Finally, let

- $I^F = \{i^F[ch(\mathcal{C})] \mid \mathcal{C} \in \mathbb{C}, \mathcal{C} \neq k, ch(\mathcal{C}) \in F\text{-}CHAIN(\mathcal{C})\}$,
- $I^S = \{i^S[ch(\mathcal{C}, D)] \mid \mathcal{C} \in \mathbb{C}, ch(\mathcal{C}, D) \in S\text{-}CHAIN(\mathcal{C}, D)\}$,
- $I = I^F \cup I^S$.

Remark 4. Definition 4 gives the construction tools and the construction of the canonical set I that makes all of Δ true. $SUCC(\mathcal{C})$ is the set of immediate 'contrary-to-duty' successors $\mathcal{C}'$ of $\mathcal{C}$, i.e. there is some basis $\mathcal{O}_{\mathcal{C}} \in \mathbb{O}^F_{\mathcal{C}}$ with $\mathcal{C}' = \mathcal{C} \wedge \neg \mathcal{O}_{\mathcal{C}}$. As (L15) showed, each $\mathcal{C} \in \mathbb{C}$ is a successor of (a successor of ...) t, and $F\text{-}CHAIN(\mathcal{C})$ is the set of all such nestings beginning with t and ending with $\mathcal{C}$. The label ϕ is used to make any two $i^F[ch(\mathcal{C}')]$, $i^F[ch(\mathcal{C}'')]$, $\mathcal{C}'$ and $\mathcal{C}''$ being successors of (successors of...) $\mathcal{C}$, inconsistent with each other and with any $i^S[ch(\mathcal{C}, D)]$ via σ. As $\mathbb{C}$ is finite, so is the number of proposition letters introduced by π, and hence are I^F, I^S and I.

Lemma 1 (Properties of the Construction). *For all* $\mathcal{C}, \mathcal{C}' \in \mathbb{C}, D \in r(\mathrm{L}^{\delta}_{\mathrm{PL}})$,

a) if $\langle D_1, ..., D_n \rangle \in F\text{-}CHAIN(\mathcal{C})$ *or* $S\text{-}CHAIN(\mathcal{C}, D)$ *then* $\vdash_{PL} \mathcal{C}_{i+1} \to \mathcal{C}_i$ *and* $\nvdash_{PL} \mathcal{C}_i \to \mathcal{C}_{i+1}$, $1 \leq i < n$,

b) $\{i^F[ch(\mathcal{C})], i^F[ch(\mathcal{C}')]\} \vdash_{PL} k$ *or* $ch(\mathcal{C})$ *is a segment of* $ch(\mathcal{C}')$ *or vice versa,*

c) $\{i^S[ch(\mathcal{C}, D)], i^S[ch(\mathcal{C}', D')]\} \vdash_{PL} k$ *or* $ch(\mathcal{C}, D) = \langle D_1, ..., D_i \rangle$ *is a segment of* $ch(\mathcal{C}', D') = \langle D_1, ..., D_n \rangle$, $D_k = \mathcal{C} = \mathcal{C}'$, $k < i \leq n$, *or vice versa,*

d) $\{i^F[ch(\mathcal{C})], i^S[ch(\mathcal{C}', D)]\} \vdash_{PL} k$ *or* $ch(\mathcal{C}) = \langle \mathcal{C}_1, ..., \mathcal{C}_i \rangle$ *is a segment of* $\langle D_1, ..., D_n \rangle$, $D_k = \mathcal{C}'$, $1 < i \leq k < n$,

e) no $i^F[ch(\mathcal{C})]$, $i^S[ch(\mathcal{C}, D)] \in I$ *is a contradiction.*

Proof. a) follows from the constructions of *F-CHAIN* and *S-CHAIN* and (L14). b)-c) follow from the definitions of ϕ and σ. For d), note that each $i \in I$ consists of a $r(\mathrm{L}^{\delta}_{\mathrm{PL}})$-conjunct and a $[\mathrm{L}_{\mathrm{PL}} - \mathrm{L}^{\delta}_{\mathrm{PL}}]$-conjunct. For $i^F[ch(\mathcal{C})]$, the $r(\mathrm{L}^{\delta}_{\mathrm{PL}})$-conjunct is $\neg\mathcal{C}$ which is consistent since $\mathcal{C} = t$ is excluded. For $i^S[ch(\mathcal{C}, D)]$, the $r(\mathrm{L}^{\delta}_{\mathrm{PL}})$-conjunct is $\neg D$, so suppose $D = t$. Let $ch(\mathcal{C}, D) = \langle D_1, ..., D_n\rangle$: due to the construction $n \neq 1$, but then $D_1 = D_n = t$ contradicts a). For the $[\mathrm{L}_{\mathrm{PL}} - \mathrm{L}^{\delta}_{\mathrm{PL}}]$-conjuncts of $i^F[ch(\mathcal{C})]$, let $ch(\mathcal{C}) = \langle \mathcal{C}_1, ..., \mathcal{C}_n\rangle \in \text{F-}CHAIN(\mathcal{C})$:

- No conjunct $\phi(\mathcal{C}_i, \mathcal{C}_{i+1})$, $1 \leq i < n$, is a contradiction: For any $\mathcal{C}', \mathcal{C}'' \in \mathbb{C}$, $\pi(\mathcal{C}') \neq \pi(\mathcal{C}'')$, and no $\pi(\mathcal{C}')$ occurs negated and unnegated in $\phi(\mathcal{C}_i, \mathcal{C}_{i+1})$.
- If $\pi(\mathcal{C}')$ occurs unnegated in $\phi(\mathcal{C}_i, \mathcal{C}_{i+1})$ and negated in $\phi(\mathcal{C}_j, \mathcal{C}_{j+1})$, $i < j$, then $\mathcal{C}' = \mathcal{C}_{i+1}$ and $\mathcal{C}' \in SUCC(\mathcal{C}_j)$. So there is a $ch(\mathcal{C}') \in \text{F-}CHAIN(\mathcal{C}')$, $ch(\mathcal{C}') = \langle \mathcal{C}_1, ..., \mathcal{C}_i, \mathcal{C}', \mathcal{C}_{i+2}, ..., \mathcal{C}_j, \mathcal{C}'\rangle$. which violates a).
- If $\pi(\mathcal{C}')$ occurs negated in $\phi(\mathcal{C}_i, \mathcal{C}_{i+1})$, and unnegated in $\phi(\mathcal{C}_j, \mathcal{C}_{j+1})$, $i < j$, then $\mathcal{C}' \in SUCC(\mathcal{C}_i)$ and $\mathcal{C}' = \mathcal{C}_{j+1}$. From $\vdash_{PL} \mathcal{C}_{j+1} \rightarrow \mathcal{C}_{i+1}$ we obtain $\vdash_{PL} \mathcal{C}' \rightarrow \mathcal{C}_{i+1}$. So there are $\mathcal{O}_{\mathcal{C}_i}, \mathcal{O}^*_{\mathcal{C}_i} \in \mathbb{O}^F_{\mathcal{C}_i}$ with $\mathcal{C}' = \mathcal{C}_i \wedge \neg\mathcal{O}_{\mathcal{C}_i}$, $\mathcal{C}_{i+1} = \mathcal{C}_i \wedge \neg\mathcal{O}^*_{\mathcal{C}_i}$, and $\vdash_{PL} (\mathcal{C}_i \wedge \neg\mathcal{O}_{\mathcal{C}_i}) \rightarrow (\mathcal{C}_i \wedge \neg\mathcal{O}^*_{\mathcal{C}_i})$. Then $\vdash_{PL} \mathcal{O}^*_{\mathcal{C}_i} \rightarrow (\mathcal{C}_i \rightarrow \mathcal{O}_{\mathcal{C}_i})$, and with (CExt) $\vdash_{PL} \mathcal{O}^*_{\mathcal{C}_i} \rightarrow \mathcal{O}_{\mathcal{C}_i}$. By minimality $\mathcal{O}^*_{\mathcal{C}_i} = \mathcal{O}_{\mathcal{C}_i}$ and $\mathcal{C}' = \mathcal{C}_{i+1}$, but $\phi(\mathcal{C}_i, \mathcal{C}_{i+1})$ left $\pi(\mathcal{C}_{i+1})$ unnegated.

For the $[\mathrm{L}_{\mathrm{PL}} - \mathrm{L}^{\delta}_{\mathrm{PL}}]$-conjuncts of $i^S[ch(\mathcal{C}, D)]$, the case that $\pi(\mathcal{C}')$ occurs unnegated in $\phi(\mathcal{C}_i, \mathcal{C}_{i+1})$ and negated in $\sigma(\mathcal{C})$ is done like the second case above.

Lemma 2 ('Coincidence Lemma'). *For all* $A, B \in r(\mathrm{L}^{\delta}_{\mathrm{PL}})$*:*

$$\begin{array}{lll} (a) & I \models O^F(A/B) & \textit{iff} \quad O^F(A/B) \in \Delta \\ (b) & I \models P^F(A/B) & \textit{iff} \quad P^F(A/B) \in \Delta \\ (c) & I \models O^S(A/B) & \textit{iff} \quad O^F(A/B) \in \Delta \\ (d) & I \models P^S(A/B) & \textit{iff} \quad P^F(A/B) \in \Delta \end{array}$$

Proof. I give the right-to-left directions only, the others hold due to (DDef).

Case a) Assume $O^F(A/B) \in \Delta$, so some $\mathcal{O}_B \in \mathbb{O}^F_B$ derives A. By (L9) there is a $\mathcal{C}_B \in \mathbb{C}$, $\mathcal{O}_{\mathcal{C}_B} \in \mathbb{O}^F_{\mathcal{C}_B}$ such that $\vdash_{PL} ((\mathcal{C}_B \rightarrow \mathcal{O}_{\mathcal{C}_B}) \wedge B) \leftrightarrow \mathcal{O}_B$. By (L12) $(\mathcal{C}_B \wedge \neg\mathcal{O}_{\mathcal{C}_B}) \in \mathbb{C}$, so for some $ch(\mathcal{C}_B \wedge \neg\mathcal{O}_{\mathcal{C}_B}) \in \text{F-}CHAIN(\mathcal{C}_B \wedge \neg\mathcal{O}_{\mathcal{C}_B})$ we have $i^F[ch(\mathcal{C}_B \wedge \neg\mathcal{O}_{\mathcal{C}_B})] \in I$. If $\{i^F[ch(\mathcal{C}_B \wedge \neg\mathcal{O}_{\mathcal{C}_B})], B\} \vdash_{PL} k$ this must be due to its $r(\mathrm{L}^{\delta}_{\mathrm{PL}})$-conjunct $\mathcal{C}_B \rightarrow \mathcal{O}_{\mathcal{C}_B}$, since the others are consistent (Lemma 1 e) and not relevant for a derivation of -B. But if $\{\mathcal{C}_B \rightarrow \mathcal{O}_{\mathcal{C}_B}, B\} \vdash_{PL} k$ then $\mathcal{O}_B = k$ which contradicts $P^F(t/B) \in \Delta$ by (DD^F) and the DDFS-consistency of Δ. So $\{i^F[ch(\mathcal{C}_B \wedge \neg\mathcal{O}_{\mathcal{C}_B})], B\} \nvdash_{PL} k$, so for some $I' \in I\perp\neg B$: $I' \cup \{B\} \vdash_{PL} A$.

Case b) Assume $P^F(A/B) \in \Delta$ and for r.a.a. suppose that there is some $I' \in I\perp\neg B : I' \cup \{B\} \vdash_{PL} \neg A$. If $I' \cap I^F \neq \varnothing$, then there is a $i^F[ch(\mathcal{C})] \in I'$ such that $ch(\mathcal{C})$ is an initial segment of any $ch(\mathcal{C}')$ or $ch(\mathcal{C}', D)$ with $i^F[ch(\mathcal{C}')]$ or $i^S[ch(\mathcal{C}', D)] \in I'$ (Lemma 1 b, d). Concerning the $r(\mathrm{L}^{\delta}_{\mathrm{PL}})$-conjuncts $\neg\mathcal{C}$ of $i^F[ch(\mathcal{C})]$ and $\neg D$ of any other $i \in I'$, we have $\vdash_{PL} \neg\mathcal{C} \rightarrow \neg D$ by Lemma 1 a), so $\{\neg\mathcal{C}\} \cup \{B\} \vdash_{PL} \neg A$ since no $[\mathrm{L}_{\mathrm{PL}} - \mathrm{L}^{\delta}_{\mathrm{PL}}]$-conjunct is relevant, $\mathcal{C} \neq t$ by the construction, so $\neg\mathcal{C} = (\mathcal{C}'' \rightarrow \mathcal{O}_{\mathcal{C}''})$ for some $\mathcal{C}'' \in \mathbb{C}$. $\{\mathcal{C}'' \rightarrow \mathcal{O}_{\mathcal{C}''}\} \cup \{B\} \nvdash_{PL} k$ since else $i^F[ch(\mathcal{C})]$ could not be in I', so $\exists\mathcal{O}_B \in \mathbb{O}^F_B : \vdash_{PL} \mathcal{O}_B \rightarrow (\mathcal{C}'' \rightarrow \mathcal{O}_{\mathcal{C}''})$ due to (L10). Since $\vdash_{PL} \mathcal{O}_B \rightarrow B$ we obtain $\vdash_{PL} \mathcal{O}_B \rightarrow \neg A$, $O^F(\neg A/B) \in \Delta$, so Δ is *DDFS*-inconsistent. Hence $I^F \cap I' = \varnothing$. If $I^S \cap I' \neq \varnothing$

then there is a $i^S[ch(\mathcal{C},D)] \in I'$ such that $ch(\mathcal{C},D)$ is a segment of any $ch(\mathcal{C}',D')$ with $i^S[ch(\mathcal{C}',D')] \in I'$ (Lemma 1c). Regarding the $r(\mathrm{L}^{\delta}_{\mathrm{PL}})$-conjuncts $\neg D$ of $i^S[ch(\mathcal{C},D)]$ and $\neg D'$ of any other $i \in I'$, we have $\vdash_{PL} \neg D \rightarrow \neg D'$ by Lemma 1a), so again $\{\neg D\} \cup \{B\} \vdash_{PL} \neg A$ since no $[\mathrm{L}_{\mathrm{PL}} - \mathrm{L}^{\delta}_{\mathrm{PL}}]$-conjunct is relevant, $ch(\mathcal{C},D)$ is $\langle D_1,...,D_n\rangle$, $n \geq 1$, $\neg D = D_{n-1} \rightarrow \mathcal{O}^S_{D_{n-1}}$. If $D_{n-1} = \mathcal{C}$ then either $\mathcal{C} = t$, then $\vdash_{PL} B \rightarrow D_{n-1}$, or there is a $i^F[ch(\mathcal{C})] \in I^F$ such that its $r(\mathrm{L}^{\delta}_{\mathrm{PL}})$-conjunct $\neg\mathcal{C}$ derives $\neg D$ (Lemma 1a), and its $[\mathrm{L}_{\mathrm{PL}} - \mathrm{L}^{\delta}_{\mathrm{PL}}]$-conjunct is derived by that of $i^S[ch(\mathcal{C},D)]$. So if $i^F[ch(\mathcal{C})] \notin I'$ then $\{\neg\mathcal{C}\} \cup \{B\} \vdash_{PL} k$ and $\vdash_{PL} B \rightarrow D_{n-1}$. If $D_{n-1} \neq \mathcal{C}$ then there is a $i^S[ch(\mathcal{C},D_{n-1})] \notin I'$ from which $\vdash_{PL} B \rightarrow D_{n-1}$ is similarly obtained. So $O^S(B \rightarrow \neg A/D_{n-1}) \in \Delta$. $P^F(A/B) \in \Delta$ derives $P^F(\neg(B \rightarrow \neg A)/B) \in \Delta$, so $O^S((B \rightarrow \neg A) \wedge \neg B/D_{n-1}) \in \Delta$ due to (Down3). Hence $\vdash_{PL} \mathcal{O}^S_{D_{n-1}} \rightarrow \neg B$, $\vdash_{PL} B \rightarrow D$, so $i^S[ch(\mathcal{C},D)] \notin I'$, and $I' \cap I^S = \varnothing$. So $I = \varnothing$ and $\{B\} \vdash_{PL} \neg A$. With $P^F(A/B) \in \Delta$ and (CExt) we get $P^F(k/B)$, so $D = k$ due to (DN^F). But then $I \perp \neg B = \varnothing$ which completes the r.a.a.

Case c) Assume $O^S(A/B) \in \Delta$, and for r.a.a. suppose that there is some $I' \in I \perp \neg B : I' \cup \{B\} \nvdash A$. Suppose $I^F \cap I' \neq \varnothing$, so of some $i^F[ch(\mathcal{C})] \in I'$, $ch(\mathcal{C})$ is an initial segment of any $ch(\mathcal{C}')$ or $ch(\mathcal{C}',D)$ with $i^F[ch(\mathcal{C}')]$ or $i^S[ch(\mathcal{C}',D)] \in I'$ (Lemma 1 b, d). $\mathcal{C} = (\mathcal{C}'' \wedge \neg\mathcal{O}_{\mathcal{C}''})$ for some $\mathcal{C}'' \in \mathbb{C}$, $\mathcal{O}_{\mathcal{C}''} \in \mathbb{O}^F_{\mathcal{C}''}$. We have $\vdash_{PL} B \rightarrow \mathcal{C}''$: This is trivial if $\mathcal{C}'' = t$, otherwise there is a $i^F[ch(\mathcal{C}'')] \in I^F$ such that $ch(\mathcal{C}'')$ is an initial segment of $ch(\mathcal{C})$. The $r(\mathrm{L}^{\delta}_{\mathrm{PL}})$-conjunct of $i^F[ch(\mathcal{C}'')]$ is $\neg\mathcal{C}''$, which derives any $r(\mathrm{L}^{\delta}_{\mathrm{PL}})$-conjunct of $i^*[ch(\mathcal{C}')] \in I'$ (Lemma 1a). The $[\mathrm{L}_{\mathrm{PL}} - \mathrm{L}^{\delta}_{\mathrm{PL}}]$-conjuncts of $i^{F'}[ch(\mathcal{C}'')]$ derive from any such conjuncts of $i^F[ch(\mathcal{C}')]$ or $i^S[ch(\mathcal{C}',D)]$ by the construction of ϕ, σ, so if $i^F[ch(\mathcal{C}'')] \notin I'$ then $\{\neg\mathcal{C}''\} \cup \{B\} \vdash_{PL} k$, so $\vdash_{PL} B \rightarrow \mathcal{C}''$. $I' \cup \{B\} \vdash_{PL} \mathcal{C}'' \rightarrow \mathcal{O}_{\mathcal{C}''}$, so $I' \cup \{B\} \vdash_{PL} \mathcal{O}_{\mathcal{C}''}$. By (DC^{FS}) and the minimality of $\mathcal{O}_{\mathcal{C}''}$, $\vdash_{PL} \mathcal{O}_{\mathcal{C}''} \rightarrow O^S_{\mathcal{C}''}$, so $\vdash_{PL} \mathcal{O}^S_{\mathcal{C}''} \rightarrow (B \rightarrow \mathcal{O}^S_B)$ by (Up). So $I' \cup \{B\} \vdash_{PL} \mathcal{O}^S_B$ and $I' \cup \{B\} \vdash_{PL} A$, refuting the assumption. So $I^F \cap I' = \varnothing$. Suppose $I^S \cap I' \neq \varnothing$, so there is a $i^S[ch(\mathcal{C},D)] \in I'$ such that $ch(\mathcal{C},D)$ is a segment of any $ch(\mathcal{C}',D')$ with $i^S[ch(\mathcal{C}',D')] \in I'$ (Lemma 1c). $ch(\mathcal{C},D) = \langle D_1,...,D_n\rangle$, $n \geq 1$, $\{i^S[ch(\mathcal{C},D)]\} \vdash_{PL} \neg D$ and $\neg D = D_{n-1} \rightarrow \mathcal{O}^S_{D_{n-1}}$. Like in the previous case we prove that $\vdash_{PL} B \rightarrow D_{n-1}$, so $I' \cup \{B\} \vdash_{PL} \mathcal{O}^S_{D_{n-1}}$. By use of (Up) $\vdash_{PL} \mathcal{O}^S_{D_{n-1}} \rightarrow (B \rightarrow \mathcal{O}^S_B)$, but $O^S(A/B) \in \Delta$, so $I' \cup \{B\} \vdash_{PL} A$, refuting the assumption. So $I' \cap I^S = \varnothing$. For any $B \in r(\mathrm{L}^{\delta}_{\mathrm{PL}})$, there is a $\mathcal{C}_B \in \mathbb{C}$, and for any $B \neq k$ a $i^S[ch(\mathcal{C}_B,(\mathcal{C}_B \wedge \neg\mathcal{O}^S_{\mathcal{C}_B}))] \in I^S$. If $I' \cap I^S = \varnothing$ then $\{i^S[ch(\mathcal{C}_B,(\mathcal{C}_B \wedge \neg\mathcal{O}^S_{\mathcal{C}_B}))]\} \cup \{B\} \vdash_{PL} k$, and $\{\mathcal{C}_B \rightarrow \mathcal{O}^S_{\mathcal{C}_B}\} \cup \{B\} \vdash_{PL} k$ since only the $r(\mathrm{L}^{\delta}_{\mathrm{PL}})$-conjunct is relevant. So $\vdash_{PL} B \rightarrow \mathcal{O}^S_{\mathcal{C}_B}$, $\mathcal{O}^S_B = k$ by (L8), and $D = k$ due to (DD^S). But then $I \perp \neg B = \varnothing$ which completes the r.a.a.

Case d) Assume $P^S(A/B) \in \Delta$. $B \neq k$ due to (DN^S) and (CExt). Then $i^S[ch(\mathcal{C}_B,(\mathcal{C}_B \wedge \neg\mathcal{O}^S_{\mathcal{C}_B}))] \in I^S$ is in some $I' \in I \perp \neg B$ for else $B = k$ (see above). If $i^F[ch(\mathcal{C}')] \in I'$ then $ch(\mathcal{C}')$ is a segment of $ch(\mathcal{C}_B,(\mathcal{C}_B \wedge \neg\mathcal{O}^S_{\mathcal{C}_B}))$ (Lemma 1d), so $\vdash_{PL} \neg\mathcal{C}' \rightarrow \neg\mathcal{C}_B$, and since $\vdash_{PL} B \rightarrow \mathcal{C}_B$ also $\vdash_{PL} \neg\mathcal{C}' \rightarrow \neg B$. So $i^F[ch(\mathcal{C}')] \notin I'$ and $I' \cap I^F = \varnothing$. Suppose $i^S[ch(\mathcal{C}',D')] \in I'$, then $ch(\mathcal{C}_B,(\mathcal{C}_B \wedge \neg\mathcal{O}^S_{\mathcal{C}_B}))$ is the initial segment of $ch(\mathcal{C}',D')$ (Lemma 1c), so its $r(\mathrm{L}^{\delta}_{\mathrm{PL}})$-conjunct $\neg D'$ is derived by that of $i^S[ch(\mathcal{C}_B,(\mathcal{C}_B \wedge \neg\mathcal{O}^S_{\mathcal{C}_B}))]$. No other conjuncts are relevant, so if $I' \cup B \vdash_{PL} A$ then $\{\mathcal{C}_B \rightarrow \mathcal{O}^S_{\mathcal{C}_B})\} \cup \{B\} \vdash_{PL} A$, but then by (L8) $\vdash_{PL} \mathcal{O}^S_B \rightarrow A$, so $O^S(A/B) \in \Delta$ and Δ is inconsistent. So $I' \cup B \nvdash_{PL} A$.

Corollary 1. *Let* $\mathbb{P}$ *be a non-empty set of preference relations* $P \subseteq \boldsymbol{B} \times \boldsymbol{B}$ *such that each P is transitive, connected, and satisfies* (LA^{L})*: if* $\|A\| \neq \varnothing$ *then* $best_P(A) \neq \varnothing$*, where* $best_P(A) = \{v \in \|A\| \mid \forall v' \in \|A\| : vPv'\}$.

Let $\mathrm{L}_{\mathrm{DDF+S}}$ *be like* $\mathrm{L}_{\mathrm{DDFS}}$ *except that* O^{F+}, P^{F+} *replace* O^F, P^F*, and let the truth definitions for the deontic operators read:*

$\mathbb{P} \models O^{F+}(A/C)$ *iff* $\exists P \in \mathbb{P} : best_P(C) \subseteq \|A\|$
$\mathbb{P} \models P^{F+}(A/C)$ *iff* $\forall P \in \mathbb{P} : best_P(C) \cap \|A\| \neq \varnothing$
$\mathbb{P} \models O^{S}(A/C)$ *iff* $\forall P \in \mathbb{P} : best_P(C) \subseteq \|A\|$
$\mathbb{P} \models P^{S}(A/C)$ *iff* $\exists P \in \mathbb{P} : best_P(C) \cap \|A\| \neq \varnothing$

Let DDF^+S *be like DDFS except that* (DN^F) *and* (DD^F) *are replaced by*

(DN^{F+}) $O^{F+}(t/C)$
(DD^{F+}) *If* $\nvdash_{PL} \neg C$ *then* $\vdash_{DDF+S} P^{F+}(t/C)$

Then DDF^+S *is sound and (weakly) complete with respect to the above semantics.*

Proof (Sketch). To prove *soundness,* Arrow's axiom: $best_P(C \vee D) \cap \|C\| = \varnothing$ or $best_P(C) = best_P(C \vee D) \cap \|C\|$, is helpful. For weak *completeness,* use translations f^- and f^+, where $f^- : \mathrm{L}_{DDF^+S} \to \mathrm{L}_{DDFS}$ just replaces any occurrence of $O^{F+}(A/C)$ with $(P^F(k/C) \vee O^F(A/C))$ and of $P^{F+}(A/C)$ with $(O^F(t/C) \wedge P^F(A/C))$, and $f^+ : \mathrm{L}_{DDFS} \to \mathrm{L}_{DDF^+S}$ replaces any occurrence of $O^F(A/C)$ with $(P^F(t/C) \wedge O^{F+}(A/C))$ and of $P^F(A/C)$ with $(O^{F+}(k/C) \vee P^{F+}(A/C))$ (cf. [2] § 28). Now prove that

a) if $\vdash_{DDFS} A$ then $\vdash_{DDF^+S} f^+(A)$
b) if $\vdash_{DDF^+S} A$ then $\vdash_{DDFS} f^-(A)$
c) $\vdash_{DDF^+S} A$ iff $\vdash_{DDF^+S} f^+(f^-(A))$
d) $\vdash_{DDFS} A$ iff $\vdash_{DDFS} f^-(f^+(A))$

To prove that if $\models_{DDF^+S} A$ then $\vdash_{DDF^+S} A$, suppose $\nvdash_{DDF^+S} A$, so $\nvdash_{DDF^+S} f^+(f^-(A))$ by c), so $\nvdash_{DDFS} f^-(A)$ by a). Repeat the construction for the canonical set *I* as described in Def. 4. Let $\langle D_1, ..., D_n\rangle \in F\text{-}CHAIN(k)$ or $S\text{-}CHAIN(\mathcal{C}, k)$, $\mathcal{C} \in \mathbb{C}$, and let $S_{\langle D_1,...,D_n\rangle} = \langle S_1, ..., S_{n-1}\rangle$ where $S_i = (C_i \wedge \neg C_{i+1})$, $1 \leq i < n$. For any such $\langle D_1, ..., D_n\rangle$, define $P_{\langle D_1,...,D_n\rangle} \subseteq \mathbf{B} \times \mathbf{B}$ by

$vP_{\langle D_1,...,D_n\rangle}v'$ iff $v \in \|S_i\|$, $v' \in \|S_j\|$, and $i \leq j$,

and let $\mathbb{P}$ be the set of all such relations. Due to (L11), at least $P_{\langle t,k\rangle} \in \mathbb{P}$, so $\mathbb{P} \neq \varnothing$. By use (Lemma 1) it can be easily verified that each $v \in \mathbf{B}$ belongs to exactly one sphere S_i in $S_{\langle D_1,...,D_n\rangle} = \langle S_1, ..., S_{n-1}\rangle$, $1 \leq i < n$, and so $P_{\langle D_1,...,D_n\rangle} \in \mathbb{P}$ is transitive and connected and satisfies $(\mathrm{LA}^{\mathrm{L}})$. Finally

$\mathbb{P} \models O^{F+}(A/B)$ iff $f^-(O^{F+}(A/B)) \in \Delta$
$\mathbb{P} \models O^{S}(A/B)$ iff $O^S(A/B) \in \Delta$

is proved by appealing to the construction of the canonical set *I* and Lemma 2.

Remark 5. The corollary exploits the notorious parallels to Spohn's completeness proof for B. Hansson's *DSDL3.* The two deontic operators are interpreted by Hansson-type truth definitions which validate the characteristic theorems $O^*(A/A)$ for both. This is the main difference to the multiplex preference models devised by Goble [7]: there the truth of deontic operators also depends on some or all members of a non-empty set of preferences, but the truth definitions are Danielsson-type (cf. [2] p. 219).

5 Conclusion

It has been shown that the truth definitions for the monadic deontic operators O^F and O^S which were devised by van Fraassen and Horty to deal with openly conflicting norms, can be adjusted to also cover predicaments that arise in sub-ideal situations; the thus defined dyadic operators then characterize the dyadic deontic logic *DDFS*. This result may also be found interesting from the perspective of belief dynamics, since e.g. $O^F(A/C)$ is true with respect to a set I iff A derives from a maxichoice contraction of I by $\neg C$ expanded by C, and (somewhat) similar for $O^S(A/C)$ and full meet contraction. It is, insofar as I know, a new result: In particular the equivalences proven by Rott [23] between operators of theory change, systems of nonmonotonic reasoning, and choice theory only cover nonmonotonic systems that include agglomeration. – I have briefly commented above on the problem of 'consistent aggregation' for van Fraassen's monadic operator O^F (also cf. the bimodal extension in my [8]), and must leave it to future examination if and how this problem can be solved in a dyadic context. Missing in the present account are also 'proper' conditional imperatives that are only included in a set of actualized imperatives (used to define some score) if their condition is 'triggered' (cf. van Fraassen's intermediate definition in [28])[8]. To examine how the present contrary-to-duty conditionals combine with others expressing 'proper' conditional obligations must be left to future study.

References

1. Åqvist, L.: "Some Results on Dyadic Deontic Logic and the Logic of Preference", *Synthese,* **66,** 1986, 95–110.
2. Åqvist, L.: *Introduction to Deontic Logic and the Theory of Normative Systems,* Naples: Bibliopolis, 1987.
3. Alchourrón, C. E. and Bulygin, E.: "Unvollständigkeit, Widersprüchlichkeit und Unbestimmtheit der Normenordnungen", in Conte, A. G., Hilpinen, R., von Wright, G.H.: *Deontische Logik und Semantik,* Wiesbaden: Athenaion, 1977, 20–32.
4. Brink, D. O.: "Moral Conflict and Its Structure", *Philosophical Review* **103,** 1994, 215–247.
5. Donagan, A.: "Consistency in Rationalist Moral Systems", *Journal of Philosophy,* **81**, 1984, 291–309.
6. Føllesdal, D. and Hilpinen, R.: "Deontic Logic: An Introduction", in [11], 1–35.
7. Goble, L.: "Multiplex Semantics for Deontic Logic", *Nordic Journal of Philosophical Logic,* **5,** 2000, 113-134.
8. Hansen, J.: "Problems and Results for Logics about Imperatives", *Journal of Applied Logic,* to appear.
9. Hansson, B.: "An Analysis of Some Deontic Logics", *Noûs,* **3**, 1969, 373–398. Reprinted in [11], 121–147.
10. Hare, R. M.: *The Language of Morals,* Oxford: University Press, 1952.
11. Hilpinen, R. (ed): *Deontic Logic: Introductory and Systematic Readings,* Dordrecht: Reidel, 1971.

[8] Horty [12], [13], [14] has explored a van Fraasen type operator in a somewhat different setting, which models imperatives as dyadic in structure.

12. Horty, J. F.: "Moral Dilemmas and Nonmonotonic Logic", *Journal of Philosophical Logic,* **23**, 1994, 35–65.
13. Horty, J. F.: "Nonmonotonic Foundations for Deontic Logic", in Nute, D. (ed.): *Defeasible Deontic Logic,* Dordrecht: Kluwer, 1997, 17–44.
14. Horty, J. F.: "Reasoning with Moral Conflicts", *Nôus,* **37,** 2003, 557–605.
15. Jacquette, D.: "Moral Dilemmas, Disjunctive Obligations, and Kant's Principle that 'Ought' implies 'Can'", *Synthese,* **88**, 1991, 43–55.
16. Kelsen, H.: *Reine Rechtslehre,* 2^{nd} ed., Vienna: Deuticke, 1960.
17. Kelsen, H.: *Allgemeine Theorie der Normen,* Vienna: Manz, 1979.
18. Lemmon, E. J.: "Moral Dilemmas", *Philosophical Review,* **71,** 1962, 139–158.
19. Lewis, D.: "Semantic Analyses for Dyadic Deontic Logic", in Stenlund, S. (ed.): *Logical Theory and Semantic Analysis,* Dordrecht: Reidel, 1974, 1–14.
20. Marcus, R. Barcan: "Moral Dilemmas and Consistency", *Journal of Philosophy,* **77**, 1980, 121–136.
21. Rescher, N.: "An Axiom System for Deontic Logic", *Philosophical Studies,* **9,** 1958, 24–30.
22. Ross, W. D.: *The Right and the Good.* Oxford: Clarendon Press, 1930.
23. Rott, H.: *Change, Choice and Inference. A Study of Belief Revision and Nonmonotonic Logic,* Oxford: Clarendon, 2001.
24. Spohn, W.: "An Analysis of Hansson's Dyadic Deontic Logic", *Journal of Philosophical Logic,* **3,** 1975, 237–252.
25. Stenius, E.: "The Principles of a Logic of Normative Systems", *Acta Philosophica Fennica,* **16**, 1963, 247–260.
26. van der Torre, L.: *Reasoning about Obligation,* Amsterdam: Thesis Publishers, 1997.
27. van der Torre, L., Tan, Y.-H.: "Two-Phase Deontic Logic", *Logique & Analyse,* **43,** 2000, 411–456.
28. van Fraassen, B.: "Values and the Heart's Command", *Journal of Philosophy,* **70,** 1973, 5–19.
29. von Wright, G. H.: *Norm and Action,* London: Routledge & Kegan Paul, 1963.
30. von Wright, G.H.: *An Essay in Deontic Logic and the General Theory of Action,* Amsterdam: North-Holland, 1968.
31. von Wright, G. H.: "Deontic Logic – as I See It", in McNamara, P., Prakken, H.: *Norms, Logics and Information Systems* (ΔEON '98), Amsterdam: IOS, 1999, 15–25.
32. Williams, B.: "Ethical Consistency", *Proceedings of the Aristotelian Society,* supp. **39**, 1965, 103–124.

On Obligations and Abilities

Wojciech Jamroga[1], Wiebe van der Hoek[2], and Michael Wooldridge[2]

[1] Parlevink Group, University of Twente, The Netherlands
jamroga@cs.utwente.nl
[2] Department of Computer Science, University of Liverpool, UK
{wiebe,mjw}@csc.liv.ac.uk

Abstract. In this paper, we combine deontic logic with Alternating-time Temporal Logic (ATL) into a framework that makes it possible to model and reason about obligations *and* abilities of agents. The way both frameworks are combined is technically straightforward: we add deontic accessibility relations to ATL models (concurrent game structures), and deontic operators to the language of ATL (an additional operator $\mathcal{UP}$ is proposed for "unconditionally permitted" properties, similar to the "all I know" operator from epistemic logic). Our presentation is rather informal: we focus on examples of how obligations (interpreted as requirements) can be confronted with ways of satisfying them by actors of the game. Though some formal results are presented, the paper should not be regarded as a definite statement on how logics of obligation and strategic ability must be combined; instead, it is intended for stimulating discussion about such kinds of reasoning, and the models that can underpin it.

Keywords: deontic logic, alternating-time logic, multi-agent systems.

1 Introduction

In recent years, there has been increasing interest from within the computer science, logic, and game theory communities with respect to what might be called *cooperation logics:* logics that make it possible to explicitly represent and reason about the strategic abilities of coalitions of agents (human or computational) in game-like multi-agent scenarios. Perhaps the best-known example of such a logic is the Alternating-time Temporal Logic of Alur, Henzinger, and Kupferman [1]. In this paper, we propose a concept of "deontic ATL". As deontic logic focuses on obligatory behaviors of systems and agents, and Alternating-time Temporal Logic enables reasoning about abilities of agents and teams, we believe it interesting and potentially useful to combine these formal tools in order to confront system requirements (i.e., obligations) with possible ways of satisfying them by actors of the game (i.e., abilities). This paper is not intended as a definite statement on how logics of obligation and strategic ability should be combined. Rather, we intend it to stimulate discussion about such kinds of reasoning, and the models that can underlie it.

We begin by presenting the main concepts from both frameworks. Then, in section 2, their combination is defined and discussed. Three different approaches

A. Lomuscio and D. Nute (Eds.): DEON 2004, LNAI 3065, pp. 165–181, 2004.

to modeling obligations in a temporal context are discussed: global requirements on states of the system (i.e., that deem some states "correct" and some "incorrect"), local requirements on states ("correctness" may depend on the current state), and temporal obligations, which refer to paths rather than states. We investigate (in an informal way) the perspectives offered by each of these approaches, and present several interesting properties of agents and systems that can be expressed within their scope. Some preliminary formal results are given in Section 3.

1.1 Deontic Logic: The Logic of Obligations

Deontic logic is a modal logic of obligations [16], expressed with operator $\mathcal{O}\varphi$ ("it is obligatory that φ"). Models for deontic logic were originally defined as Kripke structures with deontic accessibility relation(s) [21]. A state q' such that $q\mathcal{R}q'$ is called a "perfect alternative" of state q (we can also say that q' is *acceptable* or *correct* from the perspective of q). As with the conventional semantics of modal operators we define,

$$M, q \models \mathcal{O}\varphi \text{ iff for all } q' \text{ such that } q\mathcal{R}q' \text{ we have } M, q' \models \varphi.$$

We believe that this stance still makes sense, especially when we treat deontic statements as referring to preservation (or violation) of some constraints one would like to impose on a system or some of its components (such as integrity constraints in a database). In this sense, deontic modalities may refer to *requirements* (specification requirements, design requirements, security requirements etc.), and we will interpret $\mathcal{O}\varphi$ as "φ is required" throughout the rest of the paper. This approach allows to put all *physically* possible states of the system in the scope of the model, and to distinguish the states that are "correct" with respect to some criteria, thus enabling reasoning about possible faults and fault tolerance of the system [22]. However, we will argue that ATL plus deontic logic allows to express obligations about what coalitions should or should not achieve – without specifying *how* they do achieve it (or refrain from it). We consider this issue in detail in Section 2.5.

Let us illustrate our main ideas with the following example. There are two trains: a and b; each can be inside a tunnel (propositions $\mathsf{a\text{-}in}$ and $\mathsf{b\text{-}in}$, respectively) or outside of it. The specification requires that the trains should not be allowed to be in the tunnel at the same time, because they will crash (so the tunnel can be seen as a kind of critical section): $\mathcal{F}(\mathsf{a\text{-}in} \wedge \mathsf{b\text{-}in})$ or, equivalently, $\mathcal{O}\neg(\mathsf{a\text{-}in} \wedge \mathsf{b\text{-}in})$. A model for the whole system is displayed in Figure 1A.

Locality and Individuality of Obligations. Note that the set of perfect alternatives is the same for each state q in the example from Figure 1A. Thus, the semantic representation can in fact be much simpler: it is sufficient to mark the states that *violate* the requirements with a special "violation" atom V [2, 15,14]. Then the accessibility relation $\mathcal{R}$ can be defined as: $q\mathcal{R}q'$ iff $q' \not\models V$. Using a more elaborate accessibility relation machinery makes it possible, in

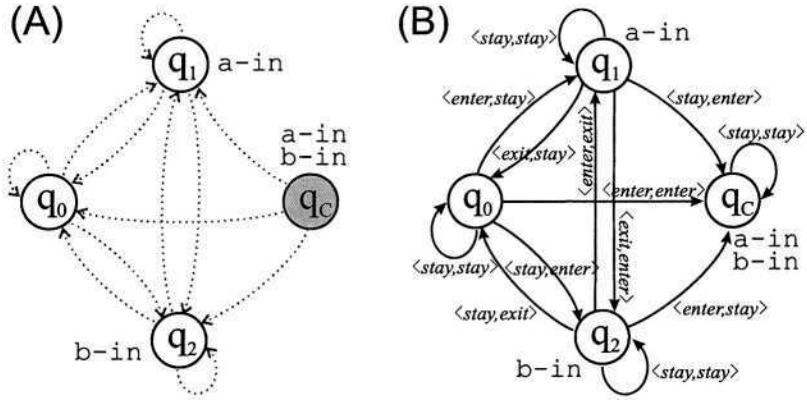


Fig. 1. (A) Critical section example: the trains and the tunnel. Dotted lines display the deontic accessibility relation. (B) The trains revisited: temporal and strategic structure

general, to model requirements that are *local* with respect to the current state. Local obligations can provide a means for specifying requirements that evolve in time. Also, they can be used to specify exception handling in situations when full recovery of the system is impossible (cf. Section 2.3).

Another dimension of classifying obligations is their *individuality*. The accessibility relation can define the requirements for the whole system, or there can be many relations, specifying different requirements for each process or agent [14].

Combining Deontic Perspective with Other Modalities. The combination of deontic logic with temporal and dynamic logics has been investigated at length in the literature [15,20,7,18]. In addition, deontic epistemic logics [5,14] and BOID ("beliefs-obligations-intentions-desires") logics [6] have also been studied. Finally, in [19], deontic and strategic perspectives were combined through applying social laws to ATL.

1.2 Strategic Ability: Alternating-Time Temporal Logic

Alternating-time Temporal Logic (ATL) [1] extends the computation tree logic (CTL) with a class of *cooperation modalities* of the form $\langle\langle A\rangle\rangle$, where A is a set of agents. The intuitive interpretation of $\langle\langle A\rangle\rangle\Phi$ is: "The group of agents A have a collective strategy to enforce Φ no matter what the other agents in the system do". The recursive definition of ATL formulas is:

$$\varphi ::= p \mid \neg\varphi \mid \varphi_1 \vee \varphi_2 \mid \langle\langle A\rangle\rangle X\varphi \mid \langle\langle A\rangle\rangle G\varphi \mid \langle\langle A\rangle\rangle \varphi_1 \mathcal{U} \varphi_2$$

The "sometime" operator F can be defined as: $\langle\langle A\rangle\rangle F\varphi \equiv \langle\langle A\rangle\rangle \top \mathcal{U} \varphi$.

Models and Semantics of ATL. *Concurrent game structures* are transition systems that are based on the collective actions of all agents involved. Formally, a *concurrent game structure* is a tuple $M = \langle \Sigma, Q, \Pi, \pi, Act, d, \delta \rangle$, where:

$\Sigma = \{a_1, ..., a_k\}$ is a (finite) set of all *agents*, Q *is* a non-empty set of *states*, Π is a set of (atomic) *propositions*, and $\pi : Q \to 2^{\Pi}$ is a *valuation* of propositions; *Act* is a set of *actions* (or choices), and $d : Q \times \Sigma \to 2^{Act}$ is a function that returns the decisions available to player a at state q. Finally, a complete tuple of decisions $\langle \alpha_1, ..., \alpha_k \rangle \subseteq d_q(a_1) \times \ldots \times d_q(a_k)$ from all the agents in state q implies a deterministic transition according to the transition function $\delta(q, \alpha_1, ..., \alpha_k)$ [1].

A *strategy* for agent a is a mapping $f_a : Q^+ \to Act$, which assigns a choice $f_a(q_0, ..., q_n) \in d_a(q_n)$ to every non-empty finite sequence of states $q_0, ..., q_n$. Thus, the function specifies a's decisions for every possible (finite) history of system transitions. A *collective strategy* for a set of agents $A \subseteq \Sigma$ is just a tuple of strategies (one for each agent in A): $F_A = \langle f_a \rangle_{a \in A}$. Now, $out(q, F_A)$ denotes the *set of outcomes* of F_A from q, i.e., the set of all (infinite) computations starting from q, in which group A has been using F_A. Let $\Lambda[i]$ denote the ith position in computation Λ. The semantics of ATL formulas follows through the clauses:

$M, q \models \langle\langle A \rangle\rangle X\varphi$ iff there exists a collective strategy F_A such that for all $\Lambda \in out(q, F_A)$ we have $M, \Lambda[1] \models \varphi$;

$M, q \models \langle\langle A \rangle\rangle G\varphi$ iff there exists a collective strategy F_A such that for all $\Lambda \in out(q, F_A)$ we have $M, \Lambda[i] \models \varphi$ for every $i \geq 0$;

$M, q \models \langle\langle A \rangle\rangle \varphi \mathcal{U} \psi$ iff there exists a collective strategy F_A such that for all $\Lambda \in out(q, F_A)$ there is $i \geq 0$ such that $M, \Lambda[i] \models \psi$ and for all j such that $0 \leq j < i$ we have $M, \Lambda[j] \models \varphi$.

Let us consider the tunnel example from a temporal (and strategic) perspective; a concurrent game structure for the trains and the tunnel is shown in Figure 1B. Using ATL, we have that $\langle\langle \Sigma \rangle\rangle F(\mathsf{a\text{-}in} \wedge \mathsf{b\text{-}in})$, so the system is physically able to display undesirable behavior. On the other hand, $\langle\langle a \rangle\rangle G\neg(\mathsf{a\text{-}in} \wedge \mathsf{b\text{-}in})$, i.e., train a can protect the system from violating the requirements. In this paper, we propose to extend ATL with deontic operator $\mathcal{O}$ in order to investigate the interplay between agents' abilities and requirements they should meet.

The Full Logic of ATL*. ATL* generalizes ATL in the same way as CTL* generalizes CTL: we release the syntactic requirement that every occurrence of a temporal operator must be preceded by exactly one occurrence of a cooperation modality. ATL* consists of *state formulas* φ and *path formulas* ψ, defined recursively below:

$$\varphi := p \mid \neg\varphi \mid \varphi_1 \vee \varphi_2 \mid \langle\langle A \rangle\rangle \psi$$
$$\psi := \varphi \mid \neg\psi \mid \psi_1 \vee \psi_2 \mid X\psi \mid \psi_1 \mathcal{U} \psi_2$$

Temporal operators F and G can be defined as: $F\psi \equiv \top \mathcal{U} \psi$ and $G\psi \equiv \neg F \neg \psi$. ATL* has strictly more expressive power than ATL, but it is also more computationally costly. Therefore ATL is more important for practical purposes. For semantics and extensive discussion, we refer the reader to [1].

[1] The definition we use here differs slightly from the original one [1], because we use symbolic labels for agents and their choices (and we do not assume finiteness of Q and *Act*). For an extensive discussion of various ATL semantics, refer to [9].

1.3 STIT Logic: The Logic of Causal Agency

It is also worth mentioning at this point a related body of work, initiated largely through the work of Belnap and Perloff, on "stit" logic – the logic of *seeing to it that* [4,3]. Such logics contain an *agentive* modality, which attempts to capture the idea of an agent *causing* some state of affairs. This modality, typically written $[i \text{ stit } \phi]$, is read as "agent i sees to it that ϕ". The semantics of stit modalities are typically given as $[i \text{ stit } \phi]$ iff i makes a choice c, and ϕ is a necessary consequence of choice c (i.e., ϕ holds in all futures that could arise through i making choice c). A distinction is sometimes made between the "generic" stit modality and the *deliberate* stit modality ("dstit"); the idea is that i deliberately sees to it that ϕ if $[i \text{ stit } \phi]$ and there is at least one future in which ϕ does not hold (the intuition being that i is then making a *deliberate choice* for ϕ, as ϕ would not necessarily hold if i did not make choice c). Such logics are a natural counterpart to deontic logics, as it clearly makes sense to reason about the obligations that an agent has in the context of the choices it makes and the consequences of these choices. Similarly, if we interpret choices as programs (cf. the strategies of ATL), then stit logics are also related to dynamic logic [12]; the main differences are that programs, which are first class entities in the object language of dynamic logic, are not present in the object language of stit logics (and of course, strategies are not present in the object language of ATL). Moreover, stit logics assert that an agent *makes* a particular choice, whereas we have no direct way of expressing this in ATL (or, for that matter, in dynamic logic). So, while stit logics embody somewhat similar concerns to ATL (and dynamic logic), the basic constructs are fundamentally different, providing (yet another) way of interpreting the dynamic choice structures that are common to these languages.

2 Deontic ATL

In this section, we extend ATL with deontic operators. We follow the definition with an informal discussion on how the resulting logic (and its models) can help to investigate the interplay between agents' abilities and requirements that the system (or individual agents) should meet.

2.1 Syntax and Semantics

The combination of deontic logic and ATL proposed here is technically straightforward: the new language consists of both deontic and strategic formulas, and models include the temporal transition function and deontic accessibility relation as two independent layers. Thus, the recursive definition of DATL formulas is:

$$\varphi := p \mid \neg\varphi \mid \varphi_1 \vee \varphi_2 \mid \mathcal{O}_A\varphi \mid \mathcal{UP}_A\varphi \mid \langle\langle A\rangle\rangle X\varphi \mid \langle\langle A\rangle\rangle G\varphi \mid \langle\langle A\rangle\rangle \varphi_1 \mathcal{U} \varphi_2$$

where $A \subseteq \Sigma$ is a set of agents. Models for Deontic ATL can be called *deontic game structures,* and defined as tuples $M = \langle \Sigma, Q, \Pi, \pi, Act, d, \delta, \mathbb{R} \rangle$, where:

- Σ is a (finite) set of all *agents,* and Q is a non-empty set of *states,*
- Π is a set of (atomic) *propositions,* and $\pi : Q \to 2^{\Pi}$ is their *valuation;*
- *Act* is a set of actions, and $d : Q \times \Sigma \to 2^{Act}$ is a function that returns the decisions available to player a at state q;
- a complete tuple of decisions $\langle \alpha_1, ..., \alpha_k \rangle \subseteq d_q(a_1) \times ... \times d_q(a_k)$ from all the agents in state q implies a deterministic transition according to the transition function $\delta(q, \alpha_1, ..., \alpha_k)$;
- finally, $\mathbb{R} : 2^{\Sigma} \to 2^{Q \times Q}$ is a mapping that returns a deontic accessibility relation $\mathcal{R}_A$ for every group of agents A.

The semantic rules for $p, \neg\varphi, \varphi \vee \psi, \langle\langle A \rangle\rangle X\varphi, \langle\langle A \rangle\rangle G\varphi, \langle\langle A \rangle\rangle \varphi \mathcal{U} \psi$ are inherited from the semantics of ATL (cf. Section 1.2), and the truth of $\mathcal{O}_A\varphi$ is defined below. We also propose a new deontic operator: $\mathcal{UP}\varphi$, meaning that "φ is unconditionally permitted", i.e., whenever φ holds, we are on the correct side of the picture (which closely resembles the "only knowing"/"all I know" operator from epistemic logic [13]).

$$M, q \models \mathcal{O}_A\varphi \text{ iff for every } q' \text{ such that } q\mathcal{R}_A q' \text{ we have } M, q' \models \varphi;$$
$$M, q \models \mathcal{UP}_A\varphi \text{ iff for every } q' \text{ such that } M, q' \models \varphi \text{ we have } q\mathcal{R}_A q'.$$

This new operator – among other things – will help to characterize the *exact* set of "correct" states, especially in the case of local requirements, where the property of a state being "correct" depends on the current state of the system.

In principle, it should be possible that the requirements on a group of agents (or processes) are independent from the requirements for the individual members of the group (or its subgroups). Thus, we will not assume any specific relationship between relations $\mathcal{R}_A$ and $\mathcal{R}_{A'}$, even if $A' \subseteq A$. We propose only that a system can be identified with the complete group of its processes, and therefore the requirements on a system as a whole can be defined as: $\mathcal{O}\varphi \equiv \mathcal{O}_\Sigma\varphi$. In a similar way: $\mathcal{UP}\varphi \equiv \mathcal{UP}_\Sigma\varphi$.

2.2 Dealing with Global Requirements

Let us first consider the simplest case, i.e., when the distinction between "good" and "bad" states is global and does not depend on the current state. Deontic game structures can in this case be reduced to concurrent game structures with "violation" atom V that holds in the states that violate requirements. Then:

$$M, q \models \mathcal{O}\varphi \text{ iff for all } q' \text{ such that } q' \not\models V \text{ we have } M, q' \models \varphi.$$

As we have both requirements and abilities in one framework, we can look at the former and then ask about the latter. Consider the trains and tunnel example from Figure 1B, augmented with the requirements from Figure 1A (let us also assume that these requirements apply to all the agents and their groups, i.e., $\mathcal{R}_A = \mathcal{R}_{A'}$ for all $A, A' \subseteq \Sigma$; we will continue to assume so throughout the rest of the paper, unless explicitly stated). As already proposed, the trains are required not to be in the tunnel at the same moment, because it would result in a crash:

$\mathcal{O}(\neg(\mathsf{a\text{-}in} \wedge \mathsf{b\text{-}in}))$. Thus, it is natural to ask whether some agent or team can prevent the trains from crashing: $\langle\langle A\rangle\rangle G\neg(\mathsf{a\text{-}in} \wedge \mathsf{b\text{-}in})$? Indeed, it turns out that both trains have this ability: $\langle\langle a\rangle\rangle G\neg(\mathsf{a\text{-}in}\wedge\mathsf{b\text{-}in}) \wedge \langle\langle b\rangle\rangle G\neg(\mathsf{a\text{-}in}\wedge\mathsf{b\text{-}in})$. On the other hand, if the goal of a train implies that it passes the tunnel, the train is unable to "safeguard" the system requirements any more: $\neg\langle\langle a\rangle\rangle\neg(\mathsf{a\text{-}in}\wedge\mathsf{b\text{-}in})\,\mathcal{U}(\mathsf{a\text{-}in}\wedge\neg\mathsf{b\text{-}in})$.

In many cases, it may be interesting to consider questions like: does an agent have a strategy to always/eventually fulfill the requirements? Or, more generally: does the agent have a strategy to achieve his goal in the way that does not violate the requirements (or so that he can recover from the violation of requirements eventually)? We try to list several relevant properties of systems and agents below:

1. the system is *stable* (with respect to model M and state q) if $M, q \models \langle\langle\varnothing\rangle\rangle G\neg V$, i.e., no agent (process) can make it crash;
2. the system is *semi-stable* (with respect to model M and state q) if it will inevitably recover from any future situation: $M, q \models \langle\langle\varnothing\rangle\rangle G\langle\langle\varnothing\rangle\rangle F\neg V$;
3. agents A form a (collective) *guardian* in model M at state q if they can protect the system from any violation of the requirements: $M, q \models \langle\langle A\rangle\rangle G\neg V$;
4. A can *repair the system* in model M at state q if $M, q \models \langle\langle A\rangle\rangle F\neg V$;
5. A is a (collective) *repairman* in model M at state q if A can always repair the system: $M, q \models \langle\langle\varnothing\rangle\rangle G\langle\langle A\rangle\rangle F\neg V$;
6. finally, another (perhaps the most interesting) property is agents' ability to eventually achieve their goal (φ) without violating the requirements. We say that agents A can *properly enforce* φ in M, q if $M, q \models \langle\langle A\rangle\rangle(\neg V)\,\mathcal{U}(\neg V \wedge \varphi)$.

We will illustrate the properties with the following example. The world is in danger, and only the Prime Minister (p) can save it through giving a speech at the United Nations session and revealing the dangerous plot that threatens the world's future. However, there is a killer (k) somewhere around who tries to murder him before he presents his speech. The Prime Minister can be hidden in a bunker (proposition pbunk), moving through the city (pcity), presenting the speech ($\mathsf{pspeaks} \equiv \mathsf{saved}$), or... well... dead after being murdered (pdead). Fortunately, the Minister is assisted by James Bond (b) who can search the killer out and destroy him (we are very sorry – we would prefer Bond to arrest the killer rather than do away with him, but Bond hardly works this way...). The deontic game structure for this problem is shown in Figure 2. The Prime Minister's actions have self-explanatory labels (*enter, exit, speak* and *nop* for "no operation" or "do nothing"). James Bond can defend the Minister (action *defend*), look for the killer *(search)* or stay idle *(nop);* the killer can either shoot at the Minister *(shoot)* or wait *(nop).* The Minister is completely safe in the bunker (he remains alive regardless of other agents' choices). He is more vulnerable in the city (can be killed unless Bond is defending him at the very moment), and highly vulnerable while speaking at the UN (the killer can shoot him to death even if Bond is defending him). James Bond can search out and destroy the killer in a while (at any moment). It is required that the world is saveable ($\mathcal{O}\langle\langle\Sigma\rangle\rangle F\mathsf{saved}$) and this is the only requirement ($\mathcal{UP}\langle\langle\Sigma\rangle\rangle F\mathsf{saved}$). Note also that the world can be saved if, and only if, the Prime Minister is alive

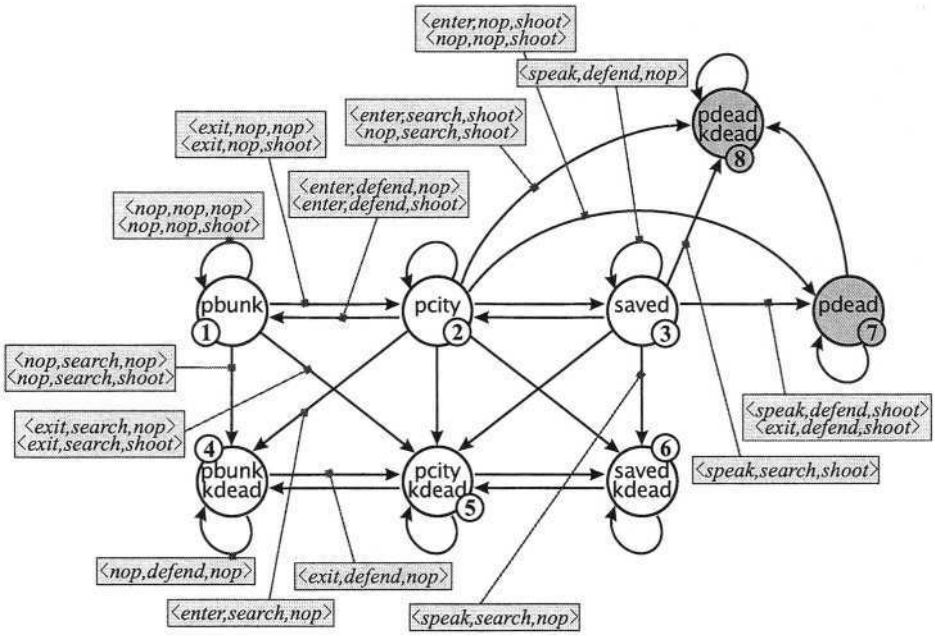


Fig. 2. James Bond saves the world. The arrows show possible transitions of the system; some of the labels are omitted to improve readability. The states that violate the requirements are marked grey

$(\langle\langle\Sigma\rangle\rangle F\mathsf{saved} \equiv \neg\mathsf{pdead})$, and the two states that violate this requirement are marked accordingly $(V \equiv \mathsf{pdead})$.

The system is neither stable nor semi-stable (the Minister can go to the UN building and get killed, after which the system has no way of recovering). Likewise, no agent can repair the system in states q_7, q_8, and hence there is no repairman. The Prime Minister is a guardian as long as he is in the bunker, because he can stay in the bunker forever: $\mathsf{pbunk} \rightarrow \langle\langle p\rangle\rangle G\neg\mathsf{pdead}$. However, if he does so, he cannot save the world: $\neg\langle\langle p\rangle\rangle(\neg\mathsf{pdead})\,\mathcal{U}(\neg\mathsf{pdead} \wedge \mathsf{saved})$. On the other hand, he can cooperate with Bond to properly save the world as long as he is initially out of the UN building: $(\mathsf{pbunk} \vee \mathsf{pcity}) \rightarrow \langle\langle p, b\rangle\rangle(\neg\mathsf{pdead})\,\mathcal{U}(\neg\mathsf{pdead} \wedge \mathsf{saved})$ – he can get to the bunker, defended by Bond, and then wait there until Bond finds the killer; then he can go out to present his speech. Incidentally, there is one more guardian in the system – namely, the killer: $(\neg\mathsf{pdead}) \rightarrow \langle\langle k\rangle\rangle G\neg\mathsf{pdead}$, and also $(\neg\mathsf{pdead}) \rightarrow \langle\langle p, k\rangle\rangle(\neg\mathsf{pdead})\,\mathcal{U}(\neg\mathsf{pdead} \wedge \mathsf{saved})$, so the Minister can alternatively pay the killer instead of employing Bond.

2.3 Local Requirements with Deontic ATL

A more sophisticated deontic-accessibility relation may be convenient for modeling dynamics of obligations, for instance when the actors of the game can negotiate the requirements (e.g., deadlines for a conference submission). Alter-

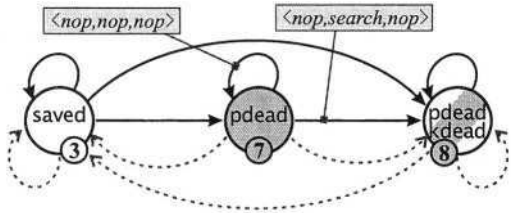


Fig. 3. "James Bond saves the world" revisited: local requirements. Dotted lines define the deontic accessibility relation. Solid lines show possible transitions of the system

natively, "localized" requirements can give a way of specifying *exception handling* in situations when a full recovery is impossible.

Consider the modified "James Bond" example from Figure 3. The Prime Minister is alive initially, and it is required that he should be protected from being shot: $q_3 \models \neg\mathsf{pdead}$ and $q_3 \models \mathcal{O}\neg\mathsf{pdead}$. On the other hand, nobody except the killer can prevent the murder: $q_3 \models \langle\langle k\rangle\rangle G\neg\mathsf{pdead} \wedge \neg\langle\langle p, b\rangle\rangle G\neg\mathsf{pdead}$; moreover, when the president is dead, there is no way for him to become alive again $(\mathsf{pdead} \rightarrow \langle\langle\varnothing\rangle\rangle G\mathsf{pdead})$. Now, when the Minister is shot, a new requirement is implemented, namely it is required that either the Minister is resurrected or the killer is eliminated: $q_7 \models \mathcal{O}(\neg\mathsf{pdead} \vee \mathsf{kdead})$. Fortunately, Bond can bring about the latter: $q_7 \models \langle\langle b\rangle\rangle F\mathsf{kdead}$. Note that q_8 is unacceptable when the Minister is alive (q_3), but it becomes the only option when he has already been shot (q_7)[2].

Similar properties of agents and systems to the ones from the previous section can be specified:

1. the system is *stable* in M, q if, given $M, q \models \mathcal{O}p \wedge \mathcal{U}\mathcal{P}p$, we have $M, q \models \langle\langle\varnothing\rangle\rangle Gp$;
2. the system is *semi-stable* in M, q if, given that $M, q \models \mathcal{O}p \wedge \mathcal{U}\mathcal{P}p$, we have $M, q \models \langle\langle\varnothing\rangle\rangle G(p \rightarrow \langle\langle\varnothing\rangle\rangle Fp)$;
3. A form a *guardian* in M, q if, given $M, q \models \mathcal{O}p \wedge \mathcal{U}\mathcal{P}p$, we have $M, q \models \langle\langle A\rangle\rangle Gp$;
4. A can *repair* the system in M, q if, given that $M, q \models \mathcal{O}p \wedge \mathcal{U}\mathcal{P}p$, we have $M, q \models \langle\langle A\rangle\rangle Fp$;
5. group A is a *repairman* in M, q if, given that $M, q \models \mathcal{O}p \wedge \mathcal{U}\mathcal{P}p$, we have $M, q \models \langle\langle\varnothing\rangle\rangle G\langle\langle A\rangle\rangle Fp$;

6a. A can *properly enforce* φ in M, q if, given that $M, q \models \mathcal{O}_A p \wedge \mathcal{U}\mathcal{P}_A p$, we have $M, q \models \langle\langle A\rangle\rangle p\mathcal{U}(p \wedge \varphi)$. Note that this requirement is individualized now;

6b. A can *properly (incrementally) enforce* φ in M, q if, given that $M, q \models \mathcal{O}_A p \wedge \mathcal{U}\mathcal{P}_A p$, we have $M, q \models p \wedge \varphi$, or $M, q \models p$ and A have a collective strategy F_A such that for every $\lambda \in out(q, F_A)$ they can properly (incrementally) enforce φ in $M, \lambda[1]$.

[2] In a way, we are making the deontic accessibility relation "serial" in a very special sense, i.e., every state has at least one *reachable* perfect alternative now.

The definitions show that many interesting properties, combining deontic and strategic aspects of systems, can be defined using semantic notions. However, at present, we do not see how they can be specified entirely in the object language.

2.4 Temporal Requirements

Many requirements have a temporal flavor, and the full language of ATL* allows to express properties of temporal paths as well. Hence, it makes sense to look at DATL*, where one specifies deontic temporal properties in terms of correct computations (rather than single states). In its simplest version, we obtain DTATL by only allowing requirements over temporal (path) subformulas that can occur within formulas of ATL:

$$\varphi := p \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid \langle\langle A\rangle\rangle\psi \mid \mathcal{O}_A\psi \mid \mathcal{UP}_A\psi$$

with the path subformulas ψ defined recursively as

$$\psi := X\varphi \mid G\varphi \mid \varphi_1 \mathcal{U} \varphi_2 \qquad (\text{where } \varphi \in \mathsf{DTATL}).$$

Properties that can be expressed in this framework are, for instance, that $\mathcal{O}F\langle\langle\Gamma\rangle\rangle G\varphi$ (it is required that sometime in the future, coalition Γ gets the opportunity to guarantee φ forever) and $\mathcal{O}F(\langle\langle\Gamma\rangle\rangle F\varphi \wedge \langle\langle\Gamma\rangle\rangle F\neg\varphi)$ (it is a requirement that eventually coalition Γ can determine φ). The latter can be strengthened to

$$\mathcal{O}G(\langle\langle\Gamma\rangle\rangle F\varphi \wedge \langle\langle\Gamma\rangle\rangle F\neg\varphi)$$

saying that it is an obligation of the system that there must always be opportunities for Γ to toggle φ as it wants. Note that the definition of DTATL straightforwardly allows to express stability properties like

$$\mathcal{O}T\psi \rightarrow \langle\langle\Gamma\rangle\rangle T\psi$$

saying that Γ can bring about the temporal requirement $T\psi$.

Semantically, rather than being a relation between states, relation $\mathcal{R}_A$ is now one between states and computations (sequences of states). Thus, for any computation λ, $q\mathcal{R}_A\lambda$ means that λ is an ideal computation, given q. The semantics of temporal obligations and unconditional permissions can be defined as:

$M, q \models \mathcal{O}_A X\varphi$ iff for every λ such that $q\mathcal{R}_A\lambda$, we have $M, \lambda[1] \models \varphi$;
$M, q \models \mathcal{O}_A G\varphi$ iff for each λ such that $q\mathcal{R}_A\lambda$, we have $M, \lambda[i] \models \varphi$ for all $i \geq 0$;
$M, q \models \mathcal{O}_A\varphi\mathcal{U}\psi$ iff for every λ such that $q\mathcal{R}_A\lambda$, there is $i \geq 0$ such that $M, \lambda[i] \models \psi$ and for all $0 \leq j < i$ we have $M, \lambda[j] \models \varphi$.

$M, q \models \mathcal{UP}_A X\varphi$ iff for every λ such that $M, \lambda[1] \models \varphi$, we have $q\mathcal{R}_A\lambda$;
$M, q \models \mathcal{UP}_A G\varphi$ iff for every λ such that $M, \lambda[i] \models \varphi$ for all $i \geq 0$, we have $q\mathcal{R}_A\lambda$;
$M, q \models \mathcal{UP}_A\varphi\mathcal{U}\psi$ iff for every λ, such that $M, \lambda[i] \models \psi$ for some $i \geq 0$ and $M, \lambda[j] \models \varphi$ for all $0 \leq j < i$, we have $q\mathcal{R}_A\lambda$.

One of the most appealing temporal constraints is that of a deadline: some property φ should be achieved within a number (say n) of steps. This could be just expressed by $\mathcal{O}X^n\varphi$ [3]: only these courses of action are acceptable, in which the deadline is met. Note that the DATL obligation $\mathcal{O}(\langle\langle\Gamma\rangle\rangle X)^n\varphi$ expresses a different property: these are Γ who *must be able* to meet the deadline.

Fairness-like properties are also a very natural area to reason about deontic constraints. Suppose we have a resource p that can only be used by one agent at the time (and as long as a is using it, p_a is true). The constraint that every agent should always be able to use the resource is expressed by $\bigwedge_{a\in\Sigma}\mathcal{O}G\langle\langle a\rangle\rangle Gp_a$ – or, if this is an obligation of a particular scheduler s, we could write $\mathcal{O}_s$ rather than $\mathcal{O}$. Finally, let $[\![\Gamma]\!]\Phi$ be the shorthand for $\neg\langle\langle\Gamma\rangle\rangle\neg\Phi$ (coalition Γ cannot prevent φ from being the case). Then, formula $\mathcal{O}G(\langle\langle\Gamma\rangle\rangle F\varphi \rightarrow [\![\Gamma]\!]G(\varphi \rightarrow \langle\langle\Gamma'\rangle\rangle F\neg\varphi))$ says that only these courses of action are acceptable in which, might coalition Γ ever have a way to enforce φ, then it must "pass the token" to Γ' and give the other agents the ability to reverse this again.

Note also that DTATL formulas $\mathcal{U}\mathcal{P}\psi$ express a kind of "the end justifies means" properties. For instance, $\mathcal{U}\mathcal{P}$ Fkdead means that *every* course of action, which yields the killer dead, is acceptable.

2.5 Deontic ATL and Social Laws

We mentioned the two main streams in deontic logic, having either states of affairs or actions as their object of constraints. In Deontic ATL, one can express deontic requirements about *who is responsible* to achieve something, without specifying how it should be achieved. The requirement $\mathcal{O}\neg\langle\langle\{a,b\}\rangle\rangle F$safe-open, for example, states that it should be impossible for a and b to bring about the disclosure of a safe in a bank. However, with c being a third employee, we might have $\mathcal{O}(\neg\langle\langle\{a,b\}\rangle\rangle F\text{safe-open} \wedge \langle\langle\{a,b,c\}\rangle\rangle G\text{safe-open})$: as a team of three, they *must* be able to do so! We can also express delegation, as in $\mathcal{O}_a\langle\langle b\rangle\rangle G\varphi$: authority a has the obligation that b can always bring about φ.

A recent paper [19] also addresses the issue of prescribed behavior in the context of ATL: behavioral constraints (specific model updates) are defined for ATL models, so that some objective can be satisfied in the updated model. The emphasis in [19] is on how the effectiveness, feasibility and synthesis problems in the area of social laws [18] can be posed as ATL model checking problems. One of the main questions addressed is: given a concurrent game structure M and a social law with objective φ (which we can loosely translate as $\mathcal{O}\varphi$), can we modify the original structure M into M', such that M' satisfies $\langle\langle\emptyset\rangle\rangle G\varphi$? In other words, we ask whether the overall system can be altered in such a way that it cannot but satisfy the requirements. [19] does not address the question whether certain coalitions are *able* to "act according to the law"; the law is *imposed* on the system as a whole. Thus, the approach of that paper is prescriptive, while our approach in this paper is rather descriptive. Moreover, [19] lacks explicit deontic notions in the object level.

[3] $\mathcal{O}X^n\varphi$ is not a DTATL formula, but the logic can be easily extended to include it.

An example of a requirement that cannot be imposed on the system as a whole (taken from [19]) is $p \wedge \langle\langle A \rangle\rangle X \neg p$: property p is obligatory, but at the same time, A should be able to achieve $\neg p$. This kind of constraints could be used to model "a-typical" situations, (such as: "it is obligatory that the emergency exit is not used, although at the same time people in the building should always be able to use it"). Putting such an overall constraint upon a system S means that S should both guarantee p and the possibility of deviating from it, which is impossible. It seems that our Deontic ATL covers a more local notion of obligation, in which $\mathcal{O}(p \wedge \langle\langle A \rangle\rangle X \neg p)$ can well be covered in a non-trivial way.

On the other hand, our "stability" requirements are rather weak: to demand that every obligation $\mathcal{O}\varphi$ is implementable by a coalition does not yet guarantee that the system *does* behave well. Rather, we might be looking for something in between the universal guarantee and a coalitional efficiency with respect to constraint φ. And it is one of the features of Deontic ATL – that one can express many various stability requirements, making explicit who is responsible for what.

3 Axioms, Model Checking and Similar Stories

Let ATL and DL be the languages for ATL and deontic logic, respectively, and let $\mathcal{ATL}$ and $\mathcal{DL}$ be their respective semantic structures. Then – if we do not have any mixing axioms relating the coalitional and the deontic operators – we obtain a logic DATL = ATL $\oplus$ DL which can be called an *independent combination* of the modal logics in question [8]. [8] gives also an algorithm for model checking such a logic, given two model checkers for each separate logics. The communication overhead for combining the two model checkers would be in the order of $m + \sum_{A \in \wp(\sigma)} m_j + n \cdot l$, where m is the number of coalitional transitions in the model, m_A is the cardinality of the deontic access of coalition A, n is the number of states and l the complexity of the formula, leaving the model checking complexity of ATL $\oplus$ DATL linear in the size of the model and the formula [8]. However, two technical remarks are in order here. First, the formal results from [8] refer to combining temporal logics, while neither ATL nor DL is a temporal logic in the strictest sense. Moreover, the algorithm they propose for model checking of an independent combination of logics assumes that the models are finite (while there is no such assumption in our case). Nevertheless, polynomial model checking of DATL is of course possible, and we show how it can be done in Section 3.2, through a reduction of the problem to ATL model checking.

3.1 Imposing Requirements through Axioms

Following a main stream in deontic logic, we can take every deontic modality to be **KD** – the only deontic property (apart from the K-axiom and necessitation for $\mathcal{O}_\Gamma$) being the D-axiom $\neg\mathcal{O}_\Gamma\bot$. An axiomatization of ATL has been recently shown in [11]. If we do not need any mixing axioms, then the axiomatization of DATL can simply consist of the axioms for ATL, plus those of DL.

Concerning the global requirements, note that endowing $\mathcal{ATL}$ with a violation atom V is semantically very easy. Evaluating whether $\mathcal{O}\varphi$ is true at state q suggests incorporating a *universal modality* (cf. [10]) although some remarks are in place here. First of all, it seems more appropriate to use this definition of global requirements in *generated models* only, i.e., those models that are generated from some initial state q_0, by the transitions that the grand coalition Σ can make. Otherwise, the obligations might be unnecesarily weakened by considering violations or their absence in *unreachable states.* As an example, suppose we have a system that has two modes: starting from q_1, the constraint is that it is a violation to drive on the left hand side of the road ℓ, and when the system originates from q_2, one should adhere to driving on the right hand side (r). Seen as a global requirement, we would have $\mathcal{O}(\ell \vee r)$, which is of course too weak; what we want is $\mathcal{O}\ell$ (for the system rooted in q_1), or $\mathcal{O}r$ (when starting in q_2). Thus, a sound definition of obligations in a system with root q_0 is, that $M, q \models \mathcal{O}\varphi$ iff $M, q_0 \models \langle\langle\emptyset\rangle\rangle G(\neg V \rightarrow \varphi)$.

Second, we note in passing that by using the global requirement definition of obligation, the $\mathcal{O}$ modality obtained in this way is a KD45 modality, which means that we inherit the properties $\mathcal{O}\varphi \rightarrow \mathcal{O}\mathcal{O}\varphi$ and $\neg\mathcal{O}\varphi \rightarrow \mathcal{O}\neg\mathcal{O}\varphi$, as was also observed in [14]. But also, we get mixing axioms in this case: every deontic subformula can be brought to the outmost level, as illustrated by the valid scheme $\langle\langle\Gamma\rangle\rangle F\mathcal{O}\varphi \leftrightarrow \mathcal{O}\varphi$ (recall that we have $M, q \models \mathcal{O}\varphi$ iff $M, q_0 \models \mathcal{O}\varphi$ iff $M, q' \models \mathcal{O}\varphi$, for all states q, q' and root q_0). Some of the properties we have mentioned earlier in this paper can constitute interesting mixing axioms as well. For instance, a minimal property for requirements might be

$$\mathcal{O}_\Gamma\varphi \rightarrow \langle\langle\Gamma\rangle\rangle F\varphi$$

saying that every coalition can achieve its obligations. Semantically, we can pinpoint such a property as follows. Let us assume that this is an axiom scheme, and the model is distinguishing (i.e., every state in the model can be characterized by some DATL formula). Then the scheme corresponds to the semantic constraint:

$$\forall q \exists F_\Gamma \forall \lambda \in out(q, F_\Gamma) : states(\lambda) \cap img(q, \mathcal{R}_\Gamma) \neq \varnothing$$

where $states(\lambda)$ is the set of all states from λ, and $img(q, R) = \{q' \mid qRq'\}$ is the image of q with respect to relation R. In other words, Γ can enforce that every possible computation goes through at least one perfect alternative of q.

3.2 Model Checking Requirements and Abilities

In this section, we present a satisfiability preserving interpretation of DATL into ATL. The interpretation is very close to the one from [9], which in turn was inspired by [17]. The main idea is to leave the original temporal structure intact, while extending it with additional transitions to "simulate" deontic accessibility links. The simulation is achieved through new "deontic" agents: they can be passive and let the "real" agents decide upon the next transition (action *pass*),

or enforce a "deontic" transition. More precisely, the "positive deontic agents" can point out a state that was deontically accessible in the original model (or, rather, a special "deontic" copy of the original state), while the "negative deontic agents" can enforce a transition to a state that was *not* accessible. The first ones are necessary to translate formulas of shape $\mathcal{O}_A\varphi$; the latter are used for the "unconditionally permitted" operator $\mathcal{UP}_A$.

As an example, let M be the deontic game structure from Figure 3, and let us consider formulas $\mathcal{O}_\Sigma$saved, $\mathcal{UP}_\Sigma$saved and $\langle\langle k,b\rangle\rangle X$pdead (note that all three formulas are true in M, q_3). We construct a new concurrent game structure M^{ATL} by adding two deontic agents: $r_\Sigma, \bar{r}_\Sigma$, plus "deontic" copies of the existing states: $q_3^{r_\Sigma}, q_7^{r_\Sigma}, q_8^{r_\Sigma}$ and $q_3^{\bar{r}_\Sigma}, q_7^{\bar{r}_\Sigma}, q_8^{\bar{r}_\Sigma}$ (cf. Figure 4). Agent r_Σ is devised to point out all the perfect alternatives of the actual state. As state q_3 has only one perfect alternative (i.e., q_3 itself), r_Σ can enforce the next state to be $q_3^{r_\Sigma}$, provided that all other relevant agents remain passive[4]. In consequence, $\mathcal{O}_\Sigma$saved translates as: $\neg\langle\langle r_\Sigma, \bar{r}_\Sigma\rangle\rangle X(\mathsf{r}_\Sigma \wedge \mathsf{saved})$. In other words, it is not possible that r_Σ points out an alternative of q_3 (while $\bar{r}_\Sigma$ obediently passes), in which saved does *not* hold.

Agent $\bar{r}_\Sigma$ can point out all the *imperfect* alternatives of the current state (for q_3, these are represented by: $q_7^{\bar{r}_\Sigma}, q_8^{\bar{r}_\Sigma}$).Now, $\mathcal{UP}_\Sigma$saved translates as $\neg\langle\langle r_\Sigma, \bar{r}_\Sigma\rangle\rangle X(\bar{\mathsf{r}}_\mathsf{A} \wedge \mathsf{saved})$: $\bar{r}_\Sigma$ cannot point out an unacceptable state in which saved holds, hence the property of saved guarantees acceptability. Finally, $\langle\langle k,b\rangle\rangle X$pdead translates as $\langle\langle k,b,r_\Sigma,\bar{r}_\Sigma\rangle\rangle X(\mathsf{act} \wedge \mathsf{pdead})$: the strategic structure of the model has remained intact, but we must make sure that both deontic agents are passive, so that a non-deontic transition (an "action" transition) is executed.

We present the whole translation below in a more formal way. An interested reader can refer to [9] for a detailed presentation of the method, and proofs of correctness.

Given a deontic game structure $M = \langle \Sigma, Q, \Pi, \pi, Act, d, \delta, \mathbb{R}\rangle$ for a set of agents $\Sigma = \{a_1, ..., a_k\}$, we construct a concurrent game structure $M^{ATL} = \langle \Sigma', Q', \Pi', \pi', Act', d', \delta'\rangle$ in the following manner:

- $\Sigma' = \Sigma \cup \Sigma^r \cup \Sigma^{\bar{r}}$, where $\Sigma^r = \{r_A \mid A \subseteq \Sigma, A \neq \varnothing\}$ is the set of "positive", and $\Sigma^{\bar{r}} = \{\bar{r}_A \mid A \subseteq \Sigma, A \neq \varnothing\}$ is the set of "negative" deontic agents;
- $Q' = Q \cup \bigcup_{A\subseteq\Sigma, A\neq\varnothing}(Q^{\mathsf{r}_\mathsf{A}} \cup Q^{\bar{\mathsf{r}}_\mathsf{A}})$. We assume that Q and all $Q^{\mathsf{r}_\mathsf{A}}, Q^{\bar{\mathsf{r}}_\mathsf{A}}$ are pairwise disjoint. Further we will be using the more general notation $S^{\mathsf{e}} = \{q^{\mathsf{e}} \mid q \in S\}$ for any $S \subseteq Q$ and proposition e;
- $\Pi' = \Pi \cup \{\mathsf{act}, ..., \mathsf{r}_\mathsf{A}, ..., \bar{\mathsf{r}}_\mathsf{A}, ...\}$, and $\pi'(p) = \pi(p) \cup \bigcup_{A\subseteq\Sigma}(\pi(p)^{\mathsf{r}_\mathsf{A}} \cup \pi(p)^{\bar{\mathsf{r}}_\mathsf{A}})$ for every $p \in \Pi$. Moreover, $\pi'(\mathsf{act}) = Q$, $\pi'(\mathsf{r}_\mathsf{A}) = Q^{\mathsf{r}_\mathsf{A}}$, and $\pi'(\bar{\mathsf{r}}_\mathsf{A}) = Q^{\bar{\mathsf{r}}_\mathsf{A}}$;
- $d'_q(a) = d_q(a)$ for $a \in \Sigma, q \in Q$: choices of the "real" agents in the original states do not change,
- $d'_q(r_A) = \{pass\} \cup img(q, \mathcal{R}_A)^{\mathsf{r}_\mathsf{A}}$, and $d'_q(\bar{r}_A) = \{pass\} \cup (Q \setminus img(q, \mathcal{R}_A))^{\bar{\mathsf{r}}_\mathsf{A}}$. Action *pass* represents a deontic agent's choice to remain passive and let other agents choose the next state. Note that other actions of deontic agents are simply labeled with the names of deontic states they point to;

[4] We can check the last requirement by testing whether the transition leads to a deontic state of r_Σ (proposition r_Σ). It can happen only if all other relevant deontic agents choose action *pass*.

- $Act' = Act \cup \bigcup_{q \in Q, A \subseteq \Sigma}(d'_q(r_A) \cup d'_q(\bar{r}_A))$;
- the new transition function for $q \in Q$ is defined as follows (we put the choices from deontic agents in any predefined order):

$$\delta'(q, \alpha_{a_1}, ..., \alpha_{a_k}, ..., \alpha_r, ...) = \begin{cases} \delta(q, \alpha_{a_1}, ..., \alpha_{a_k}) & \text{if all } \alpha_r = pass \\ \alpha_r & \text{if } r \text{ is the first active (positive or negative) deontic agent} \end{cases}$$

- the choices and transitions for the new states are exactly the same: $d'(q^{r_A}, a) = d'(q^{\bar{r}_A}, a) = d'(q, a)$, and $\delta'(q^{r_A}, \alpha_{a_1}, ..., \alpha_{r_\Gamma}, ...) = \delta'(q^{\bar{r}_A}, \alpha_{a_1}, ..., \alpha_{r_\Gamma}, ...) = \delta'(q, \alpha_{a_1}, ..., \alpha_{a_k}, ..., \alpha_{r_\Gamma}, ...)$ for every $q \in Q, a \in \Sigma', \alpha_a \in d'(q, a)$.

Now, we define a translation of formulas from DATL to ATL corresponding to the above described interpretation of DATL models into ATL models:

$$\begin{aligned} tr(p) &= p, \qquad \text{for } p \in \Pi \\ tr(\neg\varphi) &= \neg tr(\varphi) \\ tr(\varphi \vee \psi) &= tr(\varphi) \vee tr(\psi) \\ tr(\langle\langle A\rangle\rangle X\varphi) &= \langle\langle A \cup \Sigma^r \cup \Sigma^{\bar{r}}\rangle\rangle X(\mathsf{act} \wedge tr(\varphi)) \\ tr(\langle\langle A\rangle\rangle G\varphi) &= tr(\varphi) \wedge \langle\langle A \cup \Sigma^r \cup \Sigma^{\bar{r}}\rangle\rangle X\langle\langle A \cup \Sigma^r \cup \Sigma^{\bar{r}}\rangle\rangle G(\mathsf{act} \wedge tr(\varphi)) \\ tr(\langle\langle A\rangle\rangle \varphi \mathcal{U} \psi) &= tr(\psi) \vee (tr(\varphi) \wedge \langle\langle A \cup \Sigma^r \cup \Sigma^{\bar{r}}\rangle\rangle X\langle\langle A \cup \Sigma^r \cup \Sigma^{\bar{r}}\rangle\rangle \\ &\qquad (\mathsf{act} \wedge tr(\varphi))\, \mathcal{U}(\mathsf{act} \wedge tr(\psi))) \\ tr(\mathcal{O}_A\varphi) &= \neg\langle\langle \Sigma^r \cup \Sigma^{\bar{r}}\rangle\rangle X(\mathsf{r}_A \wedge \neg tr(\varphi)) \\ tr(\mathcal{UP}_A\varphi) &= \neg\langle\langle \Sigma^r \cup \Sigma^{\bar{r}}\rangle\rangle X(\bar{\mathsf{r}}_A \wedge tr(\varphi)). \end{aligned}$$

Proposition 1. *For every* DATL *formula* φ*, model* M*, and a state* $q \in Q$*, we have* $M, q \models \varphi$ *iff* $M^{ATL}, q \models tr(\varphi)$.

Proposition 2. *For every* DATL *formula* φ*, model* M*, and "action" state* $q \in Q$*, we have* $M^{ATL}, q \models tr(\varphi)$ *iff* $M^{ATL}, q^{\mathsf{e}} \models tr(\varphi)$ *for every* $\mathsf{e} \in \Pi' \setminus \Pi$.

Corollary 1. *For every* DATL *formula* φ *and model* M*,* φ *is satisfiable (resp. valid) in* M *iff* $tr(\varphi)$ *is satisfiable (resp. valid) in* M^{ATL}.

Note that the vocabulary (set of propositions Π) only increases linearly (and certainly remains finite). Moreover, for a specific DATL formula φ, we do not have to include all the deontic agents r_A and $\bar{r}_A$ in the model – only those for which $\mathcal{O}_A$ or $\mathcal{UP}_A$ occurs in φ. Also, we need deontic states only for these coalitions A. The number of such coalitions is never greater than the complexity of φ. Let m be the cardinality of the "densest" modal accessibility relation – either deontic or temporal – in M, and l the complexity of φ. Then, the "optimized" transformation gives us a model with $m' = O(lm)$ transitions, while the new formula $tr(\varphi)$ is only linearly more complex than φ[5]. In consequence, we can use the ATL model checking algorithm from [1] for an efficient model checking of DATL formulas – the complexity of such process is $O(m'l') = O(ml^2)$.

[5] The length of formulas may suffer an exponential blow-up; however, the number of *different subformulas* in the formula only increases linearly. This issue is discussed in more detail in [9].

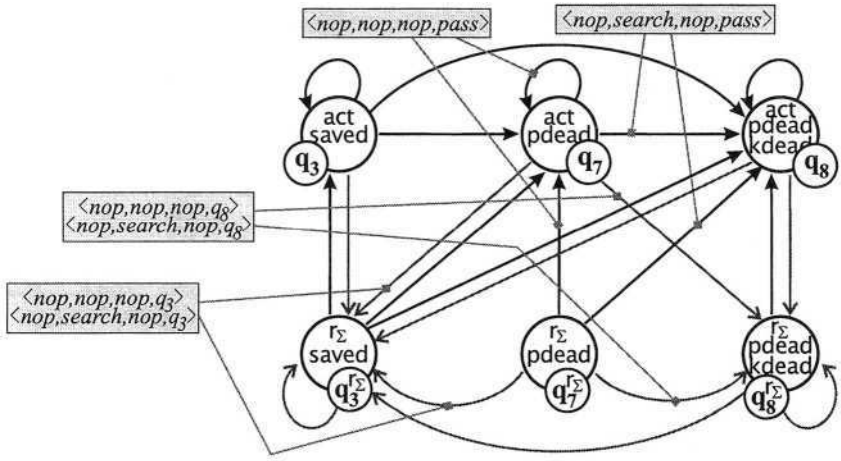


Fig. 4. ATL interpretation for the deontic game structure from Figure 3

Let us consider again the deontic game structure from Figure 3. We construct a corresponding concurrent game structure, optimized for model checking of the DATL formula $\mathcal{O}_\Sigma(\neg\mathsf{pdead} \wedge \langle\langle k\rangle\rangle X \neg\mathcal{O}_\Sigma \neg\mathsf{pdead})$: it is required that the Prime Minister is alive, but the killer is granted the ability to change this requirement. The result is shown in Figure 4. The translation of this formula is:

$$\neg\langle\langle r_\Sigma\rangle\rangle X(\mathsf{r}_\Sigma \wedge \neg(\neg\mathsf{pdead} \wedge \langle\langle k, r_\Sigma\rangle\rangle X(\mathsf{act} \wedge \neg\neg\langle\langle r_\Sigma\rangle\rangle X(\mathsf{r}_\Sigma \wedge \neg\neg\mathsf{pdead}))))$$

which holds in states q_3 and $q_3^{\mathsf{r}_\Sigma}$ of the concurrent game structure.

4 Conclusions

In this paper, we have brought obligations and abilities of agents together, enabling one to reason about what coalitions should achieve, but also to formulate principles regarding who can maintain or reinstall which ideal states or courses of action. We think the tractable model checking of DATL properties makes the approach attractive as a verification language for normative multi-agent systems.

However, as stated repeatedly in the paper, it is at the same time a report of ideas rather than of a crystallized and final analysis. We have not looked at an axiomatization of any system with non-trivial mixing axioms, nor have we yet explored some obvious routes that relate our approach in a technical sense with the work on social laws or the formal approaches that enrich ATL with an epistemic flavor, for instance. Nevertheless, we believe we have put to the force the fact that indeed DATL is a very attractive framework to incorporate abilities of agents and teams with deontic notions. We hope that the growing community, interested in norms in the computational context, can provide some feedback to help making appropriate decisions in the many design choices that we left open.

References

1. R. Alur, T. A. Henzinger, and O. Kupferman. Alternating-time temporal logic. *Journal of the ACM,* 49:672–713, 2002. Updated, improved, and extended text. Available at `http://www.cis.upenn.edu/~alur/Jacm02.pdf`.
2. A.R. Anderson. A reduction of deontic logic to alethic modal logic. *Mind,* 67:100–103, 1958.
3. N. Belnap. Backwards and forwards in the modal logic of agency. *Philosophy and Phenomenological Research,* LI(4):777–807, 1991.
4. N. Belnap and M. Perloff. Seeing to it that: a canonical form for agentives. *Theoria,* 54:175–199, 1988.
5. P. Bieber and F. Cuppens. Expression of confidentiality policies with deontic logics. In J.-J.Ch. Meyer and R.J. Wieringa, editors, *Deontic Logic in Computer Science: Normative System Specification,* pages 103–123. John Wiley & Sons, 1993.
6. J. Broersen, M. Dastani, Z. Huang, and L. van der Torre. The BOID architecture: conflicts between beliefs, obligations, intentions and desires. In *Proceedings of the Fifth International Conference on Autonomous Agents,* pages 9–16, 2001.
7. J. Fiadeiro and T. Maibaum. Temporal reasoning over deontic specifications. *Journal of Logic and Computation,* 1(3):357–396, 1991.
8. M. Franceschet, A. Montanari, and M. de Rijke. Model checking for combined logics. In *Proceedings of ICTL,* 2000.
9. V. Goranko and W. Jamroga. Comparing semantics of logics for multi-agent systems. *Synthese,* section on *Knowledge, Rationality and Action,* 2004. To appear.
10. V. Goranko and S. Passy. Using the universal modality: Gains and questions. *Journal of Logic and Computation,* 2(1):5–30, 1992.
11. V. Goranko and G. van Drimmelen. Complete axiomatization and decidability of the alternating-time temporal logic. Submitted, 2003.
12. D. Harel, D. Kozen, and J. Tiuryn. *Dynamic Logic.* MIT Press, 2000.
13. H.J. Levesque. All I know: a study in auto-epistemic logic. *Artificial Intelligence,* 42(3):263–309, 1990.
14. A. Lomuscio and M. Sergot. Deontic interpreted systems. *Studia Logica,* 75(1):63–92, 2003.
15. J.-J.Ch. Meyer. A different approach to deontic logic: Deontic logic viewed as a variant of dynamic logic. *Notre Dame Journal of Formal Logic,* 29(1): 109–136, 1988.
16. J.-J.Ch. Meyer and R.J. Wieringa. Deontic logic: A concise overview. In J.-J.Ch. Meyer and R.J. Wieringa, editors, *Deontic Logic in Computer Science: Normative System Specification,* pages 3–16. John Wiley & Sons, 1993.
17. K. Schild. On the relationship between BDI logics and standard logics of concurrency. *Autonomous Agents and Multi Agent Systems,* pages 259–283, 2000.
18. Y. Shoham and M. Tennenholz. On the synthesis of useful social laws for artificial agent societies. In *Proceedings of AAAI-92,* 1992.
19. W. van der Hoek, M. Roberts, and M. Wooldridge. Social laws in alternating time: Effectiveness, feasibility and synthesis. Submitted, 2004.
20. J. van Eck. A system of temporally relative modal and deontic predicate logic and its philosophical applications. *Logique et Analyse,* 100:249–381, 1982.
21. G.H. von Wright. Deontic logic. *Mind,* 60:1–15, 1951.
22. R.J. Wieringa and J.-J.Ch. Meyer. Applications of deontic logic in computer science: A concise overview. In J.-J.Ch. Meyer and R.J. Wieringa, editors, *Deontic Logic in Computer Science: Normative System Specification,* pages 17–40. 1993.

On Normative-Informational Positions

Andrew J.I. Jones

Department of Computer Science, King's College London, The Strand,
London WC2R 2LS, UK
ajijones@dcs.kcl.ac.uk

Abstract. This paper is a preliminary investigation into the application of the formal-logical theory of normative positions to the characterisation of *normative-informational positions,* pertaining to rules that are meant to regulate the supply of information.

1 Introduction

The theory of normative positions has provided a means of generating an exhaustive characterisation of the different types of normative status (*permitted, obligatory, forbidden,* and so on) that may be assigned to a given state of affairs. In the tradition of Kanger ([6], [5]), Lindahl [7], Jones & Sergot [4], the focus has usually been on the normative status of states of affairs of type 'agent *j* brings it about that A'; for instance, the class of *normative one-agent act positions* generated by the method described in [4] consists of the following seven positions:

(E1) $\mathrm{OE}_j A$

(E2) $\mathrm{OE}_j \neg A$

(E3) $\mathrm{O}(\neg \mathrm{E}_j A \wedge \neg \mathrm{E}_j \neg A)$

(E4) $\mathrm{PE}_j A \wedge \mathrm{PE}_j \neg A \wedge \mathrm{P}(\neg \mathrm{E}_j A \wedge \neg \mathrm{E}_j \neg A)$

(E5) $\mathrm{PE}_j A \wedge \mathrm{PE}_j \neg A \wedge \mathrm{O}(\mathrm{E}_j A \vee \mathrm{E}_j \neg A)$

(E6) $\mathrm{PE}_j A \wedge \neg \mathrm{PE}_j \neg A \wedge \mathrm{P}(\neg \mathrm{E}_j A \wedge \neg \mathrm{E}_j \neg A)$

(E7) $\neg \mathrm{PE}_j A \wedge \mathrm{PE}_j \neg A \wedge \mathrm{P}(\neg \mathrm{E}_j A \wedge \neg \mathrm{E}_j \neg A)$

There, Standard Deontic Logic (SDL) - a modal system of type KD in the Chellas classification [1] – is adopted for the logic of the obligation operator 'O', and the permission operator 'P' is the dual of 'O'; and a modal system of type ET[1] is used for the relativised action modality '$\mathrm{E}j$'.

Given these choices for the deontic and action modalities, the method yields the result that there are precisely these seven mutually exclusive normative positions for one agent vis-à-vis the state of affairs described by *'A'*. So either (E1), the agent is

[1] This means, essentially, that the action modality is closed under logical equivalence, and satisfies the T. schema, the 'success condition': $\mathrm{E}jA \rightarrow A$.

A. Lomuscio and D. Nute (Eds.): DEON 2004, LNAI 3065, pp. 182–190, 2004.

obliged to see to it that *A*, or (E2) he is obliged to see to it that ¬*A*, or (E3) he is obliged to remain passive with respect to *A*, or....and so on.

The present report starts from the assumption that it might also be of interest to investigate the normative status of another sort of state of affairs - of a type quite different from those represented by act descriptions – pertaining to the *informational* state of a given agent. By this is meant the state of affairs that an agent *j* is (or is not) *informed* that *A*, or is (or is not) *informed* that ¬*A*. Consider, for instance, the situation of an individual *j* in relation to some government agency *k* that has responsibility for controlling the flow of information concerning *A*. What is the class of possible normative positions for *k* (concerning the information *j* is permitted, forbidden, required, etc. to have about *A*) ? Or consider the situation of individual *k* in relation to some authority *j* (say, a court of Law), where *k* has certain obligations to supply information to *j,* or is permitted to withold information from *k.* In both of these contexts, among others, it would be useful to have at our disposal an exhaustive characterisation of the class of possible *normative-informational positions,* as they will here be called.

Furthermore, since the original aim of the theory of normative positions, as proposed by Kanger and others, was to provide a formal-logical framework for the articulation of Hohfeldian rights-relations, it seems natural to suppose that the development of an account of normative-informational positions, along the lines indicated above, might also provide a platform for the systematic investigation of such rights as *the right to silence, the right to know* and *the right to conceal information.* However, the potential application domain for a theory of this kind would seem not to be confined to legal analysis, but might also contribute to the formal specification of the normative status of electronic information agents, whose tasks may include the acquisition of information, and the monitoring of information flow, among others.

2 A Modality for 'Informational State'

In [2, 3] modal-logical characterisations are given of the (forms of) conventions that constitute various key types of signalling acts: *asserting, commanding, requesting, promising,....*among others. These characterisations employ several modalities, among them an ideality/optimality modality, '$\mathbf{I}^*_s$'[2], used to represent those states of affairs that would obtain if a conventional signalling system were in an optimal state, relative to its function of facilitating the transmission of reliable information. For instance, if - according to the conventions constituting signalling system *s* – the hoisting on board a ship of a particular sequence of coloured flags counts as an assertion that the ship is carrying explosives, then – when on a particular occasion those flags *are* hoisted – the signalling system *s* would be in an optimal/ideal state, relative to its function of faciltating the transmission of reliable information, only if it were then indeed the case that the ship was carrying explosives. An observer, or audience, *j,* who is familiar with the conventions governing *s* and who witnesses the hoisting of this sequence of flags, will understand the meaning of the signal in the sense that he is aware of what would now be the case, given that the signaller is telling the truth. So

[2] The '*' in the notation had no particular significance. It was introduced in the multi-modal language described in [2] merely to distinguish this particular notion of ideality from an evaluative normative modality, 'I', that also figured in the same language.

j's informational state, following his observation of the flag-raising, is represented – on this approach – by a belief whose content takes the form 'I^*_sA', where '*A*' describes the state of affairs that the ship is carrying explosives, *j*'s understanding the meaning of the signal amounts to his being aware that, were the signalling system *s* in an optimal state relative to its function of facilitating the transmission of reliable information, '*A*' would now be true. (Of course, *if j* also believes that the signaller is reliable, *j* will move on from the belief whose content is 'I^*_sA' to the belief that *A*.)

The modality 'I^*_s' was assigned the logic of the smallest normal modal system K. Closure under logical consequence would seem to be a natural assumption for this operator, given the intended interpretation. (For if signalling system *s* would be in an optimal state only if '*A*' were true, then it could be in an optimal state only if the logical consequences of '*A*' were also true.) Obviously, the T. schema

(T.I*) $\quad I^*_sA \rightarrow A$

does not accord with the indended interpretation. What of the D. schema

(D.I*) $\quad I^*_sA \rightarrow \neg I^*_s\neg A$

which is of course equivalent to

$\neg(I^*_sA \wedge I^*_s\neg A)$?

Well, the validity of (D.I*) would not be acceptable, for the simple reason that it would rule out the possibility of making inconsistent assertions. (It *would* be perfectly possible, for instance, in many circumstances, for one or more signallers to raise the flag sequence that means (according to *s*) that the ship is carrying explosives, *and* to raise the flag sequence that means (according to *s*) that the ship is not carrying explosives.)

For the purposes of the present investigation into normative-informational positions, an operator similar in interpretation to 'I^*_s' will be adopted, and will be denoted by 'I_j', where *j* is any agent. Expressions of the form 'I_jA' will be understood to mean 'were the information supplied to *j* to be true, then '*A*' would be the case', or 'according to the information supplied to *j*, '*A*' is the case'. The simpler, and perhaps less accurate, readings '*j* is told that *A*' and '*j* is informed that *A*' may also be used, for ease of expression. For reasons parallel to those mentioned for the 'I^*_s' operator, 'I_j' will also be assigned the logic of a (relativised) normal modal operator of type K.

In what follows, relativised versions of the obligation and permission modalities of Standard Deontic Logic (SDL), which is a normal modal system of type KD, will be employed to represent the normative component of the positions to be investigated. Expressions of the forms 'O_kA' and 'P_kA' will be read 'it is obligatory for *k* that *A*' and 'it is permitted for *k* that *A*', respectively. Thus the agent *k* is understood to be the bearer of the obligation/permission.

The problems associated with SDL are well documented in the literature. However, its adoption for the purposes of the present enquiry is defensible on the following three grounds:

(a) the property of the closure of the operators under logical consequence will here be exploited in ways that appear to be innocuous;

(b) conditional obligation sentences will not figure in this investigation; it is surely in connection with the treatment of conditionals that SDL's inadequacies are most fully exposed;

(c) adoption of the D schema should simply be understood as a restriction on the enquiry to those normative systems that are well-formed, or well-organised, in the sense that they do not allow conflict of obligation of the form '$O_kA \wedge O_k \neg A$'. In other words, the adoption of the D schema should not be understood as a claim to the effect that such conflicts *cannot* ever arise in any system of norms – but rather as a deliberate choice to focus on normative systems that are rationally organised, in the sense just described. It should be noted, however, that nothing in what follows presupposes that (where *k* and *j* are distinct agents) conflicts of the form '$O_kA \wedge O_j\neg A$' could not arise. This is important in the present context since, for instance, one agent might well be required to reveal information, whilst another is required not to do so.

Note, finally, as regards these comments on the selection of SDL, that the method of generation of normative positions employed below is itself independent of the particular choice of deontic logic (a point also emphasised in [4]). So anyone who remains unconvinced by the defence (a)-(c), above, can take from the shelf his favoured logic for obligation and permission, and insert that into the generation procedure instead. It goes without saying, of course, that the resultant set of consistent positions generated might well be quite different from those described in what now follows.

3 Generating Normative-Informational Positions

Given that the modality 'I_j' is assigned the logic of a (relativised) normal modality of type K, there are precisely 4 *informational positions* for *j* vis-à-vis the state of affairs described by '*A*'. These are:

(I1) $I_jA \wedge \neg I_j\neg A$

(I2) $I_j\neg A \wedge \neg I_jA$

(I3) $\neg I_jA \wedge \neg I_j\neg A$

(I4) $I_jA \wedge I_j\neg A$

It will be useful to introduce some phrases to refer to these positions:

In (I1), *j* is told *straight truth/straight lie,* depending on whether '*A*' is/is not the case.

In (I2), *j* is told *straight truth/straight lie,* depending on whether '¬*A*' is/is not the case.

In (I3), *j* is told neither '*A*' nor '¬*A*', and in this sense (13) represents the *silence* position.

In (I4), *j* is told both 'A' and '¬*A*', and in this sense (14) represents the *conflicting information* position.

In order to apply to (I1)-(I4) the method for generating *normative* positions described in [4], enclose each of (I1)-(I4) in parentheses, then prefix each with 'O_k' and

'$O_k\neg$', respectively, to form 8 obligation expressions. Then prefix each of those 8 expressions with the negation sign, and display the resulting 16 expressions as a list of 8 tautologies:

(1) $O_k\,(I_jA \wedge \neg I_j\neg A) \vee \neg O_k\,(I_jA \wedge \neg I_j\neg A)$

(2) $O_k\,(I_j\neg A \wedge \neg I_jA) \vee \neg O_k\,(I_j\neg A \wedge \neg I_jA)$

(3) $O_k\,(\neg I_jA \wedge \neg I_j\neg A) \vee \neg O_k\,(\neg I_jA \wedge \neg I_j\neg A)$

(4) $O_k\,(I_jA \wedge I_j\neg A) \vee \neg O_k\,(I_jA \wedge I_j\neg A)$

(5) $O_k\,\neg(I_jA \wedge \neg I_j\neg A) \vee \neg O_k\,\neg(I_jA \wedge \neg I_j\neg A)$

(6) $O_k\,\neg(I_j\neg A \wedge \neg I_jA) \vee \neg O_k\,\neg(I_j\neg A \wedge \neg I_jA)$

(7) $O_k\,\neg(\neg I_jA \wedge \neg I_j\neg A) \vee \neg O_k\,\neg(\neg I_jA \wedge \neg I_j\neg A)$

(8) $O_k\,\neg(I_jA \wedge I_j\neg A) \vee \neg O_k\,\neg(I_jA \wedge I_j\neg A)$

There are $2^8 = 256$ ways of selecting just one of the disjuncts from each of the disjunctions (1)-(8). That is, 256 distinct conjunctions, each of 8 conjuncts, may be generated from (1)-(8). It turns out that, of these 256 conjunctions, just 15 are logically consistent, given the logics selected for the component modalities. Each of these 15 may be simplified, to remove redundant conjuncts (i.e., conjuncts that are themseves logically implied by some other conjunct in the same conjunction). The result may be exhibited as (N1)-(N15), below:

(N1) $O_k\,(I_jA \wedge \neg I_j\neg A)$

(N2) $O_k\,(I_j\neg A \wedge \neg I_jA)$

(N3) $O_k\,(\neg I_jA \wedge \neg I_j\neg A)$

(N4) $O_k\,(\neg I_jA \vee \neg I_j\neg A) \wedge P_k\,(\neg I_jA \wedge \neg I_j\neg A) \wedge P_k\,(\neg I_jA \wedge I_j\neg A) \wedge P_k\,(I_jA \wedge \neg I_j\neg A)$

(N5) $O_k\,(I_jA \leftrightarrow \neg I_j\neg A) \wedge P_k\,(I_jA \wedge \neg I_j\neg A) \wedge P_k\,(I_j\neg A \wedge \neg I_jA)$

(N6) $\neg\, P_k\, I_j\neg A \wedge P_k\,(\neg I_j\neg A \wedge I_jA) \wedge P_k\,(\neg I_j\neg A \wedge \neg I_jA)$

(N7) $\neg\, P_k\, I_jA \wedge P_k\,(\neg I_jA \wedge I_j\neg A) \wedge P_k\,(\neg I_jA \wedge \neg I_j\neg A)$

(N8) $O_k\,(I_jA \wedge I_j\neg A)$

(N9) $O_k\, I_jA \wedge P_k\,(I_jA \wedge I_j\neg A) \wedge P_k\,(I_jA \wedge \neg I_j\neg A)$

(N10) $O_k\, I_j\neg A \wedge P_k\,(I_jA \wedge I_j\neg A) \wedge P_k\,(I_j\neg A \wedge \neg I_jA)$

(N11) $\neg\, P_k\,(I_jA \wedge \neg I_j\neg A) \wedge P_k\,(I_j\neg A \wedge \neg I_jA) \wedge P_k\,(\neg I_jA \wedge \neg I_j\neg A) \wedge P_k\,(I_jA \wedge I_j\neg A)$

(N12) $\neg\, P_k\,(I_j\neg A \wedge \neg I_jA) \wedge P_k\,(I_jA \wedge \neg I_j\neg A) \wedge P_k\,(I_jA \wedge I_j\neg A) \wedge P_k\,(\neg I_jA \wedge \neg I_j\neg A)$

(N13) $\neg\, P_k\,(\neg I_jA \wedge \neg I_j\neg A) \wedge P_k\,(I_jA \wedge I_j\neg A) \wedge P_k\,(I_jA \wedge \neg I_j\neg A) \wedge P_k\,(I_j\neg A \wedge \neg I_jA)$

(N14) $O_k\,(I_jA \leftrightarrow I_j\neg A) \wedge P_k\,(I_jA \wedge I_j\neg A) \wedge P_k\,(\neg I_jA \wedge \neg I_j\neg A)$

(N15) $P_k\,(I_jA \wedge I_j\neg A) \wedge P_k\,(I_jA \wedge \neg I_j\neg A) \wedge P_k\,(I_j\neg A \wedge \neg I_jA) \wedge P_k\,(\neg I_jA \wedge \neg I_j\neg A)$

4 Describing the Positions

For the purpose of discussing (N1)-(N15), it is convenient first to split the group into two sub-groups, consisting of (N1)-(N7) and (N8)-(N15), respectively. Each of (N1)-(N7) implies that the *conflicting information position* (vis-à-vis *j*) is not permitted for *k*. That is, each of (N1)-(N7) is incompatible with the truth of '$P_k\ (I_jA \wedge I_j\neg A)$' . By contrast, each of (N8)-(N15) implies that the *conflicting information position* (vis-à-vis *j*) *is* permitted for *k*.

Note the correspondence between (N1)-(N7) and (E1)-(E7), ((N1) to (E1), (N2) to (E2),……, and so on). The formal differences between each pair arise from the fact that the logic of the 'E_j' modality contains the T. schema, which in turn implies the D. schema:

$$E_jA \rightarrow \neg E_j\neg A$$

In other words, if the logic of the 'I_j' modality had contained the D. schema, then each of (N1)-(N7) would have been reducible to forms that correspond exactly to those of (E1)-(E7), with (of course) 'I_j' replacing 'E_j' throughout, and 'O_k'/'P_k' replacing 'O'/'P' throughout, and some re-arrangement of the order of the conjuncts.

Suppose now that *'A'* is true. Then the positions (N1)-(N7) may be described as follows:

(N1) It is obligatory for *k* that *j* is told the *straight truth.*

(N2) It is obligatory for *k* that *j* is told a *straight lie.*

(N3) It is obligatory for *k* that the *silence position* obtains.

(N4) The *conflicting information position* is forbidden for *k,* but the *silence position,* the *straight lie position* and the *straight truth position* are each permitted for *k.*

(N5) The *conflicting information position* and the *silence position* are both forbidden for *k,* but the *straight truth* and *straight lie positions* are both permitted for *k.*

(N6) It is not permitted for *k* that *j* is told a lie, but the *straight truth* and *silence positions* are both permitted for *k.*

(N7) It is not permitted for *k* that *j* is told the truth, but the *straight lie* and *silence positions* are both permitted for *k.*

If, on the other hand, it is '$\neg A$' rather than 'A' that is true, then (N1) and (N2) swap descriptions, (N6) and (N7) swap descriptions, and the descriptions of each of (N3), (N4) and (N5) remain unchanged.

It is the presence of (I4) in the list of *informational positions* that gives rise to the *normative-informational positions* (N8)-(N15). (Clearly, there is no counterpart to (N8)-(N15) in the class of *normative one-agent act positions* just because the action counterpart to (I4) is a logical contradiction.)

Supposing, first, again, that '*A*' is true, then the following descriptions may be proposed for (N8)-(N15):

(N8) It is obligatory for *k* that the *conflicting information position* obtains.

(N9) It is obligatory for *k* that *j* is told the truth; the *straight truth position* is permitted for *k,* but so is the *conflicting information position.*

(N10) It is obligatory for *k* that *j* is told a lie; the *straight lie position* is permitted for *k,* but so is the *conflicting information position.*

(N11) The *straight truth position* is forbidden (= not permitted) for *k,* but the *straight lie position,* the *silence position* and the *conflicting information position* are each permitted for *k.*

(N12) The *straight lie position* is forbidden for *k,* but the *straight truth position,* the *silence position* and the *conflicting information position* are each permitted for *k.*

(N13) The *silence position* is forbidden for *k,* but the *conflicting information position,* the *straight truth position,* the *straight lie position* are each permitted for *k.*

(N14) The *straight truth* and *straight lie* positions are both forbidden for *k,* but the *conflicting information position* and the *silence position* are both permitted for *k.*

(N15) The *conflicting information position,* the *straight truth position,* the *straight lie position* and the *silence position* are each permitted for *k.*

If, on the other hand, it is '$\neg A$' rather than 'A' that is true, then (N9) and (N10) swap descriptions, and (N11) and (N12) swap descriptions, but the descriptions of each of (N8), (N13), (N14) and (N15) remain unchanged.

5 Some Observations about the Application of the Theory

Given the choice of logics for the modalities, the generation method shows that – for any agents *k* (norm-bearer) and, *j* (informee), and for any state of affairs *'A'* – precisely one of the set of 15 normative-informational positions holds. So the set may be used as a tool in the analysis of normative-informational concepts, such as the *permission to be silent,* and the *permission to be correctly informed;* furthermore, as is clearly indicated in the existing literature on normative positions, the method provides a means of approaching the representation of more complex structures, of the kind exhibited by Hohfeldian rights-relations; in the present context, the *right to silence* and the *right to know* would be interesting candidates for investigation.

The present paper takes some preliminary steps in laying the formal-logical foundations for these kinds of conceptual analyses, but further details remain as the focus for future work. But, by way of illustration of how to proceed, consider the example *'permission to be silent'* in relation to the set of 15 positions. In fact 8 of these 15 contain or imply *k*'s permission to be silent, vis-à-vis *j,* with respect to *'A',* and these are (N3), (N4), (N6), (N7), (N11), (N12), (N14) and (N15). To define the context further, suppose that the concern is with the permission ordinarily granted to a person, under English Law, at the time of that person's arrest for an alleged criminal offence[3]. Which of the 8 cases would be the appropriate choice?

Well, it is reasonable to eliminate (N3) immediately, since *k,* the person arrested, is not under an *obligation* (as far as *j,* the arresting authority is concerned) to remain silent. There would seem to be good grounds for eliminating (N7), too, since it forbids

[3] The person arrested is ordinarily told that he/she has the *right* to remain silent, but that anything he/she says may be taken down and used in evidence against him/her. The *right* to remain silent implies (but is not implied by) the *permission* to remain silent, but the *relational* aspect, characteristic of the Hohfeldian interpretation of rights (rather than mere permissions), will be ignored for present purposes. It will most definitely figure in future work, however.

k to tell the truth, which again would not ordinarily be understood to be part of the arresting authority's intention. Similar considerations would eliminate (N11) and (N14). Then there remain the 4 positions: (N4), (N6), (N12) and (N15). Is the agent *k* *forbidden,* i.e., *not permitted,* to give conflicting information (as far as *j* is concerned), at the time of arrest (when he/she is *not,* one supposes, *under oath*),? If not, then (N4) gets eliminated, along with (N6). The final choice, between (N12) and (N15), depends on whether or not the *straight lie position* is permitted for *k.*

As a second illustration, consider the situation of a future British government, led by P.M. Bliar, which is not permitted to be silent on the burning issue of the use, by the government, of weapons of mass deception. What might here be the normative-informational position of the government (*k*), vis-à-vis the citizen (*j*), as P.M. Bliar sees it? Clearly, the 8 positions considered above in discussion of the previous example, each of which contains or implies that the *silence position* is permitted, are ruled out. The positions (N1), (N2), (N5), (N8), (N9), (N10) and (N13) remain. Given Bliar's aversion to *straight truth,* and the lack of subtlety of the *straight lie,* (N1) and (N2) are eliminated. The transmission of *conflicting information,* being a valuable strategy for the spin doctors, is hardly going to be forbidden by Bliar and his magic circle, so (N5) goes out; but then perhaps they don't want to *tie* themselves to the use of conflicting information, so (N8) is eliminated. Furthermore, supposing that they don't want to be *required* to let the truth out, or *required* to lie, (N9) and (N10) go too. So (N13) remains as the *Bliar position:* 'say what you like about 'A', so long as you say something' ![4]

6 Concluding Remarks

Two points, in conclusion: First, Marek Sergot has indicated to me that there is also another point of departure for the generation of normative-informational positions, taking not the 4 positions (I1)-(I4) as the base, but rather the 8 positions obtained by conjoining each of those positions with *'A'* or *'¬A'*. This is clearly an option worthy of further investigation, although the inclusion of '*A*'/'¬*A*' within the scope of the deontic operator, in generating the normative-informational positions from this base, will perhaps not always produce interesting results, since the question of whether or not '*A*'/'¬*A*' is itself obligatory may be quite irrelevant. By contrast, the approach taken above first generated the normative-informational positions, and afterwards considered the truth/falsity of '*A*'.

Secondly, discussion of the application of this formal framework would greatly benefit by taking a range of concrete examples – from the Law, or from other types of existing regulations – where the point of the rules is to define a policy to govern the transmission of information. The merits and shortcomings of the present formal framework could then be given a more thorough assessment, by measuring the extent to which it exposes, or not, the details and nuances exhibited by those rules.

---

[4] This example, despite being facetious, does nevertheless serve to illustrate the way in which a map of the class of possible positions might play a role in choosing (for good or ill) an 'appropriate' strategy or policy. *Prolegomena to the Theory of spin*…….?

Acknowledgements

Thanks to Marek Sergot for helpful comments and advice. Work on this paper began during the last stages of the EC project ALFEBIITE (IST-1999-10298). It has continued during the EC Working Group iTRUST, and during the EC Integrated Project TrustCoM. Grateful acknowledgement of the support of the EC is hereby given.

References

1. Chellas, B.F.: Modal Logic – an introduction, Cambridge University Press, Cambridge, UK (1980)
2. Jones, A.J.I.: A Logical Framework. In: Pitt, J. (ed.): The Open Agent Society, John Wiley & Sons, Chichester, UK (forthcoming in 2004)
3. Jones, A.J.I., Parent, X.: Conventional Signalling Acts and Conversation. In: Dignum, F. (ed): Advances in Agent Communication, Lecture Notes in Artificial Intelligence, Vol. 2922, Springer-Verlag, Berlin Heidelberg New York (2004) 1-17
4. Jones, A.J.I., Sergot, M.J.: Formal Specification of Security Requirements Using the Theory of Normative Positions. In: Deswarte, Y, *et al.,* (eds.): Computer Security-ESORICS 92, Proc. of the 2nd European Symposium on Research in Computer Security, Springer Lecture Notes in Computer Science, Vol.648, Springer-Verlag, Berlin Heidelberg New York (1992) 103-121. Reprinted as Part II of: On the Characterisation of Law and Computer Systems: the Normative Systems Perspective. In: Meyer, J.-J. Ch., Wieringa, R.J. (eds.): Deontic Logic in Computer Science: Normative System Specification, Wiley Professional Computing Series, John Wiley & Sons, Chichester, UK (1993) 275-307
5. Kanger, S.: Law and Logic, Theoria, Vol. 38, (1972)
6. Kanger, S., Kanger, H.: Rights and Parliamentarism, Theoria, Vol. 32 (1966)
7. Lindahl, L.: Position and Change – a Study in Law and Logic, Synthese Library Vol. 112, D. Reidel, Dordrecht, Holland (1977)

Quasi-matrix Deontic Logic

Andrei Kouznetsov

Department of Philosophy
Kemerovo State University, Novokuznetsk Branch
654041 Tsyolkovskogo St., 23, Novokuznetsk, Russia
andr-kusnetsov@yandes.ru

Abstract. We use non-Kripkean quasi-matrix semantics for the formalization of the systems $\mathbf{S}_{3d}$, $\mathbf{S}_{3dp}$ and $\mathbf{S}_{3dq}$ of deontic logic. The system $\mathbf{S}_{3d}$ is weaker than the standard logic **SDL**. The semantics for $\mathbf{S}_{3dp}$ represents combination of quasi-matrix semantics and the *semantics of truth value gluts,* which allows $\mathbf{S}_{3dp}$ to avoid deontic explosion $\mathbf{O}A \wedge \mathbf{O}\neg A \supset \mathbf{O}B$. The system $\mathbf{S}_{3dq}$ rejects both deontic explosion and the formula $\mathbf{O}A \wedge \mathbf{O}\neg A \supset \mathbf{O}A \wedge \neg\mathbf{O}A$, thus it allows to consider deontic dilemmas without classical contradictions.
The systems $\mathbf{S}_{5d}$, $\mathbf{S}_{5dp}$ and $\mathbf{S}_{5dq}$ in which the two types of deontic operators are used, namely, strong and weak obligation (permission), can be built as an extension of the correspondent systems $\mathbf{S}_{3d}$, $\mathbf{S}_{3dp}$ and $\mathbf{S}_{3dq}$.

1 Quasi-Matrix Semantics for Modal Logic

Since the classical paper of Georg Henrik von Wright [19] deontic logic is considered as a special branch of modal logic. The so called *standard deontic logic* **(SDL)** can be obtained from the normal modal logic **KD** (system name is taken from B.Chellas [3]) by replacing the usual operators of necessity and possibility - respectively, box and diamond - by the "normative" operators, correspondingly, **O** (is read as *it is obligatory that,* or *it ought to be that*), and **P** (*it is permissible that*).

The system **SDL** contains all tautologies of classical propositional logic and the following axioms:

(**SDL***A*1) $\mathbf{O}(p \supset q) \supset (\mathbf{O}p \supset \mathbf{O}q)$;
(**SDL***A*2) $\mathbf{O}p \supset \mathbf{P}p$;
(**SDL***A*3) $\mathbf{P}p \equiv \neg\mathbf{O}\neg p$;
Inference rules:
(**SDL***R*1) *Modus ponens*;
(**SDL***R*2) $\frac{p}{\mathbf{O}p}$.

Operator **F** (*it is forbidden that*) can be defined as $\mathbf{F}p \stackrel{def}{=} \neg\mathbf{P}p$.

Semantics of **SDL** is based on the well-known Kripke model $M = (S, \pi, R)$, where S is a set of possible worlds, π is a truth assignment function assigning truth to the primitive propositions per each world, and $R \subseteq S \times S$ is a relation relating with each world a set of alternative worlds. Given a Kripke-model M and a world $s \in S$, the modal operators are defined as follows:

A. Lomuscio and D. Nute (Eds.): DEON 2004, LNAI 3065, pp. 191–208, 2004.

$(M, s) \models \mathbf{O}p$ iff $\forall t(R(s,t) \Rightarrow (M,t) \models p)$
$(M, s) \models \mathbf{P}p$ iff $\exists t(R(s,t) \& (M,t) \models p)$
$(M, s) \models \mathbf{F}p$ iff $\forall t(R(s,t) \Rightarrow (M,t) \neg \models p)$

J.Kearns ([12]) and Yu.Ivlev ([11]) in different ways suggested an idea of modal semantics which is completely different from Kripke-style semantics in that it does not use the idea of possible worlds. Instead of the alternative worlds one deals with the alternative interpetation quasi-functions formed by the given interpretation function ([11]). This approach allows to give table definitions for modal operators; there sometime can be vagueness in evaluations of modal formulas- in that case the value of the given formula A is considered to be "undetermined", "alternative"- for instance, the value p/q means "either p, or q". Instead of an interpretation function, interpretation quasi-function takes only one single value from the fraction; in case of the value p/q there are two alternative quasi-interpretations, one in which A takes the value p, and the other in which A takes q. The formula A is true in the interpretation if and only if it is true in each alternative interpretation caused by the given interpretation.

The advantage of using quasi-matrix approach is that on its basis it is possible to consider the wide range of modal systems, which seems to include as a subset all known Kripkean modal logics and contains even more "intermediate" systems, for example the ones weaker then **K**, **T**, **B**, **S4**, etc. (some of that *four valued* modal systems were suggested in [11]). One can expect that the application of quasi-matrix approach to deontic logic can promote consideration of some additional aspects of deontic matters. Another good point of quasi-matrix semantics is that it allows to define the properties of modal (deontic) logic under construction beforehand, just in the table definitions of modal operators. In particular, in the described below systems there is an opportunity to consider separately the properties of the action sentences (acts) and to enter special deontic connectives by means of which the complex acts are formed. The properties of deontic connectives are set by table definitions in accordance with the preliminary informal discussions about the character of action of that connectives.

The considered below deontic system $\mathbf{S}_{3d}$ represents modification of three-valued quasi-matrix deontic logic suggested in [11] and is weaker than the standard deontic logic **SDL,** because the axiom **SDL***A*1 the rule **SDL***R*2 are no longer valid in $\mathbf{S}_{3d}$. In the next part we bring the intuitions on the system $\mathbf{S}_{3d}$.

2 Intuitions

First of all, the distinction between terms and formulas of deontic system $\mathbf{S}_{3d}$ has the same motivation as it was given in *logic for normative propositions* [1] which has been constructed on the basis of possible-world semantics. The reason for such distinction is that "deontic operators require as operands descriptions of actions (action sentences) but once deontic operator is applied to an action sentence the resulting deontic sentence is no longer a description of an action but a normative qualification of the action described by the action sentence

contained within. Hence the occurrence of a deontic operator within the scope of another deontic operator makes no sense" ([1], p.47).

Now let us speak about the intuitive meaning of the connectives for the acts $*$, $\otimes$, $\oplus$. The act p^* means *the abstention from the act* p. An interesting discussion of what the abstention from the act does mean could be found in the papers of Wright, but from the point of view of deontic values one could say definitely that if the act p is obligatory at some code, then the abstention from p should be forbidden by that code, and the same reasoning when p is forbidden. If p is indifferent then it seems like p^* should also be indifferent. Thus our table definition for the act p^* is consistent with the intuition.

The act $(p \otimes q)$ means *consecutive or parallel performance of actions* p *and* q. Note that the act $(p \otimes p)$ has different meaning than the act p, since it could stay for the consecutive performance of the acts p and p (or, for the recurrence of p) which in general case is not the same as p. The abstention from the act $(p \otimes q)$ intuitively would mean *the abstention from at least one of the acts* p, q. So it seems intuitively reasonable to keep such equalities as $(p \otimes q)^* = p^* \oplus q^*$, thus **the order of values** $u/v/w$ **in table definitions for** $\otimes$ **and** $\oplus$ **is important.** In other words, one can't take arbitrary combinations of quasi-matrices for $\otimes$ and $\oplus$: given the two correspondent alternative values $u_1/v_1/w_1$ and $u_2/v_2/w_2$ in tables for $\otimes$ and $\oplus$, the only three quasi-interpretations are admissible: one in which u_1 in table for $\otimes$ corresponds to u_2 in table for $\oplus$, and the other two with the same correspondences v_1 to v_2 and w_1 to w_2. In fact, the tables for the system $\mathbf{S}_{3d}$ defines **6** different quasi-interpretations.

The act $(p \oplus q)$ means *either performance of the act* p, *or* q, *or* $(p \otimes q)$. The act $(p \oplus p)$ can be understood as an abstention from the act $(q \otimes q)$, if $p = q^*$, that is to abstain from recurrence of the act q, one should abstain from the first or from the second performance of the act q.

Now let us give some intuitions for choosing the alternative truth values that are given in the table definitions.

• The term $(p \otimes q)$ takes the value $o/i/b$- "either *obligatory,* or *indifferent,* or *forbidden"* when both acts p, q are evaluated as *obligatory.* Let p and q are the two **different** acts. Intuitively, if each of the different terms p, q is obligatory, then the act $(p \otimes q)$ should also be obligatory. We assume here that the normative code must be consistent in itself and in its relation to other codes with which it submits. For, assume that an agent (suppose, a spy) would be obliged (p) *to fill her automobile with petrol at 2 o'clock* and (q) *to smoke a cigarette at 2 o'clock* (as a secret sign for somebody), but the whole act $(p \otimes q)$ is forbidden at least at filling station, thus it is forbidden in the instructions for the spy (in order not to conflict with the environment). Now, since we have $\mathbf{F}(p \otimes q) \equiv \mathbf{O}(p \otimes q)^* \equiv \mathbf{O}(p^* \oplus q^*)$, we would have an obligation for the agent *to abstain from filling her automobile or from smoking her cigarette* and at the same time the agent is *obliged to fill her automobile* and is *obliged to smoke a cigarette*! Such code would be inconsistent. Similar examples could be brought to illustrate the case when the terms p and q both take the value *obligatory,* and the term $p \otimes q$ is *indifferent* - again, it would lead to inconsistency of the considered code.

A little bit different situation occures in case of the act $(p \otimes p)$. If the act p is obligatory at some code then nothing could be said about its recurrence - suppose, a spy is obliged per the certain day to visit some agreed place. However visiting by the spy per the same day of the same place secondarily could be indifferent or even forbidden (and obligatory as well) in spy's instructions. Thus, in general case we take it that for the two obligatory acts p, q, the act $(p \otimes q)$ takes the value $o/i/b$.

• The term $(p \otimes q)$ takes the value i/b when both acts p, q are evaluated as *indifferent.* Again, let p and q are the two different acts. The case when the term $(p \otimes q)$ takes the value i when both acts p, q take this value, is trivial: let, for example, p be *to seat on the bench* and q be *to read the newspaper*- both acts are indifferent in some normative code, and so is the complex act $(p \otimes q)$. But if an agent decides *to fill her automobile with petrol at 2 o'clock* (the act p which is indifferent in some code) and *to smoke a cigarette at 2 o'clock* (the act q which in itself is also indifferent in the same code), then the act $(p \otimes q)$ is of course forbidden! Or even consider the act $(p \otimes p^*)$- *to fill the automobile with petrol at 2 o'clock and to abstain from it at 2 o'clock* - this act is impossible, thus it must take the value *forbidden* at our system. The act $(p \otimes p)$ for the value *indifferent* of p also takes the alternative value i/b. Suppose, the act p is *to take a free cup of coffee,* and the rules of some cafe allow to take the first cup of coffee for free, then the performance of the action p is indifferent while the action $(p \otimes p)$ is forbidden in the rules of that cafe. Another value is trivial.

Can the term $(p \otimes q)$ take the value o - *obligatory* - in case when both acts p, q are *indifferent*? Would it be correct to bring an example that a man is *obliged to take off his shoes* (q) *when entering the house* (p) (both acts p, q are *indifferent*)? The answer is no since that example represents not the act $(p \otimes q)$ but the conditional $(p \mapsto q)$ which **could** take the value o in case when p and q both take the value i. Otherwise we would have for the agent the obligation *to enter the house and to take off his shoes.* If the acts p, q are both indifferent at some code, then how could it be that the consecutive or parallel performance of that acts is obligatory in that code? Since we are allowed to abstain from any of the acts p, q (because they are indifferent), then how could we keep the obligation to fulfil them both? In a similar way, if p is indifferent, then how could the recurrence of p be obligatory? Thus we take it that $(p \otimes q)$ can't take the value *obligatory* when p, q are both *indifferent.*

• In a similar way we discuss the alternative value $b/i/o$ for the term $(p \oplus q)$, if both acts p and q take the value b. If p, q are the two different forbidden acts, then $(p \oplus q)$ seems to be forbidden as well, because if p and q are both forbidden, then how could it be that the act *at least one of* p *and* q is allowed? Now, consider the act $(p \oplus p)$. As we told, if $p = q^*$, then the most probable intuitive meaning for the act $(p \oplus p)$ is the abstention from recurrence of the act q, $(q \otimes q)$, otherwise $(p \oplus p)$ should have the same meaning as p. Therefore, if the recurrence of some obligatory act q is either *obligatory,* or is *indifferent,* or, at last, is *forbidden,* then the abstention from that recurrence, $(p \oplus p)$, should

be either *forbidden,* or is *indifferent,* or is *obligatory,* just when p takes the value b. So in general case the alternative value for the term $(p \oplus q)$ is $b/i/o$.

• The term $(p \oplus q)$ takes the alternative value i/o, if both acts p and q take the value i. Consider again the two different acts p, q. The case when the act $(p \oplus q)$ is indifferent if both acts p, q are indifferent, is trivial. Now suppose that the agent is obliged (by some code) *to deliver the letter to the addressee,* and suppose the agent has few alternatives to reach the addressee (by choosing the road, transport, etc.) The choice of the concrete alternative is indifferent from the point of view of the given code. Nevertheless, the whole act is obligatory for the agent. In case of the act $(p \oplus p)$, if $p = q^*$, then we have again the abstention from the act $(q \otimes q)$, which takes the value i/b, thus the value of $(p \oplus p)$ is i/o.

Can the term $(p \oplus q)$ take the value b - *forbidden* - in case when both acts p, q are *indifferent*? Again, intuitively, if both acts p, q are allowed, then how could it be that the act *at least one of* p *and* q is forbidden, or, is not allowed? Thus, we take it that in this case $(p \oplus q)$ can't take the value b.

3 Quasi-matrix Deontic Logic $\mathbf{S}_{3d}$

The semantics $\mathbf{QM}_{3d}$ is defined in a following way.

Language

The language L_{3d} of three-valued quasi-matrix logic contains:

- p, q, r,... - variables for the atomic action sentences (acts);
- $\otimes$, $\oplus$, $*$ - connectives for the acts, correspondingly is read "and", "or", "it is not the case that" ("abstention from..."");
- **O, P**- operators, correspondingly is read "it is obligatory that" and "it is permissible that";
- $\neg$, $\wedge$, $\vee$, $\supset$, $\equiv$ - logical signs for negation, conjunction, disjunction, material conditional and material biconditional;
- brackets.

Formation Rules

1. *Definition of a term:*

- Every atomic action sentence is a term;
- if p and q are terms then p^*, $(p \otimes q)$,$(p \oplus q)$ are also terms.

2. *Definition of a formula:*

- if p is a term then $\mathbf{O}p$ and $\mathbf{P}p$ are the formulas;
- if A and B are the formulas, then also are $\neg A$, $(A \wedge B)$, $(A \vee B)$, $(A \supset B)$, $(A \equiv B)$.

3. *Definitions of the connectives for the acts:*
The variables for the acts take values from the field $\{o, i, b\}$, correspondingly is read *obligatory, indifferent, forbidden.*

	$\otimes$	o	i	b	$\oplus$	o	i	b
p	o	$o/i/b^{\bullet}$	i	b	o	o	o	o
p^*	i	i	$i/b^{\bullet\bullet}$	b	i	o	$i/o^{\bullet\bullet}$	i
	b	b	b	b	b	o	i	$b/i/o^{\bullet}$

Rem. The value $u/v/w$ is read "either u, or v, or w". It means that in fact there are several alternative quasi-matrices for each of the connectives $\oplus$ and $\otimes$, in which one chooses the only one value from the "fraction". But one can't take the arbitrary combinations of values: the only admissible quasi-matrices for $\otimes$ and $\oplus$ are those with the same positions of the values in fractions marked with $\bullet$ or with $\bullet\bullet$: if, say, there is a quasi-interpretation in which $|p| = |q| = b$ iff $|p \oplus q| = b$, then the same quasi-interpretation function assigns $|p| = |q| = o$ iff $|p \otimes q| = o$.
Rem. One could define also deontic conditional $p \mapsto q$ as a term $p^* \oplus q$.

4. *Definitions of the operators* $\mathbf{O}$, $\mathbf{P}$:

p	$\mathbf{O}p$	$\mathbf{P}p$
o	t	t
i	f	t
b	f	f

The formulas take values from the field $\{t, f\}$ (*true, false*). The definitions of the connectives $\neg, \wedge, \vee, \supset, \equiv$ are given as usual.

For the formula A of the system $\mathbf{S}_{3d}$ we define an interpretation function $|\cdot|$ for the terms and for the formulas in accordance with the above described table definitions. The "alternative" value u/v of the act means that this value is fixed but undetermined ("either u, or v"). *The alternative interpretation quasi-function* $||\cdot||$, caused by the function $|\cdot|$, maps the formula (term) V to the only one single element from the fraction; thus if the "fraction" contains m alternative values then we get m alternative interpretations of V.

The formula A is *valid in the given interpretation* (we will write $|A| = t$) if and only if (iff) it takes the value t in each alternative interpretation caused by this interpretation ($||A||_k = t$ for every $1 \leq k \leq m$, m is a number of alternative interpretations); A is *satisfiable in the given interpretation* iff it takes the value t in some alternative interpretation caused by the given interpretation; A is *satisfiable* in semantics $\mathbf{QM}_{3d}$ iff it is valid in some interpretation; A is *valid* in semantics $\mathbf{QM}_{3d}$ iff it is valid in each interpretation.

The result of the formalization of the described semantics is the system $\mathbf{S}_{3d}$.

The system $\mathbf{S}_{3d}$

($\mathbf{S}_{3d}A0$) *All tautologies of classical propositional logic*;
($\mathbf{S}_{3d}A1$) $\mathbf{O}p \equiv \mathbf{O}p^{**}$;
($\mathbf{S}_{3d}A2$) $\mathbf{P}p \equiv \mathbf{P}p^{**}$;

($\mathbf{S}_{3d}A3$) $\mathbf{O}(p \otimes q) \equiv \mathbf{O}(p^* \oplus q^*)^*$;
($\mathbf{S}_{3d}A4$) $\mathbf{O}(p \oplus q) \equiv \mathbf{O}(p^* \otimes q^*)^*$;
($\mathbf{S}_{3d}A5$) $\mathbf{P}(p \otimes q) \equiv \mathbf{P}(p^* \oplus q^*)^*$;
($\mathbf{S}_{3d}A6$) $\mathbf{P}(p \oplus q) \equiv \mathbf{P}(p^* \otimes q^*)^*$;
($\mathbf{S}_{3d}A7$) $\mathbf{O}p \supset \mathbf{P}p$;
($\mathbf{S}_{3d}A8$) $\mathbf{O}p \equiv \neg\mathbf{P}p^*$;
($\mathbf{S}_{3d}A9$) $\mathbf{O}(p \otimes q) \supset \mathbf{O}p \wedge \mathbf{O}q$;
($\mathbf{S}_{3d}A10$) $\mathbf{P}(p \otimes q) \supset \mathbf{P}p \wedge \mathbf{P}q$;
($\mathbf{S}_{3d}A11$) $\mathbf{O}p \vee \mathbf{O}q \supset \mathbf{O}(p \oplus q)$;
($\mathbf{S}_{3d}A12$) $\mathbf{P}p \vee \mathbf{P}q \supset \mathbf{P}(p \oplus q)$;

Inference rules:

($\mathbf{S}_{3d}R1$) *Modus ponens* for the formulas;
($\mathbf{S}_{3d}R2$) *Substitution rule* for the terms and for the formulas.

Definitions:

• $\mathbf{F}p \stackrel{def}{=} \neg\mathbf{P}p$
($\mathbf{F}p$ is read *it is forbidden that* p);

• $\mathbf{I}p \stackrel{def}{=} \mathbf{P}p \wedge \mathbf{P}p^*$
($\mathbf{I}p$ is read *it is indifferent that* p).

Theorem 1.3. *The system* $\mathbf{S}_{3qd}$ *is sound and complete with respect to semantics* $\mathbf{QM}_{3d}$.

Sketch of the Proof:

Soundness can be proved by showing that each axiom is valid in each interpretation, since it is true in every alternative interpretation of every given interpretation. Suppose, for example, that the axiom ($\mathbf{S}_{3d}A10$) is not valid. That means that there is an interpretation function $|\cdot|_v$ in which ($\mathbf{S}_{3d}A10$) is not valid, so there is an alternative quasi-function $||\cdot||_{v'}$, derived from $|\cdot|_v$ s.t. $||\mathbf{P}(p \otimes q) \supset \mathbf{P}p \wedge \mathbf{P}q||_{v'} = f$, so $||\mathbf{P}(p \otimes q)||_{v'} = t$ and $||\mathbf{P}p \wedge \mathbf{P}q||_{v'} = f$. It follows that either $||p \otimes q||_{v'} = o$, or $||p \otimes q||_{v'} = i$, thus according to table definitions neither $||p||_{v'} = b$ nor $||q||_{v'} = b$. But at the same time, $||\mathbf{P}p||_{v'} = f$ or $||\mathbf{P}q||_{v'} = f$, so $||p||_{v'} = b$ or $||q||_{v'} = b$. Therefore, we get the contradiction.

The rules $\mathbf{S}_{3d}R1$ and $\mathbf{S}_{3d}R2$ do preserve validity of the formulas.

To prove completeness, one has to prove the two additional lemmas.

Lemma 1. *The set of formulas* Q *which is consistent with respect to* $\mathbf{S}_{3d}$ *can be extended to a maximal consistent set* T *which has the following properties:*

1) for each term p *either* $\mathbf{P}p \wedge \mathbf{P}p^* \in T$, *or* $\mathbf{O}p \in T$, *or* $\neg\mathbf{P}p \in T$;
2) if $\Gamma \vdash B$ *and* $\Gamma \subseteq T$ *then* $B \in T$;
3) $B \in T$ *or* $C \in T$ *if and only if* $B \vee C \in T$;
4) $\neg B \in T$ *or* $C \in T$ *if and only if* $B \supset C \in T$.

Let us prove for example the statement *1)*. The extension of the given set of formulas Q to a maximal consistent set T could be shown in a usual way. Let $B_1, B_2, \ldots$ be some sequence of formulas of $\mathbf{S}_{3d}$ and let Q be a set of formulas which is consistent with $\mathbf{S}_{3d}$. We construct a sequence of sets of formulas

$T_0, T_1, T_2, \ldots$ in a following way. Let $T_0 = Q$ and suppose that the set T_n is defined. If the formula B_{n+1} is not derivable from T_n, then $T_{n+1} = T_n \cup \{\neg B_{n+1}\}$; if B_{n+1} is derivable from T_n, then $T_{n+1} = T_n$. Let T be a union of all sets T_i. For the proof of consistency of T, one could show by induction the consistency of each set T_i. By construction of T, for each formula A either $A \in T$, or $\neg A \in T$. Therefore, in our system, for the given term p,

($\diamond$) **{either** $\mathbf{O}p \in T$, **or** $\neg\mathbf{O}p \in T$**} and {either** $\mathbf{P}p \in T$, **or** $\neg\mathbf{P}p \in T$**},** from which one could obtain the statement *1)* using the correspondent axioms of $\mathbf{S}_{3d}$ (the cases when the formulas represent boolean combinations of $\mathbf{O}p$ and $\mathbf{P}p$ can be easily reduced to ($\diamond$)).

Lemma 2. *There is a function* $|\cdot|_T$ *such that it possesses all the properties of an interpretation function, and for each formula A,* $|A|_T = t \Leftrightarrow A \in T$.

Consider the function $|\cdot|_T$ possessing the following properties:

- $|p|_T = o \Leftrightarrow \mathbf{O}p \in T$;
- $|p|_T = i \Leftrightarrow \mathbf{P}p \wedge \mathbf{P}p^* \in T$;
- $|p|_T = b \Leftrightarrow \neg\mathbf{P}p \in T$;
- $|A|_T = t \Leftrightarrow A \in T$ (A is a formula).

It can be shown that the function $|\cdot|_T$ possesses all the properties of an interpretation function for logic $\mathbf{S}_{3d}$. Let us show some steps of the proof of lemma, for example, let us prove that if $|p|_T = b$, then $|p^*|_T = o$. Suppose, $|p|_T = b$, thus, according to definition of $|\cdot|_T$, $\neg\mathbf{P}p \in T$. By ($\mathbf{S}_{3d}A2$) and ($\mathbf{S}_{3d}A8$), $\neg\mathbf{P}p \equiv \neg\mathbf{P}p^{**} \equiv \mathbf{O}p^*$, thus $\mathbf{O}p^* \in T$, and therefore, $|p^*|_T = o$. Now let us show that if $|p|_T = |q|_T = i$, then $|p \otimes q|_T$ is either i or b. Suppose $|p|_T = |q|_T = i$, then by definition of $|\cdot|_T$, $\mathbf{P}p \in T$, $\mathbf{P}p^* \in T$ and $\mathbf{P}q \in T$, $\mathbf{P}q^* \in T$. By ($\mathbf{S}_{3d}A8$), $\mathbf{P}p^* \vdash \neg\mathbf{O}p$, $\mathbf{P}q^* \vdash \neg\mathbf{O}q$. Thus $\neg\mathbf{O}p \in T$, $\neg\mathbf{O}q \in T$. Suppose now, that $|p \otimes q|_T = o$, then $\mathbf{O}(p \otimes q) \in T$, and, by ($\mathbf{S}_{3d}A9$), $\mathbf{O}p \in T$ and $\mathbf{O}q \in T$, and we get a contradiciton, which in its turn contradicts to the fact that the set T is consistent. Thus $|p \otimes q|_T$ is either i or b (but not both i, b, because that would make T inconsistent). The other steps could be shown in a similar way.

Let us make the last steps of the completeness proof. Suppose there is the valid formula E which is not provable in $\mathbf{S}_{3d}$. Therefore, the formula $\neg\neg E$ is also not provable in $\mathbf{S}_{3d}$. The set $\{\neg E\}$ is consistent with $\mathbf{S}_{3d}$, it can be extended to the maximal consistent set of formulas T by lemma 1. There is an interpretation $|\cdot|_T$ in which all the formulas from T are valid (lemma 2), hence $\neg E$ is also valid (in this interpretation). Thus the formula E is not valid in $|\cdot|_T$, but this contradicts to the assumption.□

Example. Let us show that the formula of **SDL,** $\neg\mathbf{O}(p \otimes p^*)$ *(No contradictory obligations),* is valid in quasi-matrix semantics. Take an arbitrary interpretation $|\cdot|_\vartheta$ and show that $|\neg\mathbf{O}(p \otimes p^*)|_\vartheta = t$. That means that $||\neg\mathbf{O}(p \otimes p^*)||_{\vartheta'} = t$ for all alternative interpretation quasi-functions, $||\cdot||_{\vartheta'}$, derived from $|\cdot|_\vartheta$. So consider any such $||\cdot||_{\vartheta'}$ and suppose for *reductio* that $||\neg\mathbf{O}(p \otimes p^*)||_{\vartheta'} \neq t$. Hence $||\mathbf{O}(p \otimes p^*)||_{\vartheta'} = t$, and so $||(p \otimes p^*)||_{\vartheta'} = o$. From this one can move

immediately to $||p||_{\vartheta'} = o$ and $||p^*||_{\vartheta'} = o$, so that $||p||_{\vartheta'} = b$, a contradiction, and we are done.

4 Semantics of Truth Value Gluts

In addition to the plausible need to allow for the possibility of moral dilemmas, which requires rejecting the formula $\neg(\mathbf{O}p \wedge \mathbf{O}\neg p)$ of **SDL,** one must also reject the thesis $(*) \vdash_{\mathbf{SDL}} (OA \wedge O\neg A) \supset OB$, since that would mean that if an agent finds herself in a moral dilemma, then she ought to do everything, which is surely going too far. Yet this formula, $(*)$, is a theorem of **SDL,** as follows. Consider the theorem $\vdash_{\mathbf{SDL}} (A \wedge \neg A) \supset B$. Application of the rule (**SDL***R*2) gives $\vdash_{\mathbf{SDL}} O((A \wedge \neg A) \supset B)$, from which by **SDL***A*1 one infers $\vdash_{\mathbf{SDL}} O(A \wedge \neg A) \supset OB$, and, combining with theorem of **SDL** $\mathbf{O}(p \wedge q) \equiv \mathbf{O}p \wedge \mathbf{O}q$, we get $(*) \vdash_{\mathbf{SDL}} (OA \wedge O\neg A) \supset OB$. Notice that this argument rests on the premise of classical logic that a contradiction implies everything.

Different approaches were suggested to avoid this and some other "undesirable" theorems in monadic deontic logic. J.Horty [9] suggests an approach based on nonmonotonic logic. Some authors consider various modifications of Kripke-semantics, for example, "two-phase approach" of L.W.N. Van der Torre [17], "preference-based deontic logic" of S.O.Hansson [8], "multiplex semantics" of L.Goble [6], non-kripkean deontic logic of P.Schotch and R.Jennings [16], semantics of FUSION logic of L.Cholvy and F.Cuppens [4].

One way to avoid the paradoxical theorem $(*)$ is to consider relevant implication instead classical one, since the theorem $(A \wedge \neg A) \rightarrow B$ is no longer valid in logic with relevance. The idea of combination deontic principles with the principles of relevant logic was discussed in different forms (for the brief overview see L.Goble [7], in which the author also constructs various systems of monadic and dyadic relevant deontic logic as an extension of the Anderson-Belnap system **R** of relevant implication).

We now propose another approach based on applying our quasi-matrix semantics for deontic logic to a basic paraconsistent logic whose semantics allows for truth value "gluts", the possibility that some formulas might be both true and false. This will produce the deontic system $\mathbf{S}_{3dp}$ which does not contain the undesirable theorem $(*)$. The approach is based on the notion of *partially defined predicates.* The idea of generalization of traditional versions of completely defined predicates and sets by considering partially defined predicates and sets was investigated in various forms in the papers of T.Scolem, G.Beman, D.Bochvar, V.Akkerman, K.Schütte, F.Fitch, S.Feferman and some others. H.Wang ([18]) suggested two systems of the calculus of partial predicates, **PP** and **EP,** propositional fragments of these systems were formalized by A.Rose ([15]) and by N.Ermolaeva ([5]).

We would be primarily interested in semantics with *truth value "gluts"* (or *paraconsistent* semantics) **TGl** and in its corresponding system **G** since we expect to obtain deontic system which is *normatively paraconsistent* in that it would not allow, for example, the theorem $(\mathbf{O}p \wedge \mathbf{O}p^*) \rightarrow \mathbf{O}q$ and at the same

time *normatively complete* such that each action sentence p has its deontic estimation. But of course there could be constructed another deontic systems, namely, *normatively pseudocomplete* but consistent, or both, normatively pseudocomplete and paraconsistent.

Let us describe the semantics **TGl.** Let v be an assignment function assigning to each formula A a subset of *{truth, false}*. We want the truth value gaps to be excluded, that can be done by requiring that $v(A)$ not be empty. We will say that the formula A *is true in* **TGl** $(v_t(A))$ if and only if $t \in v(A)$, and, similarly, A *is false in* **TGl** $(v_f(A))$ if and only if $f \in v(A)$.

The evaluations for the formulas containing $\neg$, $\wedge$, $\vee$, $\supset$ are the following:

$v_t(\neg p)$ iff $v_f(p)$ $\quad$ $v_f(\neg p)$ iff $v_t(p)$

$v_t(p \wedge q)$ iff $v_t(p)$ and $v_t(q)$ $\quad$ $v_f(p \wedge q)$ iff $v_f(p)$ or $v_f(q)$

$v_t(p \vee q)$ iff $v_t(p)$ or $v_t(q)$ $\quad$ $v_f(p \vee q)$ iff $v_f(p)$ and $v_f(q)$

$v_t(p \supset q)$ iff $v_f(p)$ or $v_t(q)$ $\quad$ $v_f(p \supset q)$ iff $v_t(p)$ and $v_f(q)$

The formula of the type $A \to B$ is *valid* in **TGl** if and only if $v_t(A) \subseteq v_t(B)$.

We note that none of the formulas which does not contain the connective $\to$ is valid.

Below is the axiomatic system which formalizes semantics **TGl.**

The System G

(**G**A1) $A \to A \vee B$;
(**G**A2) $A \vee B \to B \vee A$;
(**G**A3) $A \wedge B \to A$;
(**G**A4) $A \wedge B \to B \wedge A$;
(**G**A5) $\neg\neg A \to A$;
(**G**A6) $A \to \neg\neg A$;
(**G**A7) $A \wedge (B \vee C) \to A \wedge B \vee A \wedge C$;
(**G**A8) $\neg(A \wedge B) \to \neg A \vee \neg B$;
(**G**A9) $\neg A \vee \neg B \to \neg(A \wedge B)$;
(**G**A10) $\neg(A \vee B) \to \neg A \wedge \neg B$;
(**G**A11) $\neg A \wedge \neg B \to \neg(A \vee B)$;
(**G**A12) $B \to A \vee \neg A$

Inference Rules

GR1 *If* $A \to C$ *and* $B \to C$, *then* $A \vee B \to C$
GR2 *If* $A \to B$ *and* $A \to C$, *then* $A \to B \wedge C$
GR3 *If* $A \to B$ *and* $B \to C$, *then* $A \to C^{\bullet)}$
GR4 *If formula A contains propositional variable* v, *and formula B does not contain* $\to$, *then C is a result of substitution in A of all occurrences of* v *for B.*

•) *A, B, C* contain no connective $\to$.

Theorem 1.4. (A.Rose [15], N.Ermolaeva[5]) $\vdash_{\mathbf{G}} A \to B$ *if and only if* $v_t(A) \subseteq v_t(B)$.

5 Quasi-matrix Deontic Logic with Truth Value Gluts

We adapt this method to quasi-matrix three- valued deontic logic, here we use the language L_{3d} of system $\mathbf{S}_{3d}$. Now let υ be an assignment function assigning to each action term p a nonempty subset of $\{o, i, b\}$ and to each formula A a nonempty subset of $\{t, f\}$. Let's define the valuation function υ:

(Df 5.1) $t \in \upsilon(\mathbf{O}p)$ iff $o \in \upsilon(p)$, let us write it as $\upsilon_t(\mathbf{O}p)$ iff $\upsilon_o(p)$ (p is a term);
(Df 5.2) $\upsilon_f(\mathbf{O}p)$ iff $\upsilon_i(p)$ or $\upsilon_b(p)$;
(Df 5.3) $\upsilon_t(\mathbf{P}p)$ iff $\upsilon_o(p)$ or $\upsilon_i(p)$;
(Df 5.4) $\upsilon_f(\mathbf{P}p)$ iff $\upsilon_b(p)$;
(Df 5.5) $\upsilon_k(p^*)$ iff $\upsilon_{4-k}(p)$ ($k \in \{o, i, b\}$), here and below (1) stands for o, (2) - for i and (3)- for b;
(Df 5.6) if $\upsilon_o(p \otimes q)$ then $\upsilon_o(p)$ and $\upsilon_o(q)$;
(Df 5.7) if $\upsilon_i(p \otimes q)$ then $\{\upsilon_o(p)$ and $\upsilon_o(q)\}$ or $\{\upsilon_o(p)$ and $\upsilon_i(q)\}$ or $\{\upsilon_i(p)$ and $\upsilon_o(q)\}$ or $\{\upsilon_i(p)$ and $\upsilon_i(q)\}$;
(Df 5.8) if $\upsilon_b(p \otimes q)$ then $\upsilon_b(p)$ or $\upsilon_b(q)$ or $\{\upsilon_i(p)$ and $\upsilon_i(q)\}$;
(Df 5.9) if $\upsilon_l(p)$ and $\upsilon_m(q)$ then $\upsilon_k(p \otimes q)$, where $k = \max(l, m)$ except the cases $l = m = o$ and $l = m = i$:
if $\upsilon_o(p)$ and $\upsilon_o(q)$ then $\upsilon_o(p \otimes q)/\upsilon_i(p \otimes q)/\upsilon_b(p \otimes q)$ (either $\upsilon_o(p \otimes q)$, or $\upsilon_i(p \otimes q)$, or $\upsilon_b(p \otimes q)$);
if $\upsilon_i(p)$ and $\upsilon_i(q)$ then $\upsilon_i(p \otimes q)/\upsilon_b(p \otimes q)$;
(Df 5.10) if $\upsilon_o(p \oplus q)$ then $\upsilon_o(p)$ or $\upsilon_o(q)$ or $\{\upsilon_i(p)$ and $\upsilon_i(q)\}$or $\{\upsilon_b(p)$ and $\upsilon_b(q)\}$;
(Df 5.11) if $\upsilon_i(p \oplus q)$ then $\{\upsilon_i(p)$ and $\upsilon_b(q)\}$or $\{\upsilon_b(p)$ and $\upsilon_i(q)\}$or $\{\upsilon_i(p)$ and $\upsilon_i(q)\}$or $\{\upsilon_b(p)$ and $\upsilon_b(q)\}$;
(Df 5.12) if $\upsilon_b(p \oplus q)$ then $\upsilon_b(p)$ and $\upsilon_b(q)$;
(Df 5.13) if $\upsilon_l(p)$ and $\upsilon_m(q))$ then $\upsilon_k(p \oplus q)$, where $k = \min(l, m)$ except the cases $l = m = b$ and $l = m = i$:
if $\upsilon_b(p)$ and $\upsilon_b(q)$ then $\upsilon_b(p \oplus q)/\upsilon_i(p \oplus q)/\upsilon_o(p \oplus q)$;
if $\upsilon_i(p)$ and $\upsilon_i(q)$ then $\upsilon_i(p \oplus q)/\upsilon_o(p \oplus q)$;
and the same valuations for the formulas $\neg A$, $A \wedge B$, $A \vee B$ and $A \supset B$ as in semantics **TGl.**

The example of using the above definitions will be given in a soundness proof of the system $\mathbf{S}_{3dp}$.

Consider the alternative valuation quasi-functions υ^i, each of which takes only one single value from fractions in definitions (Df5.9) and (Df 5.13). Formula $A \to B$ is *valid* in semantics $\mathbf{SP_3D}$ if and only if, for each i, if $t \in \upsilon^i(A)$, then $t \in \upsilon^i(B)$ ($1 \leq i \leq n$, n is a number of valuation quasi-functions). We will write $\models_{\mathbf{SP_3D}} A \to B \stackrel{def}{=} \forall i(\upsilon^i_t(A) \subseteq \upsilon^i_t(B))$.

The system $\mathbf{S}_{3dp}$ represents the axiomatization of semantics $\mathbf{SP_3D}$.

The System $\mathbf{S}_{3dp}$

($\mathbf{S}_{3dp}A0$) All theorems of system **G**;
($\mathbf{S}_{3dp}A1$) $\mathbf{O}p \leftrightarrow \mathbf{O}p^{**}$;
($\mathbf{S}_{3dp}A2$) $\mathbf{P}p \leftrightarrow \mathbf{P}p^{**}$;
($\mathbf{S}_{3dp}A3$) $\mathbf{O}(p \otimes q) \leftrightarrow \mathbf{O}(p^* \oplus q^*)^*$;
($\mathbf{S}_{3dp}A4$) $\mathbf{O}(p \oplus q) \leftrightarrow \mathbf{O}(p^* \otimes q^*)^*$;
($\mathbf{S}_{3dp}A5$) $\mathbf{P}(p \otimes q) \leftrightarrow \mathbf{P}(p^* \oplus q^*)^*$;
($\mathbf{S}_{3dp}A6$) $\mathbf{P}(p \oplus q) \leftrightarrow \mathbf{P}(p^* \otimes q^*)^*$;
($\mathbf{S}_{3dp}A7$) $\mathbf{O}p \rightarrow \mathbf{P}p$;
($\mathbf{S}_{3dp}A8$) $\mathbf{O}p \leftrightarrow \neg\mathbf{P}p^*$;
($\mathbf{S}_{3dp}A9$) $\mathbf{O}(p \otimes q) \rightarrow \mathbf{O}p \wedge \mathbf{O}q$;
($\mathbf{S}_{3dp}A10$) $\mathbf{P}(p \otimes q) \rightarrow \mathbf{P}p \wedge \mathbf{P}q$;
($\mathbf{S}_{3dp}A11$) $\mathbf{O}p \vee \mathbf{O}q \rightarrow \mathbf{O}(p \oplus q)$;
($\mathbf{S}_{3dp}A12$) $\mathbf{P}p \vee \mathbf{P}q \rightarrow \mathbf{P}(p \oplus q)$;

Inference Rules

($\mathbf{S}_{3dp}R0$) Rules of system **G**;
($\mathbf{S}_{3dp}R1$) $\frac{A}{C}^{\bullet)}$.

•) Substitution rule for the terms: *if formula A contains atomic term* p, *and* q *is an arbitrary term, then C is a result of substitution in A of all occurrences of* p *by* q.

Rem. $A \leftrightarrow B$ stands for "$A \rightarrow B$ and $B \rightarrow A$".

Operators **F** and **I** (correspondingly, *it is forbidden* that and *it is indifferent that*) could be defined.

Theorem 1.5. $\models_{\mathbf{SP_3D}} A \rightarrow B \Leftrightarrow \vdash_{3dp} A \rightarrow B$.

Proof: Soundness follows from the fact that, for each axiom ($\mathbf{S}_{3dp}A0 - \mathbf{S}_{3dp}A12$) of the type $A \rightarrow B$, $v_t(A) \subseteq v_t(B)$. Since, by (Df.5.5), $v_k(p^{**})$ iff $v_k(p)$, then the axioms ($\mathbf{S}_{3dp}A1$) and ($\mathbf{S}_{3dp}A2$) are valid. Then, since in each quasi-interpretation $\sigma \in (p \otimes q)$ iff $\sigma \in ((p^* \oplus q^*)^*)$ for any deontic value σ (that follows from definitions by considering separately each quasi-interpretation), then the axioms ($\mathbf{S}_{3qd}A3$) and ($\mathbf{S}_{3qd}A5$) are valid. By the same reasons, since $\sigma \in (p \oplus q)$ iff $\sigma \in (p^* \otimes q^*)^*$, then the axioms ($\mathbf{S}_{3dp}A4$) and ($\mathbf{S}_{3dp}A6$) are valid as well. Now let's show for example the validity of the axiom ($\mathbf{S}_{3dp}A9$). Suppose, it is not valid, therefore there must be quasi-interpretation v' s.t. $t \in v'(\mathbf{O}(p \otimes q)$ would not imply $t \in v'(\mathbf{O}p \wedge \mathbf{O}q)$, or, $t \in v'(\mathbf{O}(p \otimes q)$ and $f \in v'(\mathbf{O}p \wedge \mathbf{O}q)$. Therefore, $o \in v'(p \otimes q)$ and $\{i \in v'(p)$ or $b \in v'(p)$ or $i \in v'(q)$ or $b \in v'(q)\}$. By definitions, there are quasi-interpretations v° in which $o \in v^\circ(p \otimes q)$. But $o \in v^\circ(p \otimes q)$ in such interpretations implies $o \in v^\circ(p)$ and $o \in v^\circ(p)$, and we get contradiction.

The inference rules do preserve the validity of formulas.

For the proof of completeness, assume that there is the formula $A \rightarrow B$, such that $v_t(A) \subseteq v_t(B)$, which is not a theorem of $\mathbf{S}_{3dp}$ (or $\dashv A \rightarrow B$). Since formulas A and B do not contain the connective "$\rightarrow$", one could present these formulas in the disjunctive normal form (**DNF**) by using the **G**-theorems **GA4-GA11,**

the substitution rule **GR4** and the derived rule **DR1** : *If* $F \to G$ *and* $A \leftrightarrow B$, *then* $F' \to G'$,where F', G' are formulas obtained from F and G by replacing some occurences of A by B.

Now let us have the formula $A \to B$, which is not a theorem, and at which both A and B are presented in **DNF,** namely,

$\dashv d_A^1 \vee d_A^2 \vee ... \vee d_A^n \to d_B^1 \vee d_B^2 \vee ... \vee d_B^m$.

Using the rule **GR1,** we get

$\dashv d_A^1 \to d_B^1 \vee ... \vee d_B^m$, or $\dashv d_A^2 \to d_B^1 \vee ... \vee d_B^m$, or... ,or $\dashv d_A^n \to d_B^1 \vee ... \vee d_B^m$.

Then, using the derived rule **DR2** : *If* $A \to B$, *then* $A \to B \vee C$, we get for each j $(1 \leq j \leq m)$

$\dashv d_A^1 \to d_B^j$, or $\dashv d_A^2 \to d_B^j$, or..., or $\dashv d_A^n \to d_B^j$. Let us take $\dashv d_A^k \to d_B^j$ for some fixed k.

On the other hand, by assumption, $v_t(d_A^1 \vee d_A^2 \vee ... \vee d_A^n) \subseteq v_t(d_B^1 \vee d_B^2 \vee ... \vee d_B^m)$. It follows that $v_t(d_A^i) \subseteq v_t(d_B^1 \vee d_B^2 \vee ... \vee d_B^m)$ for each i, $1 \leq i \leq n$. There could be two possibilities.

Case 1. For every i $(1 \leq i \leq n)$, there is j $(1 \leq j \leq m)$, such that $v_t(d_A^i) \subseteq v_t(d_B^j)$, and

Case 2. There is i $(1 \leq i \leq n)$, such that for every j $(1 \leq j \leq m)$, $v_t(d_A^i) \not\subseteq v_t(d_B^j)$.

Consider the **case 1.** Let us have the disjuncts d_A^k, $d_B^{j_k}$ such that $v_t(d_A^k) \subseteq v_t(d_B^{j_k})$. In accordance with the above reasoning,

$\dashv d_A^k \to d_B^{j_k}$. The disjunct d_A^k, as well as the disjunct $d_B^{j_k}$, represents conjunction of modalized formulas $\widetilde{\mathbf{O}}\, p_1 \wedge \widetilde{\mathbf{O}}\, p_2 \wedge ... \wedge \widetilde{\mathbf{O}}\, p_v \wedge \widetilde{\mathbf{P}}\, p_{v+1} \wedge \widetilde{\mathbf{P}}\, p_{v+2} \wedge ... \wedge \widetilde{\mathbf{P}}\, p_{v+w}$, where the symbol $\widetilde{\Theta}$ means that operator Θ has been taken either with, or without negation, and $p_1, p_2, ..., p_v, p_{v+1}, ..., p_{v+w}$ are arbitrary terms.

Using the theorem $\mathbf{S}_{3dp}A7$, the rule (**DR1**) and the axioms of **G**, one could present this conjunction such that each conjunct would represent either the formula $\mathbf{O}p_i$, or the formula $\mathbf{P}p_j$ $(1 \leq i, j \leq v + w)$.

Therefore, for the formula $\dashv d_A^k \to d_B^{j_k}$ we would have

$\dashv \mathbf{O}p_1 \wedge \mathbf{O}p_2 \wedge ... \wedge \mathbf{O}p_v \wedge \mathbf{P}p_{v+1} \wedge \mathbf{P}p_{v+2} \wedge ... \wedge \mathbf{P}p_{v+w} \to \mathbf{O}r_1 \wedge \mathbf{O}r_2 \wedge ... \wedge \mathbf{O}r_\vartheta \wedge \mathbf{P}r_{\vartheta+1} \wedge \mathbf{P}r_{\vartheta+2} \wedge ... \wedge \mathbf{P}r_{\vartheta+\rho}$.

Using the rule **GR2**, we get

$(\circ_{r1}) \dashv \mathbf{O}p_1 \wedge ... \wedge \mathbf{O}p_v \wedge \mathbf{P}p_{v+1} \wedge ... \wedge \mathbf{P}p_{v+w} \to \mathbf{O}r_1$, or

$(\circ_{r2}) \dashv \mathbf{O}p_1 \wedge ... \wedge \mathbf{O}p_v \wedge \mathbf{P}p_{v+1} \wedge ... \wedge \mathbf{P}p_{v+w} \to \mathbf{O}r_2$, or

.....,or

$(\circ_{r\vartheta}) \dashv \mathbf{O}p_1 \wedge ... \wedge \mathbf{O}p_v \wedge \mathbf{P}p_{v+1} \wedge ... \wedge \mathbf{P}p_{v+w} \to \mathbf{O}r_\vartheta$, or

$(\circ_{r\vartheta+1}) \dashv \mathbf{O}p_1 \wedge ... \wedge \mathbf{O}p_v \wedge \mathbf{P}p_{v+1} \wedge ... \wedge \mathbf{P}p_{v+w} \to \mathbf{P}r_{\vartheta+1}$, or

....., or

$(\circ_{r\vartheta+\rho}) \dashv \mathbf{O}p_1 \wedge ... \wedge \mathbf{O}p_v \wedge \mathbf{P}p_{v+1} \wedge ... \wedge \mathbf{P}p_{v+w} \to \mathbf{P}r_{\vartheta+\rho}$.

Let $p \sqsubseteq q$ means that either $p = q \otimes u$, or $q = p \oplus v$ (u, v are arbitrary terms). Using the derived rule **DR3:**

If the formula contains the arbitrary term r, *then this term could be replaced by any "equivalent" term* s (the equivalence means that, for the valuation function v, $v \in r$ iff $v \in s$), one could use instead of the terms $q \otimes u$ and $p \oplus v$

the "equivalent" terms. For example, instead of $q \otimes u$ one could take $q \otimes u^{**}$, or $(q^* \oplus u^*)^*$. In all such cases we also take it that $p \sqsubseteq q$. According to the axioms of $\mathbf{S}_{3dp}$, if

$(\star)$ $\overset{\xi}{\underset{i=\eta}{\otimes}} p_\eta \sqsubseteq r_m$ at $(\circ_m)$ for each m $(1 \leq m \leq \vartheta)$ and for some η, ξ, $(0 \leq \xi, \eta \leq v; \xi \geq \eta)$, and if

$(\star\star)$ $\overset{\xi}{\underset{i=\eta}{\otimes}} p_\eta \sqsubseteq r_l$ at $(\circ_l)$ for each l $(\vartheta + 1 \leq l \leq \vartheta + \rho)$ and for some η, ξ, or $\overset{\theta}{\underset{i=\delta}{\otimes}} p_\delta \sqsubseteq r_l$ for each l and for some δ, θ, $(v + 1 \leq \theta, \delta \leq v + w; \theta \geq \delta)$, then

$\vdash \mathbf{O}p_1 \wedge ... \wedge \mathbf{O}p_v \wedge \mathbf{P}p_{v+1} \wedge ... \wedge \mathbf{P}p_{v+w} \rightarrow \mathbf{O}r_m$ and

$\vdash \mathbf{O}p_1 \wedge ... \wedge \mathbf{O}p_v \wedge \mathbf{P}p_{v+1} \wedge ... \wedge \mathbf{P}p_{v+w} \rightarrow \mathbf{P}r_l$ for each l, m, which is not the case. Therefore, either condition $(\star)$, or condition $(\star\star)$ must not be fulfilled.

On the other hand, by assumption,

$\upsilon_t(\mathbf{O}p_1 \wedge ... \wedge \mathbf{O}p_v \wedge \mathbf{P}p_{v+1} \wedge ... \wedge \mathbf{P}p_{v+w}) \subseteq \upsilon_t(\mathbf{O}r_1 \wedge ... \wedge \mathbf{O}r_\vartheta \wedge \mathbf{P}r_{\vartheta+1} \wedge ... \wedge \mathbf{P}r_{\vartheta+\rho})$, it follows that

(1) $\upsilon_t(\mathbf{O}p_1 \wedge ... \wedge \mathbf{O}p_v \wedge \mathbf{P}p_{v+1} \wedge ... \wedge \mathbf{P}p_{v+w}) \subseteq \upsilon_t(\mathbf{O}r_m)$ for each m $(1 \leq m \leq \vartheta)$ and

(2) $\upsilon_t(\mathbf{O}p_1 \wedge ... \wedge \mathbf{O}p_v \wedge \mathbf{P}p_{v+1} \wedge ... \wedge \mathbf{P}p_{v+w}) \subseteq \upsilon_t(\mathbf{P}r_l)$ for each l $(\vartheta + 1 \leq l \leq \vartheta + \rho)$.

Consider expression (1). From definitions of valuation υ it follows that

$\upsilon_t(\mathbf{O}p_1) \wedge ... \wedge \upsilon_t(\mathbf{O}p_v) \wedge \upsilon_t(\mathbf{P}p_{v+1}) \wedge ... \wedge \upsilon_t(\mathbf{P}p_{v+w}) \subseteq \upsilon_t(\mathbf{O}r_m)$, so

$(\Diamond)$ *if* $\upsilon_o(p_1)$ *and...and* $\upsilon_o(p_v)$ *and* $(\upsilon_o(p_{v+1})$ *or* $\upsilon_i(p_{v+1}))$ *and...and* $(\upsilon_o(p_{v+w})$ *or* $\upsilon_i(p_{v+w}))$, *then* $\upsilon_o(r_m)$.

Comparing the last expression to the correspondent definitions of υ, one could see that to satisfy $(\Diamond)$, the condition $(\star)$ must be fulfilled. Similarly, condition (2) gives the expression

$(\Diamond\Diamond)$ *if* $\upsilon_o(p_1)$ *and...and* $\upsilon_o(p_v)$ *and* $(\upsilon_o(p_{v+1})$ *or* $\upsilon_i(p_{v+1}))$ *and... and* $(\upsilon_o(p_{v+w})$ *or* $\upsilon_i(p_{v+w}))$, *then* $\upsilon_o(r_m)$ *or* $\upsilon_i(r_m)$.

In accordance with definitions of υ , to satisfy $(\Diamond\Diamond)$, the condition $(\star\star)$ must be fulfilled. Therefore, both conditions $(\star)$ and $(\star\star)$ hold, and we get contradiction.

Now consider the **case2.** In which case it could be that, for some i $(1 \leq i \leq n)$, there is no j $(1 \leq j \leq m)$, such that $\upsilon_t(d^i_A) \subseteq \upsilon_t(d^j_B)$, but, at the same time, $\upsilon_t(d^i_A) \subseteq \upsilon_t(d^1_B \vee d^2_B \vee ... \vee d^m_B)$. By construction of **DNF,** there are the following possibilities.

Case 2.1. Among the disjuncts $d^1_B, d^2_B, ..., d^m_B$ there are disjuncts d and $\neg d$, which are both different from d^i_A. But that situation is impossible, since, in that case,

$\vdash d^i_A \rightarrow d^1_B \vee d^2_B \vee ... \vee d^m_B$ for any i $(1 \leq i \leq n)$, which contradicts the assumption.

Case 2.2. Among the disjuncts $d^1_B, d^2_B, ..., d^m_B$ there are disjuncts of the form $d^{i'}_A \wedge f$ and $d^{i'}_A \wedge \neg f$ (conjunction d^i_A is a subset of the conjunction $d^{i'}_A$). Suppose, we have the expression $(\bullet)$ $\dashv d^i_A \rightarrow d^1_B \vee d^2_B \vee ... \vee d^{i'}_A \wedge f \vee d^{i'}_A \wedge \neg f \vee ... \vee d^m_B$.

By the rule (**DR2**), from (•) we get $\dashv d^i_A \rightarrow d^{i'}_A \wedge f \vee d^{i'}_A \wedge \neg f$. Then, by the derived rule **(DR3):**

If $A \rightarrow B \wedge (C \vee D)$, *then* $A \rightarrow (B \wedge C) \vee (B \wedge D)$ (which, in turn, is obtained from **GA7** and **GR3),** $\dashv d^i_A \rightarrow d^{i'}_A \wedge (f \vee \neg f)$. According to **GR2,** either $\dashv d^i_A \rightarrow d^{i'}_A$, or $\dashv d^i_A \rightarrow f \vee \neg f$. Since $\vdash d^i_A \rightarrow f \vee \neg f$ (**GA12**), then $\dashv d^i_A \rightarrow d^{i'}_A$, which contradicts to the axiom (**GA3**).□

One of the main points to notice about the considered system $\mathbf{S}_{3dp}$ is that although $\mathbf{S}_{3dp}$ avoids deontic explosion, it does contain a principle that if there are deontic dilemmas, cases where $\mathbf{O}p$ and $\mathbf{O}p^*$ are both true, then there are true contradictions, cases where $\mathbf{O}p$ and $\neg\mathbf{O}p$ are both true. That is to say, $\mathbf{O}p \wedge \mathbf{O}p^* \rightarrow \mathbf{O}p \wedge \neg\mathbf{O}p$ is valid in the semantics $\mathbf{SP_3D}$, and is provable in $\mathbf{S}_{3dp}$ (with axioms $\mathbf{S}_{3dp}A7$, $\mathbf{S}_{3dp}A8$ and the rule $\mathbf{S}_{3dp}R1$). That $\mathbf{S}_{3dp}$ can accept deontic dilemmas, and hence real contradictions, and still avoid deontic explosion, is due to its being based on a paraconsistent logic, **G**, that rejects simple explosion, that a contradiction implies everything. However, by using quasi-matrix approach it seems possible to build deontic system in which both formulas $\mathbf{O}p \wedge \mathbf{O}p^* \rightarrow \mathbf{O}p \wedge \neg\mathbf{O}p$ and $\mathbf{O}p \wedge \mathbf{O}p^* \rightarrow \mathbf{O}q$ are no longer theorems.

6 Deontic Dilemmas without Real Contradictions

One of the possibilities to avoid both above mentioned formulas is to consider on the level of formulas of $\mathbf{S}_{3d}$ the *two valued quasi-matrix logic* instead of classical two valued logic. Let's reason as follows. If the act p is obligatory (possesses the value o), then this obligation (formula $\mathbf{O}p$) is true. The obligation of the forbidden act p is false, thus $\mathbf{O}p$ takes f. If the act p is normatively indifferent, then $\mathbf{O}p$ takes t/f. The cases where the obligation of the indifferent act is false are obvious. Now consider what is the possible source for the conflict of obligation $\mathbf{O}p \wedge \mathbf{O}p^*$? Normally, the code of norms is made so that it does not include inconsistent norms. The situation with inconsistent norms may arise when some act earlier considered as indifferent in relation to some code of norms, becomes obligatory or forbidden by that code. Suppose the two legislative bodies independently from each other make obligatory (forbidden) both indifferent acts p and p^*. Obviously, such situation is contingent, but it can take place, thus we choose t as an alternative truth value for $\mathbf{O}p$ when the act p is indifferent.

Consider the following semantics $\mathbf{QM}_{3dq}$.

Language

The language L_{3dq} is the same as the language L_{3d}.

Formation Rules

1. Terms, formulas and connectives for the acts are defined the same way as in $\mathbf{S}_{3d}$.
2. *Definitions of the operators* **O**, **P**:

p	$\mathbf{O}p$	$\mathbf{P}p$
o	t	t
i	t/f	t
b	f	f

The formulas take values from the field $\{t, f\}$. Definitions of logical connectives are usual.

Formula A is *valid in the given interpretation* iff it takes the value t in each alternative interpretation caused by this interpretation; A is *satisfiable in the given interpretation* iff it takes t in some alternative interpretation caused by the given interpretation; A is *satisfiable* in semantics $\mathbf{QM}_{3dq}$ iff it is valid in some interpretation; A is *valid* in semantics $\mathbf{QM}_{3dq}$ iff it is valid in each interpretation. The result of the formalization of the described semantics is the system $\mathbf{S}_{3dq}$.

The System $\mathbf{S}_{3dq}$

($\mathbf{S}_{3dq}A0$) *All tautologies of classical propositional logic;*
($\mathbf{S}_{3dq}A1$) $\mathbf{O}p \equiv \mathbf{O}p^{**}$;
($\mathbf{S}_{3dq}A2$) $\mathbf{P}p \equiv \mathbf{P}p^{**}$;
($\mathbf{S}_{3dq}A3$) $\mathbf{O}(p \otimes q) \equiv \mathbf{O}(p^* \oplus q^*)^*$;
($\mathbf{S}_{3dq}A4$) $\mathbf{O}(p \oplus q) \equiv \mathbf{O}(p^* \otimes q^*)^*$;
($\mathbf{S}_{3dq}A5$) $\mathbf{P}(p \otimes q) \equiv \mathbf{P}(p^* \oplus q^*)^*$;
($\mathbf{S}_{3dq}A6$) $\mathbf{P}(p \oplus q) \equiv \mathbf{P}(p^* \otimes q^*)^*$;
($\mathbf{S}_{3dq}A7$) $\mathbf{O}p \supset \mathbf{P}p$;
($\mathbf{S}_{3dq}A8$) $\neg\mathbf{P}p \supset \mathbf{O}p^*$;
($\mathbf{S}_{3dq}A9$) $\mathbf{O}(p \otimes q) \supset \mathbf{O}p \wedge \mathbf{O}q$;
($\mathbf{S}_{3dq}A10$) $\mathbf{P}(p \otimes q) \supset \mathbf{P}p \wedge \mathbf{P}q$;
($\mathbf{S}_{3dq}A11$) $\mathbf{O}p \vee \mathbf{O}q \supset \mathbf{O}(p \oplus q)$;
($\mathbf{S}_{3dq}A12$) $\mathbf{P}p \vee \mathbf{P}q \supset \mathbf{P}(p \oplus q)$;

Inference Rules

($\mathbf{S}_{3dq}R1$) *Modus ponens* for the formulas;
($\mathbf{S}_{3dq}R2$) *Substitution rule* for the terms and for the formulas.

Rem. Operator $\mathbf{F}$ (*it is forbidden that*) can be defined as $\mathbf{F}p \stackrel{def}{=} \mathbf{O}p^*$.

Theorem 1.6. *The system* $\mathbf{S}_{3dq}$ *is sound and complete with respect to semantics* $\mathbf{QM}_{3dq}$.

Sketch of the proof:

Soundness can be easily shown using the above table definitions. For the completeness proof one has to prove the two lemmas.

Lemma 1 is the same as the one for $\mathbf{S}_{3d}$.

Lemma 2. *There is a function* $|\cdot|_T$ *such that it possesses all the properties of an interpretation function, and for each formula* A, $|A|_T = t \Leftrightarrow A \in T$.

Consider the function $|\cdot|_T$ possessing the following properties:

- $|p|_T = o \Leftrightarrow \mathbf{O}p \in T$;
- $|p|_T = i \Leftrightarrow \mathbf{P}p \wedge \mathbf{P}p^* \in T$;
- $|p|_T = b \Leftrightarrow \mathbf{O}p^* \in T$;
- $|A|_T = t \Leftrightarrow A \in T$ (A is a formula).

It can be shown that the function $|\cdot|_T$ possesses all the properties of an interpretation function for logic $\mathbf{S}_{3dq}$. Let's show for example $|p|_T = o$ iff $|p^*|_T = b$. Suppose, $|p|_T = o$, then according to definition of $|\cdot|$, $\mathbf{O}p \in T$. By $\mathbf{S}_{3dq}A1$, $\mathbf{O}p^{**} \in T$, and by definition of $|\cdot|$, $|p^*|_T = b$. Now suppose that $|p^*|_T = b$, then $\mathbf{O}p^{**} \in T$, by $\mathbf{S}_{3dq}A1$, $\mathbf{O}p \in T$, and $|p|_T = o$. Another example. Let's show that if $|p|_T = |q|_T = i$, then $|p \oplus q|_T = i/o$. Suppose, $|p|_T = |q|_T = i$, therefore, $\mathbf{P}p \wedge \mathbf{P}p^* \in T$ and $\mathbf{P}q \wedge \mathbf{P}q^* \in T$. Now suppose that $|p \oplus q|_T = b$, then, by definition of $|\cdot|_T$, $\mathbf{O}(p \oplus q)^* \in T$, then, by $\mathbf{S}_{3dq}A1$, $\mathbf{S}_{3dq}A4$ and $\mathbf{S}_{3dq}A9$, $\mathbf{O}p^* \in T$ and $\mathbf{O}q^* \in T$. By lemma 1, $\mathbf{P}p^* \wedge \mathbf{P}p^{**} \notin T$ and $\mathbf{P}q^* \wedge \mathbf{P}q^{**} \notin T$, therefore, $\mathbf{P}p \wedge \mathbf{P}p^* \notin T$ and $\mathbf{P}q \wedge \mathbf{P}q^* \notin T$, a contradiction, hence $|p \oplus q|_T = i/o$.

The final steps of the completeness proof are the same as the ones for the system $\mathbf{S}_{3d}$.□

One can easily see that both formulas $\mathbf{O}p \wedge \mathbf{O}p^* \supset \mathbf{O}p \wedge \neg\mathbf{O}p$ and $\mathbf{O}p \wedge \mathbf{O}p^* \supset \mathbf{O}q$ are not valid in semantics $\mathbf{QM}_{3dq}$, thus the considered system $\mathbf{S}_{3dq}$ accepts deontic dilemmas without classical contradictions.

7 Conclusions and Further Research

We have constructed the three different deontic systems, $\mathbf{S}_{3d}$, $\mathbf{S}_{3dp}$ and $\mathbf{S}_{3dq}$, on the basis of possible world -free quasi-matrix semantics. The system $\mathbf{S}_{3d}$ contains the axioms $\mathbf{SDL}A2$ and $\mathbf{SDL}A3$ of **SDL** but does not contain $\mathbf{SDL}A1$ and the rule ($\mathbf{SDL}R2$) $\frac{p}{\mathbf{O}p}$ (for example, $\nvdash_{S_{3d}} \mathbf{O}(p \oplus p^*)$). But deontic explosion $\mathbf{O}p \wedge \mathbf{O}p^* \supset \mathbf{O}q$ is still the theorem of $\mathbf{S}_{3d}$. One of the purposes throughout our paper has been to build deontic system that allows for conflicts of obligation. The systems $\mathbf{S}_{3dp}$ and $\mathbf{S}_{3dq}$ both satisfy that task but for the different reasons. The system $\mathbf{S}_{3dp}$ represents modal extension of paraconsistent logic **G**, thus that $\mathbf{S}_{3dp}$ can accept deontic dilemmas, is due to its being based on a paraconsistent logic. In case of $\mathbf{S}_{3dq}$, the two valued quasi-matrix logic which acts on the level of formulas is not paraconsistent. The system $\mathbf{S}_{3dq}$ allows for conflicts of obligation but does not accept classical contradictions.

The considered logic $\mathbf{S}_{3d}$ can also be extended onto the case of five valued logic (Kouznetsov, [13]). Sometimes it is useful or even important to qualify the agent's acts not only in terms of the *strict* norms - what an agent is obliged or what is permissible for her - but also in terms of *weak* norms - what is *desirable* or is *undesirable* from the point of view of some code. The system $\mathbf{S}_{5d}$ can be built as an extension of $\mathbf{S}_{3d}$ on the case where the action sentences take the values from the field {*obligatory, desirable, indifferent, undesirable, forbidden*}. The language L_{5d} of $\mathbf{S}_{5d}$ differs from L_{3d} in the definitions of deontic connectives and in the definition of the operators- the two more operators are added, **D** (*it is desirable that...,* or *it is weakly obligatory that...*) and **U** (*it is undesirable that...,* or *it is weakly prohibited that...*). Obviously, the systems $\mathbf{S}_{5dp}$ and $\mathbf{S}_{5dq}$ can be built in a similar way as the correspondent systems $\mathbf{S}_{3dp}$ and $\mathbf{S}_{3dq}$.

As we mentioned, quasi-matrix approach seems to have advantages over possible world semantics in that it allows to construct a wide range of modal logics including the ones weaker then the standard Kripkean systems. We expect in

further research to build quasi-matrix deontic systems which use on the level of formulas the *four valued quasi-matrix logic.* The formulas will take the values from the field {*necessary truth, contingent truth, contingent false, necessary false*}. By varying the definitions of implication and of deontic operators in four valued logic one can obtain various deontic systems depending on the considerations that are taken into account in the given semantics.

References

1. Alchourrón, C.E.: Philosophical foundations of deontic logic and the logic of defeasible conditionals, *Deontic logic in computer science: normative system specification/* ed. by J.-J. Meyer and R. J. Wieringa, 1993, pp.43-84.
2. Bezhanishvili, M.N.: Hao Wang partial predicate calculi and their extensions allowing the iterations of implication, *Logical Studies,* Moscow, 2001 (in Russian).
3. Chellas, B.: Modal logic: An introduction, *Cambridge University Press,* 1980.
4. Cholvy, L., Cuppens, F.: Reasoning about norms provided by conflicting regulations, *Norms, logics and information systems: new studies in deontic logic and computer science /* ed. by P.McNamara and H.Prakken, Amsterdam, 1999, pp.247-262.
5. Ermolaeva, N.M.: O logikah, rodstvennyh ischisleniyu Hao Van'a, *Nauchno-technicheskaya informatsiya,* ser.2, N8, 1973, pp.34-37 (in Russian).
6. Goble, L.: Multiplex semantics for deontic logic, *Nordic Journal of Philosophical Logic,* 2000, Vol.5, N2, pp.113-134
7. Goble, L.: Deontic logic with relevance, *Norms, logics and information systems: new studies in deontic logic and computer science /* ed. by P.McNamara and H.Prakken, Amsterdam, 1999, pp.331-345.
8. Hansson, S.O.: Preference-based deontic logic (PDL), *Journal of Philosophical Logic,* 19, 1990, pp.75-93.
9. Hintikka, J. Impossible possible worlds vindicated, *Journal of Philosophical Logic,* 1975, Vol.44, pp.475-484.
10. Horty, J.F.: Deontic logic as founded on nonmonotonic logic, *Annals of Mathematics and Artificial Intelligence,* 9, pp.69-91.
11. Ivlev, Yu.V.: Modal logic, *Moscow University Publ.,* Moscow, 1991 (in Russian).
12. Kearns, J.: Modal semantics without possible worlds, *Journal of Symbolic Logic,* 1981, Vol.46, N1, pp.77-86.
13. Kouznetsov, A.M.: Quasi-matrix deontic logic, *PhD-thesis,* Moscow State University, 1998 (in Russian).
14. Kouznetsov, A.M.: N-valued quasi-functional logic, *Abstracts of the conference "Smirnov Readings",* Moscow, 2001, pp.109-110 (in Russian).
15. Rose, A. A formalization of the propositional calculus corresponding to Wang's calculus of partial predicates, *Zeitschrift fuer mathematische Logik und Grundlagen der Mathematik,* N9, 1963, pp.177-198.
16. Schotch, P.K., Jennings, R.E.: Non-kripkean deontic logic, *New studies in deontic logic/* ed. by R.Hilpinen, 1981, pp.149-162.
17. Torre, L.W.N. van der: Reasoning about obligations: defeasibility in preference-based deontic logic, *Thesis Publishers,* Amsterdam, 1997.
18. Wang, H.: The calculus of partial predicates and its extension to set theory, *Zeitschrift fuer mathematische Logik und Grundlagen der Mathematik,* N7, 1961, pp.283-288.
19. Wright, G.H. von: Deontic logic, *Mind* 60, 1951.

Delegation in a Role-Based Organization

Olga Pacheco[1] and Filipe Santos[2]

[1] Department of Informatics,
University of Minho,
Campus de Gualtar, 4710-057 Braga, Portugal
omp@di.uminho.pt
[2] Department of Information Systems, ISCTE,
Av. das Forças Armadas,
1600 Lisboa Codex, Portugal
filipe.santos@iscte.pt

Abstract. In an organizational context the norms that apply to an agent depend on the roles he holds in the organization. The deontic characterization of structural roles is defined when the organization is created. But an organization is not a static entity. Among the dynamic phenomena that occur in an organization there are interactions between agents consisting in a transference of obligations or permissions from an agent to another. These kind of interactions are called delegation. In this paper we analyze different ways in which delegation occurs in an organizational context. We argue that the concept of "agent in a role" is relevant to understand delegation. A deontic and action modal logic is used to specify this concept.

1 Introduction

In an organizational context, agents' behavior are ruled through norms defined by the organization. By norms we mean obligations, permissions, prohibitions or other deontic attributes. The norms that apply to an agent depend on the roles he holds in the organization. The deontic characterization of a role of the structure of an organization (structural role) is defined when the organization is created and is part of its identity. But an organization is not a static entity: it interacts with the external world (e.g. establishing contracts with other agents) and the agents that hold roles in its structure interact with each other. Among the dynamic phenomena that occur in an organization there are interactions between agents consisting in a transference of obligations, permissions, responsibilities, powers or other normative attributes, from an agent to another, or to be precise, from an agent in a role to other agents in roles. These transferences may be temporary or permanent, and correspond to a sort of redistribution of competences, temporary in many cases, that may change the organization way of working but do not change its identity.

These kind of interactions are usually called *delegation.* Delegation is a complex concept, having multiple interpretations depending on the context where

A. Lomuscio and D. Nute (Eds.): DEON 2004, LNAI 3065, pp. 209–227, 2004.

it is used. Several authors have addressed this issue, like [5], [12], [13], [4], [18], among others.

With this paper we want to contribute to the understanding of this concept, analyzing different ways in which delegation occurs. We focus the study in a role-based organizational context, taking organizations as normative systems (set of interacting agents whose behavior is ruled by norms). A role-based organization has a stable structure consisting of a set of roles, whose deontic characterization is described by a set of obligations, permissions or prohibitions. Within this context, agents always act in some role.

A delegation relationship may be established between agents holding roles of the organization structure or between agents inside the organization and agents outside of the organization.

We do not analyze motivations of agents to enter in a delegation relationship, nor the success or failure of delegation. We do not consider, either, informal or implicit delegation. We are interested in explicit and formal delegation relationships, where the agents involved are aware of the relationship, as well as all the agents that interact with them.

In this paper we will show how the delegation concept can be clarified in a role-based organization, using a deontic and action logic to express its different meanings.

The rest of the paper is organized as follows: we briefly summarize the deontic and action logic we will use to formally express the concepts analyzed. Next we present the formal model we adopt for organizations, based on this logic. Then we discuss the concept of delegation and how it could be expressed in the formal model proposed. We conclude with the discussion of further logical principles in order to deal with delegation in a role-based organization.

2 Action and Deontic Logic

Following the tradition initiated by Kanger ([9], [10]), Pörn ([15], [16]) and Lindahl [11], and followed by many others, of combining deontic and action logics to describe social interaction and complex normative concepts, a logical framework has been proposed by Pacheco and Carmo ([2], [14]) that tries to capture the notion of *action of an agent playing a role.* To know the role an agent is playing when he acts is crucial to analyze the deontic classification of the action (e.g. is it a permitted action?) and the effects of the action (e.g on action of other agents, or legal effects – obligations resultant from the action). It was proposed a new action operator of the form $E_{a:r}$ (for a an agent and r a role), being expressions of the form $E_{a:r}\psi$ read as *agent* a*, playing the role* r*, brings it about that* ψ. These actions operators were combined with personal deontic operators in order to express obligations and permissions of agents in roles ($O_{a:r}\psi$ – read as *agent* a *is obliged to bring about* ψ *by acting in role* r; $P_{a:r}\psi$ – read as *agent* a *is permitted to bring about* ψ *when acting in role* r). In [2] and [14] it is discussed if these operators should be primitive or derived from impersonal deontic opera-

tors and action operators (e.g. $O_{a:r}\psi \stackrel{def}{=} OE_{a:r}\psi$). Here we omit that discussion and adopt $O_{a:r}$ and $P_{a:r}$ as primitives, and define $F_{a:r}$ as $\neg P_{a:r}$.

This logic has been used as the formal support to the specification and analysis of role-based organizations. In this paper we will use it to discuss the concept of delegation in the same organizational context.

Next, we will present the main features of the logic proposed in [2] and [14], in a simplified way and omitting reference to the underlying semantics.

2.1 The Logic $\mathcal{L}_{\mathcal{DA}}$: Formal Language

$\mathcal{L}_{\mathcal{DA}}$ is a modal (deontic and action) first-order many-sorted language. The non-modal component of $\mathcal{L}_{\mathcal{DA}}$ is used to express factual descriptions, and properties and relationships between agents. It contains a finite number of sorts, not related with agents or roles, and three special sorts: *Ag* (the agent sort), *R* (the role sort) and *AgR* (the agent in a role sort).

As usual, for each of these sorts we assume an infinite number of variables, and possibly some constants. (We are not considering for the moment variables of the sort *AgR*). There may be functions between these sorts, but we do not consider any function with *Ag* as co-domain (the terms of sort *Ag* are either variables or constants.). The terms of each of these sorts are defined as usual.

$\mathcal{L}_{\mathcal{DA}}$ also contains a finite number of role generators, generically denoted by *rg*,of sort $(\rightarrow R)$. There is always a role generator, denoted by *itself.* Moreover, for each role generator *rg,* there exists a predicate (qualification predicate), denoted by *is-rg* of sort (Ag) and denotes a property that an agent may have.

The terms of the sorts *R* and *AgR* are built as follows:

(i) If *rg* is of sort $(\rightarrow R)$, then *rg*() is a term of sort *R* (we will write *rg,* instead of *rg*());
(ii) If t is a term of sort *Ag* and r is a term of sort *R,* then $t : r$ is a term of sort *AgR.*

From now on, we use $r, r_1, \ldots,$ to generically refer to roles, and $a, a_1, \ldots,$ to generically refer to a term of sort *Ag* (either a constant and a variable), and we will continue using $t, t_1, \ldots,$ to generically refer to terms of the appropriate sorts. Finally, $a : a$ is used as an abbreviation of *a:itself,* and $qual(a : rg)$ is an abbreviation of *is-rg*(a), and intuitively means that agent a is qualified to play the role *rg.*

The formulas of $\mathcal{L}_{\mathcal{DA}}$ are inductively defined as follows:

(i) if p is a predicate symbol of sort $(s_1, \ldots, s_n)$ and $t_1, \ldots, t_n$ are terms of sort $s_1, \ldots, s_n$, then $p(t_1, \ldots, t_n)$ is an atomic formula;
(ii) if B is a formula, then $\neg B$ is a formula;
(iii) if B_1 and B_2 are formulas, then $(B_1 \wedge B_2)$ is a formula;
(iv) if B is a formula and x^s is a variable of sort s, then $(\forall_{x^s})B$ is a formula;
(v) if B is a formula and $a : r$ is a term of sort AgR, then $E_{a:r}B$, $O_{a:r}B$ and $P_{a:r}B$ are formulas.

The other standard logical connectives ($\vee$, $\rightarrow$ and $\leftrightarrow$) and the existential quantifiers are introduced through the usual abbreviation rules, and parentheses may be omitted assuming the following priorities: first $\wedge$; then $\vee$; and finally $\rightarrow$ and $\leftrightarrow$. The forbidding operator is defined as follows: $F_{a:r}B \stackrel{abv}{=} \neg P_{a:r}B$

2.2 Axiomatization of $\mathcal{L}_{\mathcal{DA}}$

The logical principles satisfied by the proposed operators have been discussed and presented in [2] and [14]. Here we just list some of those principles.

Naturally, we assume that all tautologies are axioms of our logic, and that we have the rule of Modus Ponens (in the sense that the set of theorems of our logic is closed under Modus Ponens).

With respect to the first-order component, we have the general properties of quantifiers. We have the generalization rule (if $\vdash B$ then $\vdash (\forall_x)B$), and the following axioms (schema):

$(\forall_x)(B_1 \rightarrow B_2) \rightarrow ((\forall_x)B_1 \rightarrow (\forall_x)B_2)$
$B \rightarrow (\forall_x)B$, if x does not occur free in B
$(\forall_x)B \rightarrow B[xfree/t]$, for t a constant of sort s or a variable x_1 such that x does not occur free in B within the scope of $(\forall_{x_1})$.

($B[x^s free/t]$ denotes the formula we obtain when we replace (in B) the free occurrences of x^s by t.)

The formal properties of the action operator $E_{a:r}$ are described bellow:

Axioms:	
(T_E)	$E_{a:r}B \rightarrow B$
(C_E)	$E_{a:r}A \wedge E_{a:r}B \rightarrow E_{a:r}(A \wedge B)$
(Qual)	$E_{a:r}B \rightarrow qual(a:r)$
(Itself)	$(\forall_x)qual(x:itself)$
Proof rule:	
(RE_E)	If $\vdash A \leftrightarrow B$ then $\vdash E_{a:r}A \leftrightarrow E_{a:r}B$

With respect to the formal properties of the deontic operators, and of the relationships between each other and with the action operator, we consider the following axioms and proof-rules:

Axioms:	
(C_O)	$O_{a:r}A \wedge O_{a:r}B \rightarrow O_{a:r}(A \wedge B)$
$(O \rightarrow P)$	$O_{a:r}B \rightarrow P_{a:r}B$
$(O \rightarrow \neg P \neg)$	$O_{a:r}B \rightarrow \neg P_{a:r}\neg B$
$(O \wedge P)$	$O_{a:r}A \wedge P_{a:r}B \rightarrow P_{a:r}(A \wedge B)$
Proof rules:	
(RE_O)	If $\vdash A \leftrightarrow B$ then $\vdash O_{a:r}A \leftrightarrow O_{a:r}B$
(RM_P)	if $\vdash A \rightarrow B$ then $\vdash P_{a:r}A \rightarrow P_{a:r}B$
(RM_{EP})	If $\vdash E_{a_1:r_1}A \rightarrow E_{a_2:r_2}B$ then $\vdash P_{a_1:r_1}A \rightarrow P_{a_2:r_2}B$

More details can be found in the above referred papers.

3 Formal Specification of Role-Based Organizations

3.1 Organizations as Institutional Agents

Organizations are legally classified as *artificial persons.* Artificial persons are collective entities that have a real existence in human societies: they have *juridical personality,* which means that may be the subject of obligations or rights and they also have *legal qualification,* which means that they can exercise their rights and be responsible for the unfulfillment of their obligations. Based on this legal concept of artificial person, we introduced in [14] the concept of *institutional agent* to model organizations.

Institutional agents are agents. They interact in the society like any other agent: they can establish contracts or other normative relationships with other agents, they can hold roles, they may be the subject of obligations or other normative attributes, and may be responsible for the nonfulfilment of obligations or other "non ideal" situations.

An institutional agent has a structure formed by a set of roles. Each structural role is deontically characterized by a set of obligations, permissions, or other normative attributes. This abstract structure is supported by other agents: the holders of the roles. When an agent act in a role, his behavior will be evaluated according to the deontic characterization of the role he is playing.

An institutional agent is not capable of direct action. It always act through the holders of the roles of its structure. There must be defined how obligations of an institutional agent are transmitted to the roles of its structure (and indirectly to the holders of those roles), stating who is responsible for fulfilling them. It must be also defined what are the representative roles of the institutionalized agent, stating who is authorized to act on behalf of the institutional agent.

An agent may hold several roles.

3.2 Some Extensions to the Logic

Next we extend the logic in order to obtain a framework with the adequate expressive power to specify organizations as institutional agents.

Deontic characterization of roles and agents in roles. The deontic characterization of a role in an organization is part of the identity of the organization and does not depend on the agent that hold that role in a particular moment. To capture this idea, deontic notions are attached to roles, but they are actually interpreted as applied to the holders of such roles, when acting in such roles (deontic notions are only meaningful when applied to agents). Thus, we do not introduce new operators, but just new abbreviations:

$$O_r B \stackrel{abv}{=} (\forall_x)(qual(x:r) \rightarrow O_{x:r} B)$$
$$P_r B \stackrel{abv}{=} (\forall_x)(qual(x:r) \rightarrow P_{x:r} B)$$
$$F_r B \stackrel{abv}{=} (\forall_x)(qual(x:r) \rightarrow F_{x:r} B)$$

Apart from the set of obligations, permissions and prohibitions that are intrinsic to the role and characterize the identity of the organization, other obligations or permissions may be attributed to the role dynamically, resultant from the interaction of the organization with the external world. For instance, when an organization i has an obligation ψ resultant from a contract established with other agent, that obligation will have to be transmitted to specific roles of the organization's structure, stating who is responsible for its fulfillment (on behalf of the organization): $O_{i:i}\psi \rightarrow O_r\psi$ (for r a structural role).

By knowing the qualifications of an agent, that is, the roles the agent holds, we know what are the obligations, the permissions and the prohibitions that apply to him. But there are situations where the deontic characterization of an agent may be more complex.

A first case, happens when an agent establishes a contract with an organization accepting to hold a particular role of its structure. In most cases (e.g. collective labor contract) the agent, by accepting to hold a role, just inherits the deontic characterization of the role. But, in other cases, other obligations or permissions may be attributed to the agent in that role, distinct from the ones of the deontic characterization of the role. For example, an administrator may negotiate with a company to have his personal phone bills payed by the company; or an employee of a foreign company that has to work abroad, may negotiate with the company to have some compensation (e.g. take his family with him, pay for children school). These obligations will be called *personal obligations in a role,* represented by $O_{a:r}\psi$ (where a is an agent and r is a role) and are not intrinsic to the role.

Representative roles. Some roles may be classified as *representative roles* of other agents. This means that the holders of those roles may act on behalf of the represented agents within the scope of representation defined for those roles.

In order to represent this, the following notation has been introduced in [14]: $r{:}REP(a, B)$, that is read as follows: "r is a representative role of a with scope of representation B". The expression $r{:}REP(a, B)$ can be seen as an abbreviation of:

$$(\forall_x)(E_{x:r}B \rightarrow E_{a:a}B).$$

Here we extend this notation allowing the represented agent to be in a role other them the role *itself.* So we have:

$$r1 : REP(a : r2, B) \stackrel{def}{=} (\forall_x)(E_{x:r1}B \rightarrow E_{a:r2}B).$$

We can now add two properties imposing that B should be in the scope of $r2$ and in the scope of $r1$:

$$r1 : REP(a : r2, B) \rightarrow P_{r2}B$$
$$r1 : REP(a : r2, B) \rightarrow P_{r1}B$$

When an agent acts as representative of another agent he does not act on his own behalf. So, it is natural to impose that:

$E_{x:r1}B \wedge r1 : REP(a : r2, B) \rightarrow \neg E_{x:x}B$ [1]

There might exist cases where we can consider that a role $r1$ is a representative role of an agent a in a role $r2$, for everything permitted in $r1$. Using $r1 : REP(a : r2, *)$ to denote that, we can capture such situation by imposing the following axiom: $r1 : REP(a : r2, *) \wedge P_{r1}B \rightarrow r1 : REP(a : r2, B)$

Representative roles are crucial for organizations because an organization cannot act directly - it needs other agents to act on his behalf. Those agents are the titular of the representative roles.

Representative roles are not necessarily roles of the structure of an institutionalized agent. They may result from contracts or other normative relations that agents are free to establish between each other. An institutionalized agent, for instance, may also establish arbitrary representation contracts with other agents attributing to them representative roles for specific situations. Contracts are discussed below.

3.3 Contracts

Agents in a society are free to establish arbitrary normative relationships between each other. A particular kind of those relationships are *contracts.*

When two agents[2] establish a contract between each other, they attribute obligations, permissions and prohibitions to each other. They may also attribute roles (contractual roles) to each other and deontically characterize those roles (that means, they define what are the obligations, permissions or prohibitions associated to each role). Some of the roles may be classified as representative roles of one of the agents. In that case, it must be also defined in the contract, the scope of representation for that role.

Frequently, contracts also include conditional obligations (or conditional permissions). In particular, in legal contracts it is usual to include conditional obligations describing the effects of the fulfillment or violation(unfulfillment) of other obligations in the contract. For instance, besides an obligation $O_{x:r_1}A$ on x a contract $C(x, y)$ may include an obligation on y on the condition that x fulfills the previous obligation

$E_{x:r_1}A \rightarrow O_{y:r_2}B$,
or another obligation on x if he does not fulfill it
$\neg E_{x:rg_1}A \rightarrow O_{x:rg_1}B$.

Using $Ci(x, y)$ to denote (the content of) a contract between agents x and y, we may now represent some examples.

A first example is a contract where agents attribute roles to each other, define in the contract the deontic characterization of the two roles, classify one of the roles as a representative role of the other agent and define a conditional obligation on one of the agents:

[1] This does not mean that the representative agent is not responsible for "bad behavior".

[2] For simplicity reasons we only consider contracts between two agents.

$$C1(a,b) = qual(a{:}r1) \wedge qual(b{:}r2) \wedge O_{a:r1}A \wedge P_{a:r1}B \wedge O_{b:r2}C \wedge E_{a:r1}B \rightarrow O_{b:r2}D \wedge r1 : REP(b,A)$$

A second example is a titularity contract, where agent a accepts to hold role r in the organization i. The deontic characterization of the role r is not defined in this contract. It is defined in the organization and is inherited by agent a because he will become holder of r. However, the contract also assigns additional personal obligations and permissions to agent a in role r:

$$C2(a,i) = qual(a{:}r) \wedge O_{a:r}B \wedge P_{a:r}C \wedge O_{i:itself}D \wedge E_{a:r}\neg B \rightarrow O_{a:r}F$$

Finally, an example of a contract established by two agents a and b, where no roles are attributed to each other. So, the obligations and permissions are assigned to each agent in the role of itself:

$$C3(a,b) = O_{a:itself}A \wedge O_{b:itself}B \wedge (E_{a:itself}\neg A \rightarrow O_{a:itself}J) \wedge (E_{b:itself}\neg B \rightarrow O_{b:itself}G)$$

3.4 Specification of Organizations and Societies of Agents

A formal model for organizations based on the concepts we have formalized above, can now be presented.

The specification of an organization involves a name, i, and a structure: $ST_i =< R_i, DCR_i, TO_i, RER_i >$, formed by:

- R_i : a *set of roles* – structural roles of the organization. It is constituted by a finite set of atomic formulas of the form *is-role-str(r, i)*, stating that the role r is a role of the structure of the organization i.
- DCR_i : the *deontic characterization of each role* - obligations, permissions or prohibitions that are intrinsic to the role. It is a (finite) set of formulas of the form O_rA, P_rA or F_rA, where r is a structural role of the organization i.
- TO_i : *transmission of obligations* from the organization to specific roles of its structure. It is formed by a set of formulas of the form $O_{i:itself}A \rightarrow O_rA$ (for r a role of the structure of i).
- RER_i : contains information about the *representative roles* of the organization and its respective scope of representation. It is constituted by a set of formulas of the form $r : REP(i : i, B)$.

The specification of an organization i may also include other components, not considered here.

The description of $< i, ST_i >$ contains those aspects that do not change and define the identity of the organization. We need also to include in the specification of i, information describing the agents that in the present moment hold the roles r of the structure of i. Since this component corresponds to relationships between i and other agents (contracts that i establishes with each agent), we have decided to include it in component *NR* (normative relationships) of the specification of the society of agents (see below).

A society of agents, *SA,* is: $SA =< I\ A, nIA, NR, GK >$ where:

IA : Specification of each institutionalized agent (organization) of the society. So it is formed by a set of pairs $< x, ST_x >$, as explained above.

nIA : The component *nIA* contains the identification of the other agents that exist in the society.

NR : Contains normative relationships that agents have established between each other, and in particular the contracts that are actually in force. Contracts between organizations and agents, attributing to the agents titularity of roles of its structure, are also included in this component.

GK : Contains general knowledge about the society.

For more details and an example see [14].

4 Delegation

The concept of delegation appears in many different contexts having distinct interpretations. Next we discuss some possible interpretations of it, and try to express them in a precise way using the action and deontic logic presented above. This formalization process may contribute to clarify the concept of delegation.

4.1 What Do We Mean by Delegation?

We can classify as delegation a set of different situations, all having in common some kind of transference of tasks, responsibilities, permissions, obligations, powers or other normative attributes, from one agent to another. The different interpretations of the concept depend on issues like: why agents delegate, how do they delegate, what is delegated, among others.

An agent may want to delegate e.g. an obligation to other agent because he is not capable of fulfilling it (e.g. he does not have resources nor knowledge), he has not practical possibility of fulfilling it (e.g. he cannot be in two places at the same time), or any other reason.

In this paper we will not analyze *why* an agent delegates or *why* an agent accepts a delegated task, obligation, or other normative attribute. It is outside the scope of this paper to express and reason about motivations or intentions of agents involved in a delegation relationship. We also assume, without representing it explicitly, that when an agent delegates an obligation to another agent, he also transfers to that agent all the resources required to the effective fulfillment of the obligation.

Lets discuss now what may be delegated and how this delegation may occur.

Delegation as a normative relationship. First of all, we consider that delegation is a *normative relationship* between agents, or to be more precise, between agents playing roles. By this we mean that, agents do not simply delegate tasks, but delegate obligations, permissions, prohibitions, responsibilities, powers,... to do tasks. For instance, if we simply say that *the director of the Informatics Department delegates the task of writing the annual report to his secretary,* what we

usually mean is that she has the obligation of writing the annual report on behalf of the director. But that information is not explicit in the initial statement.

Another example: if we say that a *company X delegates in a specific administrator the task of selling a property,* do we mean that that administrator has permission to sell the property, or that he is obliged to sell it, or ...? The intended meaning is not clear.

On the other hand, if we say that *a company X delegates in a specific administrator the obligation to sell a property,* the meaning is clearer. Considering agents in roles instead of only agents is important, because as we will see, characterization of delegation depends on the roles agents are playing[3].

For simplicity reasons, we use only the deontic concepts of obligation and permission, in the description of the content of a delegation act[4]. In a deeper analysis other concepts like the one of power should be included. The normative concept of responsibility is only informally and indirectly referred.

Thus, we want to express statements similar to:

agent x delegates on agent y the obligation to bring about ϕ and the permission to bring about ψ.

What do we mean by this statement is not clear, yet.

A first remark that should be made, is that the agents involved in this delegation process, are part of some organization or some society. So, according to the perspective we adopt in this paper, they are playing roles (at least the role of *itself*). So we should reformulate the above statement:

agent x, playing role $r1$, delegates on agent y, which is playing role $r2$, the obligation to bring about ϕ and the permission to bring about ψ.

This new statement poses some other questions:

- *Is agent x (when acting in role $r1$), permitted to delegate the obligation to bring about ϕ and the permission to bring about ψ?*
- *Is agent y (when acting in role $r2$), permitted to accept the delegated obligation to bring about ϕ and permission to bring about ψ?*

The answer to these questions depends on the deontic characterization of the roles played by the agents.

[3] We can be more precise and say that in some situations what is delegated are obligations (permissions, powers, ...) to do some actions and in other situations what is delegated are obligations (permissions, powers...) to bring about certain states of affair, without specifying the concrete actions that should be made to achieve that state of affairs. It is a question of abstraction, that we do not address in this paper. Here, we adopt the latter version, omitting details about concrete actions.For works that use and discuss this distinction, see, for example, [1], [12], [13], [8].

[4] Moreover, it seems strange to us that agents could delegate prohibitions: they delegate obligations and/or permissions to bring about a state of affairs, and not to avoid a certain state of affairs. Formulas like $F_{i:r}A$ seen as $O\neg E_{i:r}A$, in our opinion, should not be delegated. However, formulas like $OE_{i:r}\neg A$ would be acceptable. But this issue needs further research.

It seams reasonable to require that the obligation ϕ and the permission ψ referred, should be in the *scope of* role r1 (i.e. it should be possible to infer them from the deontic characterization of role r1). According to this interpretation, we can only delegate obligations and permissions attributed to us[5]. Here, we only consider delegation cases that verify this restriction.

We also assume that an agent cannot delegate the obligations and permissions he has in a role that are not intrinsic to the role (that are not in the deontic characterization of the role), but result from the interaction of the agent with the organization, such as personal obligations or permissions negotiated in the labor contract, or obligations that result from sanctions to his behavior.

When an agent accepts a delegated obligation (permission,...), this new obligation will be "added" to the deontic characterization that applies to him resultant from the roles he holds. Therefore, his actions will be evaluated in this new deontic context. But, in the model we adopt, obligations and permissions are assigned to agents in roles. Thus, a question arises:

Are the obligations and permissions delegated to an agent attributed to that agent in the role he is playing in an organization?

In most cases yes, but in some other cases no.

In the former cases, the delegated obligations and permissions are just added to the agent in the role deontic characterization. For example, when the director of Department of Informatics delegates on his secretary the obligation to produce the annual report, this new obligation will be *added* to the deontic characterization of the person playing the secretary role ($O_{x:sec}\phi$). This new deontic attributes are not intrinsic to the secretary role, they are resultant from the interaction between the holder of secretary role and the holder of director of Department of Informatics. In this context, it seems natural to impose that the delegated obligation or permission should not enter in conflict with the deontic characterization of the role played by the agent that accepts them. By not entering in conflict we mean, in this context, that the same agent should not be under the obligation or permission to bring about A because he holds a role and, at the same time, under the obligation or permission to bring about $\neg A$, because he accepts a delegated obligation or permission. For later reference, we will call this kind of delegation *composed delegation.*

There are, however, other situations where what is delegated is a set of obligations and permissions that should not be seen as an additional characterization of the role played by the agent who accepts them, because they have a distinct nature. In this cases, what is delegated may be seen as a set of obligations and permissions – a new role, as we will see below – that is attributed to the agent

[5] There are situations where an agent has permission to "delegate" on others obligations (permissions,...) that are not attributed to him. For example, an administrator may "delegate" on an employee the obligation to perform a task that he is not obliged to perform. Although in natural language, the word "delegate" is sometimes used in similar contexts, we have doubts that this situation should be classified as a delegation case. So, we do not consider this kind of situations in this paper.

independently of other roles he is playing (although the role an agent is playing may be relevant to choose him as delegate). In this case, avoiding conflicts (of the kind mentioned above) is not relevant: agents may have conflicting obligations when playing different roles. For instance, it is possible to have $O_{x:r2}A$ and $O_{x:r3}\neg A$. As an agent can act only in a role at a time[6] he will have to decide what obligation he will fulfill and what obligation he will violate[7].

A typical example of this kind of delegation is when an agent in a role delegates some of its obligations or permissions to another agent (that may even be from outside the organization) through a contract, attributing to him a specific role. For later reference, we will call this kind of delegation *independent delegation.*

Role-based Delegation. In the role-based organizational model proposed in the previous section, roles are characterized by a set of obligations, permissions and prohibitions. So, we can aggregate the delegated obligations and permissions in a role and say that *agents delegate roles.* This delegated role may be just an artifact (a way of aggregating obligations and permissions and naming it) or, on the contrary, may be a role of the organization structure (usually part of another role of the structure), or a new role defined in a contract. In this paper we assume that the roles of the structure of an organization are composed of smaller roles corresponding to functions or competences associated to the former role. Those smaller roles may be viewed as a way of structuring the set of obligations, permissions, ... that deontically characterize the role, into units that "make sense". For example, a lecturer at some university, has obligations related with his competence of teaching, others related with research, other related with bureaucratic functions, ... We will assume, in this paper, that an holder of a role may only delegate one of its role units. The deontic characterization of each of the role units that constitute a role, is defined in the structure of the organization. So, we don't need to deontically characterize the delegated role when two agents establish a delegation relationship. We shall return to this issue later.

In order to represent this role-based delegation, we introduce the following notation $Delegate(x : r1, y : r2, r3)$, that is read as follows: "agent x playing role $r1$ delegates the role $r3$ on agent y that is playing role $r2$". Before we define the kind of formulas that correspond to the expression $Delegate(x : r1, y : r2, r3)$, we need to discuss some additional features of the delegation concept.

If we assume the properties discussed above for what is being delegated, we may say that an agent in a role may only delegate roles that are part of the role

[6] In cases where we can assume that agents may play several roles at the same time, we consider that there is some kind of composition of those roles, as discussed above.

[7] There are other kind of conflicts related, for instance, with incompatibility of goals (functions, competences,...) associated to roles. For instance, the President of BP cannot be President of GALP (BP and GALP are two known oil companies). We can express this kind of incompatibility using the relation $<>$ proposed on [14], where $r2 <> r3$ is defined as $(\forall_x)(qual(x : r2) \rightarrow \neg qual(x : r3))$, stating that the same agent cannot hold the two roles. Although in this paper we do not consider this kind of incompatibilities to restrict delegation, we intend to do it in the future.

he is playing: $Delegate(x : r1, y : r2, r3) \rightarrow r3 \leq r1$, where $r3 \leq r1$ is read "r3 is part of role r1" (the predicate symbol $\leq$ will be discussed later).

In the context of *composed delegation,* the role $r3$ that is being transferred to agent y will now be "added" to the role $r2$ he was playing, in the sense that the deontic characterization of the two roles will be joined, as if there were a new role. To capture this idea we will use the function on roles, $+ : R \times R \longrightarrow R$ (to be discussed later). We can say, then, that delegation makes y hold role $r2 + r3$. This is in fact a new role of the organization and its inclusion in the structure depends on the permanent or transitory character of the delegation. So, in the definition of the delegation relationship the following role attribution must occur:

$Delegate_c(x : r1, y : r2, r3) \stackrel{def}{=} is - r2 + r3(y) \wedge \ldots$

As we said before, in a context of *independent delegation* this role composition should not occur. The delegated role exists by itself and the agent to whom the role is delegated may act either in role $r2$ or in role $r3$. In this case we must have:

$Delegate_i(x : r1, y : r2, r3) \stackrel{def}{=} is - r3(y) \wedge \ldots$

Transfer of responsibility. Another issue that must be discussed is whether the obligations and permissions delegated to other agents, stay or not in the role played by the agent that delegates them. We will consider two situations: share of responsibilities and complete transfer of responsibilities.

Share of responsibilities. In this case, the agent that delegates obligations and permissions, also keeps them. So he shares the responsibility with the agent to whom he delegated them. This delegation case corresponds to a representation relationship.This means that when an agents delegates a sub-role to another agent, he his assigning him a representative role. We can express this in our logic through the following formulas, for composed delegation and independent delegation, respectively:

$Delegate_{cs}(x : r1, y : r2, r3) \stackrel{def}{=} is - r2 + r3(y) \wedge r3 : REP(x : r1, *)$

or

$Delegate_{is}(x : r1, y : r2, r3) \stackrel{def}{=} is - r3(y) \wedge r3 : REP(x : r1, *)$

In this case the agent y will act on behalf of the agent x.

Complete transfer of responsibilities. In this situation, when the agent delegates obligations or permissions he is no longer responsible for them. This means that those obligations and permissions are excluded from his role. The delegated role should be "subtracted" from the role held by the agent that delegates it. In this case, the role of the agent that delegates, is changed and becomes a sub-role of the initial role.

A possibility to express these role changes, would be to introduce in the language another function on roles: $- : R \times R \longrightarrow R$ (to be discussed later) and use it as follows:

$Delegate_{ct}(x : r1, y : r2, r3) \stackrel{def}{=} is - r1 - r3(x) \wedge is - r2 + r3(y)$

or

$Delegate_{it}(x : r1, y : r2, r3) \stackrel{def}{=} is - r1 - r3(x) \wedge is - r3(y)$

But this possibility needs further research[8].

For simplicity reasons, in the rest of the paper we will use $Delegate(x : r1, y : r2, r3)$ whenever it is not relevant to distinguish the particular kind of delegation used.

Forms of delegating. There are several forms of delegation. We will consider some of them: delegation by command, through a joint action, by institutional context, or implicitly.

Delegation by command. When agent x in role $r1$ has some kind of authority over agent y in role $r2$, delegation may be unilateral and have the form of a "command", which can be expressed as follows:

$E_{x:r1}Delegate(x : r1, y : r2, r3)$.

Delegation by joint action. Other frequent form of delegation is through a joint action, where both $x : r1$ and $y : r2$ decide do establish a delegation relationship. To express this joint action, we will use the action operator proposed in [14], $E_{\{a1:r1,\ldots,an:rn\}}$, indexed by a finite set of agents in roles.

Thus, we extend our logical language $\mathcal{L}_{\mathcal{DA}}$ with this operator. The formulas of the extended language ($\mathcal{L}^{+}_{\mathcal{DA}}$) are defined as follows:

- If B is a formula of $\mathcal{L}_{\mathcal{DA}}$, then B is (also) a formula of $\mathcal{L}^{+}_{\mathcal{DA}}$;
- If B is a formula of $\mathcal{L}_{\mathcal{DA}}$ and $t_1, \ldots, t_n$ $(n \geq 2)$ are terms of sort *AgR,* then $E_{\{t_1,\ldots,t_n\}}B$ is a formula of $\mathcal{L}^{+}_{\mathcal{DA}}$ (a joint action formula);
- Boolean combinations (through $\neg$ and $\wedge$) of formulas of $\mathcal{L}^{+}_{\mathcal{DA}}$ and universal quantifications of formulas of $\mathcal{L}^{+}_{\mathcal{DA}}$, are also formulas of $\mathcal{L}^{+}_{\mathcal{DA}}$.

We consider that each joint action operator $E_{\{a_1:r_1,\ldots,a_n:r_n\}}$ is of type ETC, and the qualification axiom extends naturally to joint action formulas as follows:

$E_{\{a_1:r_1,\ldots,a_n:r_n\}}B \rightarrow qual(a_1 : r_1) \wedge \ldots \wedge qual(a_n : r_n)$

We are now in position to express delegation established through a joint action:

$E_{\{x:r1,y:r2\}}Delegate(x : r1, y : r2, r3)$

Institutional delegation. The several delegation situations we have analyzed are just particular cases of *contracts* between agents. So, another claim could be that: to delegate is to establish a contract with the particularities we have discussed. But this is not accurate. There are cases where delegation does not correspond to a relationship between the agent that delegates and the agent that accepts the delegated role.

[8] Examples of open questions are: *Should we cancel the qualification of* x *to play role* $r1$*?*, or *What is the meaning of having* $r - r$*?*.

Consider for instance a situation where an agent x that plays role $r1$ in an organization is absent. It is usual that another agent z (e.g. his boss, that plays role r) delegates on other agent (y that plays role $r2$) the role $r1$, until x returns to the organization. We can express this situation by

$E_{z:r}Delegate(x : r1, y : r2, r1)$

Sometimes these situations are predefined in an organization, and agent z might be the institutional agent itself.

Informal delegation. We do not consider implicit delegation, in the sense of informal delegation relationships that agents may define between each other. Those relationships have no normative effects, in the sense that, if something fails, responsibilities could not be attributed to the agents involved.

4.2 Examples

So, we conclude this section presenting some examples of different types of delegation.

Example 1. *One of the functions of the Director of Department of Informatics (ddi), a, is to produce an annual report (wr). Associated to this sub-role (wr) there is the obligation to write the annual report of the Department (ψ) and the permission to use the director's computer(ϕ). He delegates on his Secretary (sdi), c, this role wr.*

$E_{a:ddi}\ \ Delegate_{cs}(a : ddi, c : sdi, wr) \stackrel{abv}{=}$
$E_{a:ddi}\ \ is - sdi + wr(c) \wedge wr : REP(a : ddi, *)$

The role delegated is called *wr* and is a part-of role *ddi.* Role *wr* is characterized as: $P_{wr}\phi \wedge O_{wr}\psi$ The secretary writes the report on behalf of the director (i.e. her action will count as an action of the Director). He is still responsible for the report. Notice that this delegation has the form of a command (due to the authority the Director has over the Secretary).

Example 2. *The Director of Department of Informatics (ddi), a, has the obligation to convoke the General Assembly of the Department (ϕ), once a year. He delegates this task, permanently, on the Assistant-Director (addi).*

$E_{\{a:ddi,b:addi\}}\ \ Delegate_{ct}(a : ddi, b : addi, cga) \stackrel{abv}{=}$
$E_{\{a:ddi,b:addi\}}\ \ is - addi + cga(b) \wedge is - ddi - cga(a)$

The role delegated is called *cga* and is part of role *ddi* and is characterized by $O_{cga}\phi$. This role is added to the role *addi* and subtracted from the role *ddi.* So, from now on, the agents a and b will hold different roles.

Example 3. *The Director of Department of Informatics (ddi), a, will be absent and delegates in the Assistant-Director (addi), b, all his competencies.*

$E_{\{a:ddi,b:addi\}}\ \ Delegate_{cs}(a : ddi, b : addi, ddi) \stackrel{abv}{=}$
$E_{\{a:ddi,b:addi\}}\ \ is - ddi + addi(b) \wedge ddi : REP(a : ddi, *)$

The role delegated to b is now the whole role *ddi*. In the absence of the Director of Department a, b will act as his representative , for everything the Director would have to do or would be permitted to do.

Another possible interpretation of this situation would be to say that, during that period, b is the director of the Department of Informatics. The main difference between this interpretation and the one presented before, is that in this case, b would be the only responsible for his actions as Director of Department. While in the previous case, the responsibility also goes to a. In this latter case we would have:

$E_{\{a:ddi,b:addi\}} \quad Delegate(a : ddi, b : addi, ddi) \stackrel{abv}{=}$
$E_{\{a:ddi,b:addi\}} \quad is - ddi(b)$

5 Extensions to the Formal Specification of Role-Based Organizations

In order to include the previous role-based delegation on the formal specification of organizations we need to consider further deontic logical principles related with the concepts of part-of-role, joining roles and role subtraction mentioned before.

With respect to the concept of *part-of-role* we introduce in the language a new predicate symbol $\leq$ where $r1 \leq r2$ expresses the fact that $r1$ is part of role $r2$, which means that all obligations and permissions that characterize $r1$ also characterize $r2$. Thus the logical principles:

$r1 \leq r2 \rightarrow (O_{r1}\phi \rightarrow O_{r2}\phi)$ and
$r1 \leq r2 \rightarrow P_{r1}\phi \rightarrow P_{r2}\phi$.

follow intuitively from what we want to express.

The joining roles function $+ : R \times R \longrightarrow R$ also brings the need for the following logical principles:

$O_r A \rightarrow O_{r+s} A$
$O_{r+s} A \rightarrow (O_r A \vee O_s A)$
$P_r A \rightarrow P_{r+s} A$
$P_{r+s} A \rightarrow (P_r A \vee P_s A)$

that follow the idea that a composite role also inherits all the obligations and permissions of its role composites.

On the other hand, the subtracting role function $- : R \times R \longrightarrow R$, introduces the need for the following logical principles:

$O_s A \rightarrow \neg O_{r-s} A$
$P_s A \rightarrow \neg P_{r-s} A$

following the idea that role obligations and permissions no longer apply when this role is subtracted from another role (the main role). Note however that the other obligations and permissions remain in the main role, i.e.

$O_r A \wedge \neg O_s A \rightarrow O_{r-s} A$
$O_{r-s} A \rightarrow O_r A$

Over and above the deontic principles we naturally assume role composition expressed by functions + and – as a way of expressing part-of-roles:

$r \leq r + s$
$r - s \leq r$

Finally,

$E_{x:r} A \wedge is - r + s(x) \rightarrow E_{r+s} A$

may help to characterize obligation fulfillment in this new deontic context.

Concerning the specification of organizations, our main idea is to specify organizations by roles r fragmented into smaller roles $r1, r2, \ldots, rn$ and express this by $r = r1 + r2 + \ldots + rn$. Intuitively, each part *ri* of role r may correspond to a particular function (or competence) associated to role r. We will assume that an holder of role r may only delegate one of this units.

Introducing a predicate symbol = of sort *(R,R),* reflexive, symmetric and transitive, we foresee the need for the following principles:

$r1 + r2 = r2 + r1$
$(r1 + r2) + r3 = r1 + (r2 + r3)$
$r + r = r$
$r = s \rightarrow r \leq s \; (r - s) + s = r$
$r = s \rightarrow (t - s = t - r)$
$r = s \rightarrow (t + s = t + r)$
$r = s \rightarrow (O_r A = O_s A)$

Due to space limitations it is not possible to present a full example of an organization and of a society of agents that includes that organizations. Examples can be found in [14]. We conclude this section with some comments about the inclusion of delegations cases in the formal specification of organizations. We need to include a component in the structure of the organization describing the units that compose each role. Delegations that occur in an organization are included in the *NR* component of a society of agents that contains normative relations of the agents of the society.

6 Conclusion and Future Work

In this paper we discussed the concept of delegation in an organizational context. We considered role-based organizations – organizations structured by a set of roles, which are held by agents. Moreover, organizations are seen as normative systems – a set of interacting agents whose behavior is ruled by norms (obligations, permissions and prohibitions) resultant from the deontic characterization of the roles the agents' hold. In that context, delegation is classified as a normative relation between agents, where agents transfer some (or all) of their deontic qualifications to other agents. A deontic and action modal logic has been

used to express the different interpretations of the delegation concept. This is a preliminary approach to delegation. Many open questions remain.

One of the questions we intend to address is how to express the fact that not everything (obligations, permissions, powers of a role) may be object of delegation. There are some cases where an obligation (permission) must be fulfilled directly by a particular agent in a role. A possible approach would be to distinguish direct and indirect action, using, for example, a direct action operator like the one proposed in [17], D_x, and adapted in [2] to direct actions of agents in roles. Using this operator we may express obligations that may be delegated from others that may not, using expressions similar to $OE_{a:r}A$ for the former and $OD_{a:r}$ for the latter. If this approach is adopted, we have to use impersonal obligations and permissions instead of the personal ones we have used in this paper.

Another issue that needs further research is composition of roles. A formal study of the functions referred in the paper must be done.

Other deontic concepts must be included in the characterization of delegation, specially the concept of power and representation. See [4], [15], for work on this issues.

We are aware that, given its static nature, the type of logic proposed so far is not fully adequate to characterize the dynamic aspects referred in the previous delegation examples. There are in fact two relevant snapshots in the delegation process: before and after delegation. Before a delegation $Delegate(x : r1, y : r2, r3)$ we naturally expect that agents x and y hold roles r1 and r2 respectively, i.e. $is{-}r1(x)$ and $is{-}r2(y)$. However, after delegation, the deontic qualification of both agents may change and as a consequence it may happen that $\neg is-r1(x)$ or $\neg is-r2(y)$. Obviously, these formulas together introduce a logical inconsistency. To overcome, this problem a possible approach would be to introduce temporal operators.

References

1. A.Artikis, J.Pitt and M. Sergot: "Animated Specifications of Computational Societies", Proceedings of the first International Joint Conference on Autonomous Agents and Multi-agent Systems (AAMAS'02), pp. 1053-1061, ACM, 2002.
2. J. Carmo and O. Pacheco: "Deontic and action logics for organized collective agency, modeled through institutionalized agents and roles", *Fundamenta Informaticae,* Vol.48 (No. 2,3), pp. 129-163, IOS Press, November, 2001.
3. B. J. Chellas: *Modal Logic - an Introduction,* Cambridge University Press, 1980.
4. J. Gelati and G. Sartor: "Normative co-ordination and the delegation of agency: declarative power, representation and mandate", 2002.
5. G. Governatori, J. Gelati, A. Rotolo and G. Sartor: "Actions, Institutions, Powers. Preliminary Notes", *International Workshop on Regulated Agent-Based Social Systems (RASTA'02),* Fachbereich Informatik, Universität Hamburg, pp. 131-147, 2002.
6. A.G. Hamilton: *Logic for Mathematicians,* Cambridge University Press, 1988.
7. R. Hilpinen (ed.), *Deontic Logic: Introductory and Sistematic Readings,* Dordrecht: D.Reidel, 1971.

8. A. J. I. Jones and M. J. Sergot: "A Formal Characterization of Institutionalized Power", *Journal of the IGPL ,* **4(3)**, pp.429-445, 1996. Reprinted in E. Garzón Valdés, W. Krawietz, G. H. von Wright and R. Zimmerling (eds.), *Normative Systems in Legal and Moral Theory,* (Festschrift for Carlos E. Alchourrón and Eugenio Bulygin), Berlin: Duncker & Humblot, pp.349-369, 1997.
9. S. Kanger: *New Foundations for Ethical Theory,* Stockholm, 1957. (Reprinted in [7].)
10. S. Kanger: "Law and Logic", *Theoria,* **38**, 1972.
11. L. Lindahl: *Position and Change - A Study in Law and Logic,* Synthese Library **112**, Dordrecht:D. Reidel, 1977.
12. T.J. Norman and C. Reed: "Delegation and Responsibility", In *Intelligent Agents VII,* volume 1986 of LNAI, Springer-Verlag, 2001.
13. T.J. Norman and C. Reed: "Group Delegation and Responsibility", Proceedings of the first International Joint Conference on Autonomous Agents and Multi-agent Systems (AAMAS'02), ACM, pp. 1053-1061, 2002.
14. O. Pacheco and J. Carmo: " A Role Based Model for the Normative Specification of Organized Collective Agency and Agents Interaction", *Journal of Autonomous Agents and Multi-Agent Systems,* Vol. 6, Issue 2, pp. 145-184, Kluwer, March 2003.
15. I. Pörn: *The Logic of Power.* Oxford : Blackwell, 1970.
16. I. Pörn: *Action Theory and Social Science: Some Formal Models,* Synthese Library, **120**, Dordrecht : D. Reidel, 1977.
17. F. Santos and J. Carmo: " Indirect Action, Influence and Responsibility", in M. Brown and J.Carmo (eds.), *Deontic Logic, Agency and Normative Systems,* Springer, Workshops in Computing Series, 194-215, 1996.
18. K. Vender and M. Olivier: "The delegation authorization model: a model for the dynamic delegation of authorization rights in a secure workflow management system", in ISSA2002, Muldersdrift, South Africa, 2002. Published electronically in http://mo.co.za.

Automatic Verification of Deontic Properties of Multi-agent Systems

Franco Raimondi and Alessio Lomuscio

Department of Computer Science
King's College London
London, UK
{franco,alessio}@dcs.kcl.ac.uk

Abstract. We present an algorithm and its implementation for the verification of correct behaviour and epistemic states in multiagent systems. The verification is performed via model checking techniques based on OBDD's. We test our implementation by means of a communication example: the bit transmission problem with faults.

1 Introduction

In the last two decades, the paradigm of multiagent systems (MAS) has been employed successfully in several fields, including, for example, philosophy, economics, and software engineering. One of the reasons for the use of MAS formalism in such different fields is the usefulness of ascribing autonomous and social behaviour to the components of a system of agents. This allows to *abstract* from the details of the components, and to focus on the *interaction* among the various agents.

Besides *abstracting* and *specifying* the behaviour of a complex system by means of MAS formalisms based on logic, recently researchers have been concerned with the problem of *verifying* MAS, i.e., with the problem of certifying formally that a MAS satisfies its specification.

Formal verification has its roots in software engineering, where it is used to verify whether or not a system behaves as it is supposed to. One of the most successful formal approaches to verification is *model checking*. In this approach, the system S to be verified is represented by means of a logical model M_S representing the computational traces of the system, and the property P to be checked is expressed via a logical formula φ_P. Verification via model checking is defined as the problem of establishing whether or not $M_S \models \varphi_P$. Various tools have been built to perform this task automatically, and many real-life scenarios have been tested.

Unfortunately, extending model checking techniques for the verification of MAS does not seem to be an easy task. This is because model checking tools consider standard reactive systems, and do not allow for the representation of the social interaction and the autonomous behaviour of agents. Specifically, traditional model checking tools assume that M is "simply" a *temporal* model, while MAS need more complex formalisms. Typically, in MAS we want to reason about epistemic, deontic, and doxastic properties of agents, and their temporal evolution. Hence, the logical models required are richer than the temporal model used in traditional model checking.

A. Lomuscio and D. Nute (Eds.): DEON 2004, LNAI 3065, pp. 228–242, 2004.

Various ideas have been put forward to verify MAS. In [20], M. Wooldridge et al. present the MABLE language for the specification of MAS. In this work, non-temporal modalities are translated into nested data structures (in the spirit of [1]). Bordini et al. [2] use a modified version of the AgentSpeak(L) language [18] to specify agents and to exploit existing model checkers. Both the works of M. Wooldridge et al. and of Bordini et al. translate the MAS specification into a SPIN specification to perform the verification. In this line, the attitudes for the agents are reduced to predicates, and the verification involves only the temporal verification of those. In [8] a methodology is provided to translate a deontic interpreted system into SMV code, but the verification is limited to static deontic and epistemic properties, i.e. the temporal dimension is not present, and the approach is not fully symbolic. The works of van der Meyden and Shilov [12], and van der Meyden and Su [13], are concerned with the verification of temporal and epistemic properties of MAS. They consider a particular class of interpreted systems: synchronous distributed systems with perfect recall. An automata-based algorithm for model checking is introduced in the first paper using automata. In [13] an example is presented, and [13] suggests the use of OBDD's for this approach, but no algorithm or implementation is provided.

In this paper we introduce an algorithm to model check MAS via OBDD's. In particular, in this work we investigate the verification of epistemic properties of MAS, and the verification of the "correct" behaviour of agents.

Knowledge is a fundamental property of the agents, and it has been used for decades as key concept to reason about systems[5]. In complex systems, reasoning about the "correct" behaviour is also crucial. As an example, consider a client-server interaction in which a server fails to respond as quickly as it is supposed to a client's requests. This is an unwanted behaviour that may, in certain circumstances, crash the client. It has been shown[14] that correct behaviour can be represented by means of deontic concepts: as we show in this paper, model checking deontic properties can help in establishing the extent in which a system can cope with failures. We give an example of this in Section 5.2, where two possible "faulty" behaviours are considered in the bit transmission problem[5], and key properties of the agents are analysed under these assumptions. In one case, the incorrect behaviour does not cause the whole system to fail; in the second case, the incorrect behaviour invalidates required properties of the system. We use this as a test example, but we feel that similar situations can arise in many areas, including database management, distributed applications, communication scenarios, etc.

The rest of the paper is organised as follows. In Section 2 we review the formalism of deontic interpreted systems and model checking via OBDD's. In Section 3 we introduce an algorithm for the verification of deontic interpreted systems. An implementation of the algorithm is then discussed in Section 4. In Section 5 we test our implementation by means of an example: the bit transmission problem with faults. We conclude in Section 6.

2 Preliminaries

In this section we introduce the formalisms and the notation used in the rest of the paper. In Section 2.1 we review briefly the formalism of interpreted systems as presented in [5]

to model a MAS, and its extension to reason about the correct behaviour of some of the agents as presented in [9]. In Section 2.2 we review some model checking methodologies.

2.1 Deontic Interpreted Systems and Their Temporal Extension

An interpreted system [5] is a semantic structure representing a system of agents. Each agent in the system i ($i \in \{1, \ldots, n\}$) is characterised by a set of *local states* L_i and by a set of actions Act_i that may be performed. Actions are performed in compliance with a protocol $P_i : L_i \rightarrow 2^{Act_i}$ (notice that this definition allows for non-determinism). A tuple $g = (l_1, \ldots, l_n) \in L_1 \times \ldots, L_n$, where $l_i \in L_i$ for each i, is called a *global state* and gives a description of the system at a particular instance of time. Given a set I of *initial global states,* the evolution of the system is described by n evolution functions t_i (this definition is equivalent to the definition of a single evolution function t as in [5]): $t_i : L_1 \times \ldots \times L_n \times Act_1 \times \ldots \times Act_n \rightarrow L_i$. In this formalism, the environment in which agents "live" is usually modelled by means of a special agent E; we refer to [5] for more details. The set I, the functions t_i, and the protocols P_i generate a set of *computations* (also called *runs*). Formally, a computation π is a sequence of global states $\pi = (g_0, g_1, \ldots)$ such that $g_0 \in I$ and, for each pair $(g_j, g_{j+1}) \in \pi$, there exists a set of actions a enabled by the protocols such that $t(g_j, a) = g_{j+1}$. $G \subseteq (L_1 \times \ldots \times L_n)$ denotes the set of *reachable* global states.

In [9] the notion of *correct behaviour* of the agents is incorporated in this formalism. This is done by dividing the set of local states into two disjoint sets: a non-empty set G_i of allowed (or "green") states, and a set R_i of disallowed (or faulty, or "red") states, such that $L_i = G_i \cup R_i$, and $G_i \cap R_i = \emptyset$. Given a countable set of propositional variables $\mathcal{P} = \{p, q, \ldots\}$ and a valuation function for the atoms $\mathcal{V} : \mathcal{P} \rightarrow 2^G$, a deontic interpreted systems is a tuple $DIS = (G, \{\sim_i\}_{i \in \{1,\ldots,n\}}, \{\prec_i^O\}_{i \in \{1,\ldots,n\}}, R_t, \mathcal{V})$. The relations $\sim_i$ are epistemic accessibility relations defined for each agent i by: $g \sim_i g'$ iff $l_i(g) = l_i(g')$, i.e. if the local state of agent i is the same in g and in g' (notice that this is an equivalence relation). The relations $\prec_i^O$ are accessibility relations defined by $g \prec_i^O g'$ iff $l_i(g') \in G_i$, i.e. if the local state of i in g' is a "green" state. We refer to [9] for more details. The relation R_t is a temporal relation between two global states. Deontic interpreted systems can be used to evaluate formulae involving various modal operators. Besides the standard boolean connectives, the language considered in [9] includes:

- A deontic operator $O_i\varphi$, denoting the fact that *under all the correct alternatives for agent i, φ holds.*
- An epistemic operator $K_i\varphi$, whose meaning is *agent i knows φ.*
- A particular form of knowledge $\widehat{K}_i^j \varphi$ denoting the knowledge about a fact φ that an agent i has *on the assumption that agent j is functioning correctly.*

We extend this language by introducing the following temporal operators: $EX(\varphi), EG(\varphi), E(\varphi U \psi)$. Formally, the language we use is defined as follows:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \vee \varphi \mid EX\varphi \mid EG\varphi \mid E(\varphi U \varphi) \mid K_i(\varphi) \mid O_i(\varphi) \mid \widehat{K}_i^j(\varphi)$$

We now define the semantics for this language. Given a deontic interpreted system *DIS,* a global state g, and a formula φ, satisfaction is defined as follows:

$DIS, g \models p$ iff $g \in \mathcal{V}(p)$,
$DIS, g \models \neg\varphi$ iff $g \not\models \varphi$,
$DIS, g \models \varphi_1 \vee \varphi_2$ iff $g \models \varphi_1$ or $g \models \varphi_2$,
$DIS, g \models EX\varphi$ iff there exists a computation π such that $\pi_0 = g$ and $\pi_1 \models \varphi$,
$DIS, g \models EG\varphi$ iff there exists a computation π such that $\pi_0 = g$ and $\pi_i \models \varphi$ for all $i \geq 0$.
$DIS, g \models E(\varphi U \psi)$ iff there exists a computation π such that $\pi_0 = g$ and a $k \geq 0$ such that $\pi_k \models \psi$ and $\pi_i \models \varphi$ for all $0 \leq i < k$,
$DIS, g \models K_i\varphi$ iff $\forall g' \in G, g \sim_i g'$ implies $g' \models \varphi$
$DIS, g \models O_i\varphi$ iff $\forall g' \in G, g \prec_i^O g'$ implies $g' \models \varphi$
$DIS, g \models \hat{K}_i^j \varphi$ iff $\forall g' \in G, g \sim_i g'$ and $g \prec_j^O g'$ implies $g' \models \varphi$

In the definition above, π_j denotes the global state at place j in computation π. Other temporal modalities can be derived, namely *AX, EF, AF, AG, AU.* We refer to [5,9, 15] for more details.

2.2 Model Checking Techniques

The problem of model checking can be defined as establishing whether or not a model M satisfies a formula φ ($M \models \varphi$). Though M could be a model for any logic, traditionally the problem of building tools to perform model checking automatically has been investigated almost only for *temporal* logics [4,7].

The model M is usually represented by means of a dedicated programming language, such as PROMELA[6] or SMV[11]. The verification step avoids building the model M explicitly from the program; instead, various techniques have been investigated to perform a *symbolic* representation of the model and the parameters needed by verification algorithms. Such techniques are based on automata [6], *ordered binary decision diagrams* (OBDD's, [3]), or other algebraic structures. These approaches are often referred to as *symbolic model checking* techniques. For the purpose of this paper, we review briefly symbolic model checking using OBDD's.

It has been shown that OBDD's offer a compact representation of boolean functions. As an example, consider the boolean function $a \wedge (b \vee c)$. The truth table of this function would be 8 lines long. Equivalently, one can evaluate the truth value of this function by representing the function as a directed graph, as exemplified on the left-hand side of Figure 1 . As it is clear from the picture, under certain assumptions, this graph can be simplified into the graph pictured on the right-hand side of Figure 1. This "reduced" representation is called the OBDD of the boolean function. Besides offering a compact representation of boolean functions, OBDD's of different functions can be composed efficiently. We refer to [3,11] for more details.

The key idea of model checking temporal logics using OBDD's is to represent the model M and all the parameters needed by the algorithms by means of boolean functions. These boolean functions can then be encoded as OBDD's, and the verification step can operate directly on these. The verification is performed using fix-point characterisation of the temporal logics operators. We refer to [7] for more details. Using this technique, systems with a state space in the region of 10^{40} have been verified.

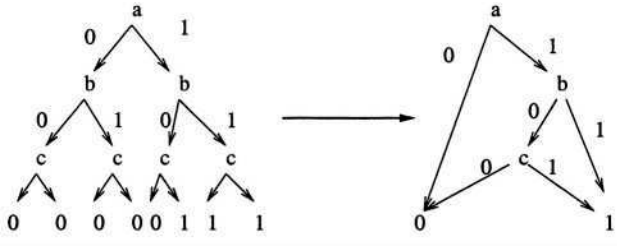


Fig. 1. OBDD representation for $a \wedge (b \vee c)$.

3 Model Checking Deontic Properties of Interpreted Systems

In this section we present an algorithm for the verification of deontic, epistemic, and temporal modalities of MAS, extending with deontic modalities the work that appeared in [17]. Our approach is similar, in spirit, to the traditional model checking techniques for the logic CTL. Indeed, we start in Section 3.1 by representing the various parameters of the system by means of boolean formulae. In Section 3.2, we provide and algorithm based on this representation for the verification step. The whole technique uses deontic interpreted systems as its underlying semantics.

3.1 From Deontic Interpreted Systems to Boolean Formulae

In this section we translate a deontic interpreted system into boolean formulae. As boolean formulae are built using boolean variables, we begin by computing the required number of boolean variables. To encode local states of an agent, the number of boolean variables required is $nv(i) = \lceil log_2 |L_i| \rceil$. To encode actions, the number of variables required is $na(i) = \lceil log_2 |Act_i| \rceil$. Hence, given $N = \sum_i nv(i)$, a global state can be encoded by means N boolean variables: $g = (v_1, \ldots, v_N)$. Similarly, given $M = \sum_i na(i)$, a joint action can be encoded as $a = (a_1, \ldots, a_M)$.

Having encoded local states, global states, and actions by means of boolean variables, all the remaining parameters can be expressed as boolean functions as follows. The protocols relate local states to set of actions, and can be expressed as boolean formulae. The evolution functions can be translated into boolean formulae, too. Indeed, the definition of t_i in Section 2.1 can be seen as specifying a list of *conditions* $c_{i,1}, \ldots, c_{i,k}$ under which agent i changes the value of its local state. Each $c_{i,j}$ has the form "if [conditions on global state and actions] then [value of "next" local state for i]". Hence, t_i is expressed as a boolean formula as follows: $t_i = c_{i,1} \oplus \ldots \oplus c_{i,k}$ where $\oplus$ denotes exclusive-or. We assume that the last condition $c_{i,k}$ of t_i prescribes that, if none of the conditions on global states and actions in $c_{i,j} (j < k)$ is true, then the local state for i does not change. This assumption is key to keep compact the description of the system, so that only the conditions causing a change in the configuration of the system need to be listed. The evaluation function $\mathcal{V}$ associates a set of global states to each propositional atom, and so it can be translated into a boolean function.

In addition to these parameters, the algorithm presented in Section 3.2 requires the definition of a boolean function $R_t(g, g')$ representing a temporal relation between g and g'. $R_t(g, g')$ can be obtained from the evolution functions t_i as follows. First, we introduce a *global* evolution function t:

$$t = \bigwedge_{i \in \{1,\ldots,n\}} t_i = \bigwedge_{i \in \{1,\ldots,n\}} (c_{i,1} \oplus \ldots \oplus c_{i,k_i})$$

Notice that t is a boolean function involving two global states and a joint action $a = (a_1, \ldots, a_M)$. To abstract from the joint action and obtain a boolean function relating two global states only, we can define R_t as follows:
$R_t(g, g')$ iff $\exists a \in Act : t(g, a, g')$ is true and each local action $a_i \in a$ is enabled by the protocol of agent i in the local state $l_i(g)$.

The quantification over actions above can be translated into a propositional formula using a disjunction (see [11,4] for a similar approach to boolean quantification):

$$R_t(g, g') = \bigvee_{a \in Act} [(t(g, a, g') \wedge P(g, a)]$$

where $P(g, a)$ is a boolean formula imposing that the joint action a must be consistent with the agents' protocols in global state g. The relation R_t gives the desired boolean relation between global states.

3.2 The Algorithm

In this section we present the algorithm $SAT(\varphi)$ to compute the set of global states in which a formula φ holds. The following are the parameters needed by the algorithm:

- the boolean variables $(v_1, \ldots, v_N)$ and $(a_1, \ldots, a_M)$ encoding global states and joint actions;
- n boolean functions $P_i(v_1, \ldots, v_N, a_1, \ldots, a_M)$ encoding the protocols of the agents;
- the function $\mathcal{V}(p)$ returning the set of global states in which the atomic proposition p holds. We assume that the global states are returned encoded as a boolean function of $(v_1, \ldots, v_N)$;
- the set of initial states I, encoded as a boolean function;
- the set of reachable states G. This can be computed as the fix-point of the operator $\tau = (I(g) \vee \exists g'(R_t(g', g) \wedge Q(g'))$ where $I(g)$ is true if g is an initial state and Q denotes a set of global states. The fix-point of τ can be computed by iterating $\tau(\emptyset)$ by standard procedure (see [11]);
- the boolean function R_t encoding the temporal transition;
- n boolean functions R_i encoding the accessibility relations $\sim_i$ (these functions are defined using equivalence on local states of G);
- n boolean functions R_i^O encoding the deontic accessibility relations $\prec_i^O$.

The algorithm is as follows:

$SAT(\varphi)$ {
  φ is an atomic formula: return $\mathcal{V}(\varphi)$;
  φ is $\neg\varphi_1$: return $G \setminus SAT(\varphi_1)$;
  φ is $\varphi_1 \wedge \varphi_2$: return $SAT(\varphi_1) \cap SAT(\varphi_2)$;
  φ is $EX\varphi_1$: return $SAT_{EX}(\varphi_1)$;
  φ is $E(\varphi_1 U \varphi_2)$: return $SAT_{EU}(\varphi_1, \varphi_2)$;
  φ is $EG\varphi_1$: return $SAT_{EG}(\varphi_1)$;
  φ is $K_i\varphi_1$: return $SAT_K(\varphi_1, i)$;
  φ is $O_i\varphi_1$: return $SAT_O(\varphi_1, i)$;
  φ is $\widehat{K}_i^j\ \varphi_1$: return $SAT_{KH}(\varphi_1, i, j)$;
}

In the algorithm above, SAT_{EX}, SAT_{EG}, SAT_{EU} are the standard procedures for CTL model checking [7], in which the temporal relation is R_t and, instead of temporal states, global states are considered. The procedures $SAT_K(\varphi, i)$, $SAT_O(\varphi, i)$ and $SAT_{KH}(\varphi, i, j)$ return a set of states in which the formulae $K_i\varphi$, $O_i\varphi$ and $\widehat{K}_i^j\ \varphi$ are true. Their implementation is presented below.

$SAT_K(\varphi, i)$ {
  X = $SAT(\neg\varphi)$;
  Y = $\{g \in G | \exists g' \in X$ and $R_i(g, g')\}$
  return $\neg$Y;
}

$SAT_O(\varphi, i)$ {
  X = $SAT(\neg\varphi)$;
  Y = $\{g \in G | \exists g' \in X$ and $R_i^O(g, g')\}$
  return $\neg$Y;
}

$SAT_{KH}(\varphi, \Gamma)$ {
  X = $SAT(\varphi)$;
  Y = $\{g \in G | \exists g' \in X$ and $R_i(g, g')$ and $R_j^O(g, g')\}$
  return $\neg$Y;
}

Notice that all the parameters can be encoded as OBDD's. Moreover, all the operations in the algorithms can be performed on OBDD's.

The algorithm presented here computes the set of states in which a formula holds, but we are usually interested in checking whether or not a formula holds in the whole model. $SAT(\varphi)$ can be used to verify whether or not a formula φ holds in a model by comparing two set of states: the set $SAT(\varphi)$ and the set of reachable states G. As sets of states are expressed as OBDD's, verification in a model is reduced to the comparison of the two OBDD 's for $SAT(\varphi)$ and for G.

4 Implementation

In this section we present an implementation of the algorithm introduced in Section 3. In Section 4.1 we define a language to encode deontic interpreted systems symbolically, while in Section 4.2 we describe how the language is translated into OBDD's and how the algorithm is implemented. The implementation is available for download[16].

4.1 How to Define a Deontic Interpreted System

To define a deontic interpreted system it is necessary to specify all the parameters presented in Section 2.1. In other words, for each agent, we need to represent:

- a list of local states, and a list of "green" local states;
- a list of actions;
- a protocol for the agent;
- an evolution function for the agent.

In our implementation, the parameters listed above are provided via a text file. The formal syntax of a text file specifying a list of agents is as follows:

```
agentlist ::= agentdef |
              agentlist agentdef
agentdef ::= "Agent" ID
                 LstateDef;
                 LgreenDef;
                 ActionDef;
                 ProtocolDef;
                 EvolutionDef;
             "end Agent"
LstateDef ::= "Lstate = {" IDLIST "}"
LgreenDef ::= "Lgreen = {" IDLIST "}"
ActionDef ::= "Action = {" IDLIST "}"
ProtocolDef ::= "Protocol"
                  ID ": {" IDLIST "}";
                  ...
                "end Protocol"
EvolutionDef ::= "Ev:"
                   ID "if" BOOLEANCOND;
                   ...
                 "end Ev"
IDLIST ::= ID |
           IDLIST "," ID
ID ::= [a-zA-Z][a-zA-Z0-9_]*
```

In the definition above, `BOOLEANCOND` is a string expressing a boolean condition; we omit its description here and we refer to the source code for more details. To complete the specification of a deontic interpreted system, it is also necessary to define the following parameters:

- an evaluation function;
- a set of initial states (expressed as a boolean condition);
- a list of subsets of the set of agents to be used for particular group modalities

The syntax for this set of parameters is as follows:

```
EvaluationDef ::= "Evaluation"
                   ID "if" BOOLEANCOND;
                   ...
                  "end Evaluation"
InitstatesDef ::= "InitStates"
                   BOOLEANCOND;
                  "end InitStates"
GroupDef ::= "Groups"
                   ID " = {" IDLIST " }";
                   ...
             "end Groups"
```

Due to space limitations we refer to the files available online for a full example of specification of an interpreted system.

Formulae to be checked are specified using the following syntax

```
formula ::= ID |
            formula "AND" formula |
            "NOT" formula |
            "EX(" formula ")" |
            "EG(" formula ")" |
            "E(" formula "U" formula ")" |
            "K(" ID "," formula ")" |
            "O(" ID "," formula ")" |
            "KH(" ID ", " ID ", " formula ")"
```

Above, `K` denotes knowledge of the agent identified by the string `ID`; `O` is the deontic operator for the agent identified by `ID`. To represent the knowledge of an agent under the assumption of correct behaviour of another agent we use the operator `KH` followed by an identifier for the first agent, followed by another identifier for the second agent, and a formula.

4.2 Implementation of the Algorithm

Figure 2 lists the main components of the software tool that we have implemented. Steps 2 to 6, inside the dashed box, are performed automatically upon invocation of the tool. These steps are coded mainly in C++ and can be summarised as follows:

- In step 2 the input file is parsed using the standard tools Lex and Yacc. In this step various parameters are stored in temporary lists; such parameters include the agents' names, local states, actions, protocols, etc.
- In step 3 the lists obtained in step 2 are traversed to build the OBDD's for the verification algorithm. These OBDD's are created and manipulated using the CUDD library [19]. In this step the number of variables needed to represent local states

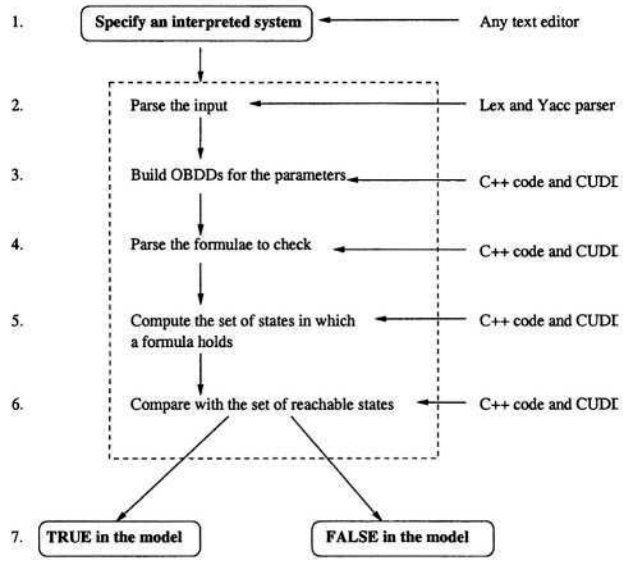


Fig. 2. Software structure.

and actions are computed; following this, all the OBDD's are built by translating the boolean formulae for protocols, evolution functions, evaluation, etc. Also, the set of reachable states is computed using the operator presented in Section 3.2.

- In steps 4 the formulae to be checked are read from a text file, and parsed.
- In step 5 the verification is performed by implementing the algorithm of Section 3.2. At the end step 5, an OBDD representing the set of states in which a formula holds is computed.
- In step 6, the set of reachable states is compared with the OBDD corresponding to each formula. If the two sets are equivalent, the formula holds in the model and the tools produce a positive output. If the two sets are not equivalent, the tool produces a negative output.

5 An Example: The Bit Transmission Problem with Faults

In this section we test our implementation by verifying temporal, epistemic and deontic properties of a communication example: the bit transmission problem [5].

The bit-transmission problem involves two agents, a *sender S,* and a *receiver R,* communicating over a faulty communication channel. The channel may drop messages but will not flip the value of a bit being sent. *S* wants to communicate some information (the value of a bit) to *R.* One protocol for achieving this is as follows. *S* immediately starts sending the bit to *R,* and continues to do so until it receives an acknowledgement from *R. R* does nothing until it receives the bit; from then on it sends acknowledgements of receipt to *S. S* stops sending the bit to *R* when it receives an acknowledgement.

This scenario is extended in [10] to deal with failures. In particular, here we assume that R may not behave as intended perhaps as a consequence of a failure. There are different kind of faults that we may consider for R. Following [10], we discuss two examples; in the first, R may fail to send acknowledgements when it receives a message. In the second, R may send acknowledgements even if it has not received any message.

In Section 5.1, we give an overview of how these scenarios can be encoded in the formalism of deontic interpreted systems. This section is taken from [10]. In Section 5.2 we verify some properties of this scenario with our tool, and we give some quantitative results about its performance.

5.1 Deontic Interpreted Systems for the Bit Transmission Problem

It is possible to represent the scenario described above by means of the formalism of deontic interpreted systems, as presented in [10,8]. To this end, a third agent called E (environment) is introduced, to model the unreliable communication channel. The local states of the environment record the possible combinations of messages that have been sent in a round, either by S or R. Hence, four possible local states L_E are taken for the environment: $L_E = \{(.,.), (sendbit,.), (.,sendack), (sendbit,sendack)\}$, where '.' represents configurations in which no message has been sent by the corresponding agent. The actions Act_E for the environment correspond to the transmission of messages between S and R on the unreliable communication channel. It is assumed that the communication channel can transmit messages in both directions simultaneously, and that a message travelling in one direction can get through while a message travelling in the opposite direction is lost. The set of actions Act_E for the environment is: $Act_E = \{S{-}R, S{\rightarrow}, {\leftarrow}R, -\}$. "$S{-}R$" represents the action in which the channel transmits any message successfully in both directions, "$S{\rightarrow}$" that it transmits successfully from S to R but loses any message from R to S, "${\leftarrow}R$" that it transmits successfully from R to S but loses any message from S to R, and "–" that it loses any messages sent in either direction. We assume the following constant function for the protocol of the environment, P_E:

$$P_E(l_E) = Act_E = \{S{-}R, S{\rightarrow}, {\leftarrow}R, -\}, \quad \text{for all } l_E \in L_E.$$

The evolution function for E is reported in Table 1.

Table 1. Transition conditions for E.

Final state	Transition condition
$(.,.)$	$Act_S = \lambda$ and $Act_R = \lambda$
$(sendbit,.)$	$Act_S = sendbit(0)$ and $Act_R = \lambda$ or $Act_S = sendbit(1)$ and $Act_R = \lambda$
$(.,sendack)$	$Act_R = \lambda$ and $Act_R = sendack$
$(sendbit,sendack)$	$Act_S = sendbit(0)$ and $Act_R = sendack$ or $Act_S = sendbit(1)$ and $Act_R = sendack$

We model sender S by considering four possible local states. They represent the value of the bit S is attempting to transmit, and whether or not S has received an acknowledgement from R: $L_S = \{0, 1, (0, ack), (1, ack)\}$. The set of actions Act_S for S

is: $Act_S = \{sendbit(0),\ sendbit(1), \lambda\}$, where λ denotes a null action. The protocol for S is defined as follows:

$$P_S(0) = sendbit(0), \quad P_S(1) = sendbit(1),$$
$$P_S((0, ack)) = P_S((1, ack)) = \lambda.$$

The transition conditions for S are listed in Table 2.

Table 2. Transition conditions for S.

Final state	Transition condition
$(0, ack)$	$L_S = 0$ and $Act_R = sendack$ and $Act_E = S{-}R$ or
	$L_S = 0$ and $Act_R = sendack$ and $Act_E = {\leftarrow}R$
$(1, ack)$	$L_S = 1$ and $Act_R = sendack$ and $Act_E = S{-}R$ or
	$L_S = 1$ and $Act_R = sendack$ and $Act_E = {\leftarrow}R$

We now consider two possible faulty behaviours for R, that we model below.

Faulty receiver –1. In this case we assume that R may fail to send acknowledgements when it is supposed to. To this end, we introduce the following local states for R: $L'_R = \{0, 1, \epsilon, (0, f), (1, f)\}$. The state "$\epsilon$" is used to denote the fact that R did not receive any message from S; "0" and "1" denote the value of the received bit. The states "(i, f)" ($i = \{0, 1\}$) are *faulty* or *red* states denoting that, at some point in the past, R received a bit but failed to send an acknowledgement. The set of allowed actions for R is: $Act_R = \{sendack, \lambda\}$. The protocol for R is the following:

$$P'_R(\epsilon) = \lambda, P'_R(0) = P'_R(1) = \{sendack, \lambda\}, P'_R((0, f)) = P'_R((1, f)) = \{sendack, \lambda\}.$$

The transition conditions for R are listed in Table 3.

Table 3. Transition conditions for R.

Final state	Transition condition
0	$Act_S = sendbit(0)$ and $L_R = \epsilon$ and $Act_E = S{-}R$ or
	$Act_S = sendbit(0)$ and $L_R = \epsilon$ and $Act_E = S{\rightarrow}$
1	$Act_S = sendbit(1)$ and $L_R = \epsilon$ and $Act_E = S{-}R$ or
	$Act_S = sendbit(1)$ and $L_R = \epsilon$ and $Act_E = S{\rightarrow}$
$(0, f)$	$L_R = 0$ and $Act_R = \epsilon$
$(1, f)$	$L_R = 1$ and $Act_R = \epsilon$

Faulty receiver – 2. In this second case we assume that R may send acknowledgements without having received a bit first. We model this scenario with the following set of local states L''_R for R:

$$L''_R = \{0, 1, \epsilon, (0, f), (1, f), (\epsilon, f)\}.$$

The local states "ϵ", "0", "1", "$(0, f)$" and "$(1, f)$" are as above; "(ϵ, f)" is a further *faulty* state corresponding to the fact that, at some point in the past, R sent an acknowledgement without having received a bit. The actions allowed are the same as in the previous example. The protocol is defined as follows:

$$P''_R(\epsilon) = \lambda,$$
$$P''_R(0) = P''_R(1) = sendack,$$
$$P''_R((0,f)) = P''_R((1,f)) = P''_R((\epsilon,f)) = \{sendack, \lambda\}.$$

The evolution function is reported in Table 4.

Table 4. Transition conditions for R.

Final state	Transition condition
0	$Act_S = sendbit(0)$ and $L_R = \epsilon$ and $Act_E = S{-}R$ or $Act_S = sendbit(0)$ and $L_R = \epsilon$ and $Act_E = S{\rightarrow}$
1	$Act_S = sendbit(1)$ and $L_R = \epsilon$ and $Act_E = S{-}R$ or $Act_S = sendbit(1)$ and $L_R = \epsilon$ and $Act_E = S{\rightarrow}$
(ϵ, f)	$L_R = \epsilon$ and $Act_R = sendack$
$(0, f)$	$Act_S = sendbit(0)$ and $L_R = (\epsilon, f)$ and $Act_E = S{-}R$ or $Act_S = sendbit(0)$ and $L_R = (\epsilon, f)$ and $Act_E = S{\rightarrow}$
$(1, f)$	$Act_S = sendbit(1)$ and $L_R = (\epsilon, f)$ and $Act_E = S{-}R$ or $Act_S = sendbit(1)$ and $L_R = (\epsilon, f)$ and $Act_E = S{\rightarrow}$

For both examples, we introduce the following evaluation function:

$$\mathcal{V}(\mathbf{bit} = \mathbf{0}) = \{g \in G | l_S(g) = 0 \text{ or } l_S(g) = (0, ack)\}$$
$$\mathcal{V}(\mathbf{bit} = \mathbf{1}) = \{g \in G | l_S(g) = 1 \text{ or } l_S(g) = (1, ack)\}$$
$$\mathcal{V}(\mathbf{recbit}) = \{g \in G | l_R(g) = 1 \text{ or } l_R(g) = 0\}$$
$$\mathcal{V}(\mathbf{recack}) = \{g \in G | l_S(g) = (1, ack) \text{ or } l_S(g) = (0, ack)\}$$

The evaluation function $\mathcal{V}$ and the parameters above generate two deontic interpreted systems, one for each faulty behaviour of *R;* we refer to these deontic interpreted systems as DIS_1 and DIS_2.

It is now possible to express formally properties of these scenarios by means of the language of Section 2.1.

$$A(\neg(K_S(K_R(\mathbf{bit} = \mathbf{0}) \vee K_R(\mathbf{bit} = \mathbf{1}))) \; U \; recack) \tag{1}$$
$$A(\neg(\widehat{K}_S^R(K_R(\mathbf{bit} = \mathbf{0}) \vee K_R(\mathbf{bit} = \mathbf{1}))) \; U \; recack) \tag{2}$$

Formula 1 above captures the fact that *S* will not know that *R* knows the value of the bit, until *S* receives an acknowledgement. Formula 2 expresses the same idea but by using knowledge under the assumption of correct behaviour. In the next section we will verify in an automatic way that Formula 1 holds in DIS_1 but not in DIS_2. This means that the faulty behaviour of *R* in DIS_1 does not affect the key property of the system. On the contrary, Formula 2 holds in both DIS_1 and DIS_2; hence, a particular form of knowledge is retained irrespective of the fault.

5.2 Experimental Results

We have encoded the deontic interpreted system and the formulae introduced in the previous section by means of the language defined in Section 4.1 (a copy of the code is included in the downlodable files). The two formulae were correctly verified by the tool for DIS_1, while Formula 1 failed in DIS_2 as expected.

To evaluate the performance of our tool, we first analyse the space requirements. Following the standard conventions, we define the size of a deontic interpreted system as $|DIS| = |S| + |R|$, where $|S|$ is the size of the state space and $|R|$ is the size of the relations. In our case, we define $|S|$ as the number all the possible combinations of local states and actions. In the example above, there are 4 local states and 3 actions for *S, S* (or 6) local states and 2 actions for *R,* and 4 local states and 4 actions for *E.* In total we have $|S| \approx 2 \cdot 10^3$. To define $|R|$ we must take into account that, in addition to the temporal relation, there are also the epistemic and deontic relations. Hence, we define $|R|$ as the sum of the sizes of temporal, epistemic, and deontic relations. We approximate $|R|$ as $|S|^2$, hence $|M| = |S| + |R| \approx |S|^2 \approx 4 \cdot 10^6$.

To quantify the memory requirements we consider the maximum number of nodes allocated for the OBDD's. Notice that this figure over-estimates the number of nodes required to encode the state space and the relations. Further, we report the total memory used by the tool (in MBytes). The formulae of both examples required a similar amount of memory and nodes. The average experimental results are reported in Table 5.

Table 5. Memory requirements.

$\|M\|$	OBDD's nodes	Memory (MBytes)
$\approx 4 \cdot 10^6$	$\approx 10^3$	≈ 4.5

In addition to space requirements, we carried out some test on time requirements. The running time is the sum of the time required for building all the OBDD's for the parameters and the actual running time for the verification. We ran the tool on a 1.2 GHz AMD Athlon with 256 MBytes of RAM, running Debian Linux with kernel 2.4.20. The average results are listed in Table 6.

Table 6. Running time (for one formula).

Model construction	Verification	Total
0.045sec	<0.01sec	0.05sec

We see these as very encouraging results. We have been able to check formulae with nested temporal, epistemic and deontic modalities in less than 0.1 seconds on a standard PC, for a non-trivial model. Also, the number of OBDD's nodes is orders of magnitude smaller than the size of the model. Therefore, we believe that our tool could perform reasonably well even in much bigger scenarios.

6 Conclusion

In this paper we have extended a major verification technique for reactive systems — symbolic model checking via OBDD's — to verify temporal, epistemic, and deontic properties of multiagent systems. We provided an algorithm and its implementation, and we tested our implementation by means of an example: the bit transmission problem with faults. The results obtained are very encouraging, and we estimate that our tool could be used in bigger examples. For the same reason, we see as feasible an extension of the tool to include other modal operators.

We gratefully acknowledge Dr Bozena Wozna and the anonymous referees for their comments and suggestions.

References

1. M. Benerecetti, F. Giunchiglia, and L. Serafini. Model checking multiagent systems. *Journal of Logic and Computation,* 8(3):401–423, June 1998.
2. R. H. Bordini, M. Fisher, C. Pardavila, and M. Wooldridge. Model checking AgentSpeak. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'03),* July 2003.
3. R. E. Bryant. Graph-based algorithms for boolean function manipulation. *IEEE Transaction on Computers,* pages 677–691, August 1986.
4. E. M. Clarke, O, Grumberg, and D. A. Peled. *Model Checking*. The MIT Press, Cambridge, Massachusetts, 1999.
5. R. Fagin, J. Y. Halpern, Y. Moses, and M. Y. Vardi. *Reasoning about Knowledge*. The MIT Press, Cambridge, Massachusetts, 1995.
6. G. J. Holzmann. The model checker spin. *IEEE transaction on software engineering,* 23(5), May 1997.
7. M. R. A. Huth and M. D. Ryan. *Logic in Computer Science: Modelling and Reasoning about Systems*. Cambridge University Press, Cambridge, England, 2000.
8. A. Lomuscio, F. Raimondi, and M. Sergot. Towards model checking interpreted systems. In *Proceedings of MoChArt,* Lyon, France, August 2002.
9. A. Lomuscio and M. Sergot. On multi-agent systems specification via deontic logic. In J.-J Meyer, editor, *Proceedings of ATAL 2001,* volume 2333. Springer Verlag, July 2001.
10. A. Lomuscio and M. Sergot. Violation, error recovery, and enforcement in the bit transmission problem. In *Proceedings of DEON'02,* London, May 2002.
11. K. L. McMillan. *Symbolic model checking: An approach to the state explosion problem.* Kluwer Academic Publishers, 1993.
12. R. van der Meyden and N. V. Shilov. Model checking knowledge and time in systems with perfect recall. *FSTTCS: Foundations of Software Technology and Theoretical Computer Science,* 19, 1999.
13. R. van der Meyden and K. Su. Symbolic model checking the knowledge of the dining cryptographers. Submitted, 2002.
14. J.-J. Meyer and R. Wieringa, editors. *Deontic Logic in Computer Science,* Chichester, 1993.
15. W. Penczek and A. Lomuscio. Verifying epistemic properties of multi-agent systems via model checking. *Fundamenta Informaticae,* 55(2): 167–185, 2003.
16. F. Raimondi and A. Lomuscio. A tool for verification of deontic interpreted systems. http://www.dcs.kcl.ac.uk/pg/franco/mcdis-0.1.tar.gz.
17. F. Raimondi and A. Lomuscio. Verification of multiagent systems via ordered binary decision diagrams: an algorithm and its implementation. Submitted, 2004.
18. A. S. Rao. AgentSpeak(L): BDI agents speak out in a logical computable language. *Lecture Notes in Computer Science,* 1038:42–52, 1996.
19. F. Somenzi. CU Decision Diagram Package - Release 2.3.1. http://vlsi.colorado.edu/~fabio/CUDD/cuddIntro.html.
20. M. Wooldridge, M. Fisher, M.P. Huget, and S. Parsons. Model checking multi-agent systems with MABLE. In M. Gini, T. Ishida, C. Castelfranchi, and W. Lewis Johnson, editors, *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'02),* pages 952–959. ACM Press, July 2002.

Specifying Multiagent Organizations

Leendert van der Torre[1,2,*], Joris Hulstijn[3], Mehdi Dastani[3], and Jan Broersen[3]

[1] CWI Amsterdam
[2] Delft University of Technology
[3] University of Utrecht

Abstract. In this paper we investigate the specification and verification of information systems with an organizational structure. Such systems are modelled as a normative multiagent system. To this end we use $\mathbf{KBDIO}_{\mathbf{CTL}}$, an extension of $\mathbf{BDI}_{\mathbf{CTL}}$ in which obligations and permissions are represented by directed modal operators. We illustrate how the logic can be used by introducing and discussing various properties of normative systems and individual agents which can be represented in the logic. In particular we discuss the enforcement of norms.

1 Introduction

Normative computer systems are computer systems which involve obligations, prohibitions and permissions [1]. The traditional applications can be found in computer security, for example to regulate access to file systems or libraries. Other applications have been studied in electronic commerce, in legal expert systems and in databases. See [2] for a survey on these applications. More recently, normative systems have been used to regulate virtual communities in the context of the (semantic) web. To support the development of such systems, several agent architectures have been proposed that incorporate obligations, prohibitions and permissions.

In this paper we investigate the formalization of regulations such as the widely discussed library regulations, parking regulations, copier regulations, cottage regulations, et cetera. Such examples are characterized by sometimes complicated normative systems, as well as organizational structures. Moreover, in contrast to earlier investigations, we not only consider the case in which humans interact with a normative computer system, but we also consider cases in which computers interact with other computer systems, that is, we consider multiagent systems. In particular, we consider the formalization of properties involving normative multiagent systems in an extension of Schild's $\mathbf{BDI}_{\mathbf{CTL}}$ [3–5], which is itself a variant of Rao and Georgeff's $\mathbf{BDI}_{\mathbf{CTL}}$ [6]. Such an extension consists of an extension of the logic and an extension of the properties expressed in the logic. Obligations are motivational attitudes, just like desires, but they are also related to organizational issues.

First, obligation is formalized as a directed modality [7–11]. Thus, whereas we may say that agent a desire to prepare a report, we say that the agent a is obliged to prepare a report *towards another agent* b. Moreover, as explained in more detail in Section 5, whereas desires and intentions remain in force as motivational attitude until the agent

* Supported by the ArchiMate research project.

A. Lomuscio and D. Nute (Eds.): DEON 2004, LNAI 3065, pp. 243–257, 2004.

believes they have been achieved or are no longer achievable, obligations remain in force until the agent *knows* they have been fulfilled or they are no longer achievable. We introduce an extension of $\mathbf{BDI}_{\mathrm{CTL}}$ called $\mathbf{KBDIO}_{\mathrm{CTL}}$, that makes the distinction between desires and obligations explicit, as well as the distinction between beliefs and knowledge.

Second, we provide organizational concepts such as roles, role relations, and groups in order to specify inter-agent relations that hold in organizations. The organizational concepts are interpreted as follows.

A *role* is a set of related constraints that should be satisfied when an agent enacts the role. For example, the role of project manager puts constraints on the expertise, capabilities, responsibilities, goals, obligations and permissions of the agent that enacts the role. Note that various different definitions of the concept of a role have been proposed. Our definition follows [12–15]. The definition of a role is always related to some organizational activity, which determines its scope. For example, the role of chairman only makes sense during a meeting. Agents may only enact a role provided they are *qualified,* i.e., meet the basic requirements for the role.

Role relations, also known as *dependencies* or *channels,* are constraints put on a relation between roles. Examples of a role relations are *supervisor_of* and the producer-consumer relation. Role relations coordinate the behavior of different agents, similar to the way channels coordinate components in software architectures [16]. One role can be enacted by many agents. Consider for example several postmen in a district. Moreover, one agent can enact many roles. Consider for example a lecturer who is also a conference reviewer.

A *group* is a set of roles that share a group characteristic. For example, roles involved in selling goods in an organization form a group often called the selling department.

The motivation of our work is to develop a specification and verification language for normative multiagent systems with organizational structure. We therefore focus on properties of regimentation, which formalize whether norms can be violated, on deadlines, and on definitions of organizational structure. Due to the fact that we not only consider humans, but also artificial agents interacting with normative computer systems, new issues and properties arise.

For example, for agent systems it is common practice to design agents that cannot violate norms, or agents that are benevolent and will always first try to fulfill the obligations or goals of other agents, before trying to achieve their own desires. Therefore it is useful to have a specification language that can express such properties too. As these properties cannot be programmed in human agents, such properties have not made sense previously, and consequently we believe that they have not been addressed in the literature. We acknowledge the criticism on such properties, but such criticism is beyond the scope of this paper.

The layout of the paper is as follows. In section 2 we describe an example specification domain. In Section 3 we extend Schild's logic with obligations, prohibitions, permissions and organizational concepts. In the remainder, we discuss properties which can be expressed in the logic, and which can be used to specify the running example.

2 Multiagent Organizations: The Running Example

In this section we exemplify the type of specification properties we are interested in. Shorter specification examples in subsequent sections will also apply to the domain described here. Our example domain is concerned with the different ways an organizational norm can be implemented in a multiagent system. A multiagent system developer has a choice of options to operationalize a norm. In each case, a number of assumptions about the mental attitudes and reasoning capabilities of the subjects of the norm, the individual agents, are necessary. A system developer can leave it up to the individual agents to respect the norm. In that case, he assumes that agents are benevolent or norm-abiding, and proving that the system comforms to the norm presupposes that this assumptions is formalized. By contrast, the system developer can hardwire the norm into the environment, making it physically impossible for agents to violate it. In that case, no additional assumptions on agents are needed. We believe that a rich logic like $\mathbf{KBDIO}_{\mathrm{CTL}}$ is suitable to express this kind of notions and assumptions.

The example is derived from an observation concerning different ticket policies of public transport networks [17]. Suppose ticket policies are specified as a multi-agent system. Using these specifications, one can formulate the consequences of such policies as logical properties, and verify them with respect to the system specifications.

Compare the Paris metro with a French train. On the entrance of a platform of the Paris metro, the authorities have placed a high barrier with gates that will only open when a valid ticket is inserted. Without a valid ticket, it is physically impossible to pass the barrier and use the metro. By contrast, it is possible to board a French train without a ticket. The authorities rely on personal benevolence, on social pressure, and on a sanctioning system of ticket inspection and fines, to persuade passengers to buy a valid ticket. Looking at other travel systems we find yet other solutions to the same problem: under which assumptions can we conclude that all passengers will pay for the ride? We can phrase the norm as follows:

When travelling by public transport, one should have paid for the trip.

This norm is an instance of a much more general pattern occurring in situations in which humans interact with normative computer systems, and also in multiagent systems such as virtual communities or web services. For example, an agent has to access a resource offered by another agent. To regulate such access, there is an organizational structure, that may contain roles, but also more complicated normative constructs such as authorization and delegation mechanisms. In this paper we restrict ourselves to the norm above.

We consider the following ways to implement this norm in a multi-agent system. Each possibility relies on some specific assumptions about the environment, or about the agents inhabiting the system.

1. **Implementing a norm in the environment** The norm is enforced with gates on the platform. No assumptions on the mental attitudes of agents are needed, only assumptions about their physical ability.

2. **Implementing a norm by designing benevolent or norm abiding agents.** All agents can be designed to be sincere. If they tell you they have paid, you can trust them. This removes the need for tickets as evidence of payment. Moreover, agents can be designed to be either benevolent, or norm abiding. If a benevolent agent understands why the norm is a good norm, for example to maintain a good quality of public transport, it will internalize the norm and make it a personal goal. A norm abiding agent will simply obey the norm, no matter how this relates to its own goals.
3. **Implementing a norm by relying on rationality.** Here tickets are introduced as evidence of payment, and hence as a right to travel. No sincerity assumption is needed. If an agent is caught travelling without a valid ticket, it is subject to a sanction: to pay a fine. This assumes that agents are rational decision makers, in the economic sense of maximizing expected utility. An agent will display the behavior corresponding to the norm, if a ticket is cheaper than the fine multiplied by the perceived chance of being caught. Authorities can affect this way of decision making by increasing the fine, or by making the agents believe that the chance of being caught has increased.
4. **Implementing a norm by relying on social control.** Here again tickets are used as evidence of payment. Being caught without a ticket leads to social embarrassment and a loss of reputation. Like in item 3 above, this solution assumes that agents are subject to sanctions, and moreover, that embarrassment counts as an effective sanction. Embarrassment typically only comes up if all other passengers can observe that the passenger does not pay.
5. **Implementing a norm by relying on a combination of mechanisms.** In most actual situations a mixture of these types of norm enforcement is in place. For example, a fine system is used to remind agents of the noble purpose behind the norm. Social embarrassment comes on top of the fine. That means that in practice, fines do not have to be as high as would be required for socially unaffected citizens.

Note that the above categories not only occur in human society, but also in multiagent systems. Implementing a norm in the environment is also the typical case used in web services: if an agent has not paid for the service, it simply cannot access it. Implementing a norm by norm abiding agents is not possible in human organizations, but frequently occurs in multiagent organizations. Human and multiagent systems often depend on rationality, for example in the context of electronic commerce. Finally, many human organizations rely on social control, and there are examples of multiagent systems containing social agents [18].

Obligations are motivational attitudes, just like desires, but they also have organizational aspects. First, obligations are always directed. Obligations can be directed towards abstract entities like 'the company' or 'the system', towards other agents, or towards the agents themselves. Second, the organizational structure is represented by the sets of roles, groups, and their interactions, as indicated above. Group membership and the assignment of agents to roles changes over time, as role relations are established or disconnected. The 'social fact' of an agent enacting a role is distinguished from the satisfaction of the requirements that go with the role. For example, although a passenger does not have a ticket while he is in the metro, even if he does not satisfy the requirements set by the role, he remains a passenger.

3 $\mathbf{KBDIO_{CTL}}$, a Logic for Specifying Multiagent Organizations

We use a version of $\mathbf{BDI_{CTL}}$ presented by Schild [3], which we extend with operators for knowledge and directed obligation. The syntax of $\mathbf{KBDIO_{CTL}}$ involves a modal operator K_a for knowledge of agent a, an operator B_a for belief, D_a for desire, I_a for intention, and $O_{a,b}$ for an obligation of agent a towards agent b [7,8,10,11]. Knowledge, belief, desire and intention are internal to the agent and thus not directed. The temporal operators of the language are imported from CTL. To specify organizational structure, special propositions '$g(a)$', '$r(a)$', and '$a\ ch\ b$' are introduced for 'agent a is a member of group g', 'agent a enacts role r', and 'agent a and b stand in role relation *ch*', respectively. Higher order relations can be defined analogously. We assume that roles, groups and role relations are all primitive, though in certain systems they have been defined in terms of each other. For example, a group can be defined as the role of being a member of the group. Also, a group can be defined as a role relation between all members of the group, or between the group members and the group leader.

Definition 1 (Syntax $\mathbf{KBDIO_{CTL}}$). *Given a finite set A of agent names, a finite set G of group names, a finite set R of role names, a finite set C of role relations, and a countable set P of primitive proposition names, which includes '$g(a)$', '$r(a)$', and '$a\ ch\ b$' for all $a, b \in A$, $g \in G$, $r \in R$, and $ch \in C$, the admissible formulae of $\mathbf{KBDIO_{CTL}}$ are recursively defined by:*

S1 *Each primitive proposition in P is a state formula.*
S2 *If α and β are state formulae, then so are $\alpha \wedge \beta$ and $\neg\alpha$.*
S3 *If α is a path formula, $E\alpha$ and $A\alpha$ are state formulae.*
S4 *If α is a state formula and $a, b \in A$, then $K_a(\alpha), B_a(\alpha), D_a(\alpha), I_a(\alpha), O_{a,b}(\alpha)$ are state formulae as well.*
P *If α and β are state formulae, then $X\alpha$ and $\alpha U \beta$ are path formulae.*

We assume the following abbreviations:

disjunction	$\alpha \vee \beta \equiv_{def} \neg(\neg\alpha \wedge \neg\beta)$	implication	$\alpha \rightarrow \beta \equiv_{def} \neg\alpha \vee \beta$
future	$F(\alpha) \equiv_{def} \top U \alpha$	globally	$G(\alpha) \equiv_{def} \neg F(\neg\alpha)$
permission	$P_{a,b}(\alpha) \equiv_{def} \neg O_{a,b}(\neg\alpha)$	prohibition	$F_{a,b}(\alpha) \equiv_{def} \neg P_{a,b}(\alpha)$
undirected	$O_a(\alpha) \equiv_{def} O_{a,a}(\alpha)$.		

The semantics of $\mathbf{KBDIO_{CTL}}$ involves two dimensions. The truth of a formula is evaluated relative to a world w and a temporal state s. A pair $\langle w, s\rangle$ is called a situation. The relation between situations is traditionally called an accessibility relation (for beliefs) or a successor relation (for time).

Definition 2 (Situation structure $\mathbf{KBDIO_{CTL}}$). *Assume a finite set A of agent names. A structure $M = \langle \Delta, \mathcal{R}, \mathcal{K}, \mathcal{B}, \mathcal{D}, \mathcal{I}, \mathcal{O}, L\rangle$ forms a situation structure if Δ is a set of situations, $\mathcal{R} \subseteq \Delta \times \Delta$ is a binary relation such that $w = w'$ whenever $\langle w, s\rangle \mathcal{R} \langle w', s'\rangle$, $Z(a) \subseteq \Delta \times \Delta$ for the functions $Z \in \{\mathcal{K}, \mathcal{B}, \mathcal{D}, \mathcal{I}\}$ and $a \in A$, and $\mathcal{O}(a, b) \subseteq \Delta \times \Delta$ with $a, b \in A$ are binary relations such that $s = s'$ whenever $\langle w, s\rangle Z(a) \langle w', s'\rangle$ or $\langle w, s\rangle \mathcal{O}(a, b) \langle w', s'\rangle$, and L an interpretation function that assigns a particular set of situations to each primitive proposition. $L(p)$ contains all those situations in which p holds.*

A speciality of CTL is that some formulae – called path formulae– are not interpreted relative to a particular situation. What is relevant here are full paths. The reference to M is omitted whenever it is understood. Note that $\alpha U \beta$ is true if α is true until the last moment before the first one in which β is true (alternative definitions are used in the literature too).

Definition 3 (Semantics KBDIO$_{CTL}$). *Given a set A of agent names. A full path in situation structure M is a sequence $\chi = \delta_0, \delta_1, \delta_2, \ldots$ such that for every $i \geq 0$, δ_i is an element of Δ and $\delta_i \mathcal{R} \delta_{i+1}$, and if χ is finite with δ_n its final situation, then there is no situation δ_{n+1} in Δ such that $\delta_n \mathcal{R} \delta_{n+1}$. We say that a full path starts at δ iff $\delta_0 = \delta$. If $\chi = \delta_0, \delta_1, \delta_2, \ldots$ is a full path in M, then we denote δ_i by χ^i $(i \geq 0)$.*

Let M be a situation structure, δ a situation, χ a full path and $a, b \in A$ two agents. The semantic relation $\models$ for KBDIO$_{CTL}$ is then defined as follows:

S1 $\delta \models p$ *iff* $\delta \in L(p)$ *and* p *is a primitive proposition*
S2 $\delta \models \alpha \wedge \beta$ *iff* $\delta \models \alpha$ *and* $\delta \models \beta$
$\delta \models \neg\alpha$ *iff* $\delta \models \alpha$ *does not hold*
S3 $\delta \models E\alpha$ *iff for some full path* χ *in M starting at* δ, *we have* $\chi \models \alpha$
$\delta \models A\alpha$ *iff for each full path* χ *in M starting at* δ, *we have* $\chi \models \alpha$
S4 $\delta \models K_a(\alpha)$ *iff for every* $\delta' \in \Delta$ *such that* $\delta \mathcal{K}(a) \delta', \delta' \models \alpha$
$\delta \models B_a(\alpha)$ *iff for every* $\delta' \in \Delta$ *such that* $\delta \mathcal{B}(a) \delta', \delta' \models \alpha$
$\delta \models D_a(\alpha)$ *iff for every* $\delta' \in \Delta$ *such that* $\delta \mathcal{D}(a) \delta', \delta' \models \alpha$
$\delta \models I_a(\alpha)$ *iff for every* $\delta' \in \Delta$ *such that* $\delta \mathcal{I}(a) \delta', \delta' \models \alpha$
$\delta \models O_{a,b}(\alpha)$ *iff for every* $\delta' \in \Delta$ *such that* $\delta \mathcal{O}(a, b) \delta', \delta' \models \alpha$
P $\chi \models X\alpha$ *iff* $\chi^1 \models \alpha$
$\chi \models \alpha U \beta$ *iff there is an* $i \geq 0$ *such that* $\chi^i \models \beta$ *and for all* $j (0 \leq j < i), \chi^j \models \alpha$

Like Rao and Georgeff, we use standard interpretations of these operators. $O_{a,b}$ is interpreted as a standard deontic operator KD [19], B as KD45, K as S5, and D, I as KD modal logic operators. The properties discussed in this paper characterize the relation between mental attitudes of a single agent. Properties can always be expressed at two levels. First, we can express that *all* obligations of an agent towards an agent satisfy a property. In that case, the obligations are characterized by this property. Second, properties may hold for one particular obligation only. In that case we may say that this particular obligation satisfies the property, but it does not characterize the agent's obligations in general. In this paper, we follow the convention that properties expressed using α are axioms, and thus α can be substituted by any prepositional formula.

However, it is important to notice that all properties expressed relative to a group or role, such as $r(a) \rightarrow K_a\alpha$, can only be expressed as formulas, not as an axiom. The reason is, roughly, a condition like $g(a)$ or $r(a)$ should not be substituted by another proposition. For example, if $r(a) \rightarrow K_a\alpha$ is an axiom, then so is $q \rightarrow K_a\alpha$. An alternative way to formalize organizational structure in Rao and Georgeff's logic is to index modal operators by groups and roles, and thus write the above property as an axiom $K_{g(a)}\alpha$. The reason we made this choice of formalizing organizational structure in propositions is that the expressive power of the alternative representation is limited. The loss of relativized axioms is considered to be less severe, as the status of interaction axioms in this logic is problematic anyway, as discussed in Section 9.

4 Specification of Organizational Structure

Organizational structure is typically specified in terms of roles and role relations. When agent $x \in A$ plays the role $p \in R$ of passenger, and has not paid before travel started, then he or she is obliged to pay a fine to the public transport company s. This can be specified by the following set of formulas, for all $x, y \in A$. Note that sanctions are modelled as obligations too, and that the violation condition is expressed using the until operator of CTL.

$$(p(x) \wedge (\neg \text{paid}_x U \text{travel}_x)) \rightarrow O_{x,s}\text{fine}_x$$

If the public transport company $s \in A$ has delegated the power to collect fines to the ticket controller, $c \in R$, we get $(p(x) \wedge c(y) \wedge (\neg \text{paid}_x U \text{travel}_x)) \rightarrow O_{x,y}\text{fine}_x$. Such so-called delegation relations can become complex and are not further discussed in this paper. See for example, [20, 21].

In general, obligations are created by interaction. For example, in an electronic market where agents are buying and selling goods, a confirmation to buy creates a obligation to pay for the buyer and an obligation of shipping the goods for the seller. Obligations may also be created by the way a social system is designed. A social system typically contains stable relationships between roles, which affect the obligations of the agents in those roles. In particular, obligations can be based on the known or believed mental attitudes of agents standing in a role relation. For example, the role relation ***adopts*** $\in C$ between agent $a \in A$ and agent $b \in A$ can be characterized by the following axiom, which says that agent a adopts all obligations of agent b towards some other agent $c \in A$. The following formula schema can be instantiated for all agents $a, b, c \in A$, and proposition letter q. Obviously we have an analogous property when we replace knowledge *(K)* by belief *(B)*.

$$(a \textit{ adopts } b \wedge K_a O_{b,c} q) \rightarrow O_{a,c} q$$

We can further specify obligation adoption with additional formulas. For example, the formula schema $(r(a) \wedge r(b)) \rightarrow a \textit{ adopts } b$ specifies that agent a adopts the obligations of agent b when they play the same role r in the organization. In a similar way, $K_a D_b \alpha \rightarrow O_{a,b}\alpha$ characterizes that agent a adopts the known desires of agent b as its obligations. Take a client-server system for example. When the server s believes that its client c desires a piece of information, then we can specify that the server s is obliged to see to it that client c gets this information. The following axiom schema characterizes the ***slave_of*** $\in C$ or "your wish is my command" role relation, which says that the desires or intentions of master $m \in A$ become the obligations of slave $s \in A$.

$$(s \textit{ slave\_of } m \wedge K_s I_m q) \rightarrow O_{s,m} q$$

For example, reconsider the running example and assume that the passenger has not paid. Now we need a detection mechanism to make sure that the sanction is applied. A ticket controller has the institutional power to make a passenger without a ticket pay a fine. However, the controller does not have the power to make any passenger pay a fine. There must be a pretext. This can be specified as a restricted instance of the master-slave principle listed above.

$$(p(x) \wedge c(y) \wedge K_x K_y \neg \text{have\_ticket}_x \wedge K_x I_y \text{fine}_x) \rightarrow O_{x,y}\text{fine}_x$$

For violation detection, we first still have to specify that not having a ticket counts as evidence of not having paid. How to formalize such constitutive norms is an open problem in deontic logic, see for example [22–24]. A very simple specification in our specification language is $(g(x) \wedge \neg K_x \text{have\_ticket}_x) \rightarrow K_x(\neg \text{paid}_x U \text{travel}_x)$, for all member agents x of some suitable group $g \in G$.

We can further extend the logic with new group related concepts to specify requirements on groups of agents. For example, the first axiom schemata for $x, y \in A$ characterizes the property that all members of group g must know each other and they must be able to have the role relation *ch* that they can communicate with each other. This is called acquaintance among members of a group. Groups and roles can also be combined. For example, for any organization it is important that agents recognize the roles that other agents are enacting. In human society, uniforms, location (behind a desk) or badges are used to this purpose. A group $g \in G$ in which a role $r \in R$ of an agent $a \in A$ is known to all agents is called *transparent.*

$$(g(x) \wedge g(y)) \rightarrow (K_x g(y) \wedge (x \; ch \; y)) \qquad (g(a) \wedge g(b) \wedge r(a) \rightarrow K_b r(a))$$

Related to transparency of roles is the property of delegation transparency, which states that agents must know of other agents on behalf of whom they are acting. So if some agent a delegates a job to b, a's role as a principal must be known. Verifying delegation chains is particularly important for legal applications, because the principle remains legally accountable.

A promising issue in the specification of multiagent organizations is the definition of a set of patterns for groups, roles and role relations. Patterns have proven to be very useful in several areas of software engineering. For example, assume that we wish to define a pattern for the role relation ***leader*** $\in C$ as the property that the agent fulfilling the role is able to communicate with the group members and vice versa. Also, a group leader must be able to delegate tasks to the group members and persuade them to have certain beliefs. In addition, the obligations of members of a group are the obligations of the group leader (a failure to satisfy an obligation by a group member is a failure to satisfy the obligation of the group leader), and the members should be committed to the task delegated to them. The following schemata characterize such a group leader. Let $a, x \in A$, $g \in G$, *leader* and *com* $\in C$ be role relations that represent 'leader of and 'able to communicate', respectively.

a leader $x \rightarrow$

$K_a(a \; com \; x) \wedge K_x(x \; com \; a) \wedge$	(ability to communicate)
$D_a AFI_x q \rightarrow AFI_x q \wedge$	(task delegation)
$I_a B_x q \rightarrow AXB_x q \wedge$	(persuading members)
$O_{x,a} q \rightarrow O_{a,a} q \wedge$	(obligation inheritance)
$I_x AFq \rightarrow A(I_x AFq \cup (B_a q \vee \neg B_a EFq)))$	(committed to delegated tasks)

An interesting question for further research is how standard patterns used in business modelling or software engineering can be formalized in our specification language. In this paper we do not consider this question, but we return to our running example.

5 Formalizing the Norm of the Running Example

The public transport norm can be phrased as follows: any agent in the role of passenger travelling by public transport, should have paid for the trip. We choose to describe this norm in terms of a so called 'deadline obligation': "if $x \in A$ is playing the role of passenger $p \in R$, then x is obliged towards society s to see to it that there is no history in which x does not pay until x travels".

$$p(x) \rightarrow O_{x,s}\neg E((\neg \text{paid}_x \wedge \neg \text{travel}_x) U \text{travel}_x)$$

The concept of deadline obligation is rather complex, as several alternative definitions can be given [25]. The concept depends on the particular interpretation of the until operator. The formula states that the obligation applies to any agent in the role of passenger. However, this formula does not describe behavior. The following formula, without the obligation, does:

$$p(x) \rightarrow \neg E((\neg \text{paid}_x \wedge \neg \text{travel}_x) U \text{travel}_x)$$

Our definition of deadline obligations is inspired by Rao and Gerogeff's formalizations of commitment strategies. The main axioms discussed in temporal extensions of BDI logic are realism properties and commitment strategies, in particular in $\mathbf{BDI}_{\mathrm{LTL}}$ by Cohen and Levesque [26] and in $\mathbf{BDI}_{\mathrm{CTL}}$ by Rao and Georgeff [27,6].

Realism puts a constraint on desires, with respect to what the agent believes about the state of the world. Some examples of realism properties are $B_a\alpha \rightarrow D_a\alpha$, for 'overcommitted realism' as defined in [26], $D_a\alpha \rightarrow \neg B_a\neg\alpha$ for 'weak realism' as defined in [27,6], and $D_aEF\alpha \rightarrow B_aEF\alpha$ for 'strong realism'.

Commitment strategies are constraints on the process of intention reconsideration: under what circumstances is it allowed to drop an intention? Examples of commitment strategies are $I_aAF\alpha \rightarrow A(I_aAF\alpha\ UB_a\alpha)$, for 'blind commitment', and the more interesting $I_aAF\alpha \rightarrow A(I_aAF\alpha\ U(B_a\alpha \vee \neg B_aEF\alpha))$, called 'single minded commitment'. Whereas realism properties are static, commitment strategies are dynamic in the sense that they specify the temporal evolution of intentions. In the remainder of this section we define static and dynamic properties that involve obligations.

Rao and Georgeff's commitment strategies are examples of interactions of motivational attitudes and tune. Such interactions also occur for desires and obligations. Cohen and Levesque [26] distinguish 'achievement goals' and 'maintenance goals'. Their definition in $\mathbf{BDI}_{\mathrm{LTL}}$ can be adapted to $\mathbf{KBDIO}_{\mathrm{CTL}}$ as the definition of $O^A_{a,b}$ below on the left. Cohen and Levesque do not give a definition for maintenance goals, but they characterize the difference as follows: "Achievement goals are those the agent believes to be false; maintenance goals are those the agent already believes to be true". This suggests that we can give a formula $O^M_{a,b}\alpha$ to express a maintenance obligation: $O^M_{a,b}\alpha \equiv_{def} B_a\alpha \wedge O_{a,b}AF\alpha$. Alternatively, we could define a maintenance obligation by the restriction that the goal or obligation should be maintained all the time.

$$O^A_{a,b}\alpha \equiv_{def} B_a\neg\alpha \wedge O_{a,b}AF\alpha \qquad O^M_{a,b}\alpha \equiv_{def} B_a\alpha \wedge O_{a,b}AG\alpha$$

Another issue are the conditions that may discharge an obligation. Obligations typically persist until a deadline, e.g., deliver the goods before noon, or they persist forever,

e.g., don't kill. We denote a deadline obligation by $O_{a,b}(\alpha, d)$, where achievement of the proposition d is the deadline for the obligation to achieve α. A deadline obligation $O_{a,b}(\alpha, d)$ persists until it is fulfilled or becomes obsolete because the deadline is reached.

$$O_{a,b}(\alpha, d) \equiv_{def} A((O_{a,b}\alpha)U(\alpha \vee d))$$

A deadline obligation $O_{a,b}(\alpha, \alpha)$, for which the only deadline is the achievement of the obligation itself, is called a 'dischargeable obligation'. The definition simplifies to $O_{a,b}(\alpha, \alpha) \equiv_{def} A((O_{a,b}\alpha)U\alpha)$. Alternatively, we may characterize the property that obligations from agent a to agent b are dischargeable by the axiom $O_{a,b}\alpha \leftrightarrow A((O_{a,b}\alpha)U\alpha)$. Analogously we can also define dischargeable desires. For example, an agent may desire a receipt until it gets one. However, a drawback of the axiom is that it is expressed in terms of facts, which are not accessible to agents. We therefore replace the occurrence of α without a preceding modal operator by $K_a\alpha$. Moreover, again we believe that dischargeable obligations and dischargeable desires obey different discharging conditions. An obligation can only be discharged by the *knowledge* that the obliged condition is fulfilled. A desire can already be discharged by the *belief* that this is the case. Consequently, the property that obligations from agent a towards agent b are dischargeable, and analogously the property that desires from agent a are dischargeable, are characterized by the following two axioms, respectively.

$$O_{a,b}\alpha \leftrightarrow A((O_{a,b}\alpha)UK_a\alpha) \qquad D_a\alpha \leftrightarrow A((D_a\alpha)UB_a\alpha)$$

We can characterize that $O_{a,b}\alpha$ persists forever, i.e., that it is a 'non-dischargeable obligation', by $O_{a,b}\alpha \leftrightarrow AGO_{a,b}\alpha$. We can also combine the definitions, such that agents for instance have non-dischargeable achievement obligations, or dischargeable maintenance obligations.

As we now have specified the norm, we finally specify the four ways to realize that the norm is fulfilled. First we regiment the norm into the environment, such that agents cannot violate the norm. Then we define agents which are designed such that they cannot violate norms. Finally we discuss formalizations that rely on rationality or social control. In the formalization, we distinguish between assumptions about societies, ticket policies, individual agents and the environment. These assumptions are either formalized as formulas or as axioms. The difference is roughly that axioms are true in any world of the model, and for axioms we can substitute the propositions by other propositions. The norm itself – the first formula above – can be part of those assumptions. We want to verify whether the second property follows from this. Γ_{ins} is a set of formulas representing assumed properties of the institution, in this case the public transport network, $\Gamma_{r_1}, ..., \Gamma_{r_n}$ are sets of formulas that represent the assumed properties for the various roles $r_1, ..., r_n$ in the institution, like passenger or ticket collector, Γ_{env} is a set of formulas representing the assumed properties of the behavior of the environment, and Δ represents the property to be shown. As usual we use the weakest version of modal entailment, i.e., $\varphi \models \psi$ holds if and only if it is the case that when φ is satisfied in some state of a model, than also ψ is satisfied.

6 Implementing a Norm in the Environment

An important question when developing a normative system is whether the norms can be violated or not, i.e., whether the norms are soft or hard constraints. In the latter case, the norms are said to be regimented. Regimented norms correspond to preventative control systems in computer security [17]. For example, in the metro example it is not possible to travel without a ticket, because there is a preventative control system, whereas it is possible to travel without a ticket on the French trains, because there is a detective control system. Norm regimentation for agent a is characterized by the following axiom.

$$O_{a,b}\alpha \rightarrow \alpha$$

The following example illustrates the specification of regimentation in the running example. It also illustrates that regimentation can be specified at different levels of abstraction. At the detailed level, it is specified precisely how the norm is implemented in the environment. At a more abstract level, the norm is given as an axiom, and it is specified that the norm is regimented - but not *how* it is regimented.

Example 1 (norm enforcement by imposing a restrictive environment). The set of agents is $A = \{x, s\}$, the set of roles is $R = \{p\}$, the set of groups and role relations is $G = ch = \emptyset$, and the set of propositions is $P = \{\text{travel}_x, \text{have\_ticket}_x, \text{paid}_x, \text{climbed\_barrier}_x, \text{pass\_barrier}_x\}$. The following formulas represent assumptions. (1) Having a ticket is the evidence for having paid. (2) Passengers cannot climb the barrier. (3) To travel, a passenger must have passed the barrier. (4) To pass the barrier, a passenger must have paid, or must have climbed it.

1. $\Gamma_{\text{ins}} = \{p(x) \rightarrow AG((\text{have\_ticket}_x \rightarrow \text{paid}_x))\}$,
2. $\Gamma_{\text{passenger}} = \{AG(\neg\text{climb\_barrier}_x)\}$,
3. $C = \{\neg E(\neg\text{pass\_barrier}_x U \text{travel}_x)$,
4. $AG(\text{pass\_barrier}_x \leftrightarrow (\text{have\_ticket}_x \vee \text{climb\_barrier}_x))\}$

We now show that $p(x) \rightarrow \neg E((\neg\text{paid}_x \wedge \neg\text{travel}_x))U\text{travel}_x)$ follows from the above set. Suppose no passenger is travelling; then the behavior is trivially satisfied. Now suppose a passenger is travelling. That means that she passed the barrier (3). That means she has a ticket, or else she climbed the barrier (4). This last option is ruled out by assumption (2). So she has a ticket, which means she paid (1).

Instead, we can specify the system at a higher level of abstraction by specifying the norm and specifying that the norm is regimented.

1. $\Gamma_{\text{ins}} = \{p(x) \rightarrow O_{x,s}\neg E((\neg\text{paid}_x \wedge \neg\text{travel}_x)U\text{travel}_x)\}$,
2. $\Gamma_{\text{passenger}} = \{O_{x,s}\alpha \rightarrow \alpha\}$,
3. $\Gamma_{\text{env}} = \{\}$

Note that the first formula is an ordinary assumption, whereas the second formula is an axiom of the logic. The desired consequence $p(x) \rightarrow \neg E((\neg\text{paid}_x \wedge \neg\text{travel}_x) U\text{travel}_x)$ follows directly from the assumptions.

7 Implementing a Norm by Designing Norm Abiding Agents

A drawback of the regimentation property in the previous section is that it is not expressed in terms of mental concepts, and thus agents cannot reason about it. Therefore we strengthen it to the case in which not only α is the case, but the agent also knows that this is the case. The property that the obligations of agent a towards agent b are regimented is characterized by the following axiom.

$$O_{a,b}\alpha \rightarrow K_a\alpha$$

Note that since we have the axiom $K_a\alpha \rightarrow \alpha$, we have that $O_{a,b}\alpha \rightarrow K_a\alpha$ implies $O_{a,b}\alpha \rightarrow \alpha$. This strong property can be weakened in various directions. First, we can weaken it in the sense that it is not necessarily a fact that the obligation is obeyed, but that at least the opposite is not the case, $O_{a,b}\alpha \rightarrow \neg K_a\neg\alpha$. Second, it can be weakened such that agents *believe* that the obligation is not violated: $O_{a,b}\alpha \rightarrow B_a\alpha$ and $O_{a,b}\alpha \rightarrow \neg B_a\neg\alpha$. Third, the time of compliance to the obligation can be weakened: $O_{a,b}\alpha \rightarrow K_aAF\alpha$, or e.g., $O_{a,b}\alpha \rightarrow K_aAX\alpha$, etc.

At the most abstract level, the formalization of the running example remains nearly the same, we replace the regimentation axiom by the epistemic variant above. Moreover, the logic can specify the decision making of agents at more detailed levels. In particular, the logic can specify when desires or obligations lead to intentions, and when intentions lead to actions. That is, the regimentation axiom $O_{a,b}\alpha \rightarrow K_a\alpha$ is decomposed into the following two axioms.

$$O_{a,b}\alpha \rightarrow I_a\alpha \qquad I_a\alpha \rightarrow K_a\alpha$$

Furthermore, there are many variants on these two axioms. For example, a variant of regimentation concerns conditionality with respect to a conflict between an agent's internal and external motivations. For example, 'if an agent is obliged to buy a ticket, but desires to spend no money, then he intends to buy the ticket anyway, because he is a 'social' agent that does not let his own desires overrule his obligations'. The property that agent a is strongly or weakly respectful with respect to agent b is characterized by the following two axioms. The second formula is implied by the first one if the D axiom $\neg(I_a\alpha \wedge I_a\neg\alpha)$ holds for modality I_a.

$$(O_{a,b}\alpha \wedge D_a\neg\alpha) \rightarrow I_a\alpha \qquad (O_{a,b}\alpha \wedge D_a\neg\alpha) \rightarrow \neg I_a\neg\alpha$$

Finally, the intention of achieving a state can interact with obligations to satisfy the conditions for achieving that state. In such a case, new intentions are implied. The interaction between intention and norms and the creation of intentions can be formulated as the following benevolent axiom:

$$I_x\alpha \wedge O_{x,s}\neg E(\neg\beta U\alpha) \rightarrow I_x\beta$$

The specification of rational agents is one of the main issues studied in agent theory, and these results can be reused in $\mathbf{KBDIO}_{\mathrm{CTL}}$. However, it is also well known that modal logic has to be extended in several ways to make detailed agent models. For example, to specify agents that maximize expected utility $\mathbf{BDI}_{\mathrm{CTL}}$ has to be extended in various ways [28].

8 Implementing a Norm by Relying on Rationality or Social Control

The first way in which norms can be implemented, is to rely on agent rationality and impose fines on norm violations. As mentioned above, the logic can specify when desires or obligations lead to intentions, and when intentions lead to actions. In particular, in the previous section the regimentation axiom $O_{a,b}\alpha \rightarrow K_a\alpha$ is decomposed into $O_{a,b}\alpha \rightarrow I_a\alpha$ and $I_a\alpha \rightarrow K_a\alpha$. In this section, we make sure that the agent *desires* to fulfill the obligation. That is, the regimentation axiom $O_{a,b}\alpha \rightarrow K_a\alpha$ is decomposed into the following three axioms.

$$O_{a,b}\alpha \rightarrow D_a\alpha \qquad D_{a,b}\alpha \rightarrow I_a\alpha \qquad I_a\alpha \rightarrow K_a\alpha$$

We thus interpret the first axiom as the specification that the system is such that it is desired to fulfill the obligation. However, there are several ways in which the axiom can be interpreted. The first explanation is that the agent is norm abiding and *internalizes* its obligations in the sense that they turn into desires. For example, if an agent is obliged to buy a ticket, then it also desires to buy a ticket. The axiom can be weakened to the condition that at least the agent cannot decide to violate the obligation, e.g., at least it cannot desire not to buy a ticket: $O_{a,b}\alpha \rightarrow \neg D_a\neg\alpha$. Instead of respectful, agents may also be egocentric, which can be characterized by similar properties like $(O_{a,b}\alpha \wedge D_a\neg\alpha) \rightarrow I_a\neg\alpha$ and $(O_{a,b}\alpha \wedge D_a\neg\alpha) \rightarrow \neg I_a\alpha$.

The second interpretation of $O_{a,b}\alpha \rightarrow D_a\alpha$ is that the obligation turns into a desire, because violating the desire implies a fine. We already discussed fines in Section 4. The following example is a simplified version, that illustrates how the desire not to be fined can lead to the desire to fulfill obligations, desires. Note that in this formalization the derived desire may also be interpreted as a goal, which is often the case in $\mathbf{BDI}_{\mathrm{CTL}}$ specifications.

$$O_{a,b}\mathrm{paid} \rightarrow K_a(\neg\mathrm{paid} \rightarrow \mathrm{fine}) \qquad (K_a(\neg\mathrm{paid} \rightarrow \mathrm{fine}) \wedge D_a\neg\mathrm{fine}) \rightarrow D_a\mathrm{paid}$$

The third interpretation of $O_{a,b}\alpha \rightarrow D_a\alpha$ is that violating the obligation leads to social embarrassment. This can be specified analogously to fines.

9 Related Work

Despite the popularity of Roa and Georgeff's logic in agent theory to specify and verify multiagent systems, the logical analysis of their logic is still in its infancy. Rao and Georgeff did not present a full axiomatization of their logic, which was only presented much more recently by Schild's reduction to the μ calculus. Moreover, the axiomatization is restricted to the logic without any interaction axioms. In the meantime, logicians have restricted themselves to small fragments of their logic, for example to study the interaction between knowledge and time, or to study the interaction between beliefs and obligations.

Within deontic logic in computer science, our work is most closely related to dynamic deontic logic, extensions of dynamic logic with modalities for obligations and

permissions. In multiagent systems, recently norms and normative systems are discussed, but their specification or verification has has not been addressed. In action programs in IMPACT, there is a discussion on whether obligations can be violated, i.e., on norm regimentation [29]. We have addressed this issue in the context of the BOID project, see http://boid.info. The present paper extends our short paper [30].

10 Summary

The motivation of our work is how such normative computer systems can be specified. This problem breaks down as follows:

1. How to develop a logic for specification of normative computer systems?
2. Which kind of properties can be expressed in the specification logic?
3. How to apply the specification logic to application domains?

Our methodology is to specify properties involving obligations in an extension of Rao and Georgeff's $\mathbf{BDI}_{\mathrm{CTL}}$ [6,3–5]. Such an extension consists of an extension of the logic and an extension of the properties expressed in the logic. Obligations are motivational attitudes, just like desires, but they also have organizational aspects. This can be represented, for example, by introducing roles and by formalizing obligation as a directed modality. Thus, whereas we may say that agent a desires to prepare a report, we say that the agent a is obliged to prepare a report *towards another agent* b. We accomplish our extension of $\mathbf{BDI}_{\mathrm{CTL}}$ with obligations in the following steps:

- The introduction of an extension of $\mathbf{BDI}_{\mathrm{CTL}}$ called $\mathbf{KBDIO}_{\mathrm{CTL}}$, that makes the distinction between desires and obligations explicit, as well as the distinction between beliefs and knowledge. We extend $\mathbf{BDI}_{\mathrm{CTL}}$ with directed obligations [7–11] and roles.
- We introduce various single agent and multiagent properties. These properties can be used in a high-level design language for normative computer systems.
- We apply the logic and the properties to the implementation of an organizational norm.

References

1. Meyer, J., Wieringa, R.: Deontic Logic in Computer Science: Normative System Specification. John Wiley and Sons (1993)
2. Wieringa, R., Meyer, J.: Applications of deontic logic in computer science: A concise overview. In: Deontic Logic in Computer Science. John Wiley & Sons, Chichester, England (1993) 17–40
3. Schild, K.: On the relationship between BDI-logics and standard logics of concurrency. Autonomous agents and multi-agent systems **3** (2000) 259–283
4. Dastani, M., van der Torre, L.: An extension of BDICTL with functional dependencies and components. In: Procs. of LPAR'02. LNCS 2514, Springer (2002) 115–129
5. Dastani, M., van der Torre, L.: Specifying the merging of desires into goals in the context of beliefs. In: Procs. of EurAsia ICT 2002. LNCS 2510, Springer (2002) 824–831

6. Rao, A.S., Georgeff, M.P.: Decision procedures for BDI logics. Journal of Logic and Computation **8** (1998) 293–343
7. Herrestad, H., Krogh, C.: Obligations directed from bearers to counterparties. In: Procs of ICAIL'95, New York (1995) 210 – 218
8. Dignum, F.: Autonomous agents with norms. Artificial Intelligence and Law **7(1)** (1999) 69–79
9. Singh, M.P.: An ontology for commitments in multiagent systems: toward a unification of normative concepts. Artificial Intelligence and Law **7** (1999) 97–113
10. Broersen, J., Dastani, M., Huang, Z., van der Torre, L.: Trust and commitment in dynamic logic. In: Procs. of EurAsia ICT 2002. LNCS 2510, Springer (2002) 677–684
11. Tan, Y., Thoen, W.: Modeling directed obligations and permissions in trade contracts. In: Procs of HICCS'98. (1998) 166–175
12. Ferber, J., Gutknecht, O.: A meta-model for the analysis and design of organizations in multiagent systems. In: Procs. of ICMAS'98, IEEE Press (1998) 128–135
13. Carmo, J., Pacheco, O.: A role based model for the normative specificatiuon of organized collective agency and agents interaction. Autonomous Agents and Multi-Agent Systems **6** (2003) 145–184
14. Wooldridge, M., Jennings, N., Kinny, D.: The Gaia methodology for agent-oriented analysis and design. Autonomous Agents and Multi-Agent Systems **3** (2000) 285–312
15. Dastani, M., Dignum, V., Dignum, F.: Role assignment in open agent societies. In: Procs. of AAMAS'03, ACM (2003) 489–496
16. Arbab, F., de Boer, F., Bonsangue, M., Scholten, J.G.: A channel-based coordination model for components. Technical Report SEN-R0127, CWI, Amsterdam (2001)
17. Firozabadi, B.S., van der Torre, L.: Towards an analysis of control systems. In: Procs. of ECAI'98. (1998) 317–318
18. Castelfranchi, C.: Modelling social actions for AI agents. Artificial Intelligence **103** (1998) 157–182
19. Wright, G.v.: Deontic logic. Mind **60** (1951) 1–15
20. Firozabadi, B.S., Sergot, M.J.: Revocation schemes for delegated authorities. In: Procs. of POLICY'02. (2002) 210–213
21. Bandmann, O., Firozabadi, B.S., Dam, M.: Constrained delegation. In: Procs. of IEEE Symposium on Security and Privacy 2002. (2002) 131–140
22. Searle, J.: The Construction of Social Reality. The Free Press, New York (1995)
23. Jones, A., Sergot, M.: A formal characterisation of institutionalised power. Journal of IGPL **3** (1996) 427–443
24. Boella, G., van der Torre, L.: Regulative and constitutive norms in normative multiagent systems. In: Procs. KR'04, Whistler, CA (2004)
25. Broersen, J., Dignum, F., Dignum, V., Meyer, J.J.: Designing a deontic logic of deadlines. In: Procs. of DEON'04. LNCS, Springer (2004) This volume.
26. Cohen, P., Levesque, H.: Intention is choice with commitment. Artificial Intelligence **42** (1990) 213–261
27. Rao, A., Georgeff, M.: Modeling rational agents within a BDI-architecture. In Allen, J., Fikes, R., Sandewall, E., eds.: Procs. of KR'91, Morgan Kaufmann Publishers (1991) 473–484
28. Rao, A.S., Georgeff, M.P.: Deliberation and its role in the formation of intentions. In: Procs. of UAI-91.(1991)
29. T. Eiter, V.S. Subrahmanian, G.P.: Heterogeneous active agents, I: Semantics. Artificial Intelligence **108** (1999) 179–255
30. Broersen, J., Dastani, M., van der Torre, L.: BDIO_CTL: Properties of obligation in agent specification languages. In: Procs. of IJCAI'03. (2003) 1389–1390

Maintaining Obligations on Stative Expressions in a Deontic Action Logic*

Adam Zachary Wyner

King's College London
London, UK
wyner@dcs.kcl.ac.uk

Abstract. We consider the logical representation of obligations on stative expressions such as *The yard must be clean* in the context of legal contract formation, execution, and monitoring (cf. Wyner ([28])). In a contract, the expression may understood as an obligation to *maintain* a property. We use a *Deontic Action Logic* to represent obligations over the course of time (Khosla and Maibaum ([13]) and Meyer ([17])). Our analysis is in contrast to d'Altan, Meyer, and Wieringa ([6]), who reduce deontic operators to an Alethic Logic plus a violation proposition (Anderson and Moore ([1])), which has no temporal component. In addition, they use a Deontic Action Logic to represent obligations on actions. We claim the Alethic component of the logic is redundant for the purposes of representing obligations on stative expressions in a contract. In the course of the analysis, we introduce *polynormativity,* which contrasts with the *binormativity* of standard DAL or alethic logic plus a violation proposition. We discuss the advantages of polynormativity in reasoning from violations and fulfillments.

1 Introduction

We consider the logical representation of maintaining obligations on stative expressions such as *The yard must be clean,* particularly in the context of legal contract formation, execution, and monitoring (cf. Wyner ([28]) for discussion of the application). Ought-to-be statements may be understood as *system invariants,* those 'normal' properties which must be true in every state of a model. The problem with such system invariants is what happens when the normal property does not hold, in which case, a violation may arise such as may appear in systems of fault tolerance. A deontic logic on properties is useful in defining how to handle violations. We also have deontically specified actions, which we analyze with Deontic Action Logics (Khosla and Maibaum ([13]) and Meyer ([17])). This paper provides an analysis of ought-to-be and ought-to-do expressions in a

* This work was prepared while the author was a postgraduate student at King's College London under the supervision of Tom Maibaum and funded by a studentship from Hewlett-Packard. It has benefitted from discussions with Tom Maibaum, Andrew Jones, and comments from two anonymous reviewers. The author thanks Tom, Andrew, and HP for their support and advice. Errors rest with the author.

A. Lomuscio and D. Nute (Eds.): DEON 2004, LNAI 3065, pp. 258–274, 2004.

Deontic Action Logic which is suitable for legal contract modelling, execution, and modelling; for reasons of space, we largely largely focus on obligations, not permissions or prohibitions.

The layout of the paper is as follows. In the next section, we discuss the problem and some framework assumptions concerning agents, actions, action negation, deontic logic 'paradoxes', and polynormativity. Following this, we present our analysis of ought-to-be expressions in terms of maintaining the property (similar to notions discussed in Khosla and Maibaum ([13]), Sergot and Richards ([22]), and Hilpinen ([9]). Then we compare our analysis to the presentation of ought-to-be expressions found in d'Altan, Meyer, and Wieringa ([6]), which provides an overview of the issues as well as a particular analysis. Our principle objections are that their critique of the 'classic' analysis of ought-to-be expressions, where one is forbidden to undo the property, does not eliminate the analysis. In addition, they introduce a logic in which ought-to-do expressions are represented using Deontic Action Logic and ought-to-be expressions are represented using Alethic logic plus a violation atomic proposition. We believe this system is more complex than need be and that the alethic component can be eliminated. Our representation also has the advantage that it defines what follows should a violation occur.

2 Initial Discussion

The following examples express obligations on actions and states respectively.

Example 1. Bill must leave.

Example 2. The yard must be clean.

In (1), the agent is obligated to do a leaving action, while in (2), it would appear that the obligation that the yard is clean holds irrespective of an agent. Should either of these obligations not be met, that is, should the agent not leave or the yard not be clean, then the obligation is violated. Consequences may follow from this violation. For instance, if the agent does not leave or if the yard is not clean, then the agent may incur another obligation, say to pay a penalty.

Initially, we might analyze (1) as in (3) and (2) as in (4), using the same deontic operator, here given as **Obligated.**

Example 3. **Obligated**(leave(bill)), where *bill* is the agent of the action predicate *leave,* and (leave(bill)) holds or not of a state.

Example 4. **Obligated**(clean(the yard)), where *clean(the yard)* is a property which holds or not of a state.

In such an analysis, the obligation operator applies equally to stative expressions and action expressions. The first question is whether such a uniform analysis is accurate; that is, whether **Obligated** indeed applies equally to any sort of expression as schematized in (5).

Example 5. **Obligated**(P), where is any sort of property of a state.

It is clear that though the natural language deontic expressions *obligated, must,* and *ought* appear with both stative and action expressions, the implications from each case are very distinct, and in particular, with respect to how violations arise. For example, with respect to the ought-to-do expressions, if Bill does not do what he is obligated to do, but performs some other action instead, then he is in violation. In contrast, with respect to ought-to-be expressions, a violation arises if the yard is not clean, irrespective of an action being performed or not.

It would appear that having one deontic operator on both sorts of expressions is intuitively unacceptable and that we must instead make some sortal distinction between the expressions which the deontic operators apply to and also among the deontic operators themselves. Suppose we do sort expressions into action sorts and stative sorts; furthermore, we suppose two different deontic operators, one on actions and another on statives, which we express with **ObligatedAction** and **ObligatedState**

Example 6. **ObligatedAction**(leave(bill))

Example 7. **ObligatedState**(clean(the yard))

We assume for current purposes that we can sort expressions into action and stative sorts[1]. The question is, then, exactly what does (6) imply in comparison and contrast to (7).

Given a clear analysis of their similarities and differences, the next question is what is the best analysis. The space of alternative analyses (besides the one we dismissed above) appears to be as follows:

1. We have a logic comprised of two distinct 'sublogics', where one sublogic defines the ought-to-be and the other which defines ought-to-do. There are no implicational relations between the sublogics. We do not reduce one operator to the other.
2. We have a logic comprised of two distinct 'sublogics', where one sublogic defines the ought-to-be and the other which defines ought-to-do. There are implicational relations between the sublogics. Yet we do not reduce one operator to the other.
3. We have a homongeneous logic in which we define both the operators, but we do not reduce one operator to the other.
4. We reduce one operator to another.

The first position is untenable largely because of the close similarities between the operators; besides the similarities of form, they both imply that if something does not hold (or is not done), a violation is incurred. The second position is

[1] See Katz ([12]) and references therein for extensive discussion of intuitions and formal semantic analysis which distinguish action and stative expressions. In general, more research remains to be done on exactly what define action and stative expressions, and just which sorts the deontic operators may apply to. Such research is crucial for any logical analysis of actual contracts.

advocated by d'Altan, Meyer, and Wieringa ([6]). We argue for the third position. d'Altan, Meyer, and Wieringa ([6]) argue persuasively against the fourth position.

In the following subsections, we touch on a range of topics relating to the assumptions and simplifications in which we make our proposal. Each topic is, in and of itself, the subject of significant research; our intention is touch just on those elements which relate to the core of our proposal.

2.1 A Dynamic Logic

We represent our contractual expressions in a Dynamic Logic rather than Standard Deontic Logic[2], for in a Dynamic Logic, we can represent changes in the values of deontically specified expressions with respect to time and the actions of agents. In our domain of application, which is automated contracting, we must account for change of states over time and as a result of actions by agents. In particular, we adopt a Deontic Action Logic (Khosla and Maibaum ([13]) and Meyer ([17])), which is based on Dynamic Logic (Harel, Kozen, and Tiuryn ([8])), here providing a very brief review of basic concepts.

A Dynamic Logic is a logic of actions, where actions are state transitions given properties of precondition (before the performance of the action) and postcondition (after the performance of the action) states. We do not consider complex actions such as those formed by choice, simultaneous, negation, or sequence operators. We have a set of atomic action names $\{\alpha, \beta, \ldots\}$, a set of agent names

[2] Standard Deontic Logic has deontic operators and action expressions. The deontic logic (from Kanger and Lindahl REFERENCES) is EMCP (in the Chellas classification). This is the smallest system containing prepositional logic and the following axioms and rules.

Example 8.
a. **O.RE:** If $\vdash A \leftrightarrow B$, then $\vdash OA \leftrightarrow OB$
b. **O.M:** $O(A \wedge B) \rightarrow (OA \wedge OB)$
c. **O.C:** $(OA \wedge OB) \rightarrow O(A \wedge B)$
d. **O.P:** $\neg O\bot$

The difference between EMCP and Standard Deontic Logic, which is a normal modal logic of type KD, is that SDL is EMCP together with the necessitation rule $\vdash A$, then $\vdash OA$. However, the rule of necessitation does not play any role in the discussion of normative positions, so it is left out. In this system, permission P is the dual of obligation O: $PA =_{def} \neg O\neg A$. The Action Logic differs from Dynamic Logic in that it abstracts from the temporal dimension. We have agent relativized action operators, E_a and E_b, where a and b are agents, and $E_a A$, where A is a property, is read as *agent a sees to it that A* or *agent a is responsible for it being the case that A*. Actions abide by the following axioms.

Example 9. **E.RE** If $A \leftrightarrow B$, then $E_x A \leftrightarrow E_x B$

Example 10. **E.T** $E_x A \rightarrow A$

Note that in SDL, $O(E_x F) \equiv O(F \wedge E_x F)$, while in DAL, correlated expression $[A, \overline{\alpha}]$(Violation) is not equivalent to $[A, \overline{\alpha}]$(Violation) $\wedge$ Violation.

{A, B, ... }, and a set of prepositional letters $\{\phi, \psi, \ldots\}$. The expression $[A, \alpha](\phi)$ is to be read *in every state where agent A performs action* α, ϕ *holds in the subsequent state.* Alternatively, where we read the action as a function from states to states, we may say that where the agent appears and the state satisfies the precondition properties specified by the ction, then the action maps the current state to a subsequent state which satisfies the postcondition property ϕ, perhaps along with other properties of the postcondition as specified by the action.

In a Deontic Action Logic, actions are also ascribed properites such as whether the action is obligatory or prohibited. For example, where obligation, permission, and prohibition are represented by predicates of actions **Obligated, Permitted,** and **Prohibited,** we have expressions of the form **Obligated**(A, α), **Permitted**(A, α), and **Prohibited**(A, α). We assume agents in every case and discuss this further below. In Khosla and Maibaum ([13]) and Meyer ([17]), a designated property **Violation,** read as *violation, is* introduced (see Anderson and Moore ([1]) for a precursor in Alethic Modal Logic); a state in which this property holds is understood to be in violation or to be flagged for violation, which is to say it is non-normative. Such an analysis characterizes a *binormative analysis,* for states are either normative or non-normative. For the moment, it is easier to discuss the notion of prohibition rather than our target notion of obligation. The meaning of **Prohibited**(A, α) is then given in terms of Dynamic Logic and a violation flag.

Example 11. **Prohibited**$(A, \alpha) \equiv [A, \alpha]$(Violation)

In other words, if A does α, then in the subsequent state, a violation is marked. One way to understand **Prohibited** is as an operator on actions – a function from actions to actions such that where the preconditions of the action are met and the action is performed, the postcondition state bears not only the properties ascribed by the action, but in addition, a designated violation property, which is used to signal that what was forbidden has occurred.

Obligation with respect to an action is somewhat more complex. We shall make a simplying assumption for the purposes of this presentation (see Meyer ([17]), Khosla and Maibaum ([13]), and Broersen ([2]) for discussion action negation, particularly Broersen which is similar to our view). For obligations on actions, a violation arises where some action other than the obligated action is performed; where α is the action, we may indicate an *alternative to* α with $\overline{\alpha}$, which may be understood as an element of the set of actions without α: $\overline{\alpha} \in \{\alpha, \beta, \ldots\} - \alpha$, where the set of actions is finite.

Example 12. **Obligated**$(A, \alpha) \equiv [A, \overline{\alpha}]$(Violation)

Given that actions can result in violations, the performance of a prohibited action or failure to perform an obligatory action is marked rather than ruled out by the system. In particular domains, such as fault tolerance or contract performance, we want to represent and reason with respect to what is prohibited or obligated; we cannot simply rule out such behavior, for the fact is that it does occur. Particularly in the domain of contract modelling and analysis, a Deontic

Action Logic is key, for with it we can represent and reason about the behaviors of the agents as they perform error prone, by accident or design, actions over the temporal course of the contract. As the agents perform the actions, they change states; we are interested in the deontic specification of such state changes.

In addition to these conceptual advantages, as Meyer ([17]) points out, a Deontic Action Logic avoids many of the so-called paradoxes which arise with Standard Deontic Logic, for many of the paradoxes either are not well-formed, have solutions (cf. discussion of free choice permission in Meyer, Weigand, and Wieringa ([18])), or are not worse than those suffered by other formalizations of deontic logic. In any case, the paradoxes do not create inconsistency, but are cases of overgeneration in which some of the formulas do not correlate well with our intuitive interpretation of what the expression should mean. There are a variety of ways to see to it that the logic does not overgenerate while preserving the appropriate expressions. Given that we only consider well-formed and acceptable formulas, the problem of the paradoxes does not bear on our discussion. We also have little to add to the discussion of *Normative Positions* (Sergot ([24]), Sergot ([23]), Sergot and Richards ([22]), and Jones and Sergot [11]), the aim of which is maximally consistent sets of expressions of actions and deontically specified actions in a state.

2.2 The Role of Agents

As made clear in the literature on the ought-to-be and ought-to-do distinction (cf. d'Altan, Meyer, and Wieringa ([6]), Forrester ([7]), Horty ([10]), Broersen and van der Torre ([3]), and references therein), a key element of the discussion is the presence or absence of an agent in the logical analysis as well as the distinction between personal and impersonal obligations (cf. Krogh and Herrestad ([14]). For instance, D'Altan, Meyer, and Wieringa ([6, :1]) claim that ought-to-be statements "*... express a desired state of affairs without necessarily mentioning actors and actions....*". Furthermore, D'Altan, Meyer, and Wieringa ([6, :78]) follow Castaneda (1970:452) "...in separating deontic statements into those that involve agents and actions and support imperatives (ought-to-do) and those that involve states of affairs and are agentless and have by themselves nothing to do with imperatives." We should point out that there is a difference between the absence of an actor in the linguistic form of an expression such as *Jill was pushed* and the absence of that actor in the semantic representation. However, to discuss this further would require a digression into the syntax and formal semantics of natural language such as found in Wyner ([26]) and is outside the scope of this paper.

For our purpose, which is to provide analyses of contractual terms, we may make a simplifying assumption, namely, that every deontically specified expression has an agent which bears the obligation with respect to the action or property; this agent may be explicitly given or implicit (cf. Wyner ([28]) for further discussion). The reason for this assumption is straightforward: in the cases under study, contracts are explicit agreements in which the parties agree to be bound by the terms of the contract. The parties are bearers of the obligations, which we

designate as the agents of actions and holders of properties. On the one hand, this may limit the applicability of the proposed analysis; on the other hand, it places a criteria which other analyses must satisfy in order to be of use in representing contracts, for ought-to-be expressions must include some agentive bearer of the obligation.

2.3 Refinement of the Violation Atom – Polynormativity

In the Deontic Action Logics of Khosla and Maibaum ([13]) and Meyer ([17]), deontically specified actions can lead to a subsequent state in which an atomic violation property holds or not. The purpose of introducing the violation property is to allow one to reason with violations rather than simply ruling them out. However, as argued in Wyner ([27]), it is not enough to have but one atomic violation property for reasoning about violations in contracts, for simply put, any two violations are, then the same. For example, if an agent *Jill* violates an obligation on her behavior and another agent *Bill* violates an obligation on his behavior, the violation is the same. However, the consequences of agent's action may differ; that is, Jill's violation may result in one penalty, while Bill's violation results in another penalty. We want the violation markers to be such as to differentiate among the agents and actions. This is particularly important in a legal setting where it is crucial to apply sanctions to particular individuals for particular actions. The theories of Khosla and Maibaum ([13]) and Meyer ([17]) do not sufficiently discriminate in this way, which is characteristic of *bi-normataive theories,* that is, theories which only distinguish between normative and non-normative states.

Our approach provides *fine-grained* distinctions among violations (or fulfillments) so as to support reasoning from them. We can call it a *polynormative theory* (cf. Wyner ([27])). Somewhat similar proposals appear in d'Altan, Meyer, and Wieringa ([6, :108-109]) and van der Meyden ([16]), though of more limited use. Let us first consider the general form and then a particular example. In (14), we abstract from the form of (13), where P is some predicate on agent-action pairs, and Q is some proposition which follows from performance of the action 'under' P.

Example 13. $\mathbf{Prohibited}(\mathrm{A}, \alpha) \equiv [\mathrm{A}, \alpha](\mathrm{Violation})$

Example 14. $\mathbf{P}(\mathrm{A}, \alpha) \equiv [\mathrm{A}, \alpha](\mathrm{Q})$

The predicate **P** is then defined in the logic in terms of how it alters the performance of the action α by that agent A, in this case by stipulating that after A does α Q holds, which need not have been the case where P not to predicate of the agent and action. In effect, **P** ascribes a *value* to A's performance of α, and we may call any dynamic logic which supports this schema a *Value Action Logic.* The Deontic Action Logics of Khosla and Maibaum ([13]) and Meyer ([17]) are, then, instances of a Value Action Logic, where **P** is the predicate Prohibited and Q the proposition Violation. Different logics are defined by how the values on actions change what holds after the performance of the action. Note that from

the schema in (14), we cannot judge whether A's performance of α is or is not *ideal* or of some lesser status; this is a judgement lain over the expressions, not intrinsic to the logic itself, and not clearly relevant to them (cf. comments by Meyer ([17, :126]) on ideality in deontic logic).

Instead of (13) as an instance of (14), we may have (15), where S and T are predicates of agent-action pairs.

Example 15. $\mathrm{S}(\mathrm{A}, \alpha) \equiv [\mathrm{A}, \alpha](\mathrm{T}(\mathrm{A}, \alpha))$

We suppose that $\mathrm{S}(\mathrm{A}, \alpha)$ is defined in terms of an action, while $\mathrm{T}(\mathrm{A}, \alpha)$ is a proposition which is not defined in terms of an action, but is a proposition which holds of a state. We can redefine Prohibition in these terms. $\mathrm{ProhibitedAction}(\mathrm{A}, \alpha)$ says that A's performance of α is prohibited, which means that were A to perform α, it would lead to a state marked with $\mathrm{ViolationProhibitionAction}(\mathrm{A}, \alpha)$, which indicates what was prohibited has been performed.

Example 16. $\mathrm{ProhibitedAction}(\mathrm{A}, \alpha)$
$\equiv [\mathrm{A}, \alpha](\mathrm{ViolationProhibitionAction}(\mathrm{A}, \alpha))$

What follows from $\mathrm{ViolationProhibitionAction}(\mathrm{A}, \alpha)$ may be further specified, for example, what further properties or actions are implied. To structure violations, we can define implicational relationships among them, or for that matter introduce fine-grained markers for reward (cf. Meyer ([17, :125])).

To get a flavor of the utility of this format, consider an example. Suppose Bill is obliged to leave the office at 5pm, Bill is prohibited from buying alcohol after 11pm, and Bill is prohibited from driving over 60 MPH. Jill is only prohibited from buying alcohol after 11pm. In a system with but one atomic violation marker, any violation would result in Bill's current account being debited £100. This seems unreasonable, and we would like to associate the violation with particular agents and actions. For instance, Bill's violation of leaving the office at the wrong time leads to a debit of £10, his violation of buying alcohol too late costs him £20, and his violation of driving too fast costs him £50. Finally, Jill's violation leads to a debit of Jill's account of £20. Thus, Jill's violation leads to a violation particular to Jill, and a consequent sanction; Bill's violations only lead to sanctions on Bill, and these may be cumulative, which could not be so with but one atomic violation property.

We can express the deontically specified actions as follows, where we assume actions such as *leaveOfficeAt5pm* are defined in the logic.

Example 17. ObligatedAction(Bill, leaveOfficeAt5pm)
$\equiv [\mathrm{Bill}, \overline{leaveOfficeAt5pm}]$
(ViolationObligationAction(Bill, leaveOfficeAt5pm))

Example 18. ProhibitedAction(Bill, buyAlcoholAfter 11pm)
$\equiv$ [Bill, buyAlcoholAfter11pm]
(ViolationProhibitionAction(Bill, buyAlcoholAfter11pm))

Example 19. ProhibitedAction(Bill, driveOver60mph)
$\equiv$ [Bill,driveOver60mph]
(ViolationProhibitionAction(Bill, driveOver60mph))

Example 20. ProhibitedAction(Jill, buyAlcoholAfter11pm)
≡ [Jill, buyAlcoholAfter11pm]
(ViolationProhibitionAction(Jill, buyAlcoholAfter11pm))

Furthermore, we may define the system such that from violations with respect to deontic specifications, agents, and actions, specific consequences follow, here just that additional obligations are incurred.

Example 21. ViolationObligationAction(Bill, leaveOfficeAt5pm)
→ ObligatedAction(Bill, pay £10)

Example 22. ViolationProhibitionAction(Bill, buyAlcoholAfter11pm)
→ ObligatedAction(Bill, pay £20)

Example 23. ViolationProhibitionAction(Bill, driveOver60MPH)
→ ObligatedAction(Bill, pay £50)

Example 24. ViolationProhibitionAction(Jill, buyAlcoholAfter11pm)
→ ObligatedAction(Bill, pay- £20)

By the same token, later we introduce markers for reward or fulfillment of an obligation (Meyer ([17]) has a somewhat similar basic notion.) Our system is richer and more flexible in that the deontic operators and their correlated violations or fulfillments can be related to a range of parameters and implications.

3 Ought-to-Be Operators Expressed in a Deontic Action Logic

We want a representation of the obligation *The yard must be clean* such as might appear in the context of a legal contract. By assumption, one of the contractual participants is the agent of this obligation; this in turn implies that when the state fails to hold, the agent is liable to suffer sanction. Consider that the expression appears in a rental contract as part of the reponsibilities of a tenant. Intuitively, it means that the tenant has the obligation to keep the yard clean over the period of time of the tenancy. Of course, at the beginning of the tenancy, the yard may not be clean, in which case, the tenant is obligated to clean it; not cleaning the year implies a violation. Alternatively, the yard may start out clean, but become dirty, in which case, the tenant is obligated to clean it, or again suffer a violation. We call this interpretation of an ought-to-be expression *an obligation to maintain a state,* for the agent is obligated to maintain a state or suffer violations. The tenant's satisfaction of the obligation over the course of the tenancy may be specified by the contract, for example that the yard is clean for a certain length of time, that the yard is not dirty when it is inspected, or that only a certain number of violations arise. This allows a flexible notion of satisfaction of the obligation, for it allows some violations to occur; however, we do not have space here to discuss this topic.

Our formal expression of an obligation to maintain a state is as follows. We assume deontic predicates *ObligatedState, PermittedState,* and *ProhibitedState,*

which are of type <Agt, Formula>, where Agt is an individual with the agentive properpty and Formula is a formula of first-order predicate logic. Our definition of an obligation to maintain a property with respect to an agent Agt and a formula ϕ is defined in a Deontic Action Logic. Just as we have markers to indicate violation of an obligated action, we may indicate fulfillment with FulfilledObligatedState(A, ϕ). There are constants and variables of agents and actions: Agt is a constant and $\mathbf{Agt}_x$ is a variable of type agent. We may suppose that ϕ is *The yard is clean;* Agt denotes some particular individual.

Definition 1. $ObligatedState(Agt, \phi) \equiv$
$[\phi \rightarrow [FulfilledObligatedState(Agt, \phi)$
$\wedge \exists Agt_x \exists \alpha [ProhibitedAction(Agt_x, \alpha)$
$\wedge [Agt_x, \alpha](\neg\phi \wedge Violation\text{-}ObligatedState(Agt, \phi))$
$\wedge \exists \beta [[Agt, \beta](\phi) \wedge ObligatedAction(Agt, \beta)]]]] \wedge$
$[\neg\phi \rightarrow [Violated\text{-}ObligatedState(Agt, \phi)$
$\wedge \exists \gamma [[Agt, \gamma](\phi) \wedge ObligatedAction(Agt, \gamma)]]]$

There are two main portions on the right hand side of the definition. The first portion begins where ϕ is the antecedent of the first conditional; the second portion begins with the case where $\neg\phi$ is the antecedent of the second conditional.

Suppose ϕ holds. This implies that the obligation is marked as being fulfilled. In addition, we assume that for some agent and some action, it is forbidden for that agent to do that action. The action is one which undoes the property and introduces a violation marker. In addition, an obligation is introducted to produce a subsequent state in which the property holds. We return in a moment to the import of the indefinite agent and this additional obligation.

Suppose $\neg\phi$. Therefore, the obligation is marked as being in violation, and an obligation is incurred on the original agent to do some action which results in the property holding again. For legal contracting, this seems to be a reasonable condition, for it may be the case that an agent in a contract accepts an obligation with respect to a state which ought to hold but does not hold at the time the obligation is incurred. The agent starts off on the wrong foot in that the agent already is marked for having violated the obligation. The significance of this is discussed later. There is a degree of redundancy between the consequents of the first and second portions, but it is worth it to deal with such initial states where the obligated property does not hold.

This formulation of an obligation to maintain a state implies that there are no obligations with respect properties where an agent cannot undo the property. Suppose we were to have an expression in a contract such as ObligatedState(A, $P \vee \neg P$). Since it is always true in every state that $P \vee \neg P$, then the agent has fulfilled the obligation. But in addition, there must be some agent who performs some action such that were the agent to perform the action, $P \vee \neg P$ is false. Since there cannot be any such action, the consequent of this portion is false, making the whole expression false. For the representation of contracts, this seems reasonable: no contract will include some obligation with respect to a property which has a truth value which cannot be altered in the model.

Consider a weaker case, whether an agent can bear an obligation that the yard is clean, but not have the capacity to perform an action to either undo the property or to redo it as needed (cf. d'Altan, Meyer, and Wieringa ([6, :80])). The analysis suggests that such an agent can bear such an obligation, depending on what it means for the agent to have the capacity. If the property is true, then the agent fulfills the obligation; were some other agent to undo the property, then the agent would be in violation. Where the property is false, then the agent is again in violation. Where the agent is in violation, the agent is obligated to do something to bring about the property holding. What actions may count towards fulfillment of this latter obligation depend on circumstances; for example, the agent bearing the obligation may perform an action which designates someone else to perform the action. If it is the case that the agent has no capacity whatsoever, say the agent is comatose, then it seems reasonable to say that the obligation no longer holds.

We believe these claims are reasonable for contracts (cf. Wyner ([28])). Suspensions or reinterpretations of contractual obligations are common in those portions of a contract having to do with exceptions such as should the country be at war, or where there is a natural disaster, or where the agents are somehow incapacitated with respect to performance requirements. In such cases, contractual obligations can be suspended because the agents cannot perform the actions needed to satisfy them. Along these lines, we may distinguish ways in which the agent is incapacitated and subsequently vary the violations. One distinction might be between a natural disempowerment, a self-induced disempowerment, and disempowerment induced by another. Whether or not the obligation is maintained as well as what penalties follow may vary. For example, in case of a horrendous natural disaster, one's debt obligations may be 'forgiven' and no violations follow. But, should one induce one's own poverty such that debt obligations cannot be met, then one might bear a violation which introduces subsequent obligations; Chapter 11 Bankruptcy law in the United States does not obviate the obligations and violations, but marks the violation and replaces the original debt obligation with other obligations. Finally, if one is robbed so cannot meet one's debts, one might still have the same obligations as before; the insurance industry is built around the notion of protecting oneself from natural disaster or disaster caused by others so as to meet one's obligations. The analysis we have presented so far could express such differences by marking agents and actions with respect to their properties and associating them with different violations.

We should return to consider the indefinite agent and action of the first portion. We could, if we wanted, be more specific, say for a specific obligated property, we can define which agent and which action must be here. It is an advantage to leave the definition underspecified on this point. Moreover, it expresses an interesting notion as it is. Suppose Bill is the one who is obliged with respect to the property of the yard being clean. It implies that he should not do anything to make the yard unclean. Furthermore, should any agent do something to make the yard unclean, then Bill bears the violation. Say the wind blows leaves into the yard, Bill is in some sense still responsible with respect

to the obligation to keep the yard clean. True, Bill did not do anything wrong himself, but he does bear responsibility, which means here that he bears the consequences of the property not holding. That he bears responsibilty in this way might motivate him to do what he can to prevent other agents from inducing the violation, say by cutting down all the trees in a 10 mile radius.

Both portions (one where ϕ holds and the second where $\neg\phi$ holds) introduce obligations to perform an action which returns a state in which the property holds. For the first portion, this obligation is incurred only where the property has become undone. In some cases, it is not possible to perform an action such that ϕ holds again. For instance, suppose one is obligated to maintain some real estate forever wild; once this has been violated, say by constructing a highway through it, no action can return it to it formerly pristine wild state. Instead, a compensatory obligation may be incurred. For example, we might say the following for some specific properties ψ and π, where ψ is *A piece of real estate is mid* and π is *pay compensation.* We see in the following that where the initial property is violated, some compensatory action is obligated. Of course, where $\psi = \pi$, Definition (2) is equivalent to Definition (1).

Definition 2. $ObligatedState(Agt, \psi) \equiv$
$[\psi \rightarrow [Fulfilled\text{-}ObligatedState(Agt, \psi)$
$\wedge \exists Agt_x \exists\alpha\ [ProhibitedAction(Agt_x, \alpha)$
$\wedge [Agt_x, \alpha](\neg\psi \wedge Violation\text{-}ObligatedState(Agt, \phi))$
$\wedge \exists\beta\ [[Agt, \beta](\pi) \wedge ObligatedAction(Agt, \beta)]]]] \wedge$
$[\neg\psi \rightarrow [Violated\text{-}ObligatedState(Agt, \psi)$
$\wedge \exists\gamma\ [[Agt, \gamma](\pi) \wedge ObligatedAction(Agt, \gamma)]]]$

Finally, notice that the definition of ObligatedState(Agt, ϕ) introduces four different ways in which violations may arise, each of which may have distinct or interrelated consequences. There is the case where the property does not hold in the state in which the obligation is given, which results in a direct violation of the obligation. Related to this case, there is the potential violation which follows from failing to perform the obligated action to bring the property about. There is the case where the property does hold in the given scenario, but in which an agent performs some action which undoes it; this induces a violation on the obligation as well. And finally, in this subordinate state, there is the obligation to perform an action which results in a scenario in which the property again holds; the failure to perform this action could result in a violation as well. These violations allow us great flexibility in defining what sorts of consequences flow from the violations. Perhaps, for example, while it is best if the agent never undoes the property, should the agent undo it, it is not punished, so long as the agent does something to bring the property to hold again. In such an instance, it is only in the case where the agent both undoes the property and does nothing to bring the property to hold again which meets the sanction. Or, alternatively, it could be the case that undoing the property meets some sanction, and failing to redo it is marked but not sanctioned. Or, that undoing the property meets some sanction and failing to redo it meets a further and worse sanction.

4 d'Altan, Meyer, and Wieringa ([6])

In this section, we discuss the analysis of ought-to-be and ought-to-do operators of d'Altan, Meyer, and Wieringa ([6]). They argue that deontic operators on properties cannot be defined as expressions in a Deontic Action Logic, but instead provide a mixed modal-dynamic system, which includes both the standard modal operators such as necessity $\Box$ and possibility $\Diamond$, as well as the action operators. The deontic operators on properties are defined with modal operators and the violation property, following Anderson and Moore ([1]). The deontic operators on actions are defined along the lines of Meyer ([17]) which is similar to Khosla and Maibaum ([13]). We claim that the alethic component is not necessary, but as we have shown, a Deontic Action Logic is enough to can capture the essential interpretation of deontic operators on properties. In the following section, we discuss some of the formal aspects of their analysis, starting with their discussion of different ways to reduce ought-to-be to ought-to-do, followed by a presentation of their analysis, and finish with some discussion.

4.1 Discussion of Potential Reductions

d'Altan, Meyer, and Wieringa ([6]) discuss four different attempts to reduce ought-to-be statements to ought-to-do. The formalizations are given in terms of state and action deontic operators, which we have indicated with ObligatedState and ObligatedAction. We have propositions ϕ and action expressions $[\alpha]\phi$ as before. Expressions such as $\text{ObligatedAction}(\alpha)$ are also understood as before except that d'Altan, Meyer, and Wieringa ([6]) do not include the agent in the representation. The target is the interpretation of $\text{ObligatedState}(\alpha)$.

The various proposed reductions are as follows, where the reduction is first given informally and then formally. In each, the attempt is to reduce the obligation on a state to some expression in a Deontic Action Logic.

First Reduction

- A property ϕ of a state-of-affairs is obligatory iff the property is a result of an obligatory action.
- $\text{ObligatedState}(\phi) =_{def}$ there is an action α such that $[\alpha]\phi \wedge \text{ObligatedAction}(\alpha)$

Second Reduction

- A property ϕ of a state-of-affairs is obligatory iff all actions that lead to the state of affairs ϕ are obligatory.
- $\text{ObligatedState}(\phi) =_{def}$ for all actions α, $[\alpha]\phi \rightarrow \text{ObligatedAction}(\alpha)$

Third Reduction

- A property ϕ of a state-of-affairs is obligatory iff it is prohibited from undoing it.
- $\text{ObligatedState}(\phi) =_{def}$ for all actions α, $[\alpha]\neg\phi \rightarrow \text{ProhibitedAction}(\alpha)$

Fourth Reduction

- A property ϕ of a state-of-affairs is obligatory iff all actions which result in ϕ are obligatory as well,
- $\text{ObligatedState}(\phi) =_{def}$ for all actions α, $\alpha \rightsquigarrow \phi \rightarrow \text{ObligatedAction}(\alpha)$, where $\alpha \rightsquigarrow \phi$ means doing α results in ϕ and not doing α results in $\neg\phi$

d'Altan, Meyer, and Wieringa ([6]) argue that each of these attempted reductions fail, and so motivate their analysis of $\text{ObligatedState}(\phi)$ in an Alethic Logic with a violation atom.

We see common themes in the first, second, and fourth reductions, namely that the obligatoriness of the property holds with respect to obligatory actions which result in states where the property holds. For example, following the first reduction, if it is obligatory to clean the house, which results in the house being clean, then it is obligatory for the house to be clean. These definitions arise from an attempt to define obligated-to-be entirely in terms of obligated-to-do, rather than, for example, defining obligated-to-be in terms of deontically specified actions along with additional properties. The only definition which breaks this pattern is the third reduction, which defines obligated-to-be in terms of prohibition on an action. d'Altan, Meyer, and Wieringa ([6]) have a range of objections against each of these proposals; we largely agree with the thrust of their comments about the first, second, and fourth. However, we do not agree with them concerning the third proposal, and accept the principle intuition it represents, so we discuss it further.

Discussion of the Third Reduction. The third analysis expresses a negative relation between the obligatory property and actions. This is the *classical* view that a property is obligatory if it is prohibited from undoing. They report no counterexample in the left to right direction of the definition, and we agree. On the other hand, in the right to left direction, the problem is that if there are no actions which lead to $\neg\phi$, then the conditional holds, and therefore the property is obligatory. This seems unreasonable, so they dismiss this analysis.

However, we do not accept their view of this case. Consider again our simple example, where it is obligatory that the yard be clean. In the case where the yard is clean, the obligation is satisfied, and in addition, one is prohibited from doing an action which would induce a state in which the yard is not clean. Alternatively, if the yard is not clean from the moment the obligation is incurred, then a violation is introduced (this is our case of starting on the wrong foot); it is not relevant whether the performance of some action has resulted in the property not holding. While it may seem unduly harsh to introduce a violation from this point, it need not be, for the implications of the violation depend on defining the consequences, which have not yet been given. For instance, it is possible to define such a violation in such a circumstance so it has no significant consequences. This is where a polynormative analysis is better than a binormative analysis, for violations can be fine-grained and support subtle implications. However, d'Altan, Meyer, and Wieringa ([6]) do not consider this interpretation, reject the third reduction, and propose a combined modal-deontic analysis.

4.2 A Sketch and Discussion of the Formal Analysis

Having rejected all potential previous proposals in which ought-to-be is defined in terms of ought-to-do, D'Altan, Meyer, and Wieringa ([6]) make a combined proposal, which we sketch here. They assume the reduction in Anderson and Moore ([1]) of deontic operators on properties to alethic modal logic plus a designated violation proposition; this is alethic modal logic of type S5, with the modality $\Box$, read as *necessarily,* and a designated *Violation* atomic proposition. D'Altan, Meyer, and Wieringa ([6]) assume the Deontic Action Logic of Meyer ([17]). In the following, we give just the axioms for $\text{ObligatedState}(\phi)$ and $\Box\alpha$. They call this system Prepositional Deontic Logic with the Anderson and Meyer's reductions.

Axiom 1 (ObligatedState). $\text{ObligatedState}(\phi) \leftrightarrow \Box(\neg\phi \rightarrow \text{Violation})$

Axiom 2 ($\Box\alpha$). $\Box\phi \rightarrow [\alpha]\phi$

Notice that should $\text{ObligatedState}(\phi)$ hold, $\Box(\neg\phi \rightarrow \text{Violation})$ holds by definition. By axiom $\Box\alpha$, this implies that $[\alpha](\neg\phi \rightarrow \text{Violation})$; that is, we have the following theorem:

Theorem 3. *$\text{ObligatedState}(\phi) \rightarrow [\alpha](\neg\phi \rightarrow \text{Violation})$.*

This says that if ϕ is obligatory, then any action α, were it performed, results in a state such that if $\neg\phi$, then Violation holds. So, if the action results in $\neg\phi$, Violation also holds. It must be the case that the expression $[\alpha](\neg\phi \rightarrow \text{Violation})$ is vacuously true in the initial state. This theorem corresponds to the first portion of our analysis and prohibitions to undo properties. The axiom **ObligatedState** implies that should the property ϕ not hold in the initial state, then there is a violation, which corresponds to the second portion of our analysis.

We see, then, that this analysis yields a similar logic, but requires the introduction of the alethic component with the violation proposition. In our analysis, this is not necessary.

Two further objections can be raised, the first relatively minor, but the other a more conceptual issue. As discussed earlier, d'Altan, Meyer, and Wieringa ([6]) suppose that obligated-to-be expressions need not be specified with respect to an agent, while actions must have agents. We can have the following expression,

Example 25. $\text{ObligatedState}(\phi) \rightarrow [\text{A}, \alpha](\neg\phi \rightarrow \text{Violation})$.

However, it is hard to see is is useful in contracting, for it essentially says that an obligation on a state implies that for every agent and every action, should they perform the action which results in a state with $\neg\phi$ then induces the violation. For our purposes, one agent may have the obligation to maintain a state, while another not, and it ought to be the case that only actions of the bearer of the obligation are relevant. We want to explicitly relate a particular agent's obligations, whether to properties or to actions, and the sanctions the agent suffers.

The second is more important. While d'Altan, Meyer, and Wieringa ([6, :112]) have a system which has both property and action deontic operators in

an integrated system, they point out that "...no specific relations between them are assumed other than those that follow immediately from both reductions to alethic modal logic." In other words, the operators are not intrinsically logically related (d'Altan, Meyer, and Wieringa [6, :112]): "In this sense, the relation between ought-to-be and ought-to-do remains rather extensional." This, we believe, is a deep conceptual flaw. In other words, according to them, it is but 'incidental' or 'arbitrary' that the two notions are so similar in form, interpretation, and use of violations. Others might use just such similarity to argue for some underlying relation between them, preferably deriving one from the other, or at least expressing them both in terms of similar basic notions, as we have. Our analysis accounts for the similarity of the operators better than d'Altan, Meyer, and Wieringa ([6]) without needing to introduce the alethic component.

5 Conclusion

We have provided an analysis of stative obligations expressed in a Deontic Action Logic. We have a simpler and more uniform way to express any sort of deontic expression, whether an action or a property. We have shown that this analysis can represent detailed and intuitively plausible logical representations, particularly as the obligation is maintained over time. Some topics which could not be covered here relate to the fulfillment and end of an obligation on a property as well as the relation of obligations on properties to permissions and prohibitions on properties.

References

1. Anderson, A., Moore, O.: The Formal Analysis of Normative Concepts. The American Sociological Review. **22** (1957) 9-17
2. Broersen, J.: Action Negation and Alternative Reductions for Dynamic Deontic Logic. To appear Journal of Applied Logic. (2004)
3. Broersen, J. and van der Torre, L.: Review of J. Horty *Agency and Deontic Logic.* Artificial Intelligence and Law **11** 45-61, (2003)
4. Carmo, J., Jones, A.: Deontic Logic and Contrary-to-duties. In D. Gabbay and Franz Guenthner (eds.) Handbook of Philosophical Logic, Dordrecht: Kluwer Academic Publishers, (2001)
5. Carmo, J., Jones, A.: Deontic Database Constraints, Violation, and Recovery. Studia Logica. **57** (1996) 139-165
6. d'Altan, P., Meyer, J.-J.Ch., Wieringa, M.: An integrated framework for ought–to–be and ought–to–do constraints. Artificial Intelligence and Law. **4** (1996) 77–111
7. Forrester, J.: Being good and being logical – Philosophical groundwork for a New Deontic Logic. M.E.Sharpe, Armonk, New York, (1996)
8. Harel, D., Kozen, D., and Tiuryn, J.: Dynamic Logic. Cambridge, MA: The MIT Press (2000)
9. Hilpinen, R.: On Action and Agency. In E. Ejerhed and S. Lindström (eds.) Logic, Action and Cognition – Essays in Philosophical Logic. Kluwer Academic Press, Dordrecht, (1997), 3-27
10. Horty, J.: Agency and Deontic Logic. Oxford University Press, Oxford, (2001)

11. Jones, A., Sergot, M.: On the Characterisation of Law and Computer Systems: the Normative Systems Perspective. In J.-J.Ch Meyer and R.J. Wieringa (eds.) Deontic Logic in Computer Science – Normative System Specification. Wiley (1993), 275-307
12. Katz, G.: Anti Neo-Davidsonianism: Against a Davidsonian Semantics for State Sentences. In E. Lang, C. Maienborn, and C. Fabricius-Hansen (Eds.), Events as Grammatical Objects (2000), Stanford, CA: CSLI Publications, 393-416
13. Khosla, S., Maibaum, T.: The Prescription and Description of State-Based Systems. In B. Banieqbal, H. Barringer, and A. Pneuli (eds.) Temporal Logic in Specification. Springer-Verlag (1987) 243-294
14. Krogh, C., Herrestad, H.: Getting Personal: Some Notes on the Relationship Between Personal and Impersonal Obligation. In M. Brown and J. Carmo (eds.) DEON: Deontic Logic, Agency and Normative Systems, DEON '96. Springer-Verlag (1996) 134-153
15. Maienborn, C.: Against a Davidsonian Analysis of Copula Sentence. In M. Kadowski and S. Kawahara (eds.) NELS 33 Proceedings. Amherst: GLSA
16. Meyden, R. v. d.: The Dynamic Logic of Permission. Journal of Logic and Computation. **6** (1996) 465-479
17. Meyer, J.-J.Ch.: A Different Approach to Deontic Logic: Deontic Logic Viewed as a Variant of Dynamic Logic. Notre Dame Journal of Formal Logic. **1** (1988) 109-136
18. Meyer, J.-J.Ch., Wieringa, R.J.: Actors, Actions, and Initiative in Normative System Specification. Annals of Mathematics and Artificial Intelligence. **7** (1993) 289-346
19. Parsons, T.: Events in the Semantics of English: A Study in Subatomic Semantics. Cambridge, MA: The MIT Press (1990)
20. Sergot, M.: A Brief Introduction to Logic Programming and its Applications in Law. In C. Walter (ed.) Computer Power and Legal Language. Quorum Books (1988) 25-39
21. Sergot, M.: The Representation of Law in Computer Computer Programs. In T.J.M. Bench-Capon (ed.) Knowledge-Based Systems and Legal Applications. Academic Press (1991) 3-67
22. Sergot, M. and Richards, R.: On the Representation of Action and Agency in the Theory of Normative Positions. Fundamenta Informaticae. **48** (2001) 273-293
23. Sergot, M.: Normative Positions. In Henry Prakken and Paul McNamara (eds.) Norms, Logics and Information Systems. New Studies in Deontic Logic and Computer Science. IOS Press (1998) 289-310
24. Sergot, M.: A Computational Theory of Normative Positions. ACM Transactions on Computational Logic. **2** (2001) 581-622
25. Wieringa, R.J., Meyer, J.: Deontic Logic in Computer Science: Normative System Specification. John Wiley and Sons (1993)
26. Wyner, A.Z.: Boolean Event Lattices and Thematic Roles in the Syntax and Semantics of Adverbial Modification. Cornell University Ph.D. Thesis (1994)
27. Wyner, A.Z.: Transfer Report. ms King's College London, Department of Computer Science, url = www.dcs.kcl.ac.uk/pg/wyner
28. Wyner, A.Z.: Informal Contract Specification. ms King's College London, Department of Computer Science, url = www.dcs.kcl.ac.uk/pg/wyner

Author Index