Ron van der Meyden
Leendert van der Torre (Eds.)

LNAI 5076

Deontic Logic in Computer Science

9th International Conference, DEON 2008
Luxembourg, Luxembourg, July 2008
Proceedings

ΔEON 2008

Springer

Lecture Notes in Artificial Intelligence 5076

Edited by R. Goebel, J. Siekmann, and W. Wahlster

Subseries of Lecture Notes in Computer Science

Ron van der Meyden
Leendert van der Torre (Eds.)

Deontic Logic in Computer Science

9th International Conference, DEON 2008
Luxembourg, Luxembourg, July 15-18, 2008
Proceedings

Springer

Volume Editors

Ron van der Meyden
The University of New South Wales
School of Computer Science and Engineering
Sydney 2052, Australia
E-mail: meyden@cse.unsw.edu.au

Leendert van der Torre
University of Luxembourg
Faculty of Sciences, Technology and Communication (FSTC)
Computer Science and Communications (CSC)
Individual and Collective Reasoning
6, Rue Richard Coudenhove - Kalergi, 1359 Luxembourg, Luxembourg
E-mail: leendert@vandertorre.com

Library of Congress Control Number: 2008930268

CR Subject Classification (1998): I.2.3, I.2.11, I.2, F.4.1

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743
ISBN-10 3-540-70524-4 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-70524-6 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springer.com

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12436781 06/3180 5 4 3 2 1 0

Preface

This volume presents the papers contributed to ΔEON 2008, the 9th International Conference on Deontic Logic in Computer Science, held in Luxembourg, July 16–18, 2008. This biennial conference series is designed to promote international cooperation amongst scholars who are interested in deontic logic and its use in computer science. The scope of the conference is interdisciplinary, and includes research that links the formal-logical study of normative concepts and normative systems with computer science, artificial intelligence, philosophy, organization theory, and law. The ΔEON website, http://www.deonticlogic.org, contains links to previous conferences and their papers. This history reveals a vibrant interdisciplinary research program.

Papers for these conferences might address such general themes as the development of formal systems of deontic logic and related areas of logic, such as logics of action and agency, or the formal analysis of all sorts of normative concepts, such as the notions of rule, role, regulation, authority, power, rights, responsibility, etc., or the formal representation of legal knowledge. They might also be concerned with applications, such as the formal specification of normative multiagent systems, the specification of systems for the management of bureaucratic processes in public or private administration, or the specification of database integrity constraints or computer security protocols, and more. Of particular interest is the interaction between computer systems and their users.

In addition to these general themes, the 2008 meeting focused also on the special topic of logical approaches to deontic notions in computer science in the area of security and trust, encompassing applications in e-commerce as well as traditional areas of computer security. Topics of interest in this special theme encompass digital rights management, electronic contracts, including service level agreements and digital media licenses, authorization, access control, security policies, privacy policies, business processes and regulatory compliance. The special theme embraced both theoretical work (formal models, representations, specifications, logics, verification) and implementation-oriented work (architectures, programming languages, design models, simulations, prototype systems) on these specific topics.

The 16 papers printed here were selected for presentation at the conference after a thorough process of review and revision of 28 submitted papers. All are original and presented here for the first time. The titles themselves demonstrate commitment to the themes of the conference. In addition to these peer-reviewed papers, we present abstracts or papers of the talks of our four invited speakers, Martin Abadi (UC Santa Cruz and Microsoft Research, USA), Ross Anderson (University of Cambridge, UK), Nuel Belnap (University of Pittsburgh, USA), and Dov Gabbay (King's College London, UK).

We are grateful to all who contributed to the success of the conference, to our invited speakers, to all the authors of the presented papers, and to all who participated in discussion. Special thanks go to the members of the Program Committee for their service in reviewing papers and advising us on the program and to the members of the Organization Committee for taking care of all the countless details that a conference like this requires, especially Gabriella Pigozzi and Martin Caminada for all local arrangements of the conference, Mathijs de Boer for setting up and maintaining the DEON 2006 website and Davide Grossi for setting up deonticlogic.org. Thanks too to Richard van de Stadt, whose CyberChairPRO system was a very great help to us in organizing the papers from their initial submission to their final publication in this volume.

The previous edition of the ΔEON conference in Utrecht had as its special topic artificial normative systems, their theory, specification and implementation, such as electronic institutions, norm-regulated multiagent systems and artificial agent societies generally. Here too the concern is both with theoretical work, such as the design of formal models and representations, and also work more oriented toward implementation, such as architectures, programming languages, design models, simulations, etc. For the first time, the ΔEON conference in Luxembourg was co-located with a workshop on normative multiagent systems. Thanks to Guido Boella (Università di Torino), Gabriella Pigozzi (University of Luxembourg), Munindar P. Singh (North Carolina State University) and Harko Verhagen (Royal Institute of Technology and Stockholm University) for organizing the NORMAS 2008 workshop in Luxembourg.

We are also very grateful to the Fonds National de Recherche du Luxembourg (FNR) and strategic priority P1 on security and reliability of the University of Luxembourg for their essential financial support. Finally, we wish to express our appreciation to Springer for publishing these proceedings in their LNCS/LNAI series. This is the third such volume in this series. We hope these volumes may continue into the future to provide a record of research in this rich and growing field.

April 2008

Ron van der Meyden
Leendert van der Torre

Conference Organization

Organization Committee

General Co-chairs

Ron van der Meyden	The University of New South Wales
Leendert van der Torre	University of Luxembourg

Local Organization Co-chairs

Martin Caminada	University of Luxembourg
Gabriella Pigozzi	University of Luxembourg

Local Committee

Patrizio Barbini	University of Luxembourg
Mathijs de Boer	University of Luxembourg
Davide Grossi	University of Luxembourg
Marija Slavkovik	University of Luxembourg
Yining Wu	University of Luxembourg

Program Committee

Co-chairs

Ron van der Meyden	The University of New South Wales
Leendert van der Torre	University of Luxembourg

Members

Paul Bartha	University of British Columbia
David Basin	ETH Zürich
Guido Boella	University of Turin
Jan Broersen	Universiteit Utrecht
Mark Brown	Syracuse University
José Carmo	University of Madeira
Frédéric Cuppens	ENST-Bretagne Rennes
Robert Demolombe	IRIT Toulouse
Frank Dignum	Universiteit Utrecht
Lou Goble	Willamette University
Carl A. Gunter	University of Illinois
Joerg Hansen	University of Leipzig
Risto Hilpinen	University of Miami
John Horty	University of Maryland

Andrew Jones	King's College London
Ninghui Li	Purdue University
Lars Lindahl	University of Lund
Alessio Lomuscio	Imperial College London
Heiko Ludwig	IBM T.J. Watson Research Center
Paul McNamara	University of New Hampshire
John-Jules Meyer	Universiteit Utrecht
John Mitchell	Stanford University
Rohit Parikh	City University of New York
Adrian Paschke	Technical University Dresden
Henry Prakken	Universiteit Utrecht / University of Groningen
Babak Sadighi Firozabadi	Swedish Institute of Computer Science
Filipe Santos	ISCTE Portugal
Giovanni Sartor	University of Bologna
Marek Sergot	Imperial College London
Carles Sierra	IIIA-CSIC
Yao-Hua Tan	Vrije Universiteit Amsterdam
Vicky Weissman	Cornell University

with the assistance of:

Matteo Baldoni
Patrizio Barbini
Francesco Belardinelli
Björn Bjurling
Pilar Dellunde
Jürgen Doser
Marc Esteva
Eduardo Fermé
Davide Grossi
Adam J. Lee
Samuel Müller
Lars E. Olson
Tomas Olsson
Olga Pacheco
David Pearce
Livio Robaldo
Benedikt Schmidt
Paolo Turrini

Sponsors

FNRL (Fonds National de la Recherche Luxembourg)
P1 Security & Reliability
Faculty of Sciences, Technology and Communication, University of Luxembourg

Table of Contents

Norms in Branching Space-Times

Nuel Belnap

University of Pittsburgh

The idea of norms presupposes agency, and agency presupposes an indeterministic causal order (so that "ought" does not imply "is"). So much can be modeled in "branching time with agents and choices" (BTAC). The seriously ontological independence of agentive choices, however, requires, as a necessary condition, a causal order permitting space-like separation of those choices in a sense definable in "branching space-times with agents and choices" (BSTAC). Let us idealize an agent, when restricted to a single space-time, as a kind of spatio-temporal "worm" in the familiar way, representing the life of the agent in that space-time. Then a representation of "the agent," since it must include representation of seriously objective choices, must look like a tree with two kinds of branching. In both kinds of branching, there is a single past-pointing worm-like representation of the past-life of the agent up to the branching, and an entire assemblage of distinct worm-like representations of the possible future-life of the agent subsequent to the branching, one for each history in which the life of the agent continues. The first kind of branching occurs at choice-points for the agent. According to BSTAC, such branching will involve a last point of agent's-choice-not-yet-made (say, a last point of deliberation), but no first point of agent's-choice-has-been-made in any possible future-life of the agent. In the second kind of branching, the agent is passive, having two or more possible future-lives due to space-like-related choices by other agents, or by metaphorical "choices" by some space-like-related element of Nature. In this case, BSTAC says that there will be no last point of the past-life of the agent, but instead a first point for each of the agent's possible future-lives.

This representation of agency presupposes continuity of each space-time. How best to represent agents-in-branching-space-times discretely, as is perhaps required for computer applications, is open for research. But just as computer representations of real arithmetic must in some way answer to "real" real arithmetic, so any such discrete representation must in the end answer to *Our World* as a BSTAC.

BSTACN (BSTAC with norms) postulates *generated* norms, that is, norms that are generated by a particular localized act of an agent. This might be the making of a promise, the laying on of an obligation, the issuing of an invitation, etc. Say that the norm has been *issued*. Such an act need not be a *speech* act, but it is technically convenient to suppose that each norm is issued as if by the use of a declarative core in direct speech. For example, at a point-event e_0, Jack promises Sarah with the following words. *I promise you as follows*:

If it rains in Chicago before x, I will see to it that I pay you \$5 before x, (1)

R. van der Meyden and L. van der Torre (Eds.): DEON 2008, LNAI 5076, pp. 1–2, 2008.

(where "x" names an event-type such that in each history to which e_0 belongs, there is a single occurrence of that type in the causal future of e_0). (1) is called the *declarative core* of the promise-event occurring at e_0.

The promise has an interesting semantics that seems to *require* indeterministic BSTACN as background. It seems essential that at the point event e_0, (1) is neither settled true nor settled false. (Don't say that it is neither true nor false; if you do, you will miss the point.) To say that the promise made at e_0 has been carried out at a certain later point event e_1 means (roughly) that the conditional (1) is true with respect to every point-history pair e_0/h, for every history h to which (not e_0) but e_1 belongs, and relative to speaker = Jack, auditor = Sarah, and point of utterance = e_0. In other words, at e_1 it is settled true that (1) was true (not settled true) at the point, e_0, at which the promise was issued. This recipe invokes what *Facing the future* calls a "double time reference," and which MacFarlane describes in branching time by saying that (1) is "true" with respect to moment of utterance e_0 and moment of assessment, e_1.

A description of the norm involved requires more than just double time references. A satisfying—though hardly unique—representation of the normative content of the promising is as a *strategy* in a world of branching space-times. Suppose the promissor has arrived at a point event e_1. If at e_1 it is settled that (1) was true at e_0, the norm has been satisfied, and there is nothing more to do. If, however, it is settled that (1) was false at e_0, then the norm has been violated, and appropriate sanctions are due. If whether (1) was true at e_0 is open (historically contingent), then the norm calls for the promissor to *do* something, at the very least to make a choice that keeps the satisfaction of the promise possible, and, if possible, to choose so as to guarantee satisfaction of the promise. (This is a "world-to-words" fit.)

Other sorts of norms call for related descriptions in branching space-times. The apparatus also suggests consideration of *joint agency* and *message passing.* These themes, common in computer science, take on an interesting flavor when put against the background of branching space-times.

Changing Legal Systems: Abrogation and Annulment Part I: Revision of Defeasible Theories

Guido Governatori[1] and Antonino Rotolo[2,⋆]

[1] School of ITEE, The University of Queensland, Australia
guido@itee.uq.edu.au
[2] CIRSFID/Law School, University of Bologna, Italy
antonino.rotolo@unibo.it

Abstract. In this paper we investigate how to model legal abrogation and annulment in Defeasible Logic. We examine some options that embed in this setting, and similar rule-based systems, ideas from belief and base revision. In both cases, our conclusion is negative, which suggests to adopt a different logical model.

1 Introduction

Mainly inspired by [1], most formal models of norm change usually focus on the dynamics of obligations and permissions. However, as rightly noted on the occasion of a recent workshop on this topic[1], "these systems did not explicitly refer to possible changes in the underlying norms [...]". In fact, "new norms may be created and old norms may need to be retracted. In this dynamic setting, it is essential to distinguish norms from obligations and permissions as studied by deontic logic, to understand the formal properties specific for the dynamics of norms, and to describe how such objects can be manipulated [...]". Unfortunately, "a formal model that captures the relevant features of norm change is still lacking".

The aim of our work is to make some steps in this direction by investigating the notion of legal modification. Legal modifications are the ways through which the law implements norm dynamics [10]. Modifications can be either explicit or implicit. In the first case, the law introduces norms whose peculiar objective is to change the system by specifying what and how other existing norms should be modified. In the second case, the legal system is revised by introducing new norms which are not specifically meant to modify previous norms, but which change in fact the system because they are incompatible with such existing norms. The most interesting case is when we deal with explicit modifications, which permit to classify a large number of modification types.

In general, we have different types of modifying norms, as their effects (the resulting modifications) may concern, for example, the text of legal provisions, their scope, or their time of force, efficacy, or applicability [10,8,9]. Derogation is an example of scope change: a norm n supporting a conclusion P and holding at the national level may be derogated by a norm n' supporting a different conclusion P' within a regional context.

⋆ Supported by the EU project ESTRELLA (IST-2004-027655).

[1] http://icr.uni.lu/normchange07/

R. van der Meyden and L. van der Torre (Eds.): DEON 2008, LNAI 5076, pp. 3–18, 2008.

Hence, derogation corresponds to introducing one or more exceptions to n. Temporal changes impact on the target norm in regard to its date of force (the time when the norm is "usable"), date of effectiveness (when the norm in fact produces its legal effects) or date of application (when conditions of norm applicability hold). An example of change impacting on time of force is when a norm n is originally in force in 2007 but a modification postpones n to 2008. Substitution replaces some textual components of a provision with other components. For example, some of its applicability conditions are replaced by other conditions.

We are interested here in studying the concepts of *abrogation* and *annulment*.

Annulment is usually seen as a kind of repeal, as it makes a norm invalid and removes it from the legal system. Its peculiar effect applies *ex tunc*: annuled norms are prevented to produce all their legal effects, independently of when they are obtained. The nature of *abrogation* is most controversial. In some cases, it is important to see whether the abrogation is the result of judicial review, legislation, or referenda. But again, despite domestic peculiarities, abrogations, too, are seen as a type of norm removal, even though they are different from annulments; the main point is usually that abrogations operate *ex nunc* and so do not cancel the effects that were obtained before the modification. If so, it seems that abrogations cannot operate retroactively. However, this is not always true. Even where retroactive abrogations are prohibited (such as in the Italian system), the problem is open in some contexts. Suppose an ordinary court is called upon to decide a case in which a norm n applies, but the court argues that n infringes some fundamental rights and so it suspends the trial proceedings referring to the constitutional court to decide on the illegitimacy and abrogation of n. Constitutional court's decision and abrogation of n is necessarily posterior to the case. Hence, what is the difference between these modifications?

Suppose that a norm n_1 in force in 2006 states that, if your annual income is less than 5,000 euros, you are a needy person and norm n_2 says that a needy person has the right to live for free in a council house. If n is retroactively *annuled* in 2007, this counts as n's removal since 2006, and all its effects are blocked. Imagine now that two norms n_3 and n_4 are added in 2007 stating that needy people's income is less than 3,000 euros and that needy people are eligible for medical aid. Even if n is retroactively *abrogated* in 2007, jurists may argue that its indirect effect (obtained via n_2: right to house) should not be estinguished in 2007, whereas the propagation of the qualification "needy person" (with an income of less than 5,000 euros) cannot propagate from 2006 to 2007, since this would make n_4 applicable. Note that, in other cases, indirect effects should propagate whereas the direct effect should be blocked, or all past effects should propagate, or, again, norm removal should apply after in 2007 and only blocking some effects retroactively holds. In fact, jurists [10] say that abrogations can at most block some, but not *all*, past effects (otherwise, we would have annulments).

To sum up, and independently of terminological issues, what we have to bear in mind is that here the law implements different reasoning patterns: in one case norms are removed with all their effects, whereas in other cases norms are removed but some or all their effects propagate if obtained before the modification.

How to model these scenarios? Clearly, a temporal representation may help, but the point is whether we can abstract from this aspect and move to a general analysis (e.g., based on theory revision) where time is not considered. We address this issue using Defeasible Logic (DL) [12,2], but analogous considerations can be extended to other nonmonotonic (sceptical) rule-based systems. Although other options are available, rule-based systems seem a natural way to represent legal systems: legal norms are usually viewed as rules specifying some applicability conditions and a legal effect.

In this paper we discuss whether it is possible to adjust belief and theory revision in DL to capture abrogation and annulment. The layout is as follows. Section 2 provides an overview of DL. Section 3 considers an immediate method to adjust revision of belief sets in DL in order to capture annulment. Section 4 examines a possible alternative in which all operations, including contraction, are captured by only adding a suitable set of new rules. Even though this second option is better for modelling abrogation and annulment, some basic problems remain unsolved. Section 5 takes advantage of some ideas from the previous section and discusses how base revision in DL can be applied to capture norm removals. However, also this approach is not fully satisfactory, which suggests to adopt a different conceptual model, whose general features are illustrated in Section 6. This is the new model we have used for our initial investigation on modelling norm changes in DL [8,9].

2 Overview of Defeasible Logic

DL is based on a logic programming-like language and it is a simple, efficient but flexible non-monotonic formalism capable of dealing with many different intuitions of non-monotonic reasoning. An argumentation semantics exists [7]. DL has a linear complexity [11] and also has several efficient implementations [3]. In addition, some preliminary works on legal modifications in DL have been recently proposed [8,9].

A *defeasible theory* D is a structure $(F,R,\succ)$ where F is a finite set of facts, R a finite set of rules, and $\succ$ an acyclic superiority relation on R. *Facts* are represented as literals and are indisputable statements. A *rule* expresses a relationship between a set of premises and a conclusion. We have in DL three types of rules conveying the strength of the relationships: strict rules, defeasible rules and defeaters. A *strict* rule has the form $A_1,\dots,A_n \rightarrow B$ and states the strongest kind of relationship since its conclusion always holds when the premises are indisputable. *Defeasible* rules have the form $A_1,\dots,A_n \Rightarrow B$ and cover the case when the conclusion normally holds when the premises tentatively hold; *defeaters* have the form $A_1,\dots,A_n \rightsquigarrow B$ and consider a situation where the premises do not warrant the conclusions: in defeaters the premises simply prevent another rule to support the opposite.

Accordingly, a conclusion can be labelled either as definite or defeasible. A definite conclusion is an indisputable conclusion, while a defeasible conclusion can be retracted if additional premises become available. DL is based on a constructive proof theory for conclusions. Hence, we can say that a derivation for a conclusion exists and that it is not possible to give a derivation for a conclusion. Based on these two ideas conclusions will be tagged according to their strength and type of derivation:

- $+\Delta B$, meaning that we have a definite proof for B (a definite proof is a proof where we use only facts and strict rules);
- $-\Delta B$, meaning that it is not possible to build a definite proof for B;

- $+\partial B$, meaning that we have a defeasible proof for B;
- $-\partial B$, meaning that it is not possible to give a defeasible proof for B.

Provability is based on the concept of a *derivation* (or proof) in $D = (F, R, \succ)$. A derivation is a finite sequence $P = (P(1), \ldots, P(n))$ of tagged literals satisfying four conditions (which correspond to inference rules for each of the four kinds of conclusion). $P(1..i)$ denotes the initial part of the sequence P of length i.

Proof conditions for strict derivations are here omitted. Strict proofs are just derivations based on detachment for strict rules. Given a strict rule $A_1, \ldots, A_n \rightarrow B$, where we have definite proofs for all A_i's, we can deduce B ($+\Delta B$).

DL is a sceptical non-monotonic formalism: with a possible conflict between two conclusions (i.e., one is the negation of the other), DL refrains to take a decision and we deem both as not provable unless we have some more pieces of information that can be used to solve the conflict. One way to solve conflicts is to use a superiority relation over rules. The superiority relation gives us a preference over rules with conflicting conclusions. In case we have a conflict between two rules we prefer the conclusion of the strongest of the two rules. The superiority relation is applied in defeasible proofs.

Some notational conventions before presenting proof conditions for defeasible derivations. Each rule is identified by a unique label. $A(r)$ denotes the set of antecedents of a rule r, while $C(r)$ denotes its consequent. If R is a set of rules, R_s is the set of all strict rules in R, R_{sd} the set in R of strict and defeasible rules, R_d the set of defeasible rules, and R_{dft} the set of defeaters. $R[B]$ denotes the set of rules in R with consequent B. If B is a literal, $\sim B$ denotes the complementary literal (if B is a positive literal C then $\sim B$ is $\neg C$; and if B is $\neg C$, then $\sim B$ is C).

Defeasible proofs proceed in three phases: we first look for an argument supporting the conclusion we want to prove (an applicable rule for the conclusion). Second, we look for arguments for the opposite of what we want to prove. Third, we rebut the counterarguments. This can be done by showing that the counterargument is not founded (i.e., some of the premises do not hold), or by defeating the counterargument, i.e., the counterargument is weaker than an argument for the conclusion we want to prove. Formally,

$+\partial$: If $P(i+1) = +\partial B$ then either
(1) $+\Delta B \in P(1..i)$ or
(2.1) $\exists r \in R_{sd}[B] \forall A \in A(r) : +\partial A \in P(1..i)$ and
(2.2) $-\Delta{\sim}B \in P(1..i)$ and
(2.3) $\forall s \in R[{\sim}B]$ either
  (2.3.1) $\exists A \in A(s) : -\partial A \in P(1..i)$ or
  (2.3.2) $\exists t \in R_{sd}[B]$ such that
    $\forall A \in A(t) : +\partial A \in P(1..i)$ and $t \succ s$.

$-\partial$: If $P(i+1) = -\partial B$ then
(1) $-\Delta B \in P(1..i)$ and
(2.1) $\forall r \in R_{sd}[B]\ \exists A \in A(r) : -\partial A \in P(1..i)$ or
(2.2) $+\Delta{\sim}B \in P(1..i)$ or
(2.3) $\exists s \in R[{\sim}B]$ such that
  (2.3.1) $\forall A \in A(s) : +\partial A \in P(1..i)$ and
  (2.3.2) $\forall t \in R_{sd}[B]$ either
    $\exists A \in A(t) : -\partial A \in P(1..i)$ or $t \not\succ s$.

3 Revising Extensions of Normative Systems

In the remainder of this paper we address the problem of how to embed in DL some ideas from belief and base revision in order to capture annulment and abrogation. We attack two different problems raised by these modifications: (i) how to block either some or all norm effects; (ii) how to model norm removals in legal systems. As we argued, even though such modifications have a temporal flavour, we move to a general analysis where time is not considered.

We assume that a defasible theory can represent the basic logical structure of a legal system [8,9]. It is a general tenet in the literature that one reason why legal reasoning is defeasible depends on the fact that, in may cases, norm conclusions can be obtained only if we do not have stronger norms attacking them [14]. DL theories consist of a set of rules (which may be defeasible), a set of facts, and a set of priorities over rules (which establish their relative strength). In this perspective, rules naturally correspond to legal norms, while priorities represent the criteria used to solve legal conflicts. Hence, a general picture like this provides a standard for capturing the basics of legal systems [13]. With this said, let us begin with our discussion on annulment and abrogation.

Approaches based on AGM usually assume that a belief set B is a theory, i.e., a set of formulas closed under a logical consequence relation, thus $B = \mathrm{Cn}(B)$. Let us consider the equivalent of this notion in DL.

Let HB_T be the Herbrand Base for a Defeasible Theory T. In [2], the extension of a Defeasible Theory T is defined as the 4-tuple

$$E(T) = (\Delta^+(T), \Delta^-(T), \partial^+(T), \partial^-(T)),$$

where $\#^{\pm}(T) = \{p | p \in HB_T, T \vdash \pm\# p\}, \# \in \{\Delta, \partial\}$.

Definition 1. *Let $T = (F, R, \succ)$ be a Defeasible Theory. We define another Defeasible Theory $T' = (\emptyset, R', \emptyset)$ such that R' is the smallest set satisfying the following conditions*

- *if $p \in \Delta^+(T)$, then $\to p \in R'$;*
- *if $p \in \partial^+(T)$, then $\Rightarrow p \in R'$;*
- *if $p \notin \Delta^+(T) \cup \Delta^-(T)$, then $p \to p \in R'_s$;*
- *if $p \in \Delta^-(T)$, then $R'_s[p] = \emptyset$;*
- *if $p \in \partial^-(T)$, then $R'_d[p] = \emptyset$;*
- *if $p \notin \partial^+(T) \cup \partial^-(T)$, then $p \Rightarrow p \in R'$.*

We will say that T' is the theory generated by the extension of T.

Proposition 1. *Let T be a defeasible theory. For every $p \in HB_T$, $T \vdash \# \pm p$ iff $T' \vdash \# \pm p$.*

The above result gives us an immediate way to define contraction for revision based on belief sets. We define $T_c^{\ominus} = T'$ such that $E(T) = (\Delta^+(T), \Delta^-(T), \partial^+(T), \partial^-(T))$ and T' is the theory generated by the extension

$$(\Delta^+(T) - \{c\}, \Delta^-(T), \partial^+(T) - \{c\}, \partial^-(T)).$$

It is easy to verify that the above way to define contraction satisfies all AGM postulates. The meaning of the result in Proposition 1 is that for every theory (and so every set of conclusions), we can generate a new equivalent theory without looking at the structure of the original theory: In fact, classically two theories are equivalent if they have the same extension (the same set of conclusions).

How can the procedure described in Definition 1 be used to cover abrogation and annulment?

Let examine *annulment*. When we annul a norm in a legal system, this means that all (direct and indirect) legal effects deriving from it must be cancelled as well. For example, if we have a normative system T containing only the rules $A \Rightarrow B$ and $B \Rightarrow C$, then the annulment of the former rule (assuming the fact A) should block both B and C. Intuition suggests that contraction is the right operation to capture annulment. Hence, the question is how to use contraction in this case. What one could do here is simply to remove the consequent of the rule. However, the (positive defeasible) extension of T (i.e., $\partial^+(T)$) is $\{A,B,C\}$,[2] and contracting B leaves C in the extension. Hence, this immediate use of contraction is not representative of legal annulment. As we said, we have to consider all consequences of the formula to be contracted. In the above example, C can only be derived if B does. Accordingly, annulment of any rule $A_1,\dots,A_n \Rightarrow B$ could be defined as follows. Let $T = (F,R,\succ)$ be a Defeasible Theory. Then

$$\begin{aligned} &T^{\ominus}_{A_1,\dots,A_n\Rightarrow B} = \begin{cases} T & \text{if } A_1,\dots,A_n \Rightarrow B \notin R \text{ or } \{A_1,\dots A_n\} \not\subseteq \partial^+ \\ (F',R',\succ') & \text{otherwise} \end{cases} \\ &\qquad \text{such that} \\ &\quad (F',R',\succ') \text{ is the theory generated by } E(T) - E(T') \\ &\qquad \text{and } T' = (F = \{B\},R,\succ). \end{aligned} \tag{1}$$

The contraction operation reflecting annulment is defined by “removing” the consequent of the rule. In addition, the theory T' generates all consequences of B with respect to T. Then $T^{\ominus}_{A\Rightarrow B}$ is the theory generated by the extension $E(T) - E(T')$. However, let us consider another example.

Example 1. Assume to work with the following theory:

$$T = (F = \{A\}, R = \{A \Rightarrow B,\ B \Rightarrow C,\ A \Rightarrow C\}, \emptyset).$$

Thus,

$$T' = (F = \{B\}, R = \{A \Rightarrow B,\ B \Rightarrow C,\ A \Rightarrow C\}, \emptyset).$$

Hence, $(\partial^+(T) = \{A,B,C\}) - (\partial^+(T') = \{B,C\}) = \{A\}$, and this leads (by applying Definition 1) to obtain that $T^{\ominus}_{A\Rightarrow B}$ corresponds to

$$T'' = (\emptyset, R = \{\Rightarrow A\}, \emptyset).$$

This procedure is not satisfactory unless more sophisticated measures are added. Example 1 shows that the procedure does not properly work, as C has *multiple causes* (B and A): with T'' we exclude $A \Rightarrow B$ by dropping B (and its consequences), but this leads to drop, too, C and so to exclude $A \Rightarrow C$, which is too much.

In addition, the above procedure requires to change the set of facts, which seems to us meaningless. Why cannot we change the set of facts? The facts of a theory are only those pieces of evidence in a case used to *apply* rules (norms) and not to *change* them: hence they should not be considered when one modifies norms. Accordingly, if norms are represented as rules, then reasoning only on the consequences of a theory is not representative of norm change. For example, the norm *HighIncome* $\Rightarrow$ *TopMarginalRate*

[2] From now on, whenever clear from the context, we will use the term ‘extension of a theory’ as either the positive defeasible extension of it or the full extension of the theory (see Definition 1).

says that if the income of a person is in excess of the threshold for high income, then the top marginal rate must be applied. If it is a fact that Nino exceeded the threshold (i.e., $\mathit{HighIncome} \in F$) then he has to pay the top marginal rate. Thus the extension is $\{\mathit{HighIncome}, \mathit{TopMarginalRate}\}$; contracting with *HighIncome* results in the theory just consisting in $\Rightarrow \mathit{TopMarginalRate}$, namely in a rule stating that, no matter what your income is, you will have to pay taxes at the top marginal rate. Thus, revising the evidence on which a case is based results in a change in the legislation, which seems a non-sense when applied to real legal systems.

The idea behind Definition 1 and (1) is that we have to generate a new normative system from the revised extension of corresponding source normative system. However, there are at least three reasons why Definition 1 and (1) do not seem satisfactory:

1. they may change the set of facts, and so do not differentiate between norms and instances of cases;
2. they revise theories regardless of the logical structure of the source theories;
3. they do not correctly account for *ex tunc* modifications, such as annulment.

Changing facts or generating new theories whose structure does not reflect the theories from which they have been obtained trivialise the concept of legal change. Indeed, it is crucial in the law to establish *what rules* generate *which effects*. Therefore, the contraction function defined in this section does not offer a suitable method for modelling annulment (and, in general, norm changes), even if it satisfies all AGM postulates.

4 Revising Normative Systems by Adding Exceptions

The difficulties under points 1 and 2 above can be alleviated by adopting in DL the approach proposed in [4] to deal with belief revision of rule-based non-monotonic formalisms, where change operators are not applied to the set of facts and are all implemented by adding new rules and changing priorities. This permits to incrementally modify the legal system, taking into account the logical structure of the source theory. Let us briefly recall the basic features of this approach.

Let us examine *expansion*. Following [6], expansion adds a formula A to $\partial^+(T)$ *only if* $\neg A \notin \partial^+(T)$. Hence, the case where $\neg A \in \partial^+(T)$ is irrelevant. However AGM decided to also add A in this case. In [4] T is kept unchanged, following [6] rather than [1]. Let $c = P_1, \dots P_n$ be the formulas to be added. Expansion can be defined as follows:

$$
\begin{aligned}
&T^+_c = \begin{cases} T & \text{if } {\sim}P_i \in \partial^+(T) \text{ or } {\sim}P_i = P_j \text{ for some } i, j \in \{1, \dots, n\} \\ (F, R', \succ') & \text{otherwise} \end{cases} \\
&\text{where} \\
&R' = R \cup \{\Rightarrow P_1, \dots, \Rightarrow P_n\} \\
&\succ' = (\succ \cup \{\Rightarrow P_i \succ r \mid i \in \{1, \dots, n\}, r \in R[{\sim}P_i]\}) - \\
&\qquad \{r \succ \Rightarrow P_i \mid i \in \{1, \dots, n\}, r \in R[{\sim}P_i]\}.
\end{aligned}
\tag{2}
$$

Thus, rules are added that prove each of the literals P_i, and it is ensured that these are strictly stronger than any possibly contradicting rules.

Let us examine *contraction*, which seems the right candidate to capture at least some aspects of abrogation and annulment[3]:

$$T_c^- = \begin{cases} T & \text{if } P_1, \ldots, P_n \notin E(T) \\ (F, R', \succ') & \text{otherwise} \end{cases}$$

$$\begin{aligned} &\text{where} \\ &R' = R \cup \{P_1, \ldots, P_{i-1}, P_{i+1}, \ldots, P_n \rightsquigarrow\ \sim P_i \mid i \in \{1, \ldots, n\}\} \\ &\succ' = \ \succ - \{s \succ r \mid r \in R' - R\}. \end{aligned} \tag{3}$$

Intuitively, (3) aims to prevent the proof of all the P_is. To achieve this it is ensured that at least one of the P_is will not be proven. The new rules in R' ensure that if all but one P_i have been proven, a defeater with head $\sim P_j$ will fire. Having made the defeaters not weaker than any other rules, the defeater cannot be "counterattacked" by another rule, and P_j will not be proven, as an inspection of the condition $+\partial$ in Section 2 shows.

This approach slightly deviates from the AGM postulates, in particular from those for contraction. The second AGM postulate states that we contract a formula only by deleting some formulas, but not by adding new ones. This postulate cannot be adopted here because it contradicts the sceptical nonmonotonic nature of DL. To see this, suppose that we know A, and we have rules $\Rightarrow B$ and $A \Rightarrow \neg B$. Then A is sceptically provable and B is not. But if we decide to contract A, B becomes sceptically provable. Note that this behaviour is not confined to DL but holds in any sceptical nonmonotonic formalism [4]. Another peculiarity of this approach is the clear distinction between facts and rules and that facts are indisputable and cannot be changed. Thus, the negation of facts correspond to contradictions, and contracted facts are still included in the extension of the theory.

The advantages of [4]'s proposal are clear, as legal systems are changed by only adding new rules. In this sense, even though it works on theory extensions (suitable new rules ensure that some literals are included in extensions, or are excluded from them), this approach seems closer to base revision (see Section 5). But, independently of this question, one problem is still open: how to adjust this approach to account for legal modifications? A legal system T is modified by selecting, as a target, one or more norms of T, whereas [4]'s proposal parametrises operations to sets of literals. Let us bear in mind these points and proceed with our discussion.

5 Revising Normative Bases

The main problem with revision based on belief sets is that this approach does not mimic how the law implements norm changes, since "new" rules are generated to reflect the changes. Legal effects of rules can be used to guide how norms should be changed, but they should not determine *what* and *how* rules are changed. Therefore the alternative to revision based on belief sets is base revision. As is well-known, base revision does not operate on the extension of a theory, but rather applies to the theory "generators" (i.e., the non-logical axioms of the theory). This idea can be naturally coupled with

[3] For space reasons, [4]'s treatment of revision is omitted.

partitioning the elements of a theory into "facts" and "rules", where the former cannot be revised (unless update is used), while the latter may be subject to revision.

Usually, belief revision operations are defined as contraction followed by expansion (according to Levi's Identity). Therefore, revision often results in some rules to be removed from the base of a theory. Base revision allows us to adopt different strategies, namely, to modify rules. In the law there are different types of norm changes: some directly correspond to the removal of rules (e.g., abrogation and annulment), while others amount to introducing new rules (e.g., derogation), and finally some are the result of partial modifications of provisions. In this perspective, assuming a rule-based representation of norms, revision on bases using modification techniques seems closer to the legal practice in so far as it allows for the conceptual distinction of these types of changes. In addition, as argued e.g. in [5], base revision results in theories that are closer to the structure of the theory to be revised.

Let us consider an example to introduce the idea of modification of bases. Suppose we want to revise a theory with a rule $r_1 : A \Rightarrow B$ and contract B when C is the case (let us say that C implies $\neg B$). The revision of the rule is $r'_1 : A, \neg C \Rightarrow B$. This means that we have modified the original rule taking into account the exception provided by C. DL has an elegant mechanism to deal with exceptions. An exception is simply implemented by a rule capturing the connection between the exceptional antecedent and the conclusion to be blocked. Thus, in the example above, instead of changing r_1 into r'_1, we may simply add a new rule such as $r_2 : C \Rightarrow \neg B$ or $r_2 : C \rightsquigarrow \neg B$, and state that $r_2 \succ r_1$. As we have seen in Section 4, this idea has been originally proposed for DL in [4], but there was still the open problem of how to set change operations in such a way to parametrise them with respect to the proper target of legal modifications, namely, legal rules.

Let us see how to adjust [4]'s definitions for norm changes, and in particular for annulment and abrogation. Let T be a theory and $A_1, \ldots, A_n \Rightarrow B$ be the rule to be removed. For *annulment*:

$$T^{annul1}_{A_1,\ldots,A_n \Rightarrow B} = T^-_B \tag{4}$$

Hence, the annulment of a rule is the contraction of the head of the rule. This solution directly applies (3) to the head of the rule to be annulled. However, (4) is too strong since it forces the removal of B from the extension (unless B is a fact). If we have two different (and independent) rules applicable at the same time and with the same head, and we just annul one of them, the other should still be able to produce its effect. But (4) affects the second rule as well. Thus, we have to give an alternative annulment operation based on a variant of the contraction operation.

$$\begin{aligned} T^{annul2}_{r:A_1,\ldots,A_n \Rightarrow B} &= \begin{cases} T & \text{if } B \notin \partial^+(T) \\ (F, R', \succ') & \text{otherwise} \end{cases} \\ &\text{where} \\ &R' = R \cup \{r' : \rightsquigarrow {\sim}B\} \\ &\succ' = \succ \cup \{(r', r)\} \cup \{(s, r') | s \in R[B] - \{r\}\} \end{aligned} \tag{5}$$

Consider the following examples.

Example 2. Let us consider the following theory:

$$T = (F = \{A\}, R = \{r_1 : A \Rightarrow B,\ r_2 : B \Rightarrow C\}, \emptyset).$$

Clearly, $\partial^+(T) = \{B, C\}$. Hence,

$$T'^{annul2}_{r_1:A\Rightarrow B} = (F, R \cup \{r'_1 :\rightsquigarrow \neg B\}, \emptyset).$$

In the resulting theory we prove $-\partial B$, which makes r_2 inapplicable, thus preventing the positive conclusion of C.

Example 3. Let us consider again the theory in Example 1:

$$T = (F = \{A\}, R = \{r_1 : A \Rightarrow B,\ r_2 : B \Rightarrow C,\ r_3 : A \Rightarrow C\}, \emptyset).$$

The annulment of r_1 still amounts to adding $r'_1 :\rightsquigarrow \neg B$ to R, which prevents the conclusion of all literals depending *only* on B. Accordingly, C will be in the extension, as it is obtained through r_3. In addition, if $r_4 : \ \Rightarrow B$ were in R, r_4 would be stronger than r^-, thus obtaining B.

In general this approach is closer to the legal practice, as it precisely focuses on modifications of norms and not on the modification of the normative positions (effects) of norms. First, it does not depend on facts. Second, it offers a seamless solution to *ex tunc* modifications. However, things can be viewed from a different perspective. Even though this approach can simulate *ex tunc* modification like annulment (since it allow us to block norm effects), the actual operation fails to remove norms. (Hence, this approach is appealing for modifications corresponding essentially to exceptions, such as derogations: see Section 1.) When a norm is annulled, it is "removed" from the legal system, whereas here we just remove the effects of the norm and its consequences.

Accordingly, we can simply remove the rule to be annulled from the set of rules:

$$T^{annul3}_r = (F, R - \{r\}, \succ) \tag{6}$$

But, then, we have another problem: How to deal with *ex nunc* modifications, such as abrogations? In this case, the modification of a rule should not necessarily prevent the derivation of its conclusions. Let us consider Example 2 and assume that the abrogation of r_1 does not prevent the derivation of B and C. This means that, if B and C were derivable before the modification, then they should remain in the extension of the revised theory. Here, we have two options. First, we can argue, as done above with annulment, that when a norm is abrogated, it is "removed" from the legal system. But, if r_1 is removed following a similar procedure to that stated in (6), the extension of the revised theory will *not* contain B as well as C, whereas abrogations can also admit cases where both conclusions should be maintained. Thus, a second option does not remove the rule, but adds a suitable set of new rules which allow to derive what should not be blocked. However, what can we do in this case if both B and C should not be dropped? It seems hard to adjust (5) in order to maintain both B and C. At most, what we can do is preventing the derivation of B and maintaining C. Only in this case, if $T = (F, R, \succ)$ is a defeasible theory, then the *abrogation* of a norm $r : A_1, \ldots, A_n \Rightarrow B$ runs as follows:

$$T^{abr}_{r:A_1,\ldots,A_n \Rightarrow B} = \begin{cases} T & \text{if } r \notin R \\ (F, R', \succ') & \text{otherwise} \end{cases}$$

where

$$\begin{aligned} R' = R &\cup \{r^- : \rightsquigarrow \neg B,\ r' : \Rightarrow B'\} \\ &\cup \{t' : (A(t) - \{B\}) \cup \{B'\} \rightarrow C(t) | t \in R_s \text{ and } B \in A(t)\} \\ &\cup \{t' : (A(t) - \{B\}) \cup \{B'\} \Rightarrow C(t) | t \in R_d \text{ and } B \in A(t)\} \\ &\cup \{t' : (A(t) - \{B\}) \cup \{B'\} \rightsquigarrow C(t) | t \in R_{dft} \text{ and } B \in A(t)\} \\ \succ' = \succ &\cup \{(w, r^-) | w \in R[B] - \{r\}\} \cup \{(t', s) | (t, s) \in \succ\} \cup \{(s, t') | (s, t) \in \succ\} \end{aligned} \tag{7}$$

where B' is a new literal not appearing in T.

Proposition 2. *Given a theory T and a rule $r : A_1, \ldots, A_n \Rightarrow B$, for every $C \in HB_T - \{B\}$, $T \vdash C$ iff $T^{abr}_r \vdash C$.*

Example 4. Consider the following theory:

$$T = (F = \{A, D\}, R = \{r : A \Rightarrow B,\ t : B \Rightarrow C,\ s : D \Rightarrow \neg C,\ w : E \Rightarrow B\}, (t, s) \in \succ)$$

Hence, according to (7), $T^{abr}_{r:A_1,\ldots,A_n \Rightarrow B}$ is as follows:

$$\begin{aligned} T^{abr}_{r:A_1,\ldots,A_n \Rightarrow B} = (F = \{A, D\} \\ R = \{r : A \Rightarrow B,\ t : B \Rightarrow C,\ s : D \Rightarrow \neg C,\ w : E \Rightarrow B \\ r^- : \rightsquigarrow \neg B,\ r' : \Rightarrow B',\ t' : B' \Rightarrow C\} \\ \succ = \{(t, s), (t', s), (w, r^-)\}) \end{aligned}$$

The fact A makes r applicable, but the introduction of r^- blocks the derivation of B using r. However, C is derived via r' and t' (which is stronger than s). Note that (7) is such that the defeater r^- attacks only r (we are abrogating rule r only): hence, if E were in F, B would be obtained from w.

In sum, we have the following possibilities:

- We omit to model annulments and abrogations as corresponding to rule removals. Hence, we represent them working only on rule conclusions and so adopt (5) and (7). However, (7) is partially satisfactory, as it blocks the derivation of the head of the abrogated rule; but an abrogation may remove only norms and not the already obtained effects of the norms to be abrogated.
- We address the issue that annulments and abrogations correspond to rule removals. Thus, (6) works for annulments, but it seems quite hard to find an adequate counterpart for abrogation.
- We do not care whether annulments and abrogations correspond to rule removals and are free to adopt, together with (7), either (5) or (6). But, as we said, (7) is problematic.

Of course, we do not exclude that the above problems can be settled. For example, some limits of (7) can avoided by combining the introduction of exceptions and the removal of the abrogated rule. This can be done by applying the idea in (6) and subsequently reinstate the conclusions that should not be blocked. This can be done by simply using expansion $^+$ as defined in (2). More precisely, suppose $c = C_1, \dots C_n$ are the consequences of the rule to be abrogated which we want to maintain.

Definition 2. *Let $T = (F, R, \succ)$ be a Defeasible Theory such that $r : A_1, \dots, A_n \Rightarrow B \in R$. Then*

$$T^{abr'}_{r:A_1,\dots,A_n\Rightarrow B} = (T')^+_c$$

such that $T' = (F, R - \{r\}, \succ)$ and $c = C_1, \dots C_n \in E(T'')$, where

- *$T'' = (F = \{B\}, R, \succ)$;*
- *for every C_k, $1 \leq k \leq n$, $C_k \notin E(T')$.*

But, even in that case, another difficulty arises when we have to deal with *retroactive* modifications: as we already mentioned, retroactivity is a typical feature of legal modifications. This problem is discussed in the following section.

6 A Temporal Model for Legal Systems and Norm Change

6.1 Revision and Retroactivity

A norm modification is an operation such that a normative system (consisting of norms and the consequences of cases) is transformed into a different normative system. Accordingly, dynamics of a normative system are described by a sequence of operations.

Suppose we have a system, let us call it T_0, in which we introduce a new rule r and subsequently we remove another rule, let us say s. The system obtained from the first operation is T_1, while the final system is T_2. Thus $T_2 = ((T_0)^+_r)^-_s$. So far so good. But let us suppose that that the removal of s is retroactive. How can we model this case? The idea is that every time we have a retroactive modification we should reconstruct the normative system at the time when the retroactive modification is effective. For example, if the modification is effective since yesterday, we have to recover the system as it was yesterday by undoing the operations leading to the normative system we have today, then we have to apply the retroactive modification and finally redo the other modifications. So, if in the example above s is a retroactive modification effective from T_0, the sequence of modifications still adds r and removes s, but the sequence of theories is $T'_1 = (T_0)^-_s$ and $T_2 = ((T_0)^-_s)^+_r$. Is this procedure in agreement with the intuition behind retroactive legal modifications? Our answer is negative. The point is that it is possible to define transformations moving from one normative system T_i to T_{i+1} where the transformation is effective at T_i itself, thus the system to be changed is not the target of the modification but the source of it. Let us consider the following example. The normative system T_0 is just the fact A. T_1 is obtained from T_0 by retroactively adding two rules $A \Rightarrow B$ and $B \Rightarrow C$ and these rules are effective in T_0. Then the next transformation, leading to T_2 is the removal of $A \Rightarrow B$ from T_0. But then we have two different versions of T_0. Analogous considerations apply when we work on rule consequences and model modifications adding defeaters.

The reason why we have multiple versions of a normative system is that norms have different temporal dimensions: the time of validity of a norm (when the norm enters in the normative system) and the time of effectiveness (when the norm can produce legal effects). Thus, if one wants to model norm modifications, then normative systems must be modelled by more complicated structures. In particular, a normative system is not just the set of norms valid in it, but it should also consider the normative systems where the norms are effective. Accordingly, a normative system is a structure $N_i = (T_i, \langle T_0, T_1, \ldots \rangle)$, where T_i is the theory modelling the set of norms/rules and facts valid in the normative system N_i, and $\langle T_0, T_1, \ldots \rangle$ is the sequence of theories encoding the effective norms for all "versions" of the normative system.

A revision of a legal system is an operation that transforms a normative system into another normative systems by 'changing' the rules in it. In particular, the operation should specify what rules are to be changed and when they are changed, and when the changes are effective. Thus a norm change can be seen as a transaction from a normative system $N_i = (T_i, \langle T_0^i, T_1^i, \ldots \rangle)$ to a normative system $N_{i+1} = (T_{i+1}, \langle T_0^{i+1}, T_1^{i+1}, \ldots \rangle)$, where there exists some j such that $T_j^{i+1} = change(T_j^i)$ for some *change* operation. For example, the abrogation of a rule r may be modeled as $T_{i+1}^{i+1} = (T_i^i)_r^{abr}$, and the retroactive annulment of r, as $T_j^{i+1} = (T_j^i)_r^{annul}$ for $j < i$. In addition, in general, once a norm has been introduced in a normative system the norm continues to be in the normative system unless it is explicitly removed. This means that the norm must be included in all theories succeeding the theory in which it has been first introduced. Accordingly, it could be very cumbersome to keep track of the changes and where the changes have to been applied. In real normative systems norms are introduced at a particular time, they are effective at a particular time, and so are changes –changes are norms themselves. Thus, to obviate the issue of keeping track of the changes, and at the same time to offer a conceptual model of norm changes, we have proposed in [8,9] an extension of DL with time, where we consider the two temporal dimensions of relevance for norm change (effectiveness and validity). This is done by labelling rules with two time values, one for the validity time of the norms, and the other for their effectiveness time; furthermore the labels indicate whether these 'changes' persist or not. The idea that changes are norms themselves is captured by the notion of meta-rule, i.e., a rule whose elements can be rules themselves and not only literals. The next section offers the conceptual background of the proposal presented in [8,9].

6.2 Inner and Outer Time of Legal Systems

The above discussion suggests that the dynamics of a legal system LS are more correctly captured by a time-series $LS(t_1), LS(t_2), \ldots, LS(t_j)$ of its versions. Each version of LS is called a *norm repository* [8,9]. The passage from one repository to another is effected by legal modifications or simply by persistence [9]. But dynamics of norm change and retroactivity need to introduce another time-line within each version of LS (see Figure 1). Clearly, retroactivity

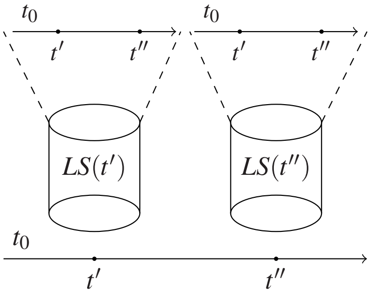
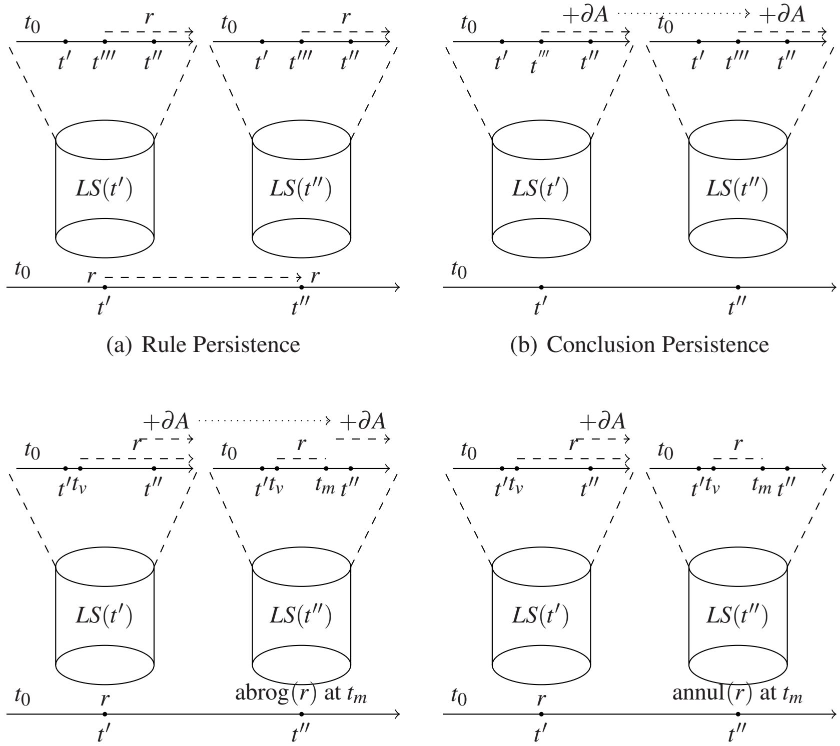


Fig. 1. Legal System at t' and t''



(a) Rule Persistence

(b) Conclusion Persistence

(c) Abrogation

(d) Annulment

does not imply that we can really change the past, but it rather requires that we have to reason on the legal system from the viewpoint of its current version as it were revised in the past: when we change some $LS(i)$ retroactively, this does not mean that we modify some $LS(k)$, $k < i$, but that we move back from the perspective of $LS(i)$. Hence, we can "travel" to the past along this inner time-line, i.e. from the viewpoint of the current version of LS where we modify norms.

Elements contained in, or derived from, theories can propagate across these time-lines. Hence, propagation concerns the derived conclusions of rules (when some consequent P holds), the rules themselves, and also derivations (i.e., queries: $+\partial P$). This introduces several options regarding how modifications affect a legal system over time:

- conclusions may persist within a certain repository or across different repositories;
- derivations may persist within a certain repository or across different repositories;
- rules may persist within a certain repository or across different repositories.

For example, Figure 2(a) shows how rule persistence works. A persistent rule r enacted at time t' and in force at t''' carries over from the legal system $LS(t')$ to the legal system $LS(t'')$, where it is still in force at t'''. Figure 2(b) illustrates conclusion persistence: a conclusion A persists from $LS(t')$ to $LS(t'')$ even if the rules used to derive it are no longer effective in $LS(t'')$. Figure 2(c) presents a case of abrogation: in $LS(t')$ rule

r, in force from t_v onwards, produces a persistent effect A. The effect carries over by persistence to $LS(t'')$ even if the rule r is abrogated at t_m and is no longer in force to produce the effect. Finally, Figure 2(d) illustrates a case of annulment: in $LS(t')$ rule r, in force since t_v, is applied and produces a persistent effect A. Since the rule is annulled in $LS(t'')$ at t_m, the effect of A must be undone as well. While the intuition in Figures 2(c) and 2(d) seems clear, its precise implementation in DL is not simple and only a partial solution was offered in [9]. The development of a complete DL temporal model for abrogation and annulment is a matter of future work.

7 Summary

In this paper we investigated how to model in DL legal abrogation and annulment. Terminology may vary from one legal system to another, but, despite this, it is possible to identify in general two different reasoning patterns: in one case norms are removed with all their effects, whereas in other cases norms are removed but all or some of their effects propagate if obtained before the modification. We examined some ways to capture these intuitions in DL using techniques from revision based on belief sets and from base revision. We concluded that abrogation and annulment can only be partially represented in these settings. In addition, we argued that it is hard, if not impossible, to simulate retroactivity, which clearly refers to the temporal dimension. Hence, we illustrated a different conceptual starting point from which the problem can be addressed.

References

1. Alchourrón, C.E., Gärdenfors, P., Makinson, D.: On the logic of theory change: Partial meet contraction and revision functions. Journal of Symbolic Logic 50, 510–530 (1985)
2. Antoniou, G., Billington, D., Governatori, G., Maher, M.J.: Representation results for defeasible logic. ACM Transactions on Computational Logic 2(2), 255–287 (2001)
3. Bassiliades, N., Antoniou, G., Vlahavas, I.: DR-DEVICE: A defeasible logic system for the Semantic Web. In: Bry, F., Henze, N., Małuszyński, J. (eds.) PPSWR 2003. LNCS, vol. 2901. Springer, Heidelberg (2004)
4. Billington, D., Antoniou, G., Governatori, G., Maher, M.J.: Revising nonmonotonic belief sets: The case of defeasible logic. In: Burgard, W., Christaller, T., Cremers, A.B. (eds.) KI 1999. LNCS (LNAI), vol. 1701. Springer, Heidelberg (1999)
5. Di Giusto, P., Governatori, G.: A New Approach to Base Revision. In: Barahona, P., Alferes, J.J. (eds.) EPIA 1999. LNCS (LNAI), vol. 1695. Springer, Heidelberg (1999)
6. Gärdenfors, P.: Knowledge in Flux: Modeling the Dynamics of Epistemic States. MIT, Cambridge (1988)
7. Governatori, G., Maher, M.J., Billington, D., Antoniou, G.: Argumentation semantics for defeasible logics. Journal of Logic and Computation 14(5), 675–702 (2004)
8. Governatori, G., Palmirani, M., Riveret, R., Rotolo, A., Sartor, G.: Norm modifications in defeasible logic. In: Proc. JURIX 2005. IOS Press, Amsterdam (2005)
9. Governatori, G., Palmirani, M., Riveret, R., Rotolo, A., Sartor, G.: Variants of temporal defeasible logic for modelling norm modifications. In: Proc. ICAIL 2007. ACM Press, New York (2007)

10. Guastini, R.: Teoria e dogmatica delle fonti, Giuffré, Milan (1998)
11. Maher, M.J.: Propositional defeasible logic has linear complexity. Theory and Practice of Logic Programming (6), 691–711 (2001)
12. Nute, D.: Defeasible reasoning. In: Proceedings of 20th HICSS. IEEE press, Los Alamitos (1987)
13. Prakken, H.: Logical Tools for Modelling Legal Argument. Kluwer, Dordrecht (1997)
14. Sartor, G.: Legal Reasoning. Springer, Dordrecht (2005)

Acting, Events and Actions

Mark A. Brown

Philosophy Department
Syracuse University
Syracuse, NY 13244
USA
mabrown@syr.edu

Abstract. A logic of action is essential for many treatments of normative concerns, but most treatments either ignore the role of agents, as in PDL, or omit all possibility of naming actions, as in various versions of *stit* theory. Moreover, most treatments of either type do not attempt to provide an account of what actions *are*, in a way that would distinguish actions from other processes or events. In this paper, I explore an account of actions as a species of events, with events interpreted against a background of the logic of branching time. This opens a new approach to exploring the relations between logics of personal action (e.g. Belnap's and Horty's *stit* theories) and impersonal logics of actions such as PDL, and offers some prospect of a deontic logic which integrates *tunsollen* (ought to do) into a system of *seinsollen* (ought to be).

1 Introduction

The literature on the logic of action provides us with two quite different styles of treatment. On the one hand, we have a family of systems—let's call them Delta-systems—in each of which some general-purpose non-normal modal operator Δ is made available[1]. In such systems a formula of the general form $\Delta_a A$ can be used to express the claim that the agent **a** sees to it (or has seen to it) that the claim expressed by A is true. On the other hand we have systems of dynamic logic[2], each with several (possibly infinitely many) designations for processes[3], several operators for combining processes to form other processes, and the capacity to form a normal modal operator $[\alpha]$ for each such process α. In such languages, any of the modal operators can be used in a formula of the general form $[\alpha]A$ to express the claim that at the completion of any application of the process named by α, the claim expressed by A will be true.

There are several conspicuous points of contrast, many of which grow out of a basic difference: Delta-systems are aimed at exploring what it means for a human being to act, while systems of dynamic logic are (at least initially) aimed at examination of

[1] This class of systems includes ones proposed by Pörn [11], Kanger, Belnap [1–3], Horty [8], [9], Carmo, and others.

[2] This class of systems is based on original work by Pratt, Fisher, and Ladner. Applications to deontic logic have been explored in a number of works, notably including [6], [7], and [10].

[3] In the original interpretation, each process would be the running of a computer program.

R. van der Meyden and L. van der Torre (Eds.): DEON 2008, LNAI 5076, pp. 19–33, 2008.

computer programs, treated as “actions”. As a result Delta-systems have designations for agents, and can express the claim that an agent has acted, or is acting, with a certain effect, but do not designate the action(s) involved; dynamic logics have designations for actions, viewed as processes, but not for agents, there being only one tacit “agent” involved, namely the computer. The Delta-systems can differentiate actions only by their results and their agents; dynamic logics directly designate processes, and thus can differentiate actions without regard to whether or not they are performed by the same agent and whether or not they have the same outcome. Delta-systems can readily consider multi-agent situations, while dynamic logics normally cannot.

Each of these kinds of system has its advantages and its disadvantages as a logic of action. Delta-systems are typically embedded in a rich temporal logic, and thus enable us to keep track of temporal relations. Moreover, Delta-systems are designed to differentiate between actions of agents and mere happenings and accidents; in contrast, in their interpretation as logics of action dynamic logics have to simply *stipulate* that all basic processes for which designations are provided are actions, but cannot differentiate the results that are genuinely due to the action from those that are accidental or unavoidable. This shows up most strikingly in the fact that if $\top$ is any logically true formula, $[\alpha]\top$ will be logically true in dynamic logic, whereas $\Delta_a\top$ will be logically false in most Delta-systems. Thus most Delta-systems specifically reject the notion that any agent is responsible for the truth of necessary truths, while dynamic systems consider all logical truths to be the doing of the tacit agent.

Because agents are explicitly represented in Delta-systems, these systems are particularly suitable for incorporation into deontic logics, where we will wish to relate agents’ actions to their responsibilities; dynamic logics can be augmented with indices for agents, but the result seems a bit unnatural. Moreover, the explicit mention of agents makes it possible for Delta-systems to consider cooperation among agents, delegation of responsibility, influence, prevention, etc., in a multi-agent setting, whereas typically in dynamic logics this is impossible.

Dynamic logics of action do have their own advantages, however, chiefly arising from the fact that they directly represent actions, not just agents and the results of their acting. Dynamic logics can express and examine the results of various operations on actions, including catenation; in contrast, Delta-systems normally have no way to express catenation of actions. Attempts to accomplish something of the sort by iterating Delta operators give us nothing satisfying: for example, $\Delta_a\Delta_aA$ is just equivalent to Δ_aA in some of the most prominent members of this class of systems, while $\Delta_aA \wedge \Delta_aB$ will report simultaneous actions, rather than successive ones. About the best we can do rather directly is to interpose a future tense operator and a possibility operator between two delta operators, to assert that the agent brings it about that it will eventually be possible for the agent to bring it about that A, but this will still fall short of expressing the claim that the agent actually does follow through with the second action. Because dynamic logics of action can represent catenations $\alpha;\beta$ of actions α and β (and other combinations, such as their nondeterministic union $\alpha \cup \beta$, corresponding to disjunction), they can be used rather naturally to examine means/ends relation. We find it more natural to think that picking up the hammer is a means to pounding the nail into the board than that the hammer’s being in my hand is a means to the nail’s being in the board.

These two kinds of treatment share a deficiency: neither attempts to give an account of what an action *is*, and derive the logic of action from the nature of this account.[4] Delta-systems content themselves with giving an account of what it means for an agent to act, or to have acted, with a certain result, but do not attempt an account of just what the action was that produced the result. They attempt to provide an account of when it is accurate to say that the agent has performed an action whose result was that the door is shut, for example, but do not attempt to give an account of what that action was. *How* did the agent shut the door? By kicking it? leaning against it? throwing something at it? …?[5] Systems of dynamic logic simply assume a stock of actions, but without giving any account of what it is about them that makes them *actions*, rather than mere events or mere processes.

Surveying the complementary advantages and disadvantages of these two styles of treatment rather naturally arouses a desire to develop some more comprehensive system that will combine the advantages of each of these approaches, and if possible overcome the defects they hold in common. Some recent work[6] has hinted at one possible way to get started towards such a goal, by introducing action names into Delta-systems. The present paper pushes that effort a bit further along: in contrast with other approaches, here we attempt to work from a Delta-system in the direction of including features like those of dynamic logic, by first introducing actions as constructs within a branching time framework.

The starting point will be to view an action as a special kind of event[7], or process, in branching time, one that intimately involves choices on the part of the agent.

2 Branching Time, Transitions and Events

The classic works on *stit* theory, [1–3], [8], [9], begin with the observation in Pörn [11] that what we *do* is not merely something that *happens*. This shows up in one way when we attribute responsibility. I am not responsible for the fact that it is raining—that just happens, and I have no choice about it. But I *am* responsible for the fact that I left my umbrella at home—that didn't just happen: I had, and made, some choice in the matter. Genuine action, as contrasted with mere occurrence of some event, requires some exercise of freedom of choice on the part of the agent involved. But the availability of genuine choices implies indeterminism about the future: if I genuinely have a choice about whether to take my umbrella with me or not, then there are details

[4] More precisely, PDL does give an account of what it means by 'action' (namely a function from states to states) but can't distinguish genuine actions from mere happenings.

[5] In his logic of the *dstit* operator, Horty [8] applies the word 'action' to any choice an agent makes. However this is at best a very limited sense of action. It will count the agent's *choosing* to *start* to shut the door as an action, but not the agent's shutting the door. The only actions recognized in this system are ones whose outcomes are instantaneous.

[6] Brown [4].

[7] Some authors use the term 'event' to denominate a momentary occurrence, unextended in time. As I use the term, I mean a process which extends through an interval of time. A concert, for example, counts as an event, in my terminology. I suggest that, as we normally use the term 'action', human actions extend over time, and thus are events in this sense.

of the future that depend on which choice I make; and when such choices are freely available, the course of the future is correspondingly undetermined.

When we reflect on how this view of the freedom of agents, we find it natural to represent time as branching into the future, with branching occurring at those moments when agents have genuine choices[8], and with different branches corresponding to the different courses which history will take, depending on the agent's choice. On one branch, having chosen to leave my umbrella behind, I get wet. On another, having chosen to bring it with me, I remain dry. Accordingly, Belnap and Horty make the logic of branching time the foundation of their work in the study of action, ability and events. So let us turn to a brief presentation of the essentials of the logic of branching time, and more specifically, to its semantics.

By a *forward-branching back-connected temporal frame*, I mean a structure $\langle \mathbf{T}, < \rangle$ such that (with $\leq$ defined in terms of $<$ in the obvious way):

(1) $\mathbf{T}$ is a non-empty set (of *moments* of time);
(2) $<$ is a strict partial ordering (antisymmetric and transitive) on $\mathbf{T}$;
(2.1) $(\forall \mathbf{m}, \mathbf{m}_1, \mathbf{m}_2 \in \mathbf{T})[\mathbf{m}_1 < \mathbf{m} \wedge \mathbf{m}_2 < \mathbf{m} \Rightarrow \mathbf{m}_1 \leq \mathbf{m}_2 \vee \mathbf{m}_2 \leq \mathbf{m}_1]$;
(2.2) $(\forall \mathbf{m}, \mathbf{m}' \in \mathbf{T})(\exists \mathbf{m}_0 \in \mathbf{T})[\mathbf{m}_0 \leq \mathbf{m} \wedge \mathbf{m}_0 \leq \mathbf{m}']$.

A *history* through a moment $\mathbf{m}$ is any complete, non-back-tracking path through $\mathbf{T}$, i.e. any subset $\mathbf{h}$ of $\mathbf{T}$ satisfying the conditions[9]

(*i*) $\mathbf{m} \in \mathbf{h}$ (through $\mathbf{m}$),
(*ii*) $(\forall \mathbf{m}_1, \mathbf{m}_2 \in \mathbf{h})[\mathbf{m}_1 \leq \mathbf{m}_2 \vee \mathbf{m}_2 \leq \mathbf{m}_1]$ (connected),
(*iii*) $\neg(\exists \mathbf{m}_\omega \in \mathbf{T})(\forall \mathbf{m} \in \mathbf{h})[\mathbf{m} < \mathbf{m}_\omega]$ (complete forwards),
(*iv*) $(\forall \mathbf{m} \in \mathbf{h})(\forall \mathbf{m}_0 \in \mathbf{T})[\mathbf{m}_0 < \mathbf{m} \Rightarrow \mathbf{m}_0 \in \mathbf{h}]$ (complete backwards).

Given any moment $\mathbf{m}$ in such a frame, and any history $\mathbf{h}$ through $\mathbf{m}$, we let $\mathbf{H}$ be the set of all histories and let $\mathbf{H}(\mathbf{m})$ be the set of all histories through $\mathbf{m}$; i.e.

$$\begin{array}{lll} \mathbf{H}(\mathbf{m}) & =_{\mathrm{df}} & \{\mathbf{h} \subseteq \mathbf{T}: \mathbf{h} \text{ satisfies constraints } i\text{–}iv \text{ above}\} \\ \mathbf{H} & =_{\mathrm{df}} & \cup_{\mathbf{m} \in \mathbf{T}} \mathbf{H}(\mathbf{m}). \end{array}$$

Formulas will be evaluated at moment/history pairs $\mathbf{m}/\mathbf{h}$ in which $\mathbf{m} \in \mathbf{h}$. This is essential[10], because at the moment at which I must choose whether or not to take my umbrella, it cannot yet be said to be true, without qualification, that I will get wet, nor yet that I won't. It depends on my choice, and thus along some histories I will get wet, while along others I won't. Without specifying the history in question, as well as the moment, we cannot expect to get an evaluation of the claim that I will get wet.

From here on, whenever I use the notation $\mathbf{m}/\mathbf{h}$ it is to be understood that $\mathbf{m} \in \mathbf{h}$.

[8] And perhaps also at certain moments at which genuinely random events, such as the radioactive decay of a radon atom, occur.

[9] No constraint is imposed on the order type of the arrangement of moments within histories. They might be discretely, densely, or continuously arranged, or they might be irregularly arranged. Because discrete orderings seem out of keeping with physics, I tend to assume in my informal thinking that the moments along a history are at least densely ordered. However, no such assumption is being built into the formalism at this juncture.

[10] See Thomason [12] for a full discussion of this point.

A frame becomes a *model* $\mathtt{M}$ when it is supplemented by a *valuation* $\mathtt{V}$ assigning, at each moment **m**, to each sentential constant S, a truth value $\mathtt{V}(\mathbf{m}, S) \in \{\mathbf{t}, \mathbf{f}\}$. (Thus the presumption is that sentential constants convey claims whose truth may change from moment to moment but which, at a given moment, do not vary in truth value from history to history. At bottom, this is the presumption that sentential constants will not be used to convey claims that in any way involve other moments than the moment of evaluation.)

Now we can give satisfaction conditions for the usual temporal operators as well as for temporal possibility and necessity operators.

Given any moment **m** and any classes **K** and **L** of moments, we extend our use of the relation < in a natural way, so that

$$\begin{array}{llll} & \mathbf{m} < \mathbf{K} & \text{iff} & (\forall \mathbf{n} \in \mathbf{K})[\mathbf{m} < \mathbf{n}], \\ & \mathbf{K} < \mathbf{m} & \text{iff} & (\forall \mathbf{n} \in \mathbf{K})[\mathbf{n} < \mathbf{m}], \\ \text{and} & \mathbf{K} < \mathbf{L} & \text{iff} & (\forall \mathbf{m} \in \mathbf{K})(\forall \mathbf{n} \in \mathbf{L})[\mathbf{m} < \mathbf{n}]. \end{array}$$

Belnap has also introduced the notion of a *transition* into the discussion of time, and Ming Xu [13] has shown ways in which this notion can be useful in our analysis of events, causation, and actions. We turn now to the development of these notions.

By a *past*, we mean any class **p** of moments such that:

- **p** is nonempty, i.e. $\mathbf{p} \neq \varnothing$;
- **p** is linear, i.e. there is some history **h** with $\mathbf{p} \subset \mathbf{h}$; and
- **p** is closed pastwards, i.e. whenever $\mathbf{m} < \mathbf{n} \in \mathbf{p}$, we have $\mathbf{m} \in \mathbf{p}$.

By an *outcome*, we mean any class **F** of moments such that:

- **F** is nonempty, i.e. $\mathbf{F} \neq \varnothing$;
- **F** is pastwards connected, i.e. whenever $\mathbf{m}, \mathbf{n} \in \mathbf{F}$,
  there is some $\mathbf{k} \in \mathbf{F}$ such that $\mathbf{k} \leq \mathbf{m}$ and $\mathbf{k} \leq \mathbf{n}$; and
- **F** is closed futurewards, i.e. whenever $\mathbf{m} > \mathbf{n} \in \mathbf{F}$, we have $\mathbf{m} \in \mathbf{F}$.

Note that an outcome will be a tree and so usually will not be linear. In general it might or might not have a first moment.

By a *transition* τ, we mean a pair $\tau = \langle \mathbf{p}, \mathbf{F} \rangle$ consisting of a past **p**, called its *prologue* and an *outcome* **F**, such that $\mathbf{p} < \mathbf{F}$, i.e. such that whenever $\mathbf{m} \in \mathbf{p}$ and $\mathbf{n} \in \mathbf{F}$, we have $\mathbf{m} < \mathbf{n}$. Note that this definition leaves room for various special cases.

immediate transitions	there is no moment **m** such that $\mathbf{p} < \mathbf{m} < \mathbf{F}$;
singleton transitions	there is a unique moment **m** such that $\mathbf{p} < \mathbf{m} < \mathbf{F}$;
extended transitions	all other transitions.

Each transition τ uniquely determines an associated interval which I shall call its *interval*, consisting of the moments between its prologue and its outcome. However the converse is not true; i.e., an interval does not in general uniquely determine an associated transition. We can have two (or more) transitions with the same interval. There certainly can be different immediate transitions, for example, even though immediate transitions will by definition determine the empty interval. This is illustrated in Figure 1, below (with earlier moments represented here as to the left of later ones).

There are less trivial ways that distinct transitions can determine the same interim interval. Figure 2 below illustrates one such. Here the transitions $\langle \mathbf{p}, \mathbf{F}_1 \rangle$ and $\langle \mathbf{p}, \mathbf{F}_2 \rangle$ both determine the interim consisting of the two moments depicted on the same horizontal line as $\mathbf{p}$. We could vary the situation to make it more exotic. $\mathbf{F}_1$ could remain as depicted, but $\mathbf{F}_2$ be altered to have no first moment; or the interval might have no first or last moment; or the prologue $\mathbf{p}$ might have no last moment; there might be even more than two transitions determining the same interim interval, etc.

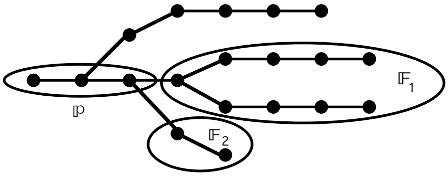


Fig. 1.

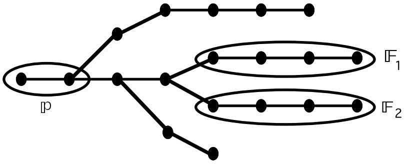


Fig. 2.

There are less trivial ways that distinct transitions can determine the same interim interval. Figure 2 above illustrates one such. Here the transitions $\langle \mathbf{p}, \mathbf{F}_1 \rangle$ and $\langle \mathbf{p}, \mathbf{F}_2 \rangle$ both determine the interim consisting of the two moments depicted on the same horizontal line as $\mathbf{p}$. We could vary the situation to make it more exotic. $\mathbf{F}_1$ could remain as depicted, but $\mathbf{F}_2$ be altered to have no first moment; or the interval might have no first or last moment; or the prologue $\mathbf{p}$ might have no last moment; there might be even more than two transitions determining the same interim interval, etc.

The fact that two transitions—even two non-immediate transitions—can determine the same interim interval makes transitions more sensitive than intervals as a means for representing processes. We might well wish to associate a given process with one transition, yet not associate it with another transition whose interval is the same, because we might want the outcome of the process to be one of its identifying features. *A fortiori*, transitions are more sensitive than moments by themselves would be.

Most events extend over a period of time. For such events, there will be associated extended transitions, with the prologue, or past, of the transition corresponding to that period of time before the event has occurred, the interval between corresponding to that portion of time during which the event is occurring, and the outcome corresponding to that portion of time by which the event has already occurred. Other events, such as choices, may well occur instantaneously, perhaps marking the boundary between times at which A is true and times at which it is not; such instantaneous events will be associated with immediate transitions.

We might consider letting events be represented by single transitions. But it often—indeed usually—seems reasonable to say that under slightly altered circumstances, the same event would have occurred. For example, I recently closed the door. Suppose I try to represent that event by a single transition, with a unique prologue leading up to my closing it, a unique interval during which I was closing it, and a unique outcome within which it could be said truly that I had closed it. But shortly before I began to close the door, someone in Brunei was exercising his free will, choosing what to have for a midnight snack, and thus affecting the course of history. Various branches of time result from various choices he might make, and some of those branches deviate from the prologue of my favored transition. Are we to say that if he had made a different choice than in fact he did, then this event of my closing the door would not have occurred? We *could* say that, of course, and maintain that although I might still have closed the door, the event of doing so would have been a different event, though of a similar (or even the same) kind.

But this seems far more fine-grained than our ordinary discourse would suggest was appropriate. Moreover, as Xu points out, we sometimes wish to say that under certain conditions an event may become inevitable, and would not want to retract this because of causally irrelevant co-occurrent circumstances. Sitting in my office, I drop a book. Under the local conditions, it becomes inevitable that the book will fall to the floor. We do not want to retract this judgment because of the possibility that a radium atom on the moon might decay just as the book begins to slip from my hand, or that someone in Brunei might at that moment be freely choosing something for a snack. If we want to say that *this* event, and not just that *an event of this kind*, was (under the circumstances) inevitable then we must be prepared to accept that various transitions would count equally as occurrences of this same event.

We say that a transition $\tau = \langle \mathbf{p}, \mathbf{F} \rangle$ *occurs* in history $\mathbf{h}$ iff $\mathbf{h}$ *runs through* τ, i.e. iff $\mathbf{h} \cap \mathbf{F} \neq \varnothing$. Note that in any such case $\mathbf{p} \subset \mathbf{h}$, but that this is not enough to assure that $\mathbf{h}$ runs through τ, since there can be other transitions with the same prologue $\mathbf{p}$.

Assume for the moment that an event η is, or is associated with, a class of transitions. For simplicity of expression, let's just speak of the event as *consisting* of the set of associated transitions. Then it is natural to say that event η **occurs** in history $\mathbf{h}$ iff $\mathbf{h}$ runs through some transition in η.

Xu [13] proposes that an event is (or at least corresponds to) a class of transitions, but that there are limitations on what classes of transitions can count as corresponding to a single event. In particular, a single event (as contrasted with events of a given kind) cannot occur twice in the same history. We can capture that requirement by saying that the collection of transitions corresponding to a single event must be such that no history runs through two different transitions of the collection. I propose to accept this for now and just identify an *event* with a class of transitions meeting this constraint. Note that this is akin to taking propositions to be classes of points of evaluation. No doubt this is too coarse-grained an approach to serve all purposes, but it is good enough to serve a surprisingly large range of purposes.

Thus far, I have simply reported work done by others, sometimes altering terminology and notation to suit my own tastes and convenience. But now let me introduce terminology for discussing special kinds of events which will be of interest later on.

When a past has a last moment, I will call it a *definite past*. Each moment $\mathbf{m}$ determines a unique past $\mathbf{p_m}$ of which it is the last moment. Similarly, by a *definite outcome*, I will mean an outcome with a first moment. Any moment $\mathbf{n}$ will determine a unique definite outcome $\mathbf{F_n}$ of which $\mathbf{n}$ is the first moment.

We will have a special interest in transitions of the form $\langle \mathbf{p_m}, \mathbf{F_n} \rangle$ for moments $\mathbf{m}$, $\mathbf{n}$. I will call such a transition a *definite transition*. A definite transition will determine a closed interval $[\mathbf{m}, \mathbf{n}]$, with $\mathbf{m} \leq \mathbf{n}$, and each such interval will uniquely determine a definite transition. We will speak of the moment $\mathbf{m}$ as the *moment of initiation* of the transition, and the moment $\mathbf{n}$ as the *outcome moment* of the transition $\langle \mathbf{p_m}, \mathbf{F_n} \rangle$. If an event consists entirely of definite transitions, I will call it a *definite event*.

It can happen that all of the transitions in a given event η have the same prologue $\mathbf{p}$. When this happens, for a given prologue $\mathbf{p}$ and a given event η, I will call η a **$\mathbf{p}$-event**. If all the transitions in a $\mathbf{p}$-event η are immediate transitions, I will call η an *immediate* $\mathbf{p}$-*event*. When the prologue in question is the prologue $\mathbf{p_m}$ determined by a moment $\mathbf{m}$, we can have an *immediate* $\mathbf{p_m}$-*event*. Since in general there can be many immediate transitions with the same prologue, a given prologue $\mathbf{p_m}$ will not usually determine a unique immediate $\mathbf{p_m}$-event. We will define $\mathbf{E}(\mathbf{p_m})$ to be the class of all immediate $\mathbf{p_m}$-events. When all the transitions in a **$\mathbf{p_m}$-event** are extended transitions, we will call it an *extended* $\mathbf{p_m}$-*event*.

Processes, unlike events, are repeatable, but each occurrence of a process is an event.[11] This makes it natural to identify processes with kinds of events, i.e. with classes of events. Probably a careful analysis of the notion of a process will reveal ways in which we should constrain this notion somewhat. For example, in order for two events to count as examples of the same process, perhaps it is important that their prologues bear some natural similarity to one another, and that their outcomes likewise have something in common. Possibly, instead, there should be constraints on the intervals associated with the member events. I suspect it will be a delicate matter to say what, if any, constraints we should impose for the final account of processes. Fortunately, I don't think it will matter much for our present work. So we will count every set of events as a process, with an occurrence of any of the member events counting as one occurrence of the process, and the occurrence of more than one member event counting as more than one occurrence of the same process.

We will then take actions to be processes of a special sort. But to characterize that sort, we will need first to look at some sample logics of action from which we can take guidance.

3 The Basic *stit* Theories of Action

To make it possible to discuss the actions of specific agents, we must augment our models for branching time in at least two ways: We need to represent the agents in the model, and we need to have a way to indicate which options each agent has at any given juncture in history. Accordingly, we now introduce the notion of a *choice frame*, gotten by inserting a class $\mathbf{A}$ of agents, and a choice-function $\mathbf{C}$ into our forward-branching back-connected temporal frames.

By a *choice frame*, then, I mean a structure $\langle \mathbf{T}, <, \mathbf{A}, \mathbf{C} \rangle$ in which $\langle \mathbf{T}, < \rangle$ is a forward-branching back-connected temporal frame, and

(3) $\mathbf{A}$ is a non-empty set (of *agents*); and

(4) $\mathbf{C}$ (the *choice function*) provides, for each agent $\mathbf{a}$ and each moment $\mathbf{m}$, a partition $\mathbf{C}(\mathbf{a}, \mathbf{m})$ of $\mathbf{H}(\mathbf{m})$ such that:

(4.1) if $\mathbf{c}$ is any function on $\mathbf{A} \times \mathbf{T}$ such that for each $\mathbf{a} \in \mathbf{A}$ and each $\mathbf{m} \in \mathbf{T}$, $\mathbf{c}(\mathbf{a}, \mathbf{m}) \in \mathbf{C}(\mathbf{a}, \mathbf{m})$, then for each $\mathbf{m} \in \mathbf{T}$: $\bigcap\{\mathbf{c}(\mathbf{a}, \mathbf{m}) : \mathbf{a} \in \mathbf{A}\} \neq \varnothing$;

[11] I construe a process as something which has occurrences, rather than as a more abstract plan, or recipe, for events. Those who may be uncomfortable with this usage can substitute 'kinds of event' where I use the term 'processes'.

$$(4.2)\quad (\forall \mathbf{m} \in \mathbf{T})(\forall \mathbf{h}_1, \mathbf{h}_2 \in \mathbf{H}(\mathbf{m}))\ [(\exists \mathbf{m}' \in \mathbf{h}_1 \cap \mathbf{h}_2)[\mathbf{m} < \mathbf{m}'] \Rightarrow (\forall \mathbf{a} \in \mathbf{A})(\forall \mathbf{c} \in \mathbf{C}(\mathbf{a},\mathbf{m}))[\mathbf{h}_1 \in \mathbf{c} \Rightarrow \mathbf{h}_2 \in \mathbf{c}]].$$

Condition 4.1 expresses the *independence of agents*, and condition 4.2 expresses that there can be *no choice between undivided histories*.

The action operator introduced by Jeff Horty, which he has called the *dstit*, or *deliberative stit* operator, deals with present actions (ones which culminate at the very moment of deliberation, i.e. the very moment of choice), rather than with actions initiated earlier and culminating now or actions initiated now and culminating latter. I will use the notation $\stackrel{d}{\triangle}_a A$ as a more compact[12] substitute for his notation [*a dstit*: *A*].

We can given satisfaction conditions for such formulas as follows:

$$\mathbf{m}/\mathbf{h}, \mathtt{M} \models \stackrel{d}{\triangle}_a A \text{ iff}$$

(*the positive condition: reliability*)

$$(+)\quad (\exists \mathbf{c} \in \mathbf{C}(\mathbf{a}, \mathbf{m})\text{: } \mathbf{h} \in \mathbf{c})(\forall \mathbf{h}' \in \mathbf{c})[\mathbf{m}/\mathbf{h}', \mathtt{M} \models A\ ]$$

and (*the negative condition: freedom*)

$$(-)\quad (\exists \mathbf{h}'' \in \mathbf{H}(\mathbf{m}))[\mathbf{m}/\mathbf{h}'', \mathtt{M} \not\models A\ ].$$

The condition +, the *positive condition*, says in effect that the choice, from among the choices available to **a** at the moment **m**, within which history **h** falls (and which may therefore be deemed the choice α actually makes at moment **m**, from the point of view of history **h**) is one which leads reliably to the truth of *A*, because *A* is true along *each* history in that choice. The condition –, the *negative condition*, says in effect that there really was a free choice about the matter, because there was at least one other choice available which would *not* have assured the truth of *A*. The negative condition is what rules out an agent's claiming credit for the fact that 2 + 2 = 4, for example, because with respect to that matter the agent had effectively no freedom of choice.

Figure 3 below gives a depiction of a typical situation in which a formula $\stackrel{d}{\triangle}_\alpha A$ turns out to be true at a point of evaluation **m/h** in a choice model.

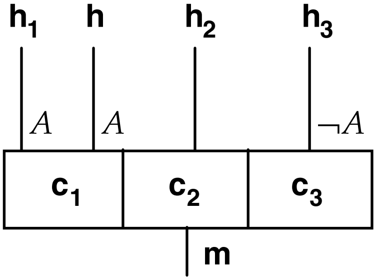


Fig. 3.

[12] This is indeed a more compact notation; but the chief reasons for preferring this notation are (1) that it presents this claim more clearly as a claim involving a parameterized sentential operator and (2) that in this notation each distinct element of the syntax makes a distinct contribution to the semantics (in contrast, for example, to the '*s*' and the '*t*' in '*astit*', not to mention the colon).

Here time is depicted as flowing upwards. At the moment $\mathbf{m}$, agent $\mathbf{a}$ has, say, three choices (two would be sufficient to illustrate my point, more would not interfere with the outcome). If $\mathbf{a}$ adopts choice $\mathbf{c_1}$, then (depending on circumstances not under $\mathbf{a}$'s control) time will either continue as in history $\mathbf{h}$, or else as in history $\mathbf{h_1}$, but in either case A is true. On the other hand, choice $\mathbf{c_3}$ is also available, and if it is adopted history will continue as in $\mathbf{h_3}$, in which case A is false. The middle choice is, for our purposes, irrelevant, and there could be additional alternative histories issuing from $\mathbf{c_3}$, without affecting matters. As long as all the histories issuing from $\mathbf{c_1}$ are ones in which A is true, the positive condition is met for ${\triangle\!\!\!\!\!\!{\scriptstyle d}\,}_a A$ to be satisfied at the point $\mathbf{m/h}$, and as long as at least one history issuing from at least one choice, such as $\mathbf{c_3}$, is one in which A is false, the negative condition is met. So this illustrates a situation in which we have

$$\mathbf{m/h}, \mathrm{M} \models {\triangle\!\!\!\!\!\!{\scriptstyle d}\,}_a A.$$

Among results which fall out fairly easily from the semantics for ${\triangle\!\!\!\!\!\!{\scriptstyle d}\,}$, we have the following:

$$\models {\triangle\!\!\!\!\!\!{\scriptstyle d}\,}_a A \to A; \qquad \text{(however:} \qquad \not\models A \to {\triangle\!\!\!\!\!\!{\scriptstyle d}\,}_a A);$$

$$\models \neg {\triangle\!\!\!\!\!\!{\scriptstyle d}\,}_a \top.$$

This last validity shows that, in the sense of action expressed by the ${\triangle\!\!\!\!\!\!{\scriptstyle d}\,}$ operator, no agent can take credit for seeing to it that logically true formulas are true.

For Belnap's version of *stit* theory, we need one further element in our models: an equivalence relation $\mathbf{I}$ on $\mathbf{T}$ whose equivalence classes are called *instants*. Each instant has exactly one element in each history, and if any moment in instant $\mathbf{i}_1$ precedes any moment in instant $\mathbf{i}_2$ then no moment in $\mathbf{i}_2$ precedes any moment in $\mathbf{i}_1$. Thus instants cut across histories, synchronizing the histories by indicating which moment in one history occurs "at the same time as" a given moment in another history.[13]

By a *synchronous choice frame*, then, I mean a structure $\langle \mathbf{T}, <, \mathbf{A}, \mathbf{C}, \mathbf{I} \rangle$ in which $\langle \mathbf{T}, <, \mathbf{A}, \mathbf{C} \rangle$ is a choice frame, and

(5) $\mathbf{I}$ (the *instant* relation) is an equivalence relation on $\mathbf{T}$ such that:

(5.1) $(\forall \mathbf{m} \in \mathbf{T})(\forall \mathbf{h} \in \mathbf{H})(\exists! \mathbf{n} \in \mathbf{h})\mathbf{Imn}$;

(5.2) $(\forall \mathbf{h}_1,\mathbf{h}_2 \in \mathbf{H})(\forall \mathbf{m}_1,\mathbf{n}_1 \in \mathbf{h}_1)(\forall \mathbf{m}_2,\mathbf{n}_2 \in \mathbf{h}_2)[\mathbf{Im}_1\mathbf{m}_2 \wedge \mathbf{In}_1\mathbf{n}_2 \wedge \mathbf{m}_1<\mathbf{n}_1 \Rightarrow \mathbf{m}_2<\mathbf{n}_2]$.

Thus in a synchronous frame moments in one history are synchronized with moments in other histories. We let $\mathbf{s}(\mathbf{m},\mathbf{h})$ be the unique moment $\mathbf{n} \in \mathbf{h}$ such that $\mathbf{Imn}$, i.e. the unique moment synchronic with $\mathbf{m}$, but in the history $\mathbf{h}$.

The action operator introduced by Belnap, which he originally called simply the *stit* operator, but later called the *astit* or *achievement stit* operator, deals with actions initiated in the past and culminating in the present, and interpreted using synchronous choice models. Belnap uses the notation [*a astit*: *A*] to express the claim that agent *a*

[13] This naturally makes one wonder whether there are problems of trans-history identity of times, corresponding to the much-discussed problems of trans-world identity of individuals and objects. However, in this case perhaps such questions can be finessed by taking sameness of clock-time in the different histories as our standard for sameness of instant.

has seen to it that A is now true, where A is any sentence of our formal language. I will use the more compact notation $\overset{\text{a}}{\triangle}_a A$ to express the same claim.

The satisfaction conditions for this operator can be given as follows:

$\mathbf{m}/\mathbf{h}, \mathtt{M} \models \overset{\text{a}}{\triangle}_a A$ iff for some moment $\mathbf{m}_0 < \mathbf{m}$:

(the positive condition: reliability)

(+) $(\exists \mathbf{c} \in \mathbf{C}(\mathbf{a}, \mathbf{m}_0){:}\ \mathbf{h} \in \mathbf{c})(\forall \mathbf{h}' \in \mathbf{c})[\mathbf{s}(\mathbf{m},\mathbf{h}')/\mathbf{h}', \mathtt{M} \models A\ ]$

and *(the negative condition: freedom)*

(–) $(\exists \mathbf{h}'' \in \mathbf{H}(\mathbf{m}_0))[\mathbf{s}(\mathbf{m},\mathbf{h}'')/\mathbf{h}'', \mathtt{M} \not\models A\ ]$.

Thus $\overset{\text{a}}{\triangle}_a A$ at moment $\mathbf{m}$, along history $\mathbf{h}$, iff there is some prior moment $\mathbf{m}_0$ at which agent $\mathbf{a}$ made a choice which included $\mathbf{h}$ and, as the positive condition, along each history within that choice A would "now" be true, but also, as the negative condition, there was at least one other history through $\mathbf{m}_0$, along which A would not have "now" been true. The positive condition assures that the agent's choice guaranteed the truth of A; the negative condition assures that there were other choices available, at least one of which would *not* have guaranteed the truth of A, so that the choice really made a difference.

4 Adding Terms for Actions

With some logics of action in place, we can begin to consider the question what it is, in these pictures, that constitutes the action involved. Formulas such as $\overset{\text{a}}{\triangle}_a A$ and $\overset{\text{d}}{\triangle}_a A$ can be considered to report the occurrence of an action, but they do not provide a name for the action reported; they merely point to the agent performing the action and the outcome of that performance.

Questions arise naturally: Would/Should we count it as the same action if it were performed by another agent? Would/Should we count it as the same action if it had a different result? It may well be that these questions have no canonical answer, i.e. that more than one of the possible combinations of answers to these questions is viable, with different combinations simply characterizing different, though closely related, notions, useful for different purposes. Let us begin, somewhat arbitrarily, by focusing attention on the narrowest of these notions, by treating both the agent and the result as essential to the identity of the action.

On this interpretation, it will still be possible for the same action to occur more than once. For example, you shut the door, I enter and inadvertently leave the door open. You shut the door again, thus performing the same action a second time. We could construe the second door-shutting as a different action of the same kind, but it seems unlikely that this would be useful, and it would be contrary to the spirit of dynamic logic, in which the same action can be performed repeatedly.

But on this account, what is the action? We begin by noting that a given performance of the action is an event, and asking ourselves what that event is? Looking at Figure 3, we see a situation in which the agent $\mathbf{a}$ has choices.[14] The choice taken

[14] The fact that, in the illustration, there are three choices is inessential. What is essential is that there at least two: one satisfying the positive condition and the other satisfying the negative.

typically includes various possible histories which may (perhaps because of simultaneous choices made by other agents) differ from one another in important respects even as regards the facts about the present moment, but which have in common a result of the action, namely the truth of A. Another choice available to $\mathbf{a}$ at that same moment includes at least one history with a different result: along that history A is false.

The natural answer to the question what event takes place here is that it is the set of immediate transitions whose histories together make up the set of histories in the choice taken. But that event counts as an occurrence of an *action* only if there is at least one other history through the same prologue $\mathbf{p}$—or, what comes to the same thing, through the same moment $\mathbf{m}$ of choice.

So we let an *immediate* $\mathbf{a}/A$ *action-event* be an immediate $\mathbf{p}_\mathbf{m}$-event η such that

- the set of all histories through η constitutes one of $\mathbf{a}$'s choices at moment $\mathbf{m}$,
- A is true at $\mathbf{m}/\mathbf{h}$ for each history $\mathbf{h}$ in η,
- A is false at $\mathbf{m}/\mathbf{h}^*$ for some other history $\mathbf{h}^*$.

Then we can define the action involved as the class of events of this type. We let $\mathbf{d}(\mathbf{a},A)$ be the class of all immediate $\mathbf{a}/A$ action-events. We can now augment our language to include an operator δ used to form expressions such as $\delta(a,A)$, to name the action $\mathbf{d}(\mathbf{a},A)$.

Whenever such an action is performed, we can now name that action. We can also introduce a (present-tense)[15] performance verb π, and provide an interpretation for it that will support the equivalence

$$\pi\delta(a,A) \quad \leftrightarrow \quad \overset{d}{\triangle}_a A.$$

We can define $\Pi(\mathbf{m},\mathbf{h})$ to be the class of immediate actions performed (by anyone, with any result) at the point $\mathbf{m},\mathbf{h}$. Then we have the satisfaction condition:

$$\mathbf{m}/\mathbf{h}, \mathsf{M} \models \pi\delta(a,A) \qquad \text{iff} \qquad \mathbf{d}(\mathbf{a},A) \in \Pi(\mathbf{m},\mathbf{h}).$$

For each action-term $\delta(a,A)$, we can now allow a modal operator $[\delta(a,A)]$. If we are to construe this in a fashion that will closely mimic a modal operator from a dynamic logic, then we will want to use this satisfaction condition:

$$\mathbf{m}/\mathbf{h}, \mathsf{M} \models [\delta(a,A)]B \qquad \text{iff} \qquad (\exists \mathbf{c} \in \mathbf{C}(\mathbf{a}, \mathbf{m})\colon \mathbf{h} \in \mathbf{c})(\forall \mathbf{h}' \in \mathbf{c})\\ [\mathbf{m}/\mathbf{h}', \mathsf{M} \models \overset{d}{\triangle}_a A \wedge B].$$

As with PDL operators, this would then be a normal modal operator, with a meaningful dual along the same lines as in dynamic logics, and supporting the schema **K** and the rule **RN**, thus assuring that the operator distributes across conditionals, but also that when B is a logical truth, $[\delta(a,A)]B$ will be, as well. We may consider this latter an unfortunate fact, but it is the price we pay for having a normal operator. We can make this palatable by adopting an appropriately careful interpretation of the operator: to say that $[\delta(a,A)]B$ is true (at $\mathbf{m}/\mathbf{h}$) is not to say that $\mathbf{a}$'s performing the action named by $[\delta(a,A)]$ *causes* B to be true, but rather simply that, *as it happens*, this action, performed now, reliably results in a state in which B is true.

[15] Other tenses can be gotten by combining this with the usual tense operators defined earlier.

Can we catenate actions, as defined above, or combine or modify them in the other ways that are characteristic of dynamic logic? Here we encounter two problems. First, since each such action is attributed to an agent, and different actions could be attributed to different agents, arbitrary catenation could lead to a sequence of events that could not be attributed to any one agent. That problem will arise in any treatment of action which permits discussion of actions by different agents and permits catenation of actions. Discussing this challenge will have to be reserved for a different paper.[16]

A second problem is specific to the use of the operator $\overset{\mathsf{d}}{\triangle}$. In the satisfaction conditions for the $\overset{\mathsf{d}}{\triangle}$ operator, the moment of evaluation is the *same* moment as the moment of choice, As a result, all the actions we define directly using this operator are *instantaneous* actions, involving sets of *immediate* transitions. Thus it makes little sense to think of catenation of such actions as their *sequential* performance. Since each such action ends at the very same moment at which it starts, the “second” action in the catenation will start at the moment at which the “first” action ends, but that is also the moment at which the “first” *starts*, so the two are really simultaneous. On reflection, this suggests that the “actions” identified this way are merely instantaneous choices, and are a degenerate case of actions more generally.

So let’s turn to the logic of the operator $\overset{\mathsf{a}}{\triangle}$, to see what guidance it might offer. Unlike the operator $\overset{\mathsf{d}}{\triangle}$, the operator $\overset{\mathsf{a}}{\triangle}$ does differentiate between the moment of choice and the moment of evaluation. This, then, will involve a non-trivial interval of time between the initiation of the action by the agent’s choice, and the culmination of the action in the truth of the formula A.

The action lying behind the truth of a claim of the form $\overset{\mathsf{a}}{\triangle}_a A$ will then be the set of all extended $\mathbf{p}_{\mathbf{m}_0}$-transitions whose outcome moments are synchronous with the moment $\mathbf{m}$ of evaluation and whose histories all lie in the same cell of $\mathbf{C}(\mathbf{a}, \mathbf{m}_0)$ as $\mathbf{h}$.

All such actions would extend over an interval of time, so it will be meaningful to speak of catenating them. But the restriction to cases in which the transitions all end at the same time seems artificial, and extremely restrictive. When I perform the action of shutting the door, various of the histories involved in my critical choice may differ from one another with respect to how quickly the door gets shut, so the defining outcome may not (indeed, most likely *will* not) occur synchronously across the histories involved. This leads to the thought that we should drop the requirement of synchrony.

5 Adding General Extended Actions

Let us then make room for the ordinary kind of action whose completion requires an interval of time, but without imposing any special requirements about when the action begins or ends. With this in mind, we now recognize that some definite events should be considered occurrences of actions. Which ones? Guided by analogy with the satisfaction conditions for the *stit* operators, we want to impose both positive and negative conditions on a definite event in order to count it as an occurrence of an action. Note that each transition τ in a definite event η will have a uniquely determined prologue.

[16] See the forthcoming [5].

Each such transition τ corresponds to a closed interval $[\mathbf{m}_\tau, \mathbf{n}_\tau]$, with the moment $\mathbf{m}_\tau$ involved being the moment at which that transition begins. Then we can state reasonable positive and negative conditions as follows:

a definite event η is an occurrence of an action iff

for each transition τ in η, there is some agent $\mathbf{a}$ such that
(*the positive condition: reliability*)
(+) there is some choice available to $\mathbf{a}$ at moment $\mathbf{m}_\tau$
such that the set of histories in that choice is the set of all
histories in all transitions of η that begin at moment $\mathbf{m}_\tau$;
and (the negative condition: freedom)
(–) there is at least one other choice available to $\mathbf{a}$ at moment $\mathbf{m}_\tau$.

Let us call such an event an *action-event*. The picture is this: there will in general be various clusters of transitions included in a definite event, each cluster corresponding to one possible occurrence of the event. For one such cluster to correspond to the occurrence of an action, it must precisely correspond to one choice available to some agent who had at least one alternative choice available at the time. To correspond precisely to a given choice, the cluster of transitions must collectively involve all and only the histories through that choice.

We can now add terms for actions to our language, without requiring that these terms be constructed from other portions of the language, as were the terms of the form $\delta(a,A)$. As before, however, actions will still be considered as *kinds* of events, rather than as single events, so that we will be able to speak of repetition of the same action. So let α, β, etc. be terms for designating actions. Semantically, let the valuation V interpret each such term by assigning it a set of action-events. Since we are no longer demanding that an action be characterized by its agent or its result, we will put no constraints on the collections of action-events. Thus we are allowing that different agents could perform the same action, and that the same action could have different, or indeed unpredictable, results. But each occurrence of an action will be an occurrence of a constituent event, and consequently will be associated with a particular set of closed intervals with the same first moment. That first moment is the moment of initiation of the occurrence of the action, and the various moments which occur as final moments of the intervals associated with these various transitions are the moments at which the results of this occurrence of the action are to be evaluated.

We are now in a position to mimic a great deal of PDL. In particular, we now have a rich enough environment that we can readily introduce normal modal operators $[\alpha]$, $[\beta]$, etc., and catenation and union of actions, in a meaningful way. In particular, catenation will now be non-trivial, because actions will in general occupy a span of time, and one can begin at the conclusion of another. The catenated action will correspond to the set of transitions with the prologue of the first action and the outcomes of the various runnings of the second that begin at terminal moments of the various outcome moments of the first.

If, however, for a given outcome moment for a transition in the occurrence of the first action, the second action doesn't contain any transitions which can be initiated at that moment, then the catenation aborts, and the catenated expression designates no action. In this respect the notion of action proposed here will differ from that examined in PDL. This seems to be an unavoidable result of the difference between human actions and executions of pieces of computer code, however. Unless the computer has crashed, the execution of any piece of code can be followed by the execution of any other, though the results may of course be garbage. But human actions will normally have pre-conditions for their very performance, not just for their useful or meaningful performance: if the door is already closed, it is simply not possible to close the door, for example; if there is no hammer at hand, I cannot swing a hammer.

Any given occurrence of an action will be attributable to at least one agent, whose array of choices was involved in its meeting the conditions for being an occurrence of an action. Consequently, we can now introduce a personalized performance operator π_a which will apply truly to an action-term α, at a given moment, along a given history, iff one of the occurrences of that action is an occurrence at that moment, by that agent, and that history falls within one of its transitions. Such an operator will indicate that the action is being initiated. Related operators will be available to indicate that the action is in progress, or that it has just finished. All of these can be combined with the usual tense operators, and the result will be a richly expressive language.

By now, I believe, we have established the outline of a promising merger between dynamic logic and Delta-systems of the logic of action. Much remains to be done, of course, in the way of exploring that promise, but that must await another occasion.

References

1. Belnap Jr., N.D.: Backwards and Forwards in the Modal Logic of Agency. Philosophy and Phenomenological Research 51, 777–807 (1991)
2. Belnap Jr., N.D.: Before Refraining: Concepts for Agency. Erkenntnis 34, 137–169 (1991)
3. Belnap Jr., N.D., Perloff, M., Xu, M.: Facing the Future: Agents and Choices in Our Indeterminist World. Oxford University Press, Oxford (2001)
4. Brown, M.A.: Acting with an End in Sight. In: Goble, L., Meyer, J.-J.C. (eds.) DEON 2006. LNCS (LNAI), vol. 4048, pp. 69–84. Springer, Heidelberg (2006)
5. Brown, M.A.: Acts and Actions. Presented at the Society for Exact Philosophy (May 2008)
6. Dignum, F., Meyer, J.-J.C., Wieringa, R.J., Kuiper, R.: A Modal Approach to Intentions, Commitments and Obligations: Intention plus Commitment Yields Obligation. In: Brown, M.A., Carmo, J. (eds.) Deontic Logic, Agency and Normative Systems (DEON 1996), pp. 80–97. Springer, Berlin (1996)
7. Grossi, D., Dignum, F., Royakkers, L.M., Meyer, J.-J.C.: Collective obligations and agency: Who gets the blame? In: Lomuscio, A., Nute, D. (eds.) DEON 2004. LNCS (LNAI), vol. 3065, pp. 129–145. Springer, Heidelberg (2004)
8. Horty, J.F.: Agency and Deontic Logic. Oxford University Press, Oxford (2001)
9. Horty, J.F., Belnap Jr., N.D.: The deliberative Stit: A Study of Action, Omission, Ability, and Obligation. J. Phil. Logic 24, 583–644 (1995)
10. 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 29.1, 109–136 (1988)
11. Pörn, I.: Some Basic Concepts of Action. In: Stedlund, S. (ed.) Logical Theory and Semantic Analysis. D. Reidel, Dordrecht (1977)
12. Thomason, R.H.: Deontic Logic and the Role of Freedom in Moral Deliberation. In: Hilpinen, R. (ed.) New Studies in Deontic Logic (Synthese Library), vol. 152, pp. 177–186. D. Reidel, Dordrecht (1981)
13. Xu, M.: Causation in Branching Time (I): Transitions, Events, and Causes. Synthese 112, 137–192 (1997)

A Tableaux System for Deontic Action Logic

Pablo F. Castro and T.S.E. Maibaum

McMaster University
Department of Computing & Software
Hamilton, Canada
castropf@mcmaster.ca,
tom@maibaum.org

Abstract. In [1] and [2] we have introduced a novel deontic action logic for reasoning about fault-tolerance. In this paper we present a tableaux method for this logic; this proof system is sound and complete, and because the logic has the usual boolean operators on actions, it also allows us to deal successfully with action complement and parallel execution of actions. Finally, we describe an example of application of this proof system which shows how the tableaux system can be used to obtain (counter-) models of specifications.

Keywords: Modal Logic, Deontic Action Logic, Tableaux Systems, Fault-tolerance, Software Specification.

1 Introduction

One of the benefits of *deontic logic* is the possibility of expressing normative predicates, and therefore differentiating between “ideal situations” and “violation states”. In particular, *deontic action logic* (or DAL for short) is a variation of deontic logic (introduced in [3] and related to dynamic deontic logic [4]) which introduces normative predicates on actions. (These systems are called *ought-to-do* logics.) As is pointed out in several works ([3], [5], [6] and [4]), deontic action logics seem to be useful for application in software specification: they allow us to express which actions are allowed (or forbidden) in a given scenario, and then to reason about the consequences of executing a particular action, whether allowed or not.

In particular, we are interested in specifying fault-tolerant systems, and therefore notions such as *violations*, *violation recovery* and *correct behaviors* are important for us. As we argued in [1] and [2], DALs are suitable in the context of fault-tolerance: the basic notions detailed above can be formalized naturally using the deontic predicates (e.g., *permission*, *obligation* and *forbidden*), some examples of application can be found in [7] and [1].

On the other hand, tableaux systems ([8]) are practical proof systems which are strongly related with automated theorem proving (see [9]). Several tableaux systems have been proposed for variants of dynamic logics and modal logics. In [10], a tableaux system for *propositional dynamic logic* is described, which

R. van der Meyden and L. van der Torre (Eds.): DEON 2008, LNAI 5076, pp. 34–48, 2008.

is shown to be more efficient than other decision methods. In [11] the method of labelled tableaux is introduced to deal with modal logic. Meanwhile, in [12], a tableaux system that incorporates some new characteristics is introduced to deal with *dynamic logic with converse*. This last system allows us to decide that logic, and in addition to find counterexamples in the case of non-valid formulae.

In this paper we introduce a tableaux system for the *deontic action logic* described in [2]; this logic uses boolean operators on actions, and therefore the system allows us to manage action complement and parallel execution of actions and the standard deontic operators. One of the main points to note is that, in case of a non-valid formula, the tableaux system can be used to build counterexamples, in such a way that the model built in this way shows which actions are performed in each step and which actions are not. We intend to use this proof system for the verification of fault-tolerant software, in such a way that it allows us to verify properties or to get critical traces of executions which yield a fault.

The paper is organized as follows. In section 2 we give a brief introduction to our *deontic action logic*. In sections 3 and 4 we present the tableaux system and some of its meta-properties (soundness and completeness). In section 5 we describe an example to illustrate the application of the ideas presented earlier. Finally, we discuss some further work and conclusions.

2 A Deontic Action Logic

In this section we present the basic definitions of the *deontic action logic* which we use in the following sections. The interested reader can find in [1], [2] a more extensive introduction to this logic, with some examples of applications.

The language of the logic is given by a vocabulary $\langle \Delta_0, \Phi_0 \rangle$, where Δ_0 is a finite set of primitive actions (we use lower case alphabet letters to denote them: $a, b, c, d, ...$), and a set Φ_0 of primitive propositions, denoted by $p, q, s, ...$. Using these two sets we can build more complicated formulae using the standard propositional connectives, the modal connectives, deontic predicates (permission and obligation) and boolean operators over actions. For example: $[a \sqcup b](p \to q)$ is a well-formed formula. The set Φ of formulae can be defined as usual by induction. The intuition behind each formula is as follows:

- $\alpha =_{act} \beta$: *actions* α *and* β *are identical.*
- $[\alpha]\varphi$: *after any possible execution of* α, φ *is true.*
- $[\alpha \sqcup \beta]\varphi$: *after the non-deterministic execution of* α *or* β, φ *is true.*
- $[\alpha \sqcap \beta]\varphi$: *after the parallel execution of* α *and* β, φ *is true.*
- $[\mathbf{U}]\varphi$: *after the non-deterministic choice of any possible action,* φ *is true.*
- $[\emptyset]\varphi$: *after executing an impossible action,* φ *becomes true.*
- $[\overline{\alpha}]\varphi$: *after executing an action different from* α, φ *is true.*
- $\mathsf{P}(\alpha)$: *every way of executing* α *is allowed.*
- $\mathsf{P}_\mathsf{w}(\alpha)$: *some way of executing* α *is allowed.*

We have two permission predicates: $\mathsf{P}_\mathsf{w}(\alpha)$ is called *weak* permission, and $\mathsf{P}(\alpha)$ is called *strong* permission. Both versions of permission are sometimes found in the deontic literature. Here we use both versions to define obligation: $\mathsf{O}(\alpha) \overset{\mathsf{def}}{\Longleftrightarrow} \mathsf{P}(\alpha) \wedge \neg \mathsf{P}_{\mathsf{w}}(\overline{\alpha})$. That is, an action is obliged if it is allowed to be performed in every context (strong permission), and every other action is forbidden (weak permission). We also define a strong relationship between the two forms of permissions. See [2]. This definition of obligation is similar to that given in [13], but here we use the two variants of permission to define the obligations. This definition avoids some paradoxes, like Ross's Paradox (see [14] for a detailed list of deontic paradoxes).

An interesting aspect of this logic is that the interpretation of deontic predicates is independent of the modal operators (in the semantic structures they have an independient relational interpretation), whereas in other works (e.g., [4] and [6]) the deontic operators are reduced to modal formulae, for example in [4] permission is defined: $\mathsf{P}(\alpha) \equiv \langle\alpha\rangle\neg V$. That is, an action is allowed if there exists a way to execute it without producing a violation. Note that, in this approach the permission of an action implies that this action can be executed. We follow the philosophy established in [15], in the sense that modal operators are used for descriptions of components (in a pre and post-condition style), while deontic operators are used for action prescription (i.e., to express when an action may, or must, occur). In our view, description and prescription are different concepts, which should be reflected in a separation of deontic and modal predicates in the logic.

We introduce briefly the semantics of the logic with some remarks; a deeper description of this can be found in [2].

Definition 1 (models). *Given a language $L = \langle \Phi_0, \Delta_0 \rangle$, an L-Structure is a tuple: $M = \langle \mathcal{W}, \mathcal{R}, \mathcal{E}, \mathcal{I}, \mathcal{P} \rangle$ where:*

- *$\mathcal{W}$, is a set of worlds.*
- *$\mathcal{R}$, is an $\mathcal{E}$-labeled relation between worlds. We require that, if $(w, w', e) \in \mathcal{R}$ and $(w, w'', e) \in \mathcal{R}$, then $w' = w''$, i.e., $\mathcal{R}$ is functional.*
- *$\mathcal{E}$, is a non-empty set of (names of) events.*
- *$\mathcal{I}$, is a function:*
  - *For every $p \in \Phi_0 : \mathcal{I}(p) \subseteq \mathcal{W}$*
  - *For every $\alpha \in \Delta_0 : \mathcal{I}(\alpha) \subseteq \mathcal{E}$.*

  In addition, the interpretation $\mathcal{I}$ has to satisfy the following properties:

  I.1 *For every $\alpha_i \in \Delta_0$: $|\mathcal{I}(\alpha_i) - \bigcup\{\mathcal{I}(\alpha_j) \mid \alpha_j \in (\Delta_0 - \{\alpha_i\})\}| \leq 1$.*

  I.2 *For every $e \in \mathcal{E}$: if $e \in \mathcal{I}(\alpha_i) \cap \mathcal{I}(\alpha_j)$, where $\alpha_i \neq \alpha_j \in \Delta_0$, then: $\cap\{\mathcal{I}(\alpha_k) \mid \alpha_k \in \Delta_0 \wedge e \in \mathcal{I}(\alpha_k)\} = \{e\}$.*

  I.3 *$\mathcal{E} = \bigcup_{\alpha_i \in \Delta_0} \mathcal{I}(\alpha_i)$.*
- *$\mathcal{P} \subseteq \mathcal{W} \times \mathcal{E}$, is a relation which indicates which event is permitted in which world.*

We can extend the function $\mathcal{I}$ to well-formed action terms and formulae, as follows:

-$\mathcal{I}(\neg\varphi) \overset{\mathsf{def}}{=} \mathcal{W} - \mathcal{I}(\varphi)$

-$\mathcal{I}(\varphi \to \psi) \overset{\mathsf{def}}{=} \mathcal{I}(\neg\varphi) \cup \mathcal{I}(\psi)$

-$\mathcal{I}(\alpha \sqcup \beta) \overset{\mathsf{def}}{=} \mathcal{I}(\alpha) \cup \mathcal{I}(\beta)$

-$\mathcal{I}(\mathbf{U}) \overset{\mathsf{def}}{=} \mathcal{E}$

-$\mathcal{I}(\alpha \sqcap \beta) \overset{\mathsf{def}}{=} \mathcal{I}(\alpha) \cap \mathcal{I}(\beta)$

-$\mathcal{I}(\overline{\alpha}) \overset{\mathsf{def}}{=} \mathcal{E} - I(\alpha)$

-$\mathcal{I}(\emptyset) \overset{\mathsf{def}}{=} \emptyset$

Note that here we do not follow the traditional approach of interpreting each action as a relation (e.g., see [16]), instead we interpret each action as a set of "*events*", the events that it "*produces*" or "*participates in*" during its execution, and then the action combinators are interpreted as the classical boolean set operators. Note that the restrictions on models (**I.1** and **I.2**) imply that we have one point sets in the family of the event sets, intuitively every "event" is produced by a combination of actions in our systems (system actions and enviromental actions). Then, if we take a maximal set of actions, the execution of this set only produces an event in our system; in other words, this set of actions is complete in the sense that they describe unambiguously one event in the system execution.

This is an "*open semantics*" approach; for example: a component A can send a message to a component B, and several factors (if the network is working correctly, no other component sending another message using the same connection, and so on) will influence the result of the action. In some sense, we adopt the view that non-determinism is caused for some reason, but sometimes we just do not know which external actors will influence our actions. In some sense our specifications will be always incomplete (we can add as many external actions as we want), but we can verify specifications modulo some hypothesis (we can restrict the number of external actions). One nice thing about this semantics is that there is a strong connection (a kind of compactness property) relating weak permission and strong permission: *when an action is weakly allowed to be performed in every context, then it is strongly allowed.* This property is a key axiom of the Hilbert-style Axiomatic system that we present in [2].

The definition of the relation $\vDash$ between worlds, models, and formulae, is as follows.

- $w, M \vDash \alpha =_{act} \beta \overset{\mathsf{def}}{\Longleftrightarrow} \mathcal{I}(\alpha) = \mathcal{I}(\beta)$
- $w, M \vDash [\alpha]\varphi \overset{\mathsf{def}}{\Longleftrightarrow}$ for all $w' \in \mathcal{W}$ and $e \in \mathcal{I}(\alpha)$ if $w \xrightarrow{e} w'$ then $w', M \vDash \varphi$.
- $w, M \vDash \mathsf{P}(\alpha) \overset{\mathsf{def}}{\Longleftrightarrow}$ for all $e \in \mathcal{I}(\alpha)$, $\mathcal{P}(w, e)$ holds.
- $w, M \vDash \mathsf{P_w}(\alpha) \overset{\mathsf{def}}{\Longleftrightarrow}$ there exists some $e \in \mathcal{I}(\alpha)$ such that $\mathcal{P}(w, e)$

For the standard formulae, the definition is as usual. Note that the interpretation of the modalities is relative to states, a difference from the relational interpretation; the semantics of the complement of an action is not the absolute complement (that is, all the pairs of states that are not related by that relationship), instead the complement is relative: only the states related with the actual state are taken into account.

3 A Tableaux System

In [11], a tableaux system which uses symbols to represent worlds of the semantic models is introduced (following the approach of [17]). The main idea behind this approach is to enrich formulae with prefixes which indicate in which worlds these formulae are true. This approach is used for many other logics, and in particular in [12], where these techniques are used to reason about dynamic logic with converse. When we adapt this techniques to our logic, deontic operators fit neatly into the system; the duality between the strong and weak permission can be used to formulate a complete (and sound) tableaux proof system. In contrast with the work cited above, we use sequences of actions as labels in the formulae; these sequences allow us to build models in the case that a formula is not valid (counter-models). The action terms used in the labels have some particular characteristics: every one of these action terms describes the occurrence of a maximal set of primitive actions, that is, each of these terms describes which primitive actions are performed and which are not.

A prefixed formula has the following structure: $\sigma : \varphi$, where σ is a label made up of a sequence of boolean (action) terms built from a given vocabulary. We use the following notation for sequences: $\langle\rangle$ (*the empty sequence*), $x \centerdot xs$ (*the sequence made of an element* x *followed by a sequence* xs).

From here on we will consider a fixed vocabulary: $V = \langle \Phi_0, \Delta_0 \rangle$. Also, we will use some axiomatization of boolean algebras, denoted by Φ_{BA}; note that there exist complete and decidable axiomatizations of boolean algebras (see [18]). We denote by Δ_0/Γ the boolean terms over Δ_0 modulo a set of axioms Γ; usually, Γ is an extension of the theory of boolean algebras, i.e., $\Phi_{BA} \subseteq \Gamma$. In this case, we say that Γ is a boolean theory. We write $\Gamma \vdash_{BA} t_1 =_{act} t_2$, if the equation $t_1 =_{act} t_2$ is provable from the boolean theory Γ using equational calculus (see [19]). This implies that our method depends on some suitable method to decide boolean algebras. Using this notation, we denote by $At(\Delta_0/\Gamma)$ the set of atoms in the boolean algebra of terms modulo Γ (note that the boolean algebra is atomic because the set of primitive action symbols is finite). In the same way we denote by $At_{\sqsubseteq\alpha}(\Delta_0/\Gamma)$ the set of atoms $\gamma \in At(\Delta_0/\Gamma)$ such that $\Gamma \vdash_{BA} \gamma \sqsubseteq \alpha$, where $\sqsubseteq$ is the order of the boolean algebra of terms (and $At_{\sqsubset\alpha}(\Delta_0/\Gamma)$ denotes the strict version of this set).

Now, we can introduce the notion of tableaux.

Definition 2 (Tableaux). *A tableaux is a (n-ary) rooted tree where nodes are labelled with prefixed formulae, and a branch is a path from the root to some leaf.*

Intuitively, a branch is a tentative model for the initial formulae (those whose negation we try to prove valid). Given a branch $\mathcal{B}$, $EQ(\mathcal{B})$ is the set of equations appearing in $\mathcal{B}$.

We introduce the following (useful) classification of formulae. (Note that in the literature α is used instead of A, and β instead of B; here we do not use Greek letters to avoid confusion with action terms.)

A	A_1	A_2	B	B_1	B_2
$\sigma : \varphi \wedge \psi$	$\sigma : \varphi$	$\sigma : \psi$	$\sigma : \varphi \vee \psi$	$\sigma : \varphi$	$\sigma : \psi$
$\sigma : \neg(\varphi \vee \psi)$	$\sigma : \neg\varphi$	$\sigma : \neg\psi$	$\sigma : \neg(\varphi \wedge \psi)$	$\sigma : \neg\varphi$	$\sigma : \neg\psi$
$\sigma : \neg\neg\varphi$	$\sigma : \varphi$				

We also introduce the less standard rules: (we follow the standard notation for modal logics (see [11])), and introduce the P and N prefixed formulae (called π and ν, respectively, in the literature).

N	$N(\gamma)$	P	$P(\gamma)$
$\sigma : [\alpha]\varphi$	$\sigma \centerdot \gamma : \varphi$	$\sigma : \langle\alpha\rangle\varphi$	$\sigma \centerdot \gamma : \varphi$
$\sigma : \neg\langle\alpha\rangle\varphi$	$\sigma \centerdot \gamma : \neg\varphi$	$\sigma : \neg[\alpha]\varphi$	$\sigma \centerdot \gamma : \neg\varphi$
		$\sigma : \neg\mathsf{P}(\alpha)$	$\sigma : \neg\mathsf{P}(\gamma)$
		$\sigma : \mathsf{P_w}(\alpha)$	$\sigma : \mathsf{P_w}(\gamma)$

where $\gamma \in At(\varDelta_0/\varGamma)$ (that is, γ is an atom in the term algebra) for some boolean theory $\varGamma$. And we introduce a new classification N_D (deontic necessity) for strong permission.

N_D	$N_D(\gamma)$
$\sigma : \mathsf{P}(\alpha)$	$\sigma : \mathsf{P}(\gamma)$
$\sigma : \neg\mathsf{P_w}(\alpha)$	$\sigma : \neg\mathsf{P_w}(\gamma)$

Using the above classification of formulae, we can introduce the rules of the tableaux method. In figure 1 the classic rules for standard formulae can be found. In figure 2 we show the rules for N and N_D formulae: rule N is standard (see [9]); it does not introduce new labels in the branch, but it adds new formulae to labels already in the branch; intuitively, for all (the states denoted by) the labels reachable from the current state, the N formula must be true. On the other hand, rule N_D for deontic necessity imposes that the corresponding action must be allowed for all the possible contexts in the actual state.

$$A : \frac{A}{\begin{array}{c} A_1 \\ A_2 \end{array}} \qquad\qquad B : \frac{B}{B_1 \mid B_2}$$

Fig. 1. Classic rules for formulae of type A and B

Rule P for modal and deontic possibility is shown in figure 3; given a P formula, this rule creates one branch for each possible execution of the front action in the formula. Note that, in each branch, an inequation saying that the action must be not impossible is added, allowing us to avoid adding labels that cannot exist in the semantics. In the same figure we can see the rule Per; this rule says that if an action which is maximal (in the sense that it cannot have different executions) is weakly allowed, then it is also strongly allowed. Note that we have not shown any rule for equality; this is because equality reasoning is implicit in our calculus. For simplicity of the presentation of the concepts, we rule out those formulae of the form: $[\alpha](\alpha =_{act} \beta)$, that is, modal formulae where equality is after a modality. This does not affect the completeness of the method since formulae of this kind are equivalent to formulae without modalities (see [2]). It is straightforward to extend the method described here to manage these kinds of formulae. Now we introduce the notions of *closed*, *boolean closed*, *deontic closed* and *open branch*. Keep in mind that a branch is a set of prefixed formulae.

$$
\begin{array}{lll}
N_D : & \begin{array}{c} \dfrac{N_D}{N_D(\gamma_1)} \\ \cdot \\ \cdot \\ \cdot \\ N_D(\gamma_n) \end{array} & \text{for all } \gamma_1, ..., \gamma_n \in At_{\sqsubset \alpha}(\Delta_0/\Gamma) \\
& & \\
N : & \begin{array}{c} \dfrac{N}{N(\gamma_1)} \\ \cdot \\ \cdot \\ \cdot \\ N(\gamma_n) \end{array} & \text{for all } \gamma_1, ..., \gamma_n \in At_{\sqsubset \alpha}(\Delta_0/\Gamma) \text{ already in the branch}
\end{array}
$$

Fig. 2. Rules for deontic and modal necessity

$$
\begin{array}{lll}
P : & \dfrac{P}{\begin{array}{c} P(\gamma_1) \\ \langle\rangle : \gamma_1 \neq \emptyset \end{array} \mid ... \mid \begin{array}{c} P(\gamma_n) \\ \langle\rangle : \gamma_n \neq \emptyset \end{array}} & \text{with } \{\gamma_1, ..., \gamma_n\} = At_{\sqsubset \alpha}(\Delta_0/\Gamma) \\
& & \\
Per : & \dfrac{\sigma : \mathsf{P_w}(\gamma)}{\sigma : \mathsf{P}(\gamma)} & \text{with } \gamma \in At(\Delta_0/\Gamma)
\end{array}
$$

Fig. 3. Rules for possibility and permission

Definition 3 (deontic closed). *Given a branch $\mathcal{B}$ and a boolean theory Γ, we say that $\mathcal{B}$ is deontic closed with respect to Γ if it satisfies some of the following items:*

- $\sigma : \mathsf{P}(\gamma) \in \mathcal{B}$ *and* $\sigma : \neg\mathsf{P}(\gamma) \in \mathcal{B}$, *for some* $\gamma \in At(\Delta_0/\Gamma)$, *and some label* σ.
- $\sigma : \mathsf{P_w}(\gamma) \in \mathcal{B}$ *and* $\sigma : \neg\mathsf{P_w}(\gamma) \in \mathcal{B}$, *for some* $\gamma \in At(\Delta_0/\Gamma)$, *and some label* σ.
- $\sigma : \neg\mathsf{P}(\gamma) \in \mathcal{B}$ *and* $\sigma : \mathsf{P_w}(\gamma) \in \mathcal{B}$, *for some* $\gamma \in At(\Delta_0/\Gamma)$, *and some label* σ.

Note that we have not included $\sigma : \mathsf{P}(\gamma)$ and $\sigma : \neg\mathsf{P_w}(\gamma)$ as being mutually contradictory; this is because they are not contradictory when $\Gamma \vdash_{BA} \gamma =_{act} \emptyset$. This fact yields the next definition.

Definition 4 (extended boolean theory). *Given a branch $\mathcal{B}$, the extended boolean theory (denoted by $EQ^*(\mathcal{B})$) of $\mathcal{B}$ is defined as follows:*

$$EQ^*(\mathcal{B}) = \{(\gamma =_{act} \emptyset) \mid \gamma \in At(\Delta_0) \wedge \sigma : \mathsf{P}(\gamma), \sigma : \neg\mathsf{P_w}(\gamma) \in \mathcal{B}\} \cup EQ(\mathcal{B})$$

It is useful for us to introduce the notion of *boolean closed* branch, intuitively these branchs are inconsistent boolean theories. We call $EQ(\mathcal{B})$ the set of equations in the set $\mathcal{B}$.

Definition 5 (boolean closed branch). *A branch $\mathcal{B}$ is boolean closed iff $EQ^*(\mathcal{B}) \cup \Phi_{BA} \vdash \emptyset =_{act} \mathbf{U}$, or $EQ^*(\mathcal{B}) \vdash \alpha =_{act} \beta$ and $\alpha \neq_{act} \beta \in \mathcal{B}$*

Definition 6 (closed branch). *A branch is closed if either it has a propositional variable $\sigma : p$ and a negation of it $\sigma : \neg p$, or it is deontic closed or boolean closed.*

An *open branch* is a branch which is not closed.

We describe an algorithm to build the tableaux of a given formula; using it we can prove if a given formula is valid or not (and we prove the completeness of the proof system in section 4.1).

Algorithm 1. *Suppose that we wish to know if φ is valid. First, we put $\langle\rangle : \neg\varphi$ at the root of the tree, and then we apply the rules given above as follows. At step* 0 *we apply rules A and B, if possible, to obtain all the equations in the formula (if it contains any). Then at step N of the algorithm: if the tableaux is closed, then stop. Also, if it is not possible to apply any rule, stop. Otherwise, take the formula $\sigma : \varphi$ which occurs as close to the root as possible, which has not been used, and then apply the following steps:*

- *If A and B rules can be applied, then we apply them and add the formulae at the end of the branch.*
- *If N_D can be applied, apply it and add the resulting formulae at the end of the branch. If this application of N_D converts the actual branch to being closed (as it can add some contradictory deontic formulae as explained above), then close the branch and start again with another branch. Otherwise continue with the next step.*
- *If we cannot apply any of the above steps, then search for a P formula, if there is one, and consider the boolean theory $\Phi_{BA} \cup EQ^*(\mathcal{B})$, where $\mathcal{B}$ is the actual branch, and extend this (new version of the) branch using the P rule.*
- *If no P rule can be applied, then if it is possible to apply a N rule, apply it and extend the actual branch at the end.*

This finishes the step $N + 1$, and then we apply the procedure again. ■

This method ensures that the set of equations underlying the formulae will be obtained before starting to break down the modalities. Note that, if we have $\sigma : \neg \mathsf{P}_{\mathsf{w}}(\gamma)$ and $\sigma : \mathsf{P}(\gamma)$ and γ occurs in an existencial formula (e.g., $\langle\alpha\rangle\varphi$), then the branch will be closed because the equations $\gamma \neq_{act} \emptyset$ and $\gamma =_{act} \emptyset$ will convert the boolean theory into a closed one.

A branch which is not closed gives us a model which is a counterexample and we shall show an example of this later on.

4 Soundness and Completeness

As usual, the soundness of the tableaux system is proved by a theorem which ensures that each rule is safe (with respect to satisfiability). Towards this goal we introduce the following definitions.

Definition 7 (Mapping). *Given a set S of prefixed formulae (being F the set of prefixes occurring in it) and a model $M = \langle \mathcal{W}, \mathcal{R}, \mathcal{E}, \mathcal{I}, \mathcal{P} \rangle$ over a vocabulary $\langle \Delta_0, \Phi_0 \rangle$, a mapping is a function $\iota : F \to W$, such that:*

- *For all σ and $\sigma \centerdot \gamma$ in F, there exists $e \in \mathcal{I}(\gamma)$ such that $\iota(\sigma) \xrightarrow{e} \iota(\sigma \centerdot \gamma)$.*

Definition 8 (SAT Branch). *A branch $\mathcal{B}$ is SAT iff there exists a model M and an interpretation ι such that, for every $\sigma : \varphi$, it is the case that $\iota(\sigma), M \vDash \varphi$.*

We say that a tableaux $\mathcal{T}$ is SAT if there exists a branch in $\mathcal{T}$ which is SAT. Let us introduce a key theorem.

Theorem 1. *If $\mathcal{T}$ is a SAT tableaux, then a tableaux $\mathcal{T}'$ obtained by an application of a tableaux rule is also SAT.*

Proof. *Suppose that a branch $\mathcal{B}$ of $\mathcal{T}$ is SAT, and let $M = \langle \mathcal{W}, \mathcal{R}, \mathcal{E}, \mathcal{I}, \mathcal{P} \rangle$ be the model and ι the interpretation for $\mathcal{B}$. We prove the theorem by induction; for the A and B rules the proof is standard.*

Rule P: Suppose $\sigma : \langle \alpha \rangle \varphi \in \mathcal{B}$. And $\iota(\sigma), M \vDash \langle \alpha \rangle \varphi$, obviously $\mathcal{I}(\alpha) \neq \emptyset$, and also:

$$\exists e \in \mathcal{I}(\alpha) : \exists w' \in W : w \xrightarrow{e} w' \wedge w', M \vDash \varphi \tag{1}$$

If $\mathcal{B} \cup \{\sigma \centerdot \gamma_i : \varphi\}$ is not SAT in M, and this means for all $\gamma_i \in At(\Delta_0/\Gamma)$:

$$\begin{array}{l} \forall e \in \mathcal{I}(\gamma_i) : \forall w' \in W : w \overset{e}{\not\to} w' \vee w', M \nvDash \varphi \\ \Rightarrow \\ \forall e \in \mathcal{I}(\gamma_1) \cup ... \cup \mathcal{I}(\gamma_n) : w \overset{e}{\not\to} w' \vee w', M \nvDash \varphi \\ \Leftrightarrow \\ \forall e \in \mathcal{I}(\gamma_1 \sqcup ... \sqcup \gamma_n) : w \overset{e}{\not\to} w' \vee w', M \nvDash \varphi \\ \Leftrightarrow \\ \forall e \in \mathcal{I}(\alpha) : w \overset{e}{\not\to} w' \vee w', M \nvDash \varphi \end{array}$$

Contradicting 1.

If $\sigma : \mathsf{P_w}(\alpha) \in \mathcal{B}$ (we must have $EQ(\mathcal{B}) \nvdash \alpha \neq_{act} \emptyset$) then $\iota(\sigma), M \vDash \mathsf{P_w}(\alpha)$, and this means:

$$\exists e \in \mathcal{I}(\alpha) : \mathcal{P}(\iota(\sigma), e) \tag{2}$$

and therefore $e \in \mathcal{I}(\gamma_i)$ for some $\gamma_i \in At(\Delta_0/\Gamma)$, and:

$$\exists e \in \mathcal{I}(\gamma_i) : \mathcal{P}(\iota(\sigma), e) \tag{3}$$

and this means: $\iota(\sigma), M \vDash \mathsf{P_w}(\alpha)$.

The cases $\neg[\alpha]\varphi \in \mathcal{B}$, $\neg\mathsf{P}(\alpha) \in \mathcal{B}$ are similar to the first and second case, respectively.

Rule N: *If $\sigma : [\alpha]\varphi \in \mathcal{B}$, then $\iota(\sigma), M \vDash [\alpha]\varphi$, this means:*

$$\forall w' \in \mathcal{W}, e \in \mathcal{I}(\alpha) : \iota(\sigma) \xrightarrow{e} w' \Rightarrow w', M \vDash \varphi \qquad (4)$$

then, since for every $\gamma_i \in At(\alpha)$: $\mathcal{I}(\gamma_i) \subseteq \mathcal{I}(\alpha)$, it is the case that:

$$\forall w' \in \mathcal{W}, e \in \mathcal{I}(\gamma_i) : \iota(\sigma) \xrightarrow{e} w' \Rightarrow w', M \vDash \varphi \qquad (5)$$

Now, if $\sigma \centerdot \gamma$ is in $\mathcal{B}$ then we have $\iota(\sigma) \xrightarrow{e} \iota(\sigma \centerdot \gamma)$ (where $e \in \mathcal{I}(\gamma_i)$) by definition, and then $\iota(\sigma \centerdot \gamma_i), M \vDash \varphi$.

Rule N_D *If $\sigma : \mathsf{P}(\alpha) \in \mathcal{B}$, then we have $\iota(\sigma), M \vDash \mathsf{P}(\alpha)$, and this means: $\forall e \in \mathcal{I}(\alpha) : \mathcal{P}(\iota(\sigma), e)$. And now if we add $\sigma : \mathsf{P}(\gamma_i)$ in $\mathcal{B}$ (for all $\gamma_i \in At(\alpha)$), then if $\iota(\sigma), M \vDash \neg\mathsf{P}(\gamma_i)$ for some i, it is not hard to see that $\iota(\sigma), M \vDash \neg\mathsf{P}(\alpha)$, which is a contradiction. The only possibility is that $w, M \vDash \neg\mathsf{P}_{\mathsf{w}}(\gamma_i)$, but this just implies that $\mathcal{I}(\gamma_i) = \emptyset$, and then $\iota(\sigma), M \vDash \mathsf{P}(\gamma_i)$.* ■

The soundness of the method follows by a standard argument.

Corollary 1. *If φ is tableaux provable (i.e. there exists a closed tableau for $\neg\varphi$) then $\vDash \varphi$.* ■

4.1 Completeness

Towards the proof of completeness we introduce the notions of *front action* and *Hintikka sets*:

Definition 9 (Front Action). *Given a N or P formula φ and considering its syntactical tree, the front action is the nearest action to the root of this tree. For example for the formula $[\alpha]([\beta]\varphi \wedge \langle\gamma\rangle\psi)$ its front action is α.*

Definition 10 (Hintikka Sets). *Let S be a set of prefixed formulae and F the set of prefixes in S. We say that S is Hintikka iff:*

- *S is not closed.*
- *If $\sigma : \mathsf{P}(\alpha)$ and $\sigma : \neg\mathsf{P}_{\mathsf{w}}(\alpha) \in S$, then $EQ(S) \vdash_{BA} \alpha =_{act} \emptyset$*
- *If $A \in S$, then $A_1 \in S$ and $A_2 \in S$.*
- *If $B \in S$, then either $B_1 \in S$ or $B_2 \in S$, or both.*
- *If $N \in S$, then $N(\gamma_i) \in S$, for all $\gamma_i \in At_{\sqsubseteq\alpha}(\Delta_0/EQ(S))$ (where α is the front action in N) or $EQ(S) \vdash_{BA} \alpha =_{act} \emptyset$.*
- *If $P \in S$, then $P(\gamma_i) \in S$, for some $\gamma_i \in At(\alpha)$ (where α is the front action in P) and $EQ(s) \nvdash_{BA} \alpha =_{act} \emptyset$.*
- *If $N_D \in S$, then $N_D(\gamma_i)$ for all $\gamma_i \in At_{\sqsubseteq\alpha}(\Delta_0/EQ(S))$, or $EQ(S) \vdash_{BA} \alpha =_{act} \emptyset$.*
- *If $\sigma : \mathsf{P}_{\mathsf{w}}(\gamma_i) \in S$ for some $\gamma_i \in At(\Delta_0/EQ(S))$, then $\sigma : \mathsf{P}(\alpha) \in S$.*

Now, we prove that any Hintikka set is SAT.

Theorem 2. *Any Hintikka set is SAT.*
Proof. *Given a Hintikka set S we define the following model:*

- $\mathcal{W} = \{\sigma \mid \sigma : \varphi \in S$, *for some formula* $\varphi\}$
- $\mathcal{E} = At(\Delta_0/\Gamma)$, *where* $\Gamma = \Phi_{BA} \cup EQ(S)$.
- $\mathcal{R} = \{\sigma \overset{[\gamma]}{\to} \sigma \centerdot \gamma \mid \sigma, \sigma \centerdot \gamma \in \mathcal{W}\}$
- $p \in \mathcal{I}(w) \Leftrightarrow (\sigma : p) \in S$
- $\mathcal{I}(\alpha) = At_{\sqsubseteq\alpha}(\Delta_0/\Gamma)$
- $\mathcal{P} = \{(\sigma, [\gamma]) \mid (\sigma : \mathsf{P}(\gamma)) \in S \wedge \gamma \in \mathcal{E}\}$

We must prove that this is a model for S. We define the mapping ι as follows: $\iota(\sigma) = \sigma$ (the identity function). Let us prove that it is really a model. The proof is by induction.
Base Case: *Obviously, if $\sigma : p \in S$ then $\sigma, M \vDash p$. We cannot have both p and $\neg p$ in S, and therefore the definition for propositional variables is correct.*

If $\sigma : \alpha =_{act} \beta \in S$, then $EQ(S) \vdash_{BA} \alpha =_{act} \beta$ and therefore $At(\alpha) = At(\beta)$.
Ind.Case: *We prove this by cases:*

A rule: *If $A \in S$ then A_1 and A_2 are both in S, and the result follows by the definition of our model.*

B rule: *Similar to the A rule case.*

N rule: *If $(\sigma : [\alpha]\varphi) \in S$, and $EQ(S) \vdash_{BA} \alpha =_{act} \emptyset)$, then $At_{\sqsubseteq\alpha}(\Delta_0/\Gamma) = \emptyset$ and therefore $\sigma, \mathcal{M} \vDash [\alpha]\varphi$. If $EQ(S) \nvdash_{BA} \alpha =_{act} \emptyset$, then $\sigma \centerdot \gamma_i : \varphi$ for all $\gamma_i \in F$ and $\gamma_i \in At_{\sqsubseteq\alpha}(\Delta_0/\Gamma)$, and therefore $\forall e \in \mathcal{I}(\alpha) : \sigma \overset{e}{\to} \sigma\gamma_i \Rightarrow \sigma \centerdot \gamma_i \vDash \varphi$.*
P rule *If $\sigma : \langle\alpha\rangle\varphi \in S$, then $\sigma \centerdot \gamma_i : \varphi$ for some $\gamma_i \in At_{\sqsubseteq\alpha}(\Delta_0/\Gamma)$, thus $\sigma \centerdot \gamma_i, \mathcal{M} \vDash \varphi$ and then $\sigma, \mathcal{M} \vDash \langle\alpha\rangle\varphi$.*

For $\sigma : \mathsf{P_w}(\alpha) \in S$, then $\sigma : \mathsf{P_w}(\gamma_i)$ for some $\gamma_i \in At_{\sqsubseteq\alpha}(\Delta_0/\Gamma)$ and then by definition of Hintikka sets $\sigma : \mathsf{P}(\gamma_i)$, which implies by definition of $\mathcal{M}$: $\sigma, \mathcal{M} \vDash \mathsf{P_w}(\alpha)$.
N_D rule: *If $\sigma : \mathsf{P}(\alpha) \in S$, then if $EQ(S) \vdash_{BA} \alpha =_{act} \emptyset$, then $At_{\sqsubset\alpha}(\Delta_0/\Gamma) = \emptyset$ and therefore $\sigma, \mathcal{M} \vDash \mathsf{P}(\alpha)$. Otherwise, $At_{\sqsubseteq\alpha}(\Delta_0/\Gamma) \neq \emptyset$ and then for all $\gamma_i \in At_{\sqsubseteq\alpha}(\Delta_0/\Gamma)$ occurrs $\sigma : \mathsf{P}(\gamma_i) \in S$, and this implies $\sigma, \mathcal{M} \vDash \mathsf{P}(\alpha)$.* ■

Let us introduce the definition of *completed branch*.

Definition 11. *A branch is completed if we have applied algorithm 1 to it.*

Note that, given an open branch $\mathcal{B}$, we can always extend it to a *completed* open branch by theorem 1 (the algorithm just applies tableaux rules).

From here we can prove completeness; first we prove that a completed open branch is a Hintikka set.

Theorem 3. *If $\mathcal{B}$ is a completed open branch, then it is a Hintikka set.*

Proof. *The result follows trivially from the definition of completed open branch and the definition of algorithm 1.* ■

Completeness follows, i.e, *every valid formula has a closed tableau.*

5 An Example

In this section we show a simple example of an application which gives some intuition about how to use the proof system in practice. We use a well-known *contrary-to-duty* paradox: the *Reagan-Gorbachov* example introduced in [20]. Contrary-to-duty reasoning is inherent in fault-tolerant systems, where we have secondary obligations arising from violations of primary norms.

One interesting aspect of deontic action logic is that several paradoxes of standard deontic logic are no longer paradoxical, or they cannot even be expressed in DAL (see [14]). The *Reagan-Gorbachov* paradox is one of them: *you ought not to tell the secret to Reagan or Gorbachov*; *if you tell the secret to Reagan, you have to tell it to Gorbachov*; *if you tell the secret to Gorbachov; you must tell it to Reagan*; *you tell the secret to Reagan*. In standard deontic logic, we get a contradiction. But in action logics the paradox vanishes, and we can formulate it as follows: $\mathsf{O}(\overline{r} \sqcap \overline{g})$, $[r]\mathsf{O}(g)$, $[g]\mathsf{O}(r)$, where g and r are the obvious actions. Note that, in the logic, we cannot express that an action is performed (we can do it in the extension given in [1]); instead we can prove that it is possible to tell the secret to Reagan or to Gorbachov (i.e., this action will not yield an inconsistency). To do that we add the sentence $\langle r \cup g \rangle \top$. If we use the tableaux with these formulae, we get a model telling us that this set of formulae is not inconsistent (i.e., the negation (of the conjunction) of these formulae is not valid).

In figure 4 we can see the tableaux for the example; we consider s the action of keeping the secret and we have added the following equations to the set of formulae: $r \sqcap g =_{act} \emptyset$ (*you cannot tell Gobarchov and Reagan at the same time*), $r \sqcap s =_{act} \emptyset$ (*you cannot tell Reagan and keep the secret*) and $g \sqcap s =_{act} \emptyset$ (*you cannot tell Gorbachov and keep the secret*). With these restrictions we have: $At(\Delta_0/\Gamma) = \{r \sqcap \overline{s} \sqcap \overline{g}, g \sqcap \overline{s} \sqcap \overline{r}, s \sqcap \overline{r} \sqcap \overline{g}\}$. Note that we do not put these equations on the tableau to save space. The tableaux is built as follows: first we put the original set of formulae in the root of the tree (labelled with the empty sequence of actions), and after that we just use the definition of obligation and the α rule, and we add lines 5 and 6. And then, using the rules of deontic necessity, we obtain the rest of the formulae in the root. Therefore, we can use rule P on line 4, obtaining the two branchs, and then we apply rule N in the formula in line 2 of the root, this gives us the second formula of the left branch, and after that we apply the definition of obligation, the α rule and the N_D rule, leaving the branch open. The same procedure is applied to the right branch.

By completeness, the existence of open branches implies that the set of formulae is satisfable, indeed the completeness proof gives us a method to build a model, and using it we can build the model shown in figure 5. In this graphic, we denote with dashed lines the forbidden actions in a given world; that is, in the state w it is forbidden to perform g or r, but if we tell the secret to Reagan then we ought to tell the secret to Gorbachov, and vice versa. This model is intuitively correct. The example is simple but it shows how the proof system can be used in practice, it is useful not only for verification of properties, but also to discover bad behaviours that can yield a faulty state.

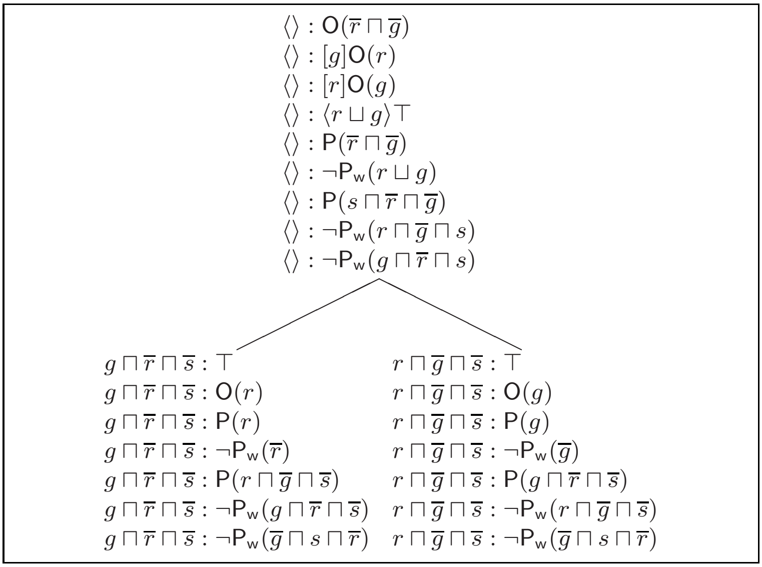

$\langle\rangle : \mathsf{O}(\overline{r} \sqcap \overline{g})$
$\langle\rangle : [g]\mathsf{O}(r)$
$\langle\rangle : [r]\mathsf{O}(g)$
$\langle\rangle : \langle r \sqcup g\rangle\top$
$\langle\rangle : \mathsf{P}(\overline{r} \sqcap \overline{g})$
$\langle\rangle : \neg\mathsf{P}_\mathsf{w}(r \sqcup g)$
$\langle\rangle : \mathsf{P}(s \sqcap \overline{r} \sqcap \overline{g})$
$\langle\rangle : \neg\mathsf{P}_\mathsf{w}(r \sqcap \overline{g} \sqcap s)$
$\langle\rangle : \neg\mathsf{P}_\mathsf{w}(g \sqcap \overline{r} \sqcap s)$
$g \sqcap \overline{r} \sqcap \overline{s} : \top$
$g \sqcap \overline{r} \sqcap \overline{s} : \mathsf{O}(r)$
$g \sqcap \overline{r} \sqcap \overline{s} : \mathsf{P}(r)$
$g \sqcap \overline{r} \sqcap \overline{s} : \neg\mathsf{P}_\mathsf{w}(\overline{r})$
$g \sqcap \overline{r} \sqcap \overline{s} : \mathsf{P}(r \sqcap \overline{g} \sqcap \overline{s})$
$g \sqcap \overline{r} \sqcap \overline{s} : \neg\mathsf{P}_\mathsf{w}(g \sqcap \overline{r} \sqcap \overline{s})$
$g \sqcap \overline{r} \sqcap \overline{s} : \neg\mathsf{P}_\mathsf{w}(\overline{g} \sqcap s \sqcap \overline{r})$
$r \sqcap \overline{g} \sqcap \overline{s} : \top$
$r \sqcap \overline{g} \sqcap \overline{s} : \mathsf{O}(g)$
$r \sqcap \overline{g} \sqcap \overline{s} : \mathsf{P}(g)$
$r \sqcap \overline{g} \sqcap \overline{s} : \neg\mathsf{P}_\mathsf{w}(\overline{g})$
$r \sqcap \overline{g} \sqcap \overline{s} : \mathsf{P}(g \sqcap \overline{r} \sqcap \overline{s})$
$r \sqcap \overline{g} \sqcap \overline{s} : \neg\mathsf{P}_\mathsf{w}(r \sqcap \overline{g} \sqcap \overline{s})$
$r \sqcap \overline{g} \sqcap \overline{s} : \neg\mathsf{P}_\mathsf{w}(\overline{g} \sqcap s \sqcap \overline{r})$

Fig. 4. Gorbachov-Reagan example

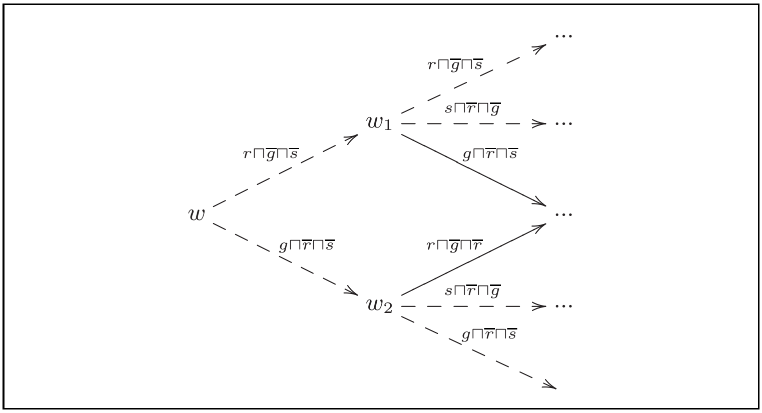

w
w_1
w_2
$r\sqcap\overline{g}\sqcap\overline{s}$
$r\sqcap\overline{g}\sqcap\overline{s}$
$s\sqcap\overline{r}\sqcap\overline{g}$
$g\sqcap\overline{r}\sqcap\overline{s}$
$g\sqcap\overline{r}\sqcap\overline{s}$
$r\sqcap\overline{g}\sqcap\overline{r}$
$s\sqcap\overline{r}\sqcap\overline{g}$
$g\sqcap\overline{r}\sqcap\overline{s}$
...
...
...
...

Fig. 5. A model for the Gorbachov-Reagan paradox

6 Conclusions and Further Work

In this paper we have presented a tableaux system for the deontic action logic presented in [2], we proposed this logic to reason about fault-tolerant systems and we have described some examples in [1] and [7]. The tableaux system presented here allows us to decide the logic (obviously the decision procedure is exponential) and also to build counter-examples. This system deals with some classic deontic predicates (*weak permission*, *strong permission* and *obligation*) and also with complement and parallel execution of actions. In the literature, some tableaux systems have been proposed for deontic logics: in [21] a tableaux system is provided for *deontic conditional systems* and in [22] a tableaux system is described for *deontic interpreted systems.* It seems that no tableaux systems have yet been proposed for *deontic action logics* (or dynamic deontic logics), perhaps because of the difficulties of dealing with parallel execution of actions and the complement of actions; in this sense the logic and the system provided here seems to be novel.

On the other hand, as is demonstrated in the example, the logic depends on the set of primitive actions considered, some properties may stop being valid when the number of actions is increased (as different scenarios might be considered); however, it is possible that, by analyzing the formulae to be proven, a bound on the number of different actions to consider could be obtained and probably this bound can be calculated taking into account only the possibility modalities in these formulae. We leave this topic to further research.

Finally, we want to extend the logic to support different sorts of obligations (in a similar way as is done in [4]), in such a way that it allows us to formalize the notions of recovery actions, which arise in contrary-to-duty reasoning and fault-tolerance. This extension can be done in a way that preserves the basic properties of the logic. Also, including temporal operators in this proof system is possible; with these extensions the proof system can be used to specify complex problems. However, a software tool must be built to apply the method in practice.

References

1. Castro, P., Maibaum, T.: An ought-to-do deontic logic for reasoning about fault-tolerance: The diarrheic philosophers. In: 5th IEEE International Conference on Software Engineering and Formal Methods. IEEE, Los Alamitos (2007)
2. Castro, P., Maibaum, T.: A complete and compact deontic action logic. In: Jones, C.B., Liu, Z., Woodcock, J. (eds.) ICTAC 2007. LNCS, vol. 4711, pp. 109–123. Springer, Heidelberg (2007)
3. Kent, S., Quirk, B., Maibaum, T.: Specifying deontic behavior in modal action logic. Technical report, Forest Research Project (1991)
4. Meyer, J.: A different approach to deontic logic: Deontic logic viewed as variant of dynamic logic. Notre Dame Journal of Formal Logic 29 (1988)
5. Fiadeiro, J., Maibaum, T.: Temporal theories as modularization units for concurrent system specification. Formal Aspects of Computing 4, 239–272 (1992)
6. Broersen, J.: Modal Action Logics for Reasoning about Reactive Systems. PhD thesis, Vrije University (2003)
7. Castro, P., Maibaum, T.: Reasoning about system-degradation and fault-recovery with deontic logic. In: Workshop on Methods, Models and Tools for Fault-Tolerance (2007)
8. Smullyian, R.: First-Order Logic. Springer, New York (1968)
9. Fitting, M.: First-Order Logic and Automated Theorem Proving. Springer, Heidelberg (1990)
10. Pratt, V.: A Practical Decision Method for Propositional Dynamic Logic. In: ACM Symposium on Theory of Computing (1978)
11. Fitting, M.: Tableau methods of proof for modal logics. Notre Dame Journal of Formal Logic XIII (April 1972)

12. Giacomo, G., Massacci, F.: Tableaux and algorithms for propositional dynamic logic with converse. In: Conference on Automated Deduction (1996)
13. Maibaum, T.: Temporal reasoning over deontic specifications. In: Sons, J.W. (ed.) Deontic Logic in Computer Science (1993)
14. Meyer, J., Wieringa, R., Dignum, F.: The paradoxes of deontic logic revisited: A computer science perspective. Technical Report UU-CS-1994-38, Utrecht University (1994)
15. Khosla, S., Maibaum, T.: The prescription and description of state-based systems. In: Banieqnal, B., Pnueli, H.A. (eds.) Temporal Logic in Computation. Springer, Heidelberg (1985)
16. Gargov, G., Passy, S.: A note on boolean logic. In: Petkov, P.P. (ed.) Proceedings of the Heyting Summerschool. Plenum Press (1990)
17. Fitch, F.B.: Tree proofs in modal logics. Journal of Symbolic Logic (1966)
18. Monk, J.: Mathematical Logic. Graduate Texts in Mathematics. Springer, Heidelberg (1976)
19. Ehrig, H., Mahr, B.: Fundamentals of Algebraic Specification 1: Equations and Initial Semantics. Springer, Heidelberg (1985)
20. Belzer, M.: Legal reasoning in 3-d. In: ICAIL, pp. 155–163 (1987)
21. Artosi, A., Governatori, G.: A tableau methodology for deontic conditional. In: DEON 1998, International Workshop on Deontic Logic in Computer Science, pp. 65–81 (1998)
22. Governatori, G., Lomuscio, A., Sergot, M.J.: A tableaux system for deontic interpreted systems. In: Gedeon, T.D., Fung, L.C.C. (eds.) AI 2003. LNCS (LNAI), vol. 2903, pp. 339–351. Springer, Heidelberg (2003)

Information Security Economics - and Beyond

Ross Anderson

University of Cambridge

The economics of information security has recently become a thriving and fast-moving discipline. As distributed systems are assembled from machines belonging to principals with divergent interests, incentives are becoming as important to dependability as technical design. The new field provides valuable insights not just into security topics such as privacy, bugs, spam, and phishing, but into more general areas such as system dependability (the design of peer-to-peer systems and the optimal balance of effort by programmers and testers), and policy (particularly digital rights management). This research program has been starting to spill over into more general security questions (such as law-enforcement strategy), and into the interface between security and the social sciences. Most recently it has started to interact with psychology, both through the psychology-and-economics tradition and in response to phishing. The promise of this research program is a novel framework for analyzing information security problems – one that is both principled and effective.

R. van der Meyden and L. van der Torre (Eds.): DEON 2008, LNAI 5076, p. 49, 2008.

Trust and Norms in the Context of Computer Security: A Logical Formalization[*]

Emiliano Lorini and Robert Demolombe

Institut de Recherche en Informatique de Toulouse (IRIT), France
lorini@irit.fr,
robert.demolombe@orange.fr

Abstract. In this paper we present a logical model of trust in which trust is conceived as an expectation of the truster about some properties of the trustee. A general typology of trust is presented. We distinguish trust in the trustee's action from trust in the trustee's disposition (motivational or normative disposition); positive trust from negative trust. A part of the paper is devoted to the formalization of security properties and to the analysis of their relationships with trust.

1 Introduction

Techniques of computer security have been mainly designed in the perspective of protecting a computer system with respect to attacks of ill-intentioned users who want, for example, to access private data. To prevent these situations techniques have been developed, like cryptography, in order to reduce risks and to make that standard users trust the computer system. However, another kind of scenario may happen where the computer system has been designed to violate some regulations about privacy. For example, private data gathered for some applications may be sold to a company for advertising without users' authorization. In these kinds of scenario, even if the computer system guarantees that ill-intentioned users have no capability to violate the norms, standard users want to trust the computer system about the fact that it will not intentionally violate the norms. In this perspective, the issue is not to trust the effectiveness of computer science techniques (like cryptography) but to trust the fact that norms are not deliberatively violated by the system. That was the initial motivation of the work presented in this paper.

Since trust is a complex mental attitude, the first step was to propose a clear definition in a logical framework which is presented in section 2. In section 3 we present a global view of trust in order to point out several refinements of this concept. In section 4 we focus on the trustee's intention to do, or not to do, a certain action for the truster. Then, in section 5, we refine this approach by analyzing the disposition of the trustee to perform a certain action for the truster which is called *willingness*. In section 6 computer security properties are defined and their normative dimension is discussed. That leads to define in section 7 the normative dispositions of the trustee toward the truster which are called *obedience* and *honesty*.

[*] This work is supported by the project "ForTrust: social trust analysis and formalization" funded by the french Agence Nationale de la Recherche (ANR).

R. van der Meyden and L. van der Torre (Eds.): DEON 2008, LNAI 5076, pp. 50–64, 2008.

2 A Logic for Trust Reasoning

The logic $\mathcal{L}$ we use to formalize the relevant concepts involved in our model of social trust is a multimodal logic which combines the expressiveness of a simple dynamic logic [13] with the expressiveness of a logic of mental attitudes [6,21] and obligations[1,3]. The syntactic primitives of the logic $\mathcal{L}$ are the following:

- a nonempty finite set of agents $AGT = \{i, j, \ldots\}$;
- a nonempty finite set of atomic actions $ACT = \{\alpha, \beta, \ldots\}$;
- a set of atomic formulas $ATM = \{p, q, \ldots\}$.

The language of $\mathcal{L}$ is the set of formulas defined by the following BNF:

$$\phi ::= p \mid \neg\phi \mid \phi \vee \phi \mid After_{i:\alpha}\phi \mid Does_{i:\alpha}\phi \mid Bel_i\phi \mid Goal_i\phi \mid Obg\phi$$

where p ranges over ATM, α ranges over ACT and i ranges over AGT.

The operators of our logic have the following intuitive meaning. $Bel_i\phi$: the agent i believes that ϕ; $After_{i:\alpha}\phi$: after agent i does α, it is the case that ϕ ($After_{i:\alpha}\bot$ is read: agent i cannot do action α); $Does_{i:\alpha}\phi$: agent i is going to do α and ϕ will be true afterward ($Does_{i:\alpha}\top$ is read: agent i is going to do α); $Goal_i\phi$: the agent i wants that ϕ holds; $Obg\phi$: it is obligatory that ϕ. The following abbreviations are given: $Can_i(\alpha) \overset{\text{def}}{=} \neg After_{i:\alpha}\bot$; $Int_i(\alpha) \overset{\text{def}}{=} Goal_i Does_{i:\alpha}\top$; $Perm\phi \overset{\text{def}}{=} \neg Obg\neg\phi$. We write $Can_i(\alpha)$ as an abbreviation of $\neg After_{i:\alpha}\bot$ in order to make explicit the fact that $\neg After_{i:\alpha}\bot$ stands for: agent i can do action α (i.e. i has the capacity/ability to do α). $Int_i(\alpha)$ stands for: the agent i intends to do α. $Perm\phi$ stands for: ϕ is permitted.

Models of our logic are tuples $M = \langle W, R, D, B, G, O, V\rangle$ where:

- W is a non empty set of possible worlds or states.
- R is a collection of binary relations $R_{i:\alpha}$ on W, one for every couple $i{:}\alpha$ where $i \in AGT$ and $\alpha \in ACT$. Given an arbitrary world $w \in W$, if $(w, w') \in R_{i:\alpha}$ then w' is a world which can be reached from world w through the occurrence of agent i's action α.
- D is a collection of binary relations $D_{i:\alpha}$ on W, one for every couple $i{:}\alpha$ where $i \in AGT$ and $\alpha \in ACT$. Given an arbitrary world $w \in W$, if $(w, w') \in D_{i:\alpha}$ then w' is the *next* world of w which will be reached from w through the occurrence of agent i's action α.
- B is a collection of binary relations B_i on W, one for every agent $i \in AGT$. Given an arbitrary world $w \in W$, if $(w, w') \in B_i$ then w' is a world which is compatible with agent i's beliefs at world w.
- G is a collection of binary relations G_i on W, one for every agent $i \in AGT$. Given an arbitrary world $w \in W$, if $(w, w') \in G_i$ then w' is a world which is compatible with agent i's goals at world w.
- O is a binary relation on W. Given an arbitrary world $w \in W$, if $(w, w') \in O$ then w' is a world which is ideal at world w.
- $V : ATM \longrightarrow 2^W$ is a valuation function.

Truth conditions for atomic formulas, negation and disjunction are entirely standard. The following are truth conditions for the modal operators introduced before.

- $M, w \models After_{i:\alpha}\phi$ iff $M, w' \models \phi$ for all w' such that $(w, w') \in R_{i:\alpha}$.
- $M, w \models Does_{i:\alpha}\phi$ iff $\exists w'$ such that $(w, w') \in D_{i:\alpha}$ and $M, w' \models \phi$.
- $M, w \models Bel_i\phi$ iff $M, w' \models \phi$ for all w' such that $(w, w') \in B_i$.
- $M, w \models Goal_i\phi$ iff $M, w' \models \phi$ for all w' such that $(w, w') \in G_i$.
- $M, w \models Obg\phi$ iff $M, w' \models \phi$ for all w' such that $(w, w') \in O$.

2.1 Properties of the Operators

The operators Bel_i, $After_{i:\alpha}$, $Does_{i:\alpha}$ $Goal_i$ and Obg are supposed to be normal modal operators satisfying standard axioms and rules of inference of system K. Operators for belief of type Bel_i are supposed to be $KD45$ modal operators, whilst every operator for goal of type $Goal_i$ is supposed to be a KD operator. Thus, we make assumptions about positive and negative introspection for beliefs and we suppose that beliefs and goals cannot be inconsistent. Operators for obligations of type Obg are also supposed to be KD as in SDL (standard deontic logic) [1].[1]

As far as actions are concerned, we assume that actions of the same agent and actions of different agents occur in parallel.

Alt$_{Act}$ $Does_{i:\alpha}\phi \rightarrow \neg Does_{j:\beta}\neg\phi$

Axiom **Alt**$_{Act}$ says that: if i is going to do α and ϕ will be true afterward, then it cannot be the case that j is going to do β and $\neg\phi$ will be true afterward.

We also suppose that the world is never static in our framework, that is, we suppose that always there exists some agent i and action α such that i is going to perform α.

Active $\bigvee_{i \in AGT, \alpha \in ACT} Does_{i:\alpha}\top$

Axiom **Active** ensures that for every world w there is a *next* world of w which is reachable from w by the occurrence of some action of some agent. This is the reason why the operator X for *next* of LTL (linear temporal logic) can be defined as follows:[2]

$$X\phi \overset{\text{def}}{=} \bigvee_{i \in AGT, \alpha \in ACT} Does_{i:\alpha}\phi$$

The following Axiom **Inc**$_{Act}$ relates the operator $Does_{i:\alpha}$ with the operator $After_{i:\alpha}$.

Inc$_{Act,PAct}$ $Does_{i:\alpha}\phi \rightarrow \neg After_{i:\alpha}\neg\phi$

According to **Inc**$_{Act,PAct}$, if i is going to do α and ϕ will be true afterward, then it is not the case that $\neg\phi$ is true after i does α.

The following axioms relating intentions with actions seem quite natural in the case of intentional actions.

IntAct1 $(Int_i(\alpha) \wedge Can_i(\alpha)) \rightarrow Does_{i:\alpha}\top$
IntAct2 $Does_{i:\alpha}\top \rightarrow Int_i(\alpha)$

[1] Semantic constraints corresponding to the axioms presented in this section are given in [12].

[2] Note that X satisfies the standard property $X\phi \leftrightarrow \neg X \neg\phi$.

According to **IntAct1**, if i has the intention to do action α and has the capacity to do α, then i is going to do α. According to **IntAct2**, an agent is going to do action α only if he has the intention to do α. In this sense we suppose that an agent's *doing* is by definition intentional. Similar axioms have been studied in [20,19] in which a logical model of the relationships between intention and action performance is proposed.

As far as beliefs and goals are concerned, we only suppose that the two kinds of mental attitudes must be compatible, that is, if an agent has the goal that ϕ he cannot believe that $\neg\phi$. Indeed, the notion of goal we characterize is a notion of an agent's *chosen goal*, i.e. a goal that an agent decides to pursue. As some authors have stressed [2], a rational agent cannot decide to pursue a certain state of affairs ϕ, if he believes that $\neg\phi$ (this is called *weak realism* hypothesis).

WR $\quad Goal_i\phi \rightarrow \neg Bel_i\neg\phi$

In this work we also assume positive and negative introspection over (chosen) goals, that is:

PIntr $\quad Goal_i\phi \rightarrow Bel_iGoal_i\phi$
NIntr $\quad \neg Goal_i\phi \rightarrow Bel_i\neg Goal_i\phi$

The following axiom relates obligations with beliefs:

BelObg $\quad Obg\phi \rightarrow Bel_iObg\phi$

This axiom is based on the assumption that every agent has complete information of what is obligatory. It is justified by the fact that if it is expected that an agent does every action which is obligatory, he must have a complete information about what is obligatory. Note that by Axiom **BelObg**, the definition of the permission operator $Perm$ and Axiom D for Bel_i, the following formula can be derived as a consequence: $Bel_iPerm\phi \rightarrow Perm\phi$. This means that in our logical framework every agent has sound information of what is permitted.

We call $\mathcal{L}$ the logic axiomatized by the axioms and rules of inference presented above. We write $\vdash \varphi$ if formula φ is a Theorem of $\mathcal{L}$.

3 A Global View of Trust

In the present logical model trust is conceived as a complex configuration of mental states in which there is a main and primary motivational component (the principal reason activating the truster's delegating behavior): the goal to achieve some state of affairs φ (the trust in the trustee is always relative to some interest, need, concern, desire of the truster); and a complex configuration of truster's beliefs about the qualities of the trustee. On this point we agree with Castelfranchi & Falcone [4,5] on the fact that a model of social trust must account for the truster's attribution process, that is, it must account for the truster's ascription of specific properties to the trustee (abilities, willingness, dispositions, etc.) and the truster's ascription of properties to the environment in which the trustee is going to act (will the environmental conditions prevent the trustee from accomplishing the task that the truster has delegated to him?). From this perspective there is a pressing need for elaborating richer models of social trust in which the

Table 1. Typology of Trust

	Trust about action	**Trust about disposition**	
		Motivational	**Normative**
Positive	*i trusts j to do* α	*i trusts j to be willing to do* α *for him*	*i trusts j to be obedient to do* α
Negative	*i trusts j not to do* α	*i trusts j to be willing not to do* α *for him*	*i trusts j to be honest to do* α

truster's expectation and its components are explicitly modeled. To this end, we present in the following sections a conceptual and logical model of social trust which shows that trust is not a unitary and simplistic notion. More precisely, we assume that i's trust in agent j necessarily involves a main and primary motivational component which is a goal of the truster. If i trusts agent j then necessarily i trusts j with respect to some of his goals. Moreover, the core of trust is a belief of the truster about some properties of the trustee, that is, if i trusts agent j then necessarily i trusts j because i has some goal and believes that j has the right properties to ensure that such a goal will be achieved. The aim of the following sections is to clarify the nature of such a belief of the truster.

We also claim that there is no unique definition of trust, but there are several types of trust depending on the kinds of properties that the truster ascribes to the trustee. The ontology of trust proposed in the following sections is organized according to two main dimensions (see Table 1). First, we distinguish between *positive trust* and *negative trust*. In positive trust i is focused on the domain of gains (goal achievements) whereas in negative trust i is focused on the domain of losses (goal frustrations). The second distinction is between *trust in the trustee's actions* and *trust in the trustee's dispositions*. In the former case, i's trust in j is based on i's belief that j will perform (resp. refrain from performing) a certain action α; whereas in the latter case i's trust in j is based on i's belief that j is disposed to perform (resp. to refrain from performing) a certain action α. By combining the previous two dimensions we characterize four general categories of trust.

- i trusts j because i believes that j can help him to achieve a certain goal by performing a certain action α and j is going to perform action α (*i's positive trust in j's action*);
- i trusts j because i believes that j is in the condition to damage him (i.e. to frustrate a goal of i) by doing a certain action α and j will refrain from performing action α (*i's negative trust in j's action*);
- i trusts j because i believes that j can help him to achieve a certain goal by performing a certain action α and j is disposed to perform action α (*i's positive trust in j's disposition*);
- i trusts j because i believes that j is in the condition to damage him (i.e. to frustrate a goal of i) by doing a certain action α and j is disposed to refrain from performing action α (*i's negative trust in j's disposition*).

We introduce a further sophistication by distinguishing between motivational dispositions and normative (or moral) dispositions of the trustee. Indeed, in the context of i's positive trust in j's disposition (resp. i's negative trust in j's disposition), j's disposition to perform a certain action α (resp. j's disposition to refrain from performing a certain action α), can be interpreted in two different ways. According to the motivational interpretation, i's belief that j is disposed to perform action α (resp. j is disposed to refrain from performing action α) stands for i's belief that j is willing to do action α for him (resp. j is willing not to do action α for him). According to the normative interpretation, i's belief that j is disposed to perform action α (resp. j is disposed to refrain from performing action α) stands for i's belief that j will obey to the obligation of doing action α (resp. will not perform action α if he has no permission to perform action α). Thus, our ontology of trust gets refined in such a way that we can distinguish two different types of i's positive trust in j's disposition and two different types of i's negative trust in j's disposition. Namely: *i's positive trust in j's motivational disposition*, *i's negative trust in j's motivational disposition*, *i's positive trust in j's moral disposition*, *i's negative trust in j's moral disposition*.

The concepts of positive and negative trust in the trustee's action are studied in section section 4. Section 5 is devoted to the analysis of positive and negative trust in the trustee's motivational disposition. The reader must wait until section 7 for positive and negative trust in the trustee's normative disposition.

3.1 Some Related Works

Our logical model of trust shares some intuitions with Castelfranchi & Falcone's conceptual and informal model of trust [4,5]. As emphasized in the previous section, we agree with them that trust should not be seen as an unitary and simplistic notion as other models implicitly suppose. For instance, there are computational models of trust in which trust is conceived as an expectation sustained by the repeated direct interactions with other agents under the assumption that iterated experiences of success strengthen the trustor's confidence [17]. More sophisticated models of social trust have been developed in which reputational information is added to information obtained via direct interaction (e.g. [14]). All these trust models are in our view over-simplified since they do not consider the indirect supports for the trust expectation. Trust is rather a complex expectation of the truster about some properties of trustee which are relevant for the achievement of goal of the truster.

Nevertheless, there are important difference between our model of trust and Castelfranchi & Falcone's model. For instance, we think that their model of trust is not sufficiently clear in distinguishing trust in the trustee's actions and trust in the trustee's willingness. This distinction is for us fundamental since it allows to capture two forms of trust which have different natures. Moreover, their model only account for positive trust and do not consider negative trust.

As far as logics of trust are concerned, we think that there is still no comprehensive logical model of this social phenomenon. Indeed, logical models of trust have been focused almost exclusively on trust in information sources (informational trust) [18,16,10,8], or they have reduced trust to a certain kind of beliefs neglecting the motivational aspects of trust [9]. In [9] trust is defined as a truster's sort of belief, called "strong belief", about some properties of the trustee. They may be epistemic properties, like sincerity or competence, dynamic properties, like ability, or deontic properties like obedience and honesty. From this perspective there is a pressing need for elaborating more general logical models of social trust in which the truster's expectation and its different components are explicitly modeled, and in which the motivational aspect of trust is taken into account.

4 Trust in the Trustee's Action

We first define the notion of positive trust in the trustee's action. Such a notion presents four different arguments: truster, trustee, truster's goal, trustee's action.

Definition 1 ***POSITIVE TRUST ABOUT ACTION.*** *i trusts j to do α with regard to his goal that ϕ if and only if i wants ϕ to be true and i believes that:*[3]

1. *j, by doing α, will ensure that ϕ AND*
2. *j has the capacity to do α AND*
3. *j intends to do α*

Condition 1 concerns the trustee's power to satisfy the truster's goal that ϕ by means of the performance of action α. Conditions 2 and 3 are about the trustee's properties which are necessary and sufficient for him to perform action α. The formal translation of Definition 1 is:

$$ATrust(i, j, \alpha, \phi) \overset{\text{def}}{=} Goal_i X\phi \wedge Bel_i(After_{j:\alpha}\phi \wedge Can_j(\alpha) \wedge Int_j(\alpha))$$

In our logic the second and third condition in the definition of positive trust are together equivalent to $Does_{j:\alpha}\top$ (by Axiom **IntAct2**), so the definition of trust can be simplified as follows:

$$ATrust(i, j, \alpha, \phi) \overset{\text{def}}{=} Goal_i X\phi \wedge Bel_i(After_{j:\alpha}\phi \wedge Does_{j:\alpha}\top)$$

$ATrust(i, j, \alpha, \phi)$ is meant to stand for: i trusts j to do α with regard to to his goal that ϕ.

The following theorem highlights the fact that if i trusts j to do α with regard to his goal that ϕ then i has a positive expectation that ϕ will be true in the next state.

Theorem 1. *Let $i, j \in AGT$ and $\alpha \in ACT$. Then:*
$\vdash ATrust(i, j, \alpha, \phi) \rightarrow Bel_i X\phi$

The dual notion of negative trust in the trustee's action is based on the fact that, by doing some action α, agent j can prevent i to reach his goal. In that case i expects that j will not intend to do α. That leads to the following definition.

Definition 2 ***NEGATIVE TRUST ABOUT ACTION.*** *i trusts j not to do α with regard to his goal ϕ if and only if i wants ϕ to be true and i believes that:*

1. *j, by doing α, will ensure that $\neg\phi$ AND*

[3] In the present paper we only focus on *full trust* involving a *certain belief* of the truster. In order to extend the present analysis to forms of *partial trust*, a notion of *graded belief* (i.e. uncertain belief) or *graded trust*, as in [11], is needed.

2. *j has the capacity to do α AND*
3. *j does not intend to do α*

The formal translation of definition 2 is given by the following abbreviation.

$$ATrust(i, j, \neg\alpha, \phi) \overset{\text{def}}{=} Goal_i X\phi \wedge Bel_i(After_{j:\alpha}\neg\phi \wedge Can_j(\alpha) \wedge \neg Int_j(\alpha))$$

$ATrust(i, j, \neg\alpha, \phi)$ stands for: i trusts j not to do α with regard to his goal that ϕ.

5 Trust in the Trustee's Disposition: The Motivational Case

The fact that agent j intends to do α may be a consequence of his willingness with regard to i's intention that j does α. That leads to define the more specific notions of positive and negative trust in the trustee's willingness. Indeed, i's trust in j does not necessarily depend on i's ascription of an actual intention to j to do a certain action α. There are forms of trust which are based on i's ascription of a potential intention to j. In these cases i attributes to j a positive disposition which is called *j's willingness*. More precisely, we suppose that j is *willing to do the action α for i* if and only if j has the conditional goal (or conditional intention) to form the intention to perform action α under the condition in which he believes that i wants him to do α. Thus, *willingness* is interpreted here as closely related to the concept of *goal adoption*. In this perspective, saying "j is willing to do everything for i" means "j wants to do whatever i wants him to do" and saying "j is willing to do action α for i" means "j wants to do α in case i wants him to do α".[4] The following abbreviation captures our notion of willingness in a formal way.

$$Will_{j,i}(\alpha) \overset{\text{def}}{=} Goal_j(Bel_j Goal_i Does_{j:\alpha}\top \rightarrow Int_j(\alpha)) \wedge$$
$$\neg Goal_j \neg Bel_j Goal_i Does_{j:\alpha}\top$$

where $Will_{j,i}(\alpha)$ stands for: j is willing to do α for i. The second condition in the definition of willingness is given in order to prevent from saying that j is willing to do α for i, when j wants not to believe that i does not want him to do α.

We define a related concept of j's willingness not to do α for i. According to our definition, *j is willing not to do the action α for i* if and only if j has the conditional goal that he will not have the intention to do action α unless he believes that i does not want him not to do α.

$$Will_{j,i}(\neg\alpha) \overset{\text{def}}{=} Goal_j(Int_j(\alpha) \rightarrow Bel_j \neg Goal_i \neg Does_{j:\alpha}\top) \wedge$$
$$\neg Goal_j Bel_j \neg Goal_i \neg Does_{j:\alpha}\top$$

[4] *Willingness* may have different natures. Agent i might be willing to do a certain action α for j since he expects that if he does α, he will get something in return by j; or i might be willing to do a certain action α for j since he expects that if he does not do α, j will do something bad for him, etc. In this work we focus on the core of the concept of willingness without investigating the more specific forms of willingness (i.e. the reasons to be willing).

where $Will_{j,i}(\neg\alpha)$ stands for: j is willing not to do α for i.[5] The following two theorems highlight some interesting properties of our concept of willingness.

Theorem 2. *Let $i, j \in AGT$ and $\alpha \in ACT$. Then:*

1. $\vdash Will_{j,i}(\alpha) \rightarrow (Bel_j Goal_i Does_{j:\alpha}\top \rightarrow Int_j(\alpha))$
2. $\vdash Will_{j,i}(\neg\alpha) \rightarrow (Int_j(\alpha) \rightarrow Bel_j \neg Goal_i \neg Does_{j:\alpha}\top)$

According to Theorem 2.1, if j is willing to do α for i and j believes that i wants him to do α, then j will adopt i's goal in such a way that he will intend to do α. In this sense Theorem 2.1 captures the *adoptive* process which leads from a j's positive disposition toward i to the situation in which j intends to do what i wants him to do. According to Theorem 2.2, if j is willing not to do α for i and intends to do action α, then he has to believe that i does not want him not to do α.

From the the concept of willingness, we can characterize the concept of i's positive trust in j's willingness.

Definition 3 ***POSITIVE TRUST ABOUT WILLINGNESS.*** *i trusts j about j's willingness to do α with regard to his goal that ϕ if and only if i wants ϕ to be true and i believes that:*

1. *j, by doing α, will ensure that ϕ AND*
2. *j has the capacity to do α AND*
3. *j is willing to do α for i*

Formally:

$$WTrust(i, j, \alpha, \phi) \stackrel{\text{def}}{=} Goal_i X\phi \wedge Bel_i(After_{j:\alpha}\phi \wedge Can_j(\alpha) \wedge Will_{j,i}(\alpha))$$

where $WTrust(i, j, \alpha, \phi)$ stands for: i trusts j about j's willingness to do α with regard to his goal that ϕ. The following theorem highlights the relationship between the notions of $ATrust(i, j, \alpha, \phi)$ and $WTrust(i, j, \alpha, \phi)$.

Theorem 3. *Let $i, j \in AGT$ and $\alpha \in ACT$. Then:*

1. $\vdash (Bel_i Bel_j Goal_i Does_{j:\alpha}\top \wedge WTrust(i, j, \alpha, \phi)) \rightarrow Bel_i Int_j(\alpha)$
2. $\vdash (Bel_i Bel_j Goal_i Does_{j:\alpha}\top \wedge WTrust(i, j, \alpha, \phi)) \rightarrow ATrust(i, j, \alpha, \phi)$

For instance, according to Theorem 3.2, if i trusts j about j's willingness to do α with regard to his goal that ϕ and i believes that j believes that i wants j to do α, then i trusts j to do α with regard to his goal that ϕ.

The concept of negative trust in the trustee's willingness can be defined as follows.

Definition 4 ***NEGATIVE TRUST ABOUT WILLINGNESS.*** *i trusts j about j's willingness not to do α with regard to his goal that ϕ if and only if i wants ϕ to be true and i believes that:*

[5] As for the definition of j's willingness to do α for i, we add the condition $\neg Goal_j Bel_j \neg Goal_i \neg Does_{j:\alpha}\top$ in order to prevent from saying that j is willing not do α for i, when j wants to believe that i does not want that he does not do action α. The same solution is adopted in section 7 for the definitions of obedience and honesty.

1. j, by doing α, will ensure that $\neg\phi$ AND
2. j has the capacity to do α AND
3. j is willing not to do α for i

Formally:

$$WTrust(i,j,\neg\alpha,\phi) \overset{\text{def}}{=} Goal_i X\phi \wedge Bel_i(After_{j:\alpha}\neg\phi \wedge Can_j(\alpha) \wedge Will_{j,i}(\neg\alpha))$$

where $WTrust(i,j,\alpha,\phi)$ stands for: i trusts j about j's willingness not to do α with regard to to his goal that ϕ.

The following theorem highlights the relationship between negative trust about willingness and negative trust about action. It says that: negative trust about willingness entails negative trust about action in the context where i believes that j does not believe that i does not want j not to do α.

Theorem 4. *Let $i,j \in AGT$ and $\alpha \in ACT$. Then:*
$\vdash (Bel_i\neg Bel_j\neg Goal_i\neg Does_{j:\alpha}\top \wedge WTrust(i,j,\neg\alpha,\phi)) \rightarrow ATrust(i,j,\neg\alpha,\phi)$

6 Norms in Computer Security

In the field of computer science the notion of security may have two different meanings: there is no computer failure, or there is no violation of norms about computer usage. In this paper we adopt the second meaning. Here agents may be human agents or software agents. In the case of software agents, we talk about their mental attitudes like beliefs or intentions and we assume that their actions are intentional actions. Moreover, we suppose that for a software agent, performing an action means executing a program, and a certain program is performed by the software agent only if the effects of its execution conform to what has been specified by the designer of the program. For instance, a software agent can inform someone about something only by performing the act *inform* which is the procedure specified by the designer as a means for inducing someone to believe something. It cannot inform someone about something by performing some sequence of *insert* actions or *delete* actions since this is not the procedure specified by the designer.

In this work the security properties that should be guaranteed are restricted to: integrity, availability and privacy [7]. For simplification, we have ignored properties like: authentication or non repudiation. As a matter of simplification we have only considered computer systems of the kind information systems (for instance a database system). A similar analysis could be done for transmission systems (for instance Internet).

In order to study security properties we extend the logic $\mathcal{L}$ with the following specific actions: $inf_j(\phi)$ (action of informing j about ϕ), $ins_j(\phi)$ (action of inserting the information ϕ in j), $del_j(\phi)$ (action of deleting the information ϕ from j), $ask_j(\alpha)$ (action of asking j to do action α). The following abbreviations are given for denoting the performance of the previous special actions by an arbitrary agent i: $Inf_{i,j}(\phi) \overset{\text{def}}{=} Does_{i:inf_j(\phi)}\top$; $Ins_{i,j}(\phi) \overset{\text{def}}{=} Does_{i:ins_j(\phi)}\top$; $Del_{i,j}(\phi) \overset{\text{def}}{=} Does_{i:del_j(\phi)}\top$; $Ask_{i,j}(\alpha) \overset{\text{def}}{=} Does_{i:ask_j(\alpha)}\top$.

The constructions $Inf_{j,i}(\phi)$, $Ins_{i,j}(\phi)$, $Del_{i,j}(\phi)$ are used to describe the interaction between an information system j and an agent i (i may be a human agent or a software agent). $Inf_{j,i}(\phi)$ means: the information system j informs agent i about ϕ. $Ins_{i,j}(\phi)$ means: agent i inserts the information ϕ in the information system j (or i makes that j believes that ϕ). $Del_{i,j}(\phi)$ means: agent i deletes the information ϕ from the information system j (or i makes that j does not believe that ϕ). For human agents or software agents the construction $Ask_{i,j}(\alpha)$ expresses that: agent i asks j to do the action α. In the following sections security properties are going to be defined.

6.1 Security Properties

Definition 5 ***The information system** j **guarantees the privacy of information** ϕ**.*** *if and only if for every agent k, if j informs k about ϕ, then it is permitted that j informs k about ϕ.*

Formally,

$$Priv_j(\phi) \stackrel{\text{def}}{=} \bigwedge_{k \in AGT} (Inf_{j,k}(\phi) \rightarrow PermInf_{j,k}(\phi))$$

where $Priv_j(\phi)$ stands for: the information system j guarantees the privacy of information ϕ.

Definition 6 ***The information system** j **guarantees the integrity of information** ϕ**.*** *if and only if for every agent k, if k inserts (resp. deletes) ϕ, then it is permitted that k inserts (resp. deletes) ϕ.*

Formally,

$$Intg_j(\phi) \stackrel{\text{def}}{=}$$

$$\bigwedge_{k \in AGT} (Ins_{k,j}(\phi) \rightarrow PermIns_{k,j}(\phi)) \wedge \bigwedge_{k \in AGT} (Del_{k,j}(\phi) \rightarrow PermDel_{k,j}(\phi))$$

where $Intg_j(\phi)$ stands for: the information system j guarantees the integrity of information ϕ.

Definition 7 ***Agent** i **guarantees the availability to do the action** α **for** j**.*** *if and only if, if i has the right to oblige j to do α and i asks j to do α, then j does α.*

Formally,

$$Avail_{i,j}(\alpha) \stackrel{\text{def}}{=} (Right_{i,j}(\alpha) \wedge Ask_{i,j}(\alpha)) \rightarrow Does_{j:\alpha}\top$$

where $Avail_{i,j}(\alpha)$ stands for: agent i guarantees the availability to do the action α for j, and

$$Right_{i,j}(\alpha) \stackrel{\text{def}}{=} Ask_{i,j}(\alpha) \rightarrow ObgDoes_{j:\alpha}\top$$

The intuitive meaning of $Right_{i,j}(\alpha)$ is that by asking j to do α i “creates” the obligation for j to do α.

7 Trust in the Trustee's Disposition: The Normative Case

In the context of computer security the fact that agent j intends to do α may be a consequence of his fulfillment of the obligation to do this action. In this case we say that j is *obedient*. In a similar way, the fact that he does not intend to do α may be a consequence of the fact that he respects the prohibition to do this action. In this case we say that j is *honest*. It is worth noting that there is a deep analogy between the fact that i's goal is that j does α (resp. it is not the case that i's goal is that j does not do α) and the fact that it is obligatory that j does α (resp. it is permitted that j does α). The justification of this analogy is that what is obligatory can be interpreted as the goal of people who institute the norms, and what is permitted as what is possible with respect to their goal. In the formal definitions below, this analogy is expressed by the fact that the definition of obedience (resp. honesty) can be obtained from the definition of willingness to do (resp. willingness not to do) given in section 5 by substituting $ObgDoes_{j:\alpha}\top$ (resp. $PermDoes_{j:\alpha}\top$) to $Goal_i Does_{j:\alpha}\top$ (resp. $\neg Goal_i \neg Does_{j:\alpha}\top$). In the following this analogy will be called "*motivational / normative analogy*".

On the one hand we suppose that j *is obedient to do the action* α if and only if, j has the conditional goal that if he believes that it is obligatory that he does α, then he intends to do α. Formally,

$$Obed_j(\alpha) \overset{\text{def}}{=} Goal_j(Bel_j ObgDoes_{j:\alpha}\top \rightarrow Int_j(\alpha)) \wedge$$

$$\neg Goal_j \neg Bel_j ObgDoes_{j:\alpha}\top$$

where $Obed_j(\alpha)$ stands for: j is obedient with regard to the obligation to do the action α.

On the other hand we suppose that j is *honest to do the action* α if and only if, j has the conditional goal that if he has the intention to do α, then he believes that it is permitted that he does α. Formally,

$$Honst_j(\alpha) \overset{\text{def}}{=} Goal_j(Int_j(\alpha) \rightarrow Bel_j PermDoes_{j:\alpha}\top) \wedge$$

$$\neg Goal_j Bel_j PermDoes_{j:\alpha}\top$$

where $Honst_j(\alpha)$ stands for: j is honest with regard to the permission to do the action α.

The following two theorems highlight some interesting properties of the concepts of obedience and honesty.

Theorem 5. *Let $j \in AGT$ and $\alpha \in ACT$. Then:*

1. $\vdash Obed_j(\alpha) \rightarrow (Bel_j ObgDoes_{j:\alpha}\top \rightarrow Int_j(\alpha))$
2. $\vdash Honst_j(\alpha) \rightarrow (Int_j(\alpha) \rightarrow Bel_j PermDoes_{j:\alpha}\top)$

According to Theorem 5.1, if j is obedient with regard to the obligation to do the action α and believes that it is obligatory to do α, then j will adopt such an obligation in such a way that he will intend to do α. Theorem 5.1, which is symmetrical to Theorem 2.1 for willingness captures the *adoptive* process which leads from j's obedience to the situation in which j intends to do what is obligatory to do. According to Theorem 5.2 (which is symmetrical to Theorem 2.2 for willingness), if j is honest with regard to the permission to do the action α and intends to do action α, then he has to believe that it is permitted to do action α.

We are now in the position to define a concept of i's trust in j's obedience which is symmetrical to the concept of i's positive trust in j's willingness given in section 5.

Definition 8 ***TRUST ABOUT OBEDIENCE.*** *i trusts j to be obedient in doing α with regard to his goal that ϕ if and only if i wants ϕ to be true and i believes that:*

1. *j, by doing α, will ensure that ϕ AND*
2. *j has the capacity to do α AND*
3. *j is obedient in doing α*

Formally,

$$OTrust(i,j,\alpha,\phi) \overset{\text{def}}{=} Goal_i X\phi \wedge Bel_i(After_{j:\alpha}\phi \wedge Can_j(\alpha) \wedge Obed_j(\alpha))$$

where $OTrust(i,j,\alpha,\phi)$ stands for: i trusts j to be obedient in doing α with regard to his goal that ϕ. The following theorems highlight the relationships between trust about obedience and positive trust about action, and between trust about obedience and the property of availability.

Theorem 6. *Let $i,j \in AGT$ and $\alpha \in ACT$. Then:*

1. $\vdash (Bel_i(Right_{i,j}(\alpha) \wedge Ask_{i,j}(\alpha)) \wedge OTrust(i,j,\alpha,\phi)) \rightarrow ATrust(i,j,\alpha,\phi)$
2. $\vdash OTrust(i,j,\alpha,\phi) \rightarrow Bel_i Avail_{i,j}(\alpha)$

The intuitive meaning of Theorem 6.1 is that trust about obedience entails positive trust about action in the context where i believes that he has the right to oblige j to do α and he exercises his right. Theorem 6.2 means that trust about obedience entails that i believes that the availability to do α is guaranteed by j. Notice that in this theorem i's goal is not $Avail_{i,j}(\alpha)$. The goal ϕ may be any situation which can be obtained by doing α. For instance, i's goal may be to know meteorological forecasts and the action α is that j informs i about these expectations. Then, in that example, the theorem 6.2 says that the consequence of i's trust in j's obedience to do α is that i believes that j guarantees the availability to inform him about meteorological forecasts.

We now define a concept of i's trust in j's honesty which is symmetrical to the concept of i's negative trust in j's willingness given in section 5.

Definition 9 ***TRUST ABOUT HONESTY.*** *i trusts j to be honest in doing α with regard to his goal that ϕ if and only if i wants ϕ to be true and i believes that:*

1. *j, by doing α, will ensure that $\neg\phi$ AND*
2. *j has the capacity to do α AND*
3. *j is honest in doing α*

Formally,

$$HTrust(i,j,\alpha,\phi) \overset{\text{def}}{=} Goal_i X\phi \wedge Bel_i(After_{j:\alpha}\neg\phi \wedge Can_j(\alpha) \wedge Honst_j(\alpha))$$

where $HTrust(i, j, \alpha, \phi)$ stands for: i trusts j to be honest in doing α with regard to his goal that ϕ.

We denote with $IAct(\psi)$ the set of all actions of informing some agent about ψ. In formal terms: $IAct(\psi) \overset{\text{def}}{=} \{inf_z(\psi) : z \in AGT\}$. Then, the following two theorems can be derived.

Theorem 7. *Let $i, j \in AGT$ and $\alpha \in ACT$. Then:*

1. $\vdash (Bel_i \neg Bel_j PermDoes_{j:\alpha} \top \wedge HTrust(i, j, \alpha, \phi)) \rightarrow ATrust(i, j, \neg\alpha, \phi)$
2. $\vdash \bigwedge_{\alpha \in IAct(\psi)} (HTrust(i, j, \alpha, \phi)) \rightarrow Bel_i Priv_j(\psi)$

Theorem 7.1, which is symmetrical to Theorem 4 for negative trust about willingness, means that trust about honesty entails negative trust about action in the context where i believes that j does not believe that he has the permission to do α. Theorem 7.2 means that i's trust in j's honesty for every action of informing an agent about ψ entails that i believes that the privacy for ψ is guaranteed by j. Like in Theorem 6.2, in Theorem 7.2 i's goal is not $Priv_j(\psi)$. A theorem similar to Theorem 7.2 can be proved for the property of integrity of an information system j ($Intg_j(\phi)$) since the set of permitted actions is explicitly defined.

8 Conclusion

The logical framework which has been presented allows to give precise definitions to several sophisticated notions of trust, going from a general one to more specific ones which are relevant to the context of computer security. In addition, theorems have been proved which give sufficient conditions about obedience or honesty to guarantee that an agent can believe that security properties hold. The benefits of the logical formalization are manyfold. It points out some facts that may look as trivialities but that may be left implicit without the help of this formal framework. For instance, consequences that an agent can infer from what he trusts are just beliefs not truth. That is inherent to the notion of trust. Also, it raises some non trivial questions.

Due to the complexity of the involved concepts we had to accept strong simplifications. The first one is that our formal definition of the concept of obligation is very crude. The second is that in some definitions entailment is formalized by a material implication in the scope of goal modalities, while some form of conditional might be more adequate. The same comment applies to the definition of right where a "counts as" conditional [15] would be more appropriate than material implication. Also, security properties have been defined for a specific proposition, while these properties are usually expected for a set of proposition about a given topic, and a more realistic notion of trust should be based on several degrees of trust. Finally, we almost ignored the temporal dimension. In many cases trust is about a trustee's property which is not contingent to the current situation, but holds for some period of time. All these issues require future investigations, but we believe that to analyze so complex problems it was better to start with simple assumptions, even if they can be seen as oversimplifications.

References

1. Åqvist, L.: Deontic logic. In: Gabbay, D.M., Geunther, F. (eds.) Handbook of Philosophical Logic. Kluwer Academic Publishers, Dordrecht (2002)
2. Bratman, M.: Intentions, plans, and practical reason. Harvard University Press (1987)
3. Carmo, J., Jones, A.: Deontic Logic and Contrary-to-Duties. In: Gabbay, D. (ed.) Handbook of Philosphical Logic (Rev. Edition), Reidel (to appear)
4. Castelfranchi, C., Falcone, R.: Principles of trust for MAS: Cognitive anatomy, social importance, and quantification. In: Proc. of the Third International Conference on Multiagent Systems (ICMAS 1998), pp. 72–79 (1998)
5. Castelfranchi, C., Falcone, R.: Social trust: A cognitive approach. In: Castelfranchi, C., Tan, Y.H. (eds.) Trust and Deception in Virtual Societies, pp. 55–90. Kluwer, Dordrecht (2001)
6. Cohen, P.R., Levesque, H.J.: Intention is choice with commitment. Artificial Intelligence 42, 213–261 (1990)
7. Cuppens, F., Demolombe, R.: A Deontic Logic for Reasoning about Confidentiality. In: Proc. of Third International Workshop on Deontic Logic in Computer Science (DEON 1996), pp. 72–79 (1996)
8. Dastani, M., Herzig, A., Hulstijn, J., van der Torre, L.: Inferring Trust. In: Leite, J.A., Torroni, P. (eds.) CLIMA 2004. LNCS (LNAI), vol. 3487, pp. 144–160. Springer, Heidelberg (2005)
9. Demolombe, R.: Reasoning about trust: a formal logical framework. In: Jensen, C., Poslad, S., Dimitrakos, T. (eds.) iTrust 2004. LNCS, vol. 2995, pp. 291–303. Springer, Heidelberg (2004)
10. Demolombe, R.: To trust information sources: a proposal for a modal logical framework. In: Castelfranchi, C., Tan, Y.-H. (eds.) Trust and Deception in Virtual Societies, pp. 111–124. Kluwer, Dordrecht (2001)
11. Demolombe, R., Liau, C.-J.: A logic of graded trust and belief fusion. In: Proc. of Fourth Workshop on Deception, Fraud and Trust in Agent Societies, pp. 13–25 (2001)
12. Demolombe, R., Lorini, E.: A logical account of trust in information sources. In: Proc. of the Eleventh International Workshop on Trust in Agent Societies (to appear)
13. Harel, D., Kozen, D., Tiuryn, J.: Dynamic Logic. MIT Press, Cambridge (2000)
14. Huynh, T.G., Jennings, N.R., Shadbolt, N.R.: An integrated trust and reputation model for open multi-agent systems. Journal of Autonomous Agent and Multi-Agent Systems 13, 119–154 (2006)
15. Jones, A.J., Sergot, M.: A formal characterisation of institutionalised power. Journal of the Interest Group in Pure and Applied Logics 4(3) (1996)
16. Jones, A.J.I., Firozabadi, B.S.: On the characterization of a trusting agent: Aspects of a formal approach. In: Castelfranchi, C., Tan, Y.H. (eds.) Trust and Deception in Virtual Societies, pp. 55–90. Kluwer, Dordrecht (2001)
17. Jonker, C.M., Treur, J.: Formal analysis of models for the dynamics of trust based on experiences. In: Garijo, F.J., Boman, M. (eds.) MAAMAW 1999. LNCS, vol. 1647, pp. 221–231. Springer, Heidelberg (1999)
18. Liau, C.J.: Belief, information acquisition, and trust in multi-agent systems: a modal logic formulation. Artificial Intelligence 149, 31–60 (2003)
19. Lorini, E., Herzig, A.: A logic of intention and attempt. Synthese (to appear)
20. Lorini, E., Herzig, A., Castelfranchi, C.: Introducing attempt in a modal logic of intentional action. In: Fisher, M., van der Hoek, W., Konev, B., Lisitsa, A. (eds.) JELIA 2006. LNCS (LNAI), vol. 4160, pp. 280–292. Springer, Heidelberg (2006)
21. Meyer, J.-J.C., van der Hoek, W., van Linder, B.: A logical approach to the dynamics of commitments. Artificial Intelligence 113(1-2), 1–40 (1999)

Specifying Intrusion Detection and Reaction Policies: An Application of Deontic Logic

Nora Cuppens-Boulahia and Frédéric Cuppens

TELECOM Bretagne, 2 rue de la châtaigneraie, 35512 Cesson Sévigné Cedex, France

Abstract. The security policy of an information system may include a wide range of different requirements. The literature has primarily focused on access and information flow control requirements and more recently on authentication and usage control requirements. Specifying administration and delegation policies is also an important issue, especially in the context of pervasive distributed systems. In this paper, we are investigating the new issue of modelling intrusion detection and reaction policies and study the appropriateness of using deontic logic for this purpose. We analyze how intrusion detection requirements may be specified to face known intrusions but also new intrusions. In the case of new intrusions, we suggest using the *bring it about* modality and specifying requirements as prohibitions to bring it about that some security objectives are violated. When some intrusions occur, the security policy to be complete should specify what happens in this case. This is what we call a reaction policy. The paper shows that this part of the policy corresponds to *contrary to duty* requirements and suggests an approach based on assigning priority to activation contexts of security requirements.

1 Introduction

Current information systems have to face many threats that attempt to exploit their vulnerabilities. Moreover, since information systems tend to be increasingly complex, specifying their security policy is a tedious and error-prone task. In this context, specifying consistent, relevant and complete security policies of information systems is a major challenge for researchers.

There are many advantages of using a formal approach to specify the policy: (1) It provides non ambiguous specification of security requirements, (2) It is possible to develop support tools to formally analyse these requirements, (3) It is also possible to develop support tools to assist the security administrator in the task of automatically deploying these requirements over a security architecture.

A security policy may actually specify very different security requirements. We first suggest a classification of these various requirements a security policy may contain. We then focus on two sub-parts of the security policy that specify (1) intrusion detection requirements (IDR) and (2) reaction requirements (RR). We investigate the relevance of deontic logic to specify such requirements.

Intrusion detection has been an active research field for more than twenty years and many intrusion detection systems (IDS) have been developed and are

R. van der Meyden and L. van der Torre (Eds.): DEON 2008, LNAI 5076, pp. 65–80, 2008.

now available. However, intrusion detection requirements enforced by IDSs are generally considered independently of the remainder of the security policy. A first contribution of this paper is to consider that IDRs are actually a sub-part of the access control policy. This approach has several advantages. A first advantage is that it is then possible to formally analyse whether IDRs are consistent with other security requirements. Another advantage is that we can integrate IDRs into a deploying process in order to automatically configure IDSs. Finally and as shown in this paper, this approach provides means to formally specify in a reaction policy what should happen in case of violation of some IDRs.

Traditionally, access control requirements are modelled using permissions and prohibitions. We show how IDRs can be specified using prohibitions. However, we consider two different types of IDRs depending on the fact these requirements apply to known or unknown attacks. In the case of a known attack, an intrusion detection requirement specifies that it is prohibited to execute the action corresponding to this attack. Notice that the attack may actually correspond to an elementary action or to a composition of elementary actions corresponding to an attack scenario. In the case of unknown attacks, the specification of IDRs is more complex. Our approach in this case is based on the specification of security objectives. An IDR then corresponds to a prohibition for any subject (or group of subjects in the case of a distributed intrusion) to *bring it about* [29] that some security objective is violated. To our best knowledge, this is the first time the bring it about operator is used in this context. We then show how different IDRs may be deployed on different Intrusion Detection Systems (IDSs) including misuse based detection systems, anomaly based detection systems or correlation based detection systems.

However, the security policy must also specify what happens when an intrusion is detected. This is what we call a reaction policy. A reaction policy is a set of deontic requirements specifying obligations, prohibitions and possibly permissions that are triggered when an intrusion is detected. We show that these requirements may be actually viewed as *contrary to duty norms* (see [30,33]).

In this paper, we do not actually develop a new deontic formalism to specify intrusion detection and reaction policies. Instead, we analyse which problems addressed in this paper may be solved using the current state of the art in deontic logic and which problems still require further investigation.

The remainder of this paper is organized as follows. In section 2, we informally introduce the concept of security policy and suggest a classification of the different requirements that may be specified in a security policy. We then focus in section 3 on a part of the security policy that corresponds to the access control policy. The access control policy should include an intrusion detection policy as explained in section 4. We show in section 4 how to express various requirements of such an intrusion detection policy. When an intrusion occurs in the information system, the security policy is violated. Thus, another part of the security policy consists in specifying a reaction policy. This is presented in section 5. In section 6, we discuss how to implement the approach suggested in this paper and list some open issues. Finally, section 7 concludes the paper.

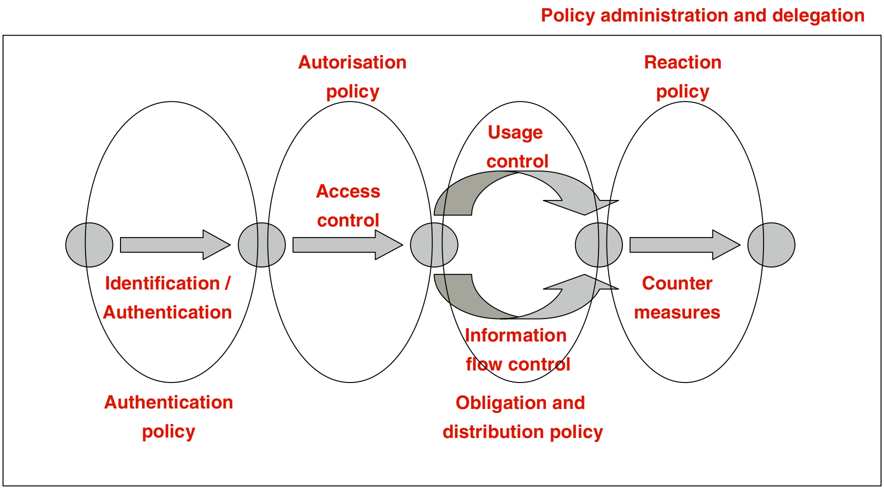


Fig. 1. General structure of a security policy

2 What Is a Security Policy?

In the following, we shall view a security policy as a set of norms corresponding to permissions, prohibitions, obligations and dispensations. In the security literature, it is generally considered that these norms apply to users or processes (called subjects) when they access to resources (called objects) in order to execute services or programs (called actions).

A security policy may be structured into several sub policies (see figure 1):

- Authentication policy. Authentication is the first step to get an access to the information system and is used to securely associate a subject with its identity. The authentication policy specifies which authentication protocol a subject is permitted, prohibited or obliged to use to get an access to the system. Many authentication protocols, using password, smart cards or biometric traits, have been defined in the literature. It is also possible to specify that mutual authentication protocols such as Kerberos must be used so that each party involved in the interaction authenticates each other. More recently, authentication has included Single Sign On (SSO) functionalities.
  The authentication policy may influence other downstream policies. For example, the access control policy may condition its decisions on the kind of authentication that has been performed by the user. The use of stronger authentication protocols (loss of availability) may allow access to more sensitive resources (increase of confidentiality).
- Access control policy. This part of the policy is also called authorization policy in the literature and applies once subjects are authenticated. It corresponds to a set of permissions and prohibitions that specify which action

a subject may or may not execute on objects. The access control policy is further discussed in section 3.

- Usage control policy. It is a set of requirements that apply once a subject gets an authorized access to a resource. The objective is to control how this subject uses the resource. A usage control policy corresponds to a set of obligations to be enforced before the access (pre usage control), during the access (on going usage control) or after the access (post usage control). Even if there are many interesting issues to investigate, this is not the purpose of this paper to further address usage control. The interesting reader may consider the following references [28,24].
- Information flow control policy. The objective here is to control how information flows in the information system, i.e. how information is transferred from subject to subject. A malicious subject, being permitted to access some information, may attempt to illegally transfer this information to another unauthorized subject. This problem is not appropriately managed by access control requirements and specific models have been defined for this purpose (see [3,21]). However, there are currently no security models that integrate both access control requirements and the different forms of information flow control requirements. Even if this is out of the scope of this paper, the interesting reader may have a look at [2] for a preliminary work in this direction.
- Reaction policy. Since some requirements of the security may be violated, a security policy would not be complete if it does not include requirements that specify what happens in case of violation. This set of requirements is what we call a reaction policy. How to specify a reaction policy is further investigated in section 5.
- Administration and delegation policy. The administration policy specifies who is permitted to define new security requirements or update existing security requirements. Delegation also corresponds to the creation of permissions and obligations but generally (1) the delegator must own the permissions or obligations he or she delegates and (2) there is a transfer of these permissions or obligations from the delegator to the delegatee (see [20] for a more detailed discussion of the differences between administration and delegation). As suggested in the literature [19], administration and delegation may be modelled using special speech acts. Another possibility suggested in [15] would be to manage administration through licenses. A license is a special object that generally represents a permission for the subject who owns this license. In that case, the administration and delegation policies are specified by permissions to create or revoke licenses. A possible extension would be to define another special object called *duty* to represent an obligation for the subject who owns a duty and specify delegation of obligations through the creation of new duties.

In this paper, we shall actually focus on intrusion detection and reaction policies. However, since we view the intrusion detection policy as a sub part of the access control policy, we shall first briefly discuss how to model an access control policy.

3 Access Control Policy

3.1 Principles of Access Control

Specifying an access control policy has been investigated for more than thirty years [23]. In many models, an access control policy is modelled as a set of permissions. There are no obligation and dispensation in an access control policy but it is possible to also specify explicit prohibitions. The issue of using prohibition in an access control policy is further investigated in section 3.2 below.

Traditionally, access control policies apply to queries corresponding to subjects that ask to execute actions on objects. When a subject formulates such a query, the access controller analyses the query to check if it is permitted, in which case the query is accepted, else this query is rejected.

Generally, it is assumed that actions to be controlled corresponds to "elementary" actions. As a consequence, access control policies are specified using first order logic using a predicate like $is\_permitted(subject, action, object)$.

More recently, it has been suggested to use the concept of role as in the RBAC (Role Based Access Control) model [32]. In this case, permissions are not directly assigned to subject but to role which may be modelled using a predicate $permission(role, action, object)$. Subjects are assigned to role using a predicate $empower(subject, role)$ and we have the following derivation rule[1]:

$$permission(Role, Action, Object) \wedge empower(Subject, Role)$$
$$\rightarrow is\_permitted(Subject, Action, Object)$$

Actually, the RBAC model does not give any formal semantics to the concept of role. This lack of formalization leads to consider in many approaches that the concept of role is a panacea to solve every access control problems. This leads to consider "strange" roles such as location dependent role [6] or temporal dependent role [5].

There were several works that attempt to use deontic logic to provide formal semantics to the role concept [10,27,18]. The central idea is that a role is an organisation dependent concept. Basically, security policies are defined for moral authorities, called organisations. In this context, most of security requirements do not directly apply to concrete and implementation dependent entities such as subject, action and object. Instead, it is more appropriate to use abstract organisation dependent concepts. As suggested by RBAC, a role is one of such concept to create organization dependent abstraction of subjects. When an organization defines roles and assigns these roles to subjects, these subjects are no longer acting as individuals but as subjects empowered in some roles.

The OrBAC model [1] suggests to proceed similarly for actions and objects. For this purpose, we respectively suggest the concepts of activity and view as abstraction of action and object (see [1] for further explanation about these concepts). The predicate $use(object, view)$ specifies that a given object is used in the organization in a given view and the predicate $consider(action, activity)$

[1] In the following, we shall assume that terms starting with a capital letter represent variables and that all free variables in formula are implicitly universally quantified.

specifies that a given action is considered in the organization as an implementation of a given activity[2].

We also need to specify access control requirements that depend on *contextual* conditions for instance:

- R1: A nurse is permitted to consult medical records *in a context of urgency*,
- R2: A physician is permitted to consult medical records *in his or her office during work hours*.

Contexts are modelled using the predicate $hold(subject, action, object, context)$ that specifies conditions to be satisfied to consider that a given subject executes a given action on a given object in some context. For instance, a context in_office is specified by the following rule:

$$subject\_office(Subject, Office) \wedge location(Subject, Office) \\ \rightarrow hold(Subject, Action, Object, in\_office)$$

Using these different concepts, an access control policy is simply defined by a set of facts having the form: $permission(Role, Activity, View, Context)$. For instance, rules R1 and R2 are respectively modelled by the two following facts:
R1: $permission(nurse, consult, med\_record, urgency)$
R2: $permission(physician, consult, med\_record, in\_office\&working\_hours)$

Notice the possibility in the second rule to combine contexts using conjunction (negation or disjunction would be also possible).

Given an access control policy (which possibly depends on contextual conditions), the objective of the access controller consists in deciding if a given subject is actually permitted to execute some action on a given object. This is modelled by the following rule:

$$permission(Role, Activity, View, Context) \wedge \\ empower(Subject, Role) \wedge consider(Action, Activity) \wedge \\ use(Object, View) \wedge hold(Subject, Action, Object, Context) \\ \rightarrow is\_permitted(Subject, Action, Object)$$

3.2 Prohibition and Management of Conflicts

In recent models, it is generally accepted that access control policies may not only specify permissions but also prohibitions [12]. For instance, in the OrBAC model, it is possible to specify organizational prohibitions using the predicate $prohibition(role, activity, view, context)$ and derive concrete prohibitions using the predicate $is\_prohibited(subject, action, object)$.

When the access control policy contains both permissions and prohibitions, a conflict occurs when it is possible to derive both $is\_permitted(s, a, o)$ and $is\_prohibited(s, a, o)$ for the same subject, action and object.

Several papers have investigated the problem of conflict detection and management (see for instance [4,12]). The solution is generally based on assigning priorities to security requirements so that when a conflict occurs between two requirements, the requirement with the higher priority takes precedence.

[2] In OrBAC, the organization is made explicit in every predicate but here, to simplify, the organization is left implicit since we consider always only one organization.

This is basically the approach suggested in the OrBAC model [12]. It consists in detecting and managing *potential* conflicts. A potential conflict exists between a permission and a prohibition if these two requirements may possibly apply to the same subject, action and object. There is no such potential conflict between two requirements if these requirements are *separated*. In OrBAC, separation between requirements is defined as follows: Two requirements are separated if their respective roles are separated, or their views are separated, or their activities are separated, or their contexts are separated. Two roles are separated if it is impossible to simultaneously assign a subject to these two roles. Separations of activities, views and contexts are similarly defined.

Since no conflict can occur between separated requirements, it is sufficient to assign priorities between every pair of non separated permission/prohibition requirements. This guarantees that all actual and potential conflicts are solved.

Notice that when using both permissions and prohibitions to specify an access control policy, it is actually possible to define two different decision procedure called *open* policy and *closed* policy. In the open policy, an access is accepted only if it is explicitly permitted by the policy, else it is rejected. In the closed policy, an access is rejected only if it is explicitly prohibited, else it is accepted.

Since an open policy is generally not equivalent to a closed policy, this means that when specifying an access control policy, it is generally assumed that a prohibition is not equivalent to a non permission[3]. Thus formalism based on Standard Deontic Logic (SDL) would not be appropriate. We need a more complex deontic formalism such as the one suggested in [9].

4 Intrusion Policy

The primary objective of computer security is actually to protect the information system against intrusions. An intrusion is an action or a sequence of actions (also called an intrusion scenario) that exploits a *vulnerability*.

Many intrusions actually correspond to known intrusions that exploit known vulnerabilities. To detect known intrusions, most intrusion detection systems (IDS) implement techniques called *misuse* detection to recognize a *signature* of the intrusion. A signature specifies evidences that actions corresponding to the intrusion have been executed. Current implementations work quite well when the intrusion corresponds to an elementary action but do not provide so good results in case of intrusion scenarios.

Detection of unknown (or new) intrusions is much more complex. Current techniques, called anomaly detection, attempt to detect abnormal behavior that would reveal an intrusion. However, the implemented techniques are far from being perfect and generate many false positives (an alert is launched whereas there is no intrusion) and also false negatives (no alert is launched whereas an intrusion actually occurs).

[3] However, conflict resolutions guarantees that prohibition *implies* not permission.

Even if an intrusion is generally defined as a violation of the security policy, there is no approach that attempts to include intrusion detection requirements in the security policy specification.

This is the purpose of this section to investigate this issue. We start investigating the case of known intrusions and then move to new intrusions.

4.1 Security Policy for Known Intrusions

Since an intrusion corresponds to a malicious behavior, it seems appropriate to specify that such malicious behaviors are prohibited. Thus an intrusion detection policy corresponds to a set of prohibition requirements. Regarding known intrusions, the action used to exploit the vulnerability can be explicitly specified and prohibited. The following example illustrates the approach.

- **Example:** The Land Attack is a known intrusion that consists in forging illegal IP packets where the IP addresses of the source is equal to the destination. The impact of this intrusion is that it may cause a denial or service on the server which receives such packets. This intrusion is taken into account in the intrusion detection policy by the following prohibition:
  R3: $prohibition(any\_host, send\_IP\_packet, same\_source\_destination, default)$ In this requirement, a subject is a network host and any_host is a role assigned to every network host, $send\_IP\_packet$ corresponds to the activity of sending packets using the IP protocol, $same\_source\_destination$ is a view that contains any IP packet with a source IP address equal to its destination IP address and $default$ is the default context which is always active.

As mentioned in the introduction, one advantage of formally specifying such prohibitions in the security policy is that it is then possible to analyze possible conflicts between other security requirements. For instance, the access control policy may include a filtering requirement specifying that hosts assigned to the role *private host* are permitted to open HTTP connection with the Internet:
R4: $permission(private\_host, open\_HTTP, to\_Internet, default)$

Since it is possible to use an HTTP connection to send a Land Attack, requirements R3 and R4 are conflicting. It is important to detect and solve such conflicts and in our example, it is likely that requirement R3 should have higher priority than R4 so that hosts from the private zone are prohibited to launch the Land Attack using HTTP connection. The approach defined in [12] provides means to detect and solve this kind of conflicts.

Another advantage is that it is possible to use a formal specification of the intrusion detection policy to automatically configure IDSs, for instance we have defined a process to automatically deploy intrusion detection requirements such as R3 onto the Snort IDS[4] [31].

Requirement R3 actually corresponds to an elementary intrusion that can be executed by a single action. Unfortunately, many intrusions require several

[4] Snort is a network intrusion detection system that uses a signature based approach.

actions in sequence or in parallel to be executed. This is called an intrusion scenario and many examples of such scenarios could be given such as worms like Nimda or distributed denial of service intrusions like Trinoo [22].

It is possible to take into account known intrusion scenarios in the intrusion detection policy by specifying prohibitions. The language presented in section 3, restricted to permissions and prohibitions that apply to elementary actions, cannot express these intrusion scenarios. However, it can be extended, and norms that apply to non elementary actions (sequence, parallelism, ...) have already been extensively investigated in the deontic logic literature (see [25] for instance).

The analysis of conflicts is more complex when the policy includes security requirements on non elementary actions corresponding to intrusion scenarios. This problem is further discussed in section 6.

Regarding the deployment of these requirements, notice that classical IDSs only manage intrusions corresponding to elementary actions. However, there are research prototypes that could be used for this purpose. For instance, the approach suggested in [26] is based on specification of chronicles to represent intrusion scenarios so that prohibitions could be translated into chronicles to automatically configure this prototype.

Notice that the enumeration of every known intrusion scenarios is a complex and fastidious task. Another problem is that it is difficult to find the appropriate level of description of the intrusion scenario. If the description is not precise enough, then false alerts could be launched (false positive). But if the description is too precise, then variants of the intrusion scenario could not be detected (false negative). This is one of the issue we attempt to address in section 4.2.

4.2 Security Policy for New Intrusions

Detection of new intrusions is still a major issue of computer security. Current approaches based on anomaly detection attempts to recognize abnormal behavior of subjects but are far from giving perfect results. In this section, we suggest an approach to specify security requirements associated with new intrusions.

This approach is based on the specification of security objectives. A security objective is a condition on the state of the information system that should be enforced. Of course, these objectives depend on the information system to be protected but it is generally considered that they can be classified into confidentiality, integrity and availability requirements. From the point of view of the defender, an attacker is a subject that attempts to violate a security objective. Thus, we call an intrusion objective the negation of a security objective. We provide examples of security objectives:

- $server(h, DNS) \wedge denial\_of\_service(h)$
  i.e. the DNS server is in denial of service.
- $get\_access(s, root, h) \wedge \neg(authorized\_access(s, root, h))$
  i.e. s illegally gets a root access on a given host h.

If S is a state formula that represents an intrusion objective, then the associated security requirement specifies that it is forbidden for any subject s to bring it about that S is true: $forbidden(E(s, S))$
where E represents the *bring it about* modality (see for instance [29])[5].

The bring it about modality makes implicit the action which is executed to get the effect S. This action may actually correspond to a new intrusion. The interest of this approach is twofold:

1. If there are sensors that could detect that some intrusion objectives are achieved, it is possible to infer that some (possibly new) intrusion occurs.
2. If there is a library of elementary actions modelled through their pre and post conditions[6], then the approach can be also used to detect new intrusion scenario. This is the approach suggested in [11] in which a mechanism is defined to correlate actions and detect when an intrusion objective can be achieved by these correlated actions. The specification of the library of elementary actions is generally easier to manage than the explicit description of entire intrusion scenarios suggested in section 4.1 with the advantage that new intrusion scenarios can be detected.

5 Reaction Policy

As its name points it out, this policy is activated to react against an intrusion. It is a set of rules that specify what happens in case of violation (or attempt of violation) of some requirements of the security policy. According to these (attempts of) violations and their impacts on the target information system, new permissions, prohibitions or obligations are activated and pushed in the appropriate security components. For instance, if an intrusion occurs, and the alert diagnosis identifies the path of the attack or the equipments targeted by this attack and used to reach the intrusion objectives, (1) some packet flows have to be rejected or at least redirected or (2) some of the vulnerable equipments used by the attack have to be stopped or at least isolated typically to contain its spread in the whole system. As suggested in [16], a first form of reaction would be to update the access control policy by activating and deploying new permissions or prohibitions. For instance, a rule:

– R5: $permission(private\_host, open\_TCP, to\_hostObelix, default)$,

might be replaced by a new one such as[7]:

– R6: $prohibition(any\_host, open\_TCP, to\_hostObelix, syn\_flooding)$.

In the second case, a reaction requirement may be specified by means of obligations. We actually consider two different kinds of obligations called server-side

[5] The $forbidden$ and $prohibition$ operators clearly refer to the same concept. In the following and to avoid confusion, we shall use $prohibition$ when we refer to an explicit action and $forbidden$ when we refer to an implicit action through the bring it about operator.

[6] The pre condition represents the condition that must be true before executing the action and the post condition represents the effect of executing the action.

[7] Syn flooding is a denial of service attack against the TCP protocol.

obligation and client-side obligation. A server-side obligation must be enforced by the security components controlled by the security server and generally corresponds to immediate obligations. R7 is an example of such rules expressed in the OrBAC model:

- R7: $obligation(mail\_daemon, stop, mailserver, imap\_threat)$

Client side obligations generally correspond to obligations that might be enforced after some delay. Several papers have already investigated this problem and suggested models to specify obligation with deadlines [7,13,8,17]. For instance, if there is an intrusion that attempts to corrupt an application server by a Trojan Horse intrusion, then this server must be quarantined by the administrator within a deadline of 10s. R8 provides a specification of this requirement:

- R8: $deadline\_obligation(administrator, quarantine,$
  $application\_server, trojan\_horse\_threat, before(10))$

where *deadline_obligation* can be used to specify one more attribute that corresponds to the deadline condition $before(10)$.

As intrusions correspond to some implicit prohibited behaviours and actions, the security requirements inferred by the need to react correspond to contrary to duty requirements. Management of contrary to duty is known to cause trouble (see “pragmatic oddity” [30]). In our approach, management of conflicts is based on classification with respect to the context of activation. In fact, we consider three different types of activation contexts: *threat*, *operational* and *minimal*.

The *operational contexts* aim at describing traditional operational policy [14]. They may correspond to temporal, geographical or provisional contexts (i.e. contexts that depend on the history of previous executed actions).

Intrusion classes are associated with *threat contexts* and for each threat a security rule is defined in the (reaction) policy. *Threat contexts* are activated when a violation of the security policy is detected and are used to specify the reaction policy. The activation of these contexts (*hold* facts, see section 3), leads to the instantiation of the policy rules in response to the considered threat. For instance, a Syn-flooding attack is reported by an alert with a classification reference equal to CVE-1999-0116 (corresponding to the CVE reference of a Syn-flooding attack), the target corresponds to some network host $Host$ and some service $Service$. Then the synflooding context is specified as follows [16]:

- IC: $hold(corp, \_, Service, Host, syn_flooding) \longleftarrow$
  $alert(Time, Source, Target, Classification),$
  $reference(Classification, 'CVE-1999-0116'),$
  $service(Target, Service), hostname(Target, Host).$

Notice that, since the intruder is spoofing (masquerading) its source address in a Syn-flooding attack, the subject corresponding to the threat origin is not instantiated in the hold predicate. When an attack occurs and a new alert is launched by the intrusion detection system, new facts *hold* are derived for threat context Ctx. So, Ctx is then active and the security rules associated with this context are triggered to react to the intrusion.

Most of reaction requirements are in conflict with other access control requirements, i.e. the access control policy may specify a permission whereas the reaction policy specifies a conflicting prohibition that applies when an intrusion is detected. For instance, HTTP is permitted when there is no intrusion but prohibited if an intrusion on the HTTP protocol is detected.

We suggest to solve these conflicts by assigning higher priority to the reaction requirement than the access control requirement. Since access control requirements are associated with operational contexts whereas reaction requirements are associated with threat contexts, we actually consider that threat contexts have higher priority than operational contexts.

However, there are some security requirements such as availability requirements that must be preserved even if an intrusion occurs. For instance, the access to the email server must be preserved even if some intrusions occur. This is modelled as a minimal requirement. *Minimal contexts* then define high priority exceptions in the policy, describing minimal operational requirements that must apply even in case of characterized threat.

So, using an algebra of contexts and priority assignment to security rules, we consider two parameters to manage conflicting situations called *criticity* and *specificity*. A criticity parameter is used to assess context priority between the three defined categories of contexts *operational*, *threats* and *minimal*. We define an operator L_c to assess the level of criticity of contexts, so that if Ctx is a set of well formed contexts: L_c: Ctx $\longrightarrow \{\text{ope, threat, min}\}$ with ope $<$ threat $<$ min. We define the criticity relation as follows: $c_1 <_c c_2 \longleftrightarrow L_{c1} < L_{c2}$. We consider also a specificity parameter that deals with inheritance and context composition, hierarchical specificity context inheritance. For instance, we say that c_2 is more specific than c_1 if $sub\_context(c_2, c_1)$. We define specificity for contexts as follows: $c1 \leq_s c_2 \longleftrightarrow sub\_context(c_2, c_1)$ and $c_1 <_s c_2 \longleftrightarrow c_1 \leq_s c_2 \wedge \neg(c_1 = c_2)$. We have then defined two strategies to assess rule priorities in case of potential conflicts and prove that they are not conflicting strategies, that is we never obtain conflicting decisions when applying them (see [16]).

6 Discussion and Open Issues

As mentioned in the introduction, one of the interest of a formal specification is that it provides means to analyze conflicts. This is not an easy task because a security policy is an heterogeneous set of requirements. It corresponds to permissions, prohibitions and obligations and some requirements apply to explicit actions whereas others correspond to unknown actions. We suggest modelling these last requirements using the bring it about modality. This is especially useful to specify security requirements associated to new intrusions.

The next problem is then to analyze conflicts when the policy combines such heterogeneous requirements. Our suggested solution consists in reformulating security requirements that apply to explicit actions into requirements that use the bring it about modality.

For this purpose, we need a formal specification of various actions used to specify the policy through their pre and post conditions. Then, let us assume that a given subject s is prohibited to execute a given action a on some object o

in a given context c and let $pre(a, o)$ and $post(a, o)$ be respectively the pre and post conditions associated with the execution of action a on object o. Then this requirement is translated into the following security requirement:
$forbidden(E(s, a, post(a, o)), c\&pre(a, o))$
where $E(s, a, p)$ means subject s brings it about p by executing action s and $forbidden(p, c)$ is a diadic modality to specify that p is forbidden in context c.

We now obtain an homogeneous set of security requirements that we can analyze to detect conflits. Defining a complete analysis still represents further work. However, we can already state the following principles:

- Let $permitted(E(s, a, p_1), c_1)$ and $forbidden(E(s, a, p_2), c_2)$ be two security requirements. These requirements conflict if p_1 implies p_2 and c_1 and c_2 are consistent.
- Let $permitted(E(s, a, p), c)$ and $forbidden(E(s, S))$ be two security requirements where S is an intrusion objective. These requirements are conflicting if p implies S.
- Let $permitted(E(s, a_1, p), c)$ and $forbidden(E(s, a_2, p), c)$ be two security requirements where a_1 and a_2 are different actions. These requirements are not necessarily conflicting. For instance, let a_1 be the action "Authentication using a credit card" and a_2 be "Authentication using a password" and p the proposition "s is authenticated". Then it is *not* conflicting to state that s is permitted to bring it about p by executing a_1 but prohibited to bring it about p by executing a_2.
- Let $permitted(E(s, a_1, p), c_1)$ and $forbidden(E(s, a_2, p), c_2)$ be two security requirements where a_1 is a non elementary scenario. These requirements are conflicting if action a_2 is part of scenario a_1 and c_1 and c_2 are consistent.

We plan to develop a complete analysis as an extension of the approach suggested in [9].

7 Conclusion

Specifying a security is a central issue when developing a secure information system. Since a security policy may include very different requirements, it is essential to use an homogeneous formal model to specify these different requirements and deontic logic provides such an adequate formalism.

Traditional security policy models only consider norms that apply to explicit elementary actions. In this paper, we first focus on intrusion detection policies and show that these traditional models are not expressive enough to specify security requirements corresponding to the detection of known non elementary intrusion scenarios. There are also not appropriate to specify security requirements that correspond to new unknown intrusions. For this last purpose, we suggest using the bring it about modality and security requirements correspond to prohibition to bring it about that some security objectives are violated.

The security policy would then not be complete if it does not include a reaction policy that specifies what happens in case of violation of some security requirement. We show that these requirements correspond to contrary to duty norms. Contrary to duty has been extensively investigated in the deontic logic literature and several proposals have been suggested to solve problems such as the Chisholm paradox or the pragmatic oddity. However, to our best knowledge, it is the first time practical applications of these works are investigated in the context of security policies. We suggest an approach to manage conflicts between reaction requirements and other security requirements based on the definition of priorities between contexts associated with the activation of these different requirements. We consider three different types of context called operational, threat and minimal contexts. Threat contexts are associated with reaction requirements and have higher priority than operational contexts. However, since reacting may have some negative side effects on the information system, we also consider minimal contexts associated with security requirements that must be preserve even when reactions are activated. Minimal contexts have higher priority than other contexts including threat contexts.

We finally address the issue of analyzing consistency of these different requirements. Since the security policy may include heterogeneous requirements corresponding to norms that apply to explicit actions and others to implicit actions, our proposal consists in translating all the requirements into norms specified using the bring it about modality. It is then possible to analyze possible conflicts between the different requirements. This approach requires further work to be validated and we plan to take our inspiration into [9] to define a complete algorithm to detect and solve conflicts in this case.

Acknowledgement. This work is supported by the European CELTIC RED (Reaction After Detection) and ANR Polux (Policy Unified Expression) projects.

References

1. Abou El Kalam, A., El Baida, R., Balbiani, P., Benferhat, S., Cuppens, F., Deswarte, Y., Miège, A., Saurel, C., Trouessin, G.: Organization Based Access Control. In: 4th IEEE Policy (June 2003)
2. Ayed, S., Cuppens-Boulahia, N., Cuppens, F.: An Integrated Model for Access Control and Information Flow Requirements. In: Cervesato, I. (ed.) ASIAN 2007. LNCS, vol. 4846, pp. 111–125. Springer, Heidelberg (2007)
3. Bell, D., LaPadula, L.: Secure Computer Systems: Unified Exposition and Multics Interpretation. Technical Report ESD-TR-75-306, MTR-2997, MITRE, Bedford, Mass (1975)
4. Benferhat, S., El Baida, R., Cuppens, F.: A Stratification-Based Approach for Handling Conflicts in Access Control. In: 8th ACM Symposium on Access Control Models and Technologies (SACMAT 2003), Lake Come, Italy (June 2003)
5. Bertino, E., Bonatti, P.A., Ferrari, E.: TRBAC: A temporal role-based access control model. ACM TISSEC 4(3), 191–233 (2001)
6. Bertino, E., Catania, B., Damiani, M.L., Perlasca, P.: Geo-rbac: a spatially aware rbac. In: 10th ACM SACMAT, June 1-3 (2005)
7. Broersen, J., Dignum, F., Meyer, J.-J., Dignum, V.: Designing a Deontic Logic of Deadlines. In: Lomuscio, A., Nute, D. (eds.) DEON 2004. LNCS (LNAI), vol. 3065. Springer, Heidelberg (2004)

8. Brunel, J., Bodeveix, J.-P., Filali, M.: A State/Event Temporal Deontic Logic. In: Goble, L., Meyer, J.-J.C. (eds.) DEON 2006. LNCS (LNAI), vol. 4048, pp. 85–100. Springer, Heidelberg (2006)
9. Cholvy, L., Cuppens, F.: Reasoning about norms provided by conflicting regulations. In: McNamara, P., Prakken, H. (eds.) Fourth International Workshop on Deontic Logic in Computer Science, Bologna, Italy (1998)
10. Cuppens, F.: Roles and Deontic Logic. In: Second International Workshop on Deontic Logic in Computer Science, Oslo, Norway (1994)
11. Cuppens, F., Autrel, F., Miège, A., Benferhat, S.: Recognizing Malicious Intention in an Intrusion Detection Process. In: HIS, Santiago, Chili (2002)
12. Cuppens, F., Cuppens-Boulahia, N., Ben Ghorbel, M.: High Level Conflict Management Strategies in Advanced Access Control Models. Electronic Notes in Theoretical Computer Science 186, 3–26 (2007)
13. Cuppens, F., Cuppens-Boulahia, N., Sans, T.: Nomad: A Security Model with Non Atomic Actions and Deadlines. In: 18th IEEE CSFW, pp. 186–196 (2005)
14. Cuppens, F., Miège, A.: Modelling Contexts in the Or-BAC Model. In: ACSAC (2003)
15. Cuppens, F., Miège, A.: Administration Model for Or-BAC. In: Computer Systems Science and Engineering (CSSE 2004), vol. 19 (May 2004)
16. Debar, H., Thomas, Y., Boulahia-Cuppens, N., Cuppens, F.: Using Contextual Security Policies for Threat Response. In: Büschkes, R., Laskov, P. (eds.) DIMVA 2006. LNCS, vol. 4064. Springer, Heidelberg (2006)
17. Demolombe, R., Bretier, P., Louis, V.: Norms with Deadlines in Dynamic Deontic Logic. In: ECAI, Riva del Garda, Italy (September 2006)
18. Demolombe, R., Louis, V.: Norms, Institutional Power and Roles: Towards a Logical Framework. In: Esposito, F., Raś, Z.W., Malerba, D., Semeraro, G. (eds.) ISMIS 2006. LNCS (LNAI), vol. 4203, pp. 514–523. Springer, Heidelberg (2006)
19. Demolombe, R., Louis, V.: Speech Acts with Institutional Effects in Agent Societies. In: Goble, L., Meyer, J.-J.C. (eds.) DEON 2006. LNCS (LNAI), vol. 4048, pp. 101–114. Springer, Heidelberg (2006)
20. Ben Ghorbel, M., Cuppens, F., Cuppens-Boulahia, N., Bouhoula, A.: Managing Delegation in Access Control Models. In: 15th ADCOM (2007)
21. Goguen, J., Meseguer, J.: Unwinding and Inference Control. In: IEEE Symposium on Security and Privacy, Oakland (1984)
22. Harrington, J.: Network Security: A Practical Approach. TheKaufmann Series in Networking (2005)
23. Harrison, M., Ruzzo, W., Ullman, J.: Protection in operating systems. CACM 19(8), 461–471 (1976)
24. Hilty, M., Pretschner, A., Basin, D.A., Schaefer, C., Walter, T.: A Policy Language for Distributed Usage Control. In: Biskup, J., López, J. (eds.) ESORICS 2007. LNCS, vol. 4734, pp. 531–546. Springer, Heidelberg (2007)
25. Meyer, J.-J.: 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)
26. Morin, B., Debar, H.: Correlation of Intrusion Symptoms: An Application of Chronicles. In: Vigna, G., Krügel, C., Jonsson, E. (eds.) RAID 2003. LNCS, vol. 2820, pp. 94–112. Springer, Heidelberg (2003)
27. Pacheco, O., Carmo, J.: A Role Based Model for the Normative Specification of Organized Collective Agency and Agents Interaction. Autonomous Agents and Multi-Agent Systems 6(3), 145–184 (2003)
28. Park, J., Sandhu, R.S.: The UCONABC usage control model. ACM Trans. Information and System Security 7(1) (2004)

29. Pörn, I.: Action Theory and Social Science; Some Formal Models. Synthese Library, vol. 120. D. Reidel, Dordrecht (1977)
30. Prakken, H., Sergot, M.: Contrary-to-duty obligations. Studia Logica 57(1), 91–115 (1996)
31. Preda, S., Cuppens-Boulahia, N., Cuppens, F., Garcia-Alfaro, J., Toutain, L.: Reliable Process for Security Policy Deployment. In: International Conference on Security and Cryptography (Secrypt 2007), Barcelona, Spain (July 2007)
32. Sandhu, R.S., Coyne, E.J., Feinstein, H.L., Youman, C.E.: Role-Based Access Control Models. Computer 29(2), 38–47 (1996)
33. van der Torre, L.W.N.: Violated Obligations in a Defeasible Deontic Logic. In: ECAI, Amsterdam, The Netherlands (1994)

Delegation of Control in Administrative Procedures

Joris Hulstijn, Jianwei Liu, and Yao-Hua Tan

Faculty of Economics and Business Administration
Vrije Universiteit, Amsterdam
jhulstijn@feweb.vu.nl, jliu@feweb.vu.nl, ytan@feweb.vu.nl

Abstract. Norms are implemented by administrative procedures. This paper addresses the delegation of control in administrative procedures. Instead of having to check all details, a controlling actor can trust the data provided by other actors, provided they can demonstrate to be 'in control'. In this paper we provide a conceptual analysis of situations in which control has been delegated. The approach is based on an analysis of the dependencies between activities performed by the actors involved and on evidence documents. To motivate and illustrate the approach, we discuss a case study about the redesign of EU customs procedures for collecting excise duties.

Keywords: administrative procedures, trust, evidence documents.

1 Introduction

Norms for a society are implemented by means of regulations, documents and administrative procedures. We can treat a norm as a kind of requirement, a desirable property of a system. In principle, there are many different implementations of the same norm, so these control procedures must be designed. This 'design perspective' on normative systems has recently attracted a lot of interest in the area of electronic institutions, e-commerce and multi-agent systems [11,12,2,4]. However, much of this work is theoretical, and based on legal or philosophical abstractions. Less in known about the way in which administrative procedures develop in practice. Only when procedures are redesigned, do we get a chance to look at the development process of control mechanisms. In the private sector, business process redesign is seen as a way to improve effectiveness and efficiency of administrative processes, often by making good use of information technology [9,6]. In government, where many administrative procedures are maintained, it is often harder to deploy redesign techniques, because processes may have other than operational objectives. In particular, complexity is added by the legal issues and issues of governance and control, required when implementing a norm.

One way to reduce complexity, is to delegate part of the control activities to another party, provided the party can demonstrate to be trustworthy. Such a redesign may provide great operational benefits, but there are huge risks too. The first risk for the legislator, has to do with assessing the trustworthiness of an actor to which procedures are delegated. The second risk has to do with the general norm that is being implemented. How can the legislator make sure that the original control objectives of the administrative procedures are at least preserved? In this paper we will focus on the latter topic.

R. van der Meyden and L. van der Torre (Eds.): DEON 2008, LNAI 5076, pp. 81–95, 2008.

> When redesigning administrative procedures by delegating control, how can we make sure that original control objectives are guaranteed or improved upon?

The possibility of delegating control may seem farfetched, but it has already found real application. Currently the European Union is redesigning customs procedures for excise goods, such as alcoholic beverages and tobacco. According to the European Commission, excise fraud regarding alcohol in the EU amounts to €1.5 billion yearly[1]. The redesign is meant to make procedures efficient, more secure, and more transparent. The measures cover three major changes to the Customs Code [8]: (i) require traders to provide customs authorities with information on goods, prior to import to or export from the European Union; (ii) provide trustworthy traders with specific trade facilities (Authorized Economic Operator, see below); (iii) introduce a mechanism for setting uniform risk-selection criteria for controls, supported by computerized systems.

Crucial to understanding both current and future customs procedures, is the role of evidence documents. The general norm is that excise duties are due in the country in which alcoholic beverages are sold. Over the years, a procedure has developed which uses a paper document, called the Accompanying Administrative Document (AAD). This document must prove to the customs office in the home country, that a shipment was indeed sold to end customers abroad, so that reimbursement of the excise duties is warranted. In practice, only about 5% of the AADs which are issued each year are checked by the customs and tax office. How can we redesign the procedures, such that the role of this evidence document is taken over by information technology?

Most trading companies have an ERP system in place, with details of their business processes, including procurement, logistics, and financial reporting. When an economic actor can demonstrate to be 'in control', customs need no longer check every transaction. Provided that certain criteria on the internal control systems, financial solvency and compliance behavior of the economic actor are met, the customs and tax office may assign it the status of AEO (Authorized Economic Operator). AEOs do not have to comply with regular customs procedures, but use simplified procedures. This provides great operational benefits, improves the image of the company and provides more certainty. In return customs retain the right to perform audits, and to access certain information extracted from the AEO's ERP systems [8,22]. This requires standardization of customs regulations, and inter-operability between information systems.

We observe a general shift in the way controls are being designed: instead of checking documents about all transactions, which is practically impossible, the authorities rely on risk monitoring based on information provided by the actors themselves. The decision which actors to trust is based on extensive auditing. In other words, there is shift from a transaction-based control model, to a relation-based control model.

In this paper we propose an approach for understanding the redesign of controls, using a so called *actor-activity-document* analysis [1,14]. For each situation, we analyze for each of the actors involved, what the dependencies are between the activities that they have as their goals, and subsequently, what their control needs are. Furthermore we analyze the required evidence documents, which may be provided by other actors to which the control is delegated.

[1] EU Commission. *EU coherent strategy against fiscal fraud*. Retrieved 18 Oct 2007 from http://europa.eu/rapid/pressReleasesAction.do?reference=MEMO/06/221

The remainder of the paper is structured as follows. In Section 2 we provide a theoretical background of the actor-activity-document analysis, based on agency theory, control theory and theories regarding trust. In section 2.4 we then describe how to perform an actor-activity-document analysis. In Section 3 we discuss the case study of customs procedures for collecting excise duties. An actor-activity-document analysis, shows that the redesigns make sense from a control perspective.

2 Delegation, Control and Trust

The approach is based on a combination of agency theory, control theory and theories of transaction trust, taken from the literature on management and organization.

2.1 Agency Theory

A well known theory in sociology and management accounting is *agency theory*, also called *principal-agent* theory. See Eisenhardt [7] for a survey. Agency theory studies the relationship between two parties: the *principal*, who delegates some activity, and the *agent*, to whom the activity is delegated. The theory argues that if (1) the principal and the agent are utility maximizers with bounded rationality and (2) there is information asymmetry in favor of the agent, the agent may behave opportunistically. Agency theory distinguishes two types of opportunistic behavior.

The first type is caused by *hidden information*: the principal can not be sure that the agent accurately presents his ability to do the work. For example, a producer (agent) generally has better information about the product he is producing, than someone who wants to buy the product (principal). The generally accepted control mechanism against hidden information is *screening*: the principal collects information about the reliability of the agent, before agreeing on a transaction.

The second type is caused by *hidden action*: the principal can not be sure whether the agent did his work according to the contract or not. For example, the producer may use low quality components to produce a product. As a result, the quality of the product is lower than agreed in the contract. The generally accepted control mechanisms against hidden action are *monitoring* the agent, and creating *incentives* to motivate the agent not to behave opportunistically [7]. The hidden action problem arises *ex-post*, after the contract is settled, but usually the contract on incentives and penalties is agreed ex-ante.

2.2 Trust

Whenever parties depend on each other, but cannot control each other, lack of trust is likely. Trust has been defined as

> “The willingness of a party to be vulnerable to the actions of another party based on the expectation that the other party will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party”[15, p.712].

Without prior trust, the party who invests in a transaction, called the *trustor*, is uncertain whether the other party, the *trustee*, will perform its part of the deal or will defect and behave opportunistically. In transaction cost economics, this kind of behavior is called 'ex-post' opportunism [23]. Legal provisions in contracts are typically meant to reduce the chances of such opportunistic behavior.

However, trust does not have to depend on the trustor and trustee alone [20]. Institutional control measures can guarantee performance according to contract. Think of a legal system, with general provisions against fraud. When there is no overarching legal system, as in international trade, parties can also arrange for control measures themselves. An example of such a control mechanism is an Escrow service [10].

To assess the effectiveness of a control from the point of view of the trustor, we can use a kind of game theoretic reasoning. We compare the behavior of the trustee in a scenario in which a control mechanism applies, with behavior in a scenario without institutional control. It is not a coincidence that the effectiveness of an Escrow service has been shown by game theoretic means [10]. This kind of reasoning can also be modeled by the qualitative game theory developed by Boella and Van der Torre [2].

2.3 Internal Control and International Trade Procedures

Over the years, practices have evolved for designing control mechanism. A control mechanism prescribes how to organize business processes in order to prevent, detect or reduce the risks posed by a control problem. A control problem is a set of related threats, identified by auditing and risk assessment. Internal control theory is concerned with administrative and organizational measures inside an organization [19,17], see also frameworks like COSO and COBIT. We have also used work on *inter-organizational trade procedures* [5,3].

Control measures generally consist of a *verification*, in which (evidence of) performance of some operational activity is compared with a standard or norm, i.e., with some claim about its legitimacy, quality or quantity [13]. Verification requires three inputs: (i) the operational activity itself, possibly represented by a *document to be verified*, (ii) a claim about the legitimacy or quality, i.e., the *standard* or *norm*, and (iii) one or more *supporting documents* containing further evidence about the operational activity. The result of a verification is usually a decision to perform some action or not, or else a document stating the decision. A template of a verification activity is shown in Figure 1. To simplify the diagrams, in the remainder of the paper, we will only show the positive outcome of a verification activity, since the negative outcome always leads to the end of the process.

For ex-ante controls like screening, verification is concerned with the previous behavior of the actor. An example is the verification of the financial solvency, compliance record and internal control system (supporting documents) in order to acquire the status of AEO, upon request (document to be verified). The norms are stated in customs regulations [22]. A classic example of an ex-post control, is the three-way reconciliation used in procurement: before paying an invoice (document to be verified), delivery of the goods at the warehouse (operational activity), is checked against the purchase order (supporting documents) [17].

In international trade, actors are not in a position to verify an activity, because of the different locations. So control must be delegated. Based on procurement practices, Chen [5] developed a list of principles about the use of evidence documents. For

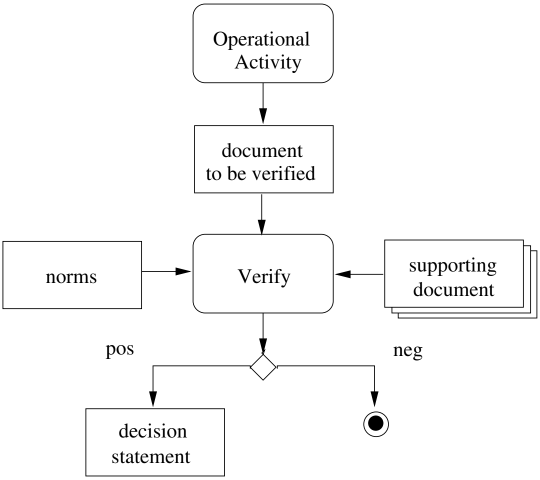


Fig. 1. Verification with required documents

example, a control activity providing evidence, must directly follow the operational activity it controls. This reduces the possibility of error or manipulation of the evidence.

By itself, verification of documents cannot provide reasonable assurance. Crucial is the control environment: the organizational system in which the controls are embedded, the system of internal control, and the dependencies between actors. General control principles can be formulated as follows [16,18].

1. *Separation of Duties*. Critical processes or activities are divided into at least two separate activities, execution and control, which should not be performed by the same person or organizational unit.
2. *Delegation*. Delegation is an important part of any working organization. Some work is better done by specialists. Delegation is the specification of responsibilities, through which a superior transfers authority downward in the organization along with the obligation to perform tasks. In case of delegation outside the organization(outsourcing), we get the control problems discussed in Section 2.1.
3. *Supervision, Review and Audit*. Supervision and review make certain that delegated activities are carried out as required. Supervision is done by a superior position. Reviewing is task-specific and does not need to be performed by a superior. Auditing must provide reasonable assurance that a control system performs its functions.

A principle that does not always hold, but that will make a control stronger, is the principle of *opposed interests*. Whenever existing evidence is used in a control, such as commercial documents, the evidence should be provided by an actor with interests opposed to those of the actor being controlled. For example, in checking excise declarations, the customs could use the invoice produced by the buyer.

2.4 Actor-Activity-Document (A-A-D)

It is easy to detect parallels in the theories above. Usually, the principal will also be the trustor, and when trust is not sufficient, additional control measures will have to be taken. In that case, the principal will also be the controlling actor, unless control has been delegated. When there is sufficient trust, the agent will also be the trustee. In case control has been delegated, the actor who produces evidence on which the control is based, the evidencing actor, will be different from the principal. Generally, evidence is provided in the form of an evidence document, which plays the role of supporting document in a verification activity.

Extracting concepts from literature mentioned above, we conclude that an effective control procedure should enable a control actor to carry out control activities by means of sufficient and independent documentary evidence. Yet, only exchanging documents between actors could not ensure a seamless control; a constraint of independence needs to be noticed. Stemming from one of the most fundamental principles of accounting practice, *segregation of duties* [19], we argue that a good design of the control procedure should include the separation of assigned duties and responsibilities in such a way that no single actor can both perpetrate and conceal errors or irregularities.

Three elements, namely *actor*, *activity* and *document* can be identified. We distinguish three types of actor: *responsible actor* (agent), *control actor* (principal), and *evidencing actor*, three types of activity: *operational activity*, *evidencing activity* and *control activity* (verification), and three kinds of document: *document to be verified*, *supporting document* and *verified document* (decision). The general idea is that by separating actors with different activities and documents, effective inter-organizational control can be designed. Because the three elements, actor, activity and document are crucial, the approach we use is called *A-A-D*[2].

A detailed description of the A-A-D components is given in Table 1. Based on the principles discussed above, we developed a checklist to help domain experts identify control problems and redesign control mechanisms. The checklist is given in Table 2.

3 Case Study: Excise Duties

The European Union is currently reshaping its customs legislation and practices, to deal with the dilemma between increasing security, financial and health requirements, and the need to reduce the administrative burden. The effort is based on three pillars: (i) extensive use of information technology (e-customs), (ii) public-private partnerships between customs and businesses, and (iii) collaboration between the national customs administrations. The concepts introduced to deal with this challenge are Authorized Economic Operators, explained above, and the vision of a *single window*, “to enable economic operators to lodge electronically and once only all the information required by customs and non-customs legislation for EU cross-border movements of goods” [21]. Such a single window should replace the multiple overlapping requests for tax and customs information, that export companies are now faced with.

[2] Note the use of ‘-’ to avoid confusion with the Administrative Accompanying Document.

Table 1. Components of the A-A-D approach

Actor An actor is a person or a group of persons, playing an organizational role, and performing activities to achieve its objectives in cooperation with other actors.

- **Responsible actor**: The actor who is responsible for performing an operational activity, or for having an operational activity performed (agent).
- **Control actor**: The actor who has a direct interest in controlling an operational activity executed by some other actor (principal).
- **Evidencing actor**: The actor who witnesses the execution of an operational activity, verifies the completeness, accuracy and compliance with applicable organizational policies, and testifies the outcome.

Activity An activity is an action or sequence of actions, mediated by resources and tools (e.g., documents).

- **Operational activity**: Basic business operations to obtain business value or achieve an operational goal.
- **Control activity**: Reconcile and verify records, documents or messages from the responsible actor and evidencing actor (verification).
- **Evidencing activity**: Witness the execution of the operational activity, verify the completeness, accuracy and accordance with organizational policies and rules and testify the outcome (witness).

Document A document contains certified information, interchanged among actors in a administrative procedure. Documents may take different forms: paper documents, records in a database, or electronic messages.

- **To-be-verified Document**. A document issued by the responsible actor to prove completion of the operational activity.
- **Supporting Document** Any document containing evidence to support the control actor, if he/she could not directly witness the performance of the operational activity.
- **Verified Document** The document stating the decision of the control actor after verifying or reconciling the to-be-verified document and supporting documents, from which a conclusion of an effective control can be drawn.

Table 2. A-A-D Checklist, largely based on audit principles of Chen [5]

P1 Does a control activity exist and directly follow the corresponding operational activity?

P2 Can the control actor directly witness the execution of the operational activity? If not, is the evidencing activity (witnessing) delegated to an evidencing actor (trusted third party)?

P3 Is there a supporting document furnishing the evidencing activity?

P4 Is the supporting document the result of a previous evidencing activity, directly witnessing the operational activity to be controlled?

P5 Is the supporting document directly transferred to the control actor from the evidencing actor who witnesses the operational activity?

P6 Is the supporting document generated by an actor independent of the actor who generates the to-be-verified document?

P7 Are the control activity and corresponding operational activity assigned to different actors?

P8 Are the actors responsible for the operational activity and its corresponding control activity socially detached?

The case study, called Beer Living Lab (BeerLL), is embedded in the ITAIDE project [1]. Partners in the project are the Dutch Tax and Customs Administration (Customs NL), Her Majesty's Revenue and Customs (Customs UK), a large international beer producer (BeerCo), and various technology and software providers. The study focuses on the redesign of export procedures for excise goods. The main principle of excise law is as follows: excise duties are due in the country in which the goods are sold to end customers. An exporting company is exempt from excise duties, provided it can be shown that the excise goods were indeed exported and sold abroad. How can this norm be implemented? First we discuss the current implementation. Then, we discuss the implementation of the norm after the redesign. Using the actor-activity-dependency approach we demonstrate that the projected situation does indeed guarantee, or even improve upon, the original control objectives.

3.1 Current Controls

Being an international brewery, BeerCo NL is exporting thousands of tons of beer every day to its counterpart BeerCo UK. The current excise control is based on physical inspections and on the AAD. The AAD performs two roles: one as export evidence document when stamped by UK Customs; the other to identify the cargo in case of a physical inspection en route. Also commercial shipping documents, like the waybill, accompany a container. The ADD is stamped by Customs UK, to certify that the goods arrived in the UK. To acquire evidence that the beer was sold to end customers, Customs UK relies on so called Excise Warehouses. These are traders with a specific status, which gives them the authority to keep a certified administration of foreign beer sold. So also in the current situation, part of the control has been delegated. After stamping the AAD, Customs UK sends the document back to the customs broker, which is usually the shipping company, who will forward it back to BeerCo NL. For the beer that BeerCo NL has claimed to have sold outside the Netherlands, excise exemption is given by default. The legitimacy can be verified afterwards by comparing excise declarations with the AADs produced, and possibly with records of physical inspections. The current exercise procedures are listed in Figure 2 and Table 3.

Table 3. Analysis of current control in BeerLL

control objective:	be exempt from excise duties provided beer is sold abroad
responsible actor:	BeerCo NL
operational activity:	export beer, without excise duties
control actor:	Customs NL
control activities:	physical inspection of shipments (random sample) verification of excise declarations against AAD, regular auditing
evidence actor:	Customs UK, Excise Warehouse
evidence activity:	testify arrival in UK, testify sale to end customers
to-be-verified document:	excise declaration
supporting document:	stamped AAD, evidence of inspections and audits
verified document:	approved excise declaration

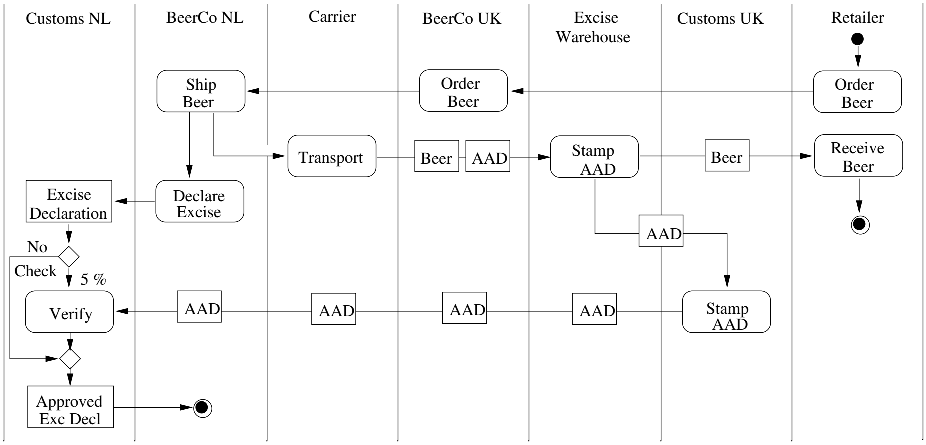


Fig. 2. Partial model of current excise procedures

The current procedure has three major disadvantages:

1. *Delay*: transferring the paper AAD can take months, so verification is done long after the fact, This harms the effectiveness of the control. In practice, often no checking is done, because it is too labor intensive. As a result, BeerCo NL will only submit AADs upon request of Customs NL.
2. *Many parties*: the AAD control involves many parties: BeerCo NL, BeerCo UK, customs broker, shipping company, and Excise Warehouse. Some parties have a financial interest in violating the control. A paper document is easily falsified. The document is transferred back along the supply chain, so each of these parties individually, or colluded, have the opportunity to alter or hide information. If the document is tampered with, it is difficult to prove where the alteration originated.
3. *Inefficiency*: physical inspections and excise declaration checks are labor intensive. In real life only about 5% of the AADs are used to check excise declarations, and even less then 5% of the containers are physically inspected at the border. Still, many parties complain about the huge administrative burden resulting from paperwork, and inspection delays.

3.2 Control Delegation Based on Inter-organizational Systems

The redesign is driven by so called inter-organizational systems (IOS), which should enable enhanced supply chain management and systems auditing. These systems are developed collaboratively, in a public-private partnership between the national customs organizations and local businesses with the status of Authorized Economic Operator (AEO). The design of the IOS relies on two innovative technologies: Tamper-Resistant Embedded Controllers (TREC), a kind of intelligent seal to detect when a container is opened[3], and Electronic Product Code Information Services (EPCIS), a shared

[3] Further information on TREC is available at http://www.zurich.ibm.com/news/05/trec.html and http://www.zurich.ibm.com/csc/process/securetradelane.html, accessed on Oct 31, 2007.

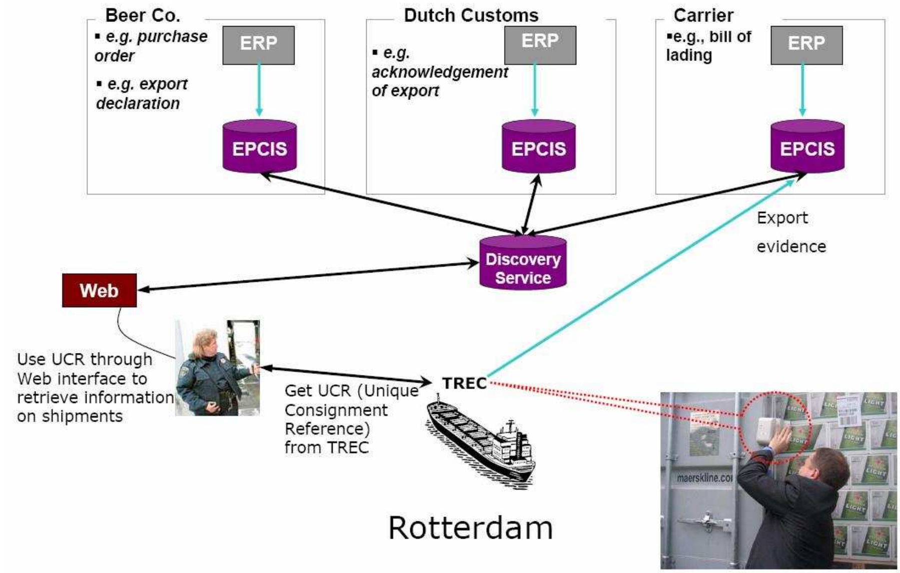

Beer Co.
▪ e.g. purchase order
▪ e.g. export declaration
ERP
EPCIS
Dutch Customs
▪ e.g. acknowledgement of export
ERP
EPCIS
Carrier
▪e.g., bill of lading
ERP
EPCIS
Export evidence
Discovery Service
Web
Use UCR through Web interface to retrieve information on shipments
Get UCR (Unique Consignment Reference) from TREC
TREC
maerskline.com
Rotterdam

Fig. 3. BeerLL IOS setting

shipment data repository. A TREC device has the following features: (1) sensors inside the container to monitor parameters including humidity, temperature, shocks and unauthorized container openings; (2) real-time container location tracing through continuous satellite connection (GPS); (3) connection to information systems during transport; and (4) ability to send messages triggered by predefined rules. For example, the TREC may send an alert when the container arrives at a geographical location, or when the temperature is lower or higher than predefined limits. TREC devices must be equipped with encryption techniques, to ensure authenticity and integrity of the messages. The shared data repository used in the supply chain makes use of the EPCIS non-proprietary standards[4]. Each partner in the supply chain can make a copy of the relevant shipment data from its own ERP system, and publish it in a shared data repository, where it becomes accessible to other supply chain partners as well as too government agencies. EPCIS will use a service oriented architecture (SOA), to minimize inter-operability problems.

In the redesigned excise procedures, BeerCo will now ship its goods in containers equipped with TREC. TREC devices can ensure shipment integrity, and enhance their security. Each TREC device has a unique key, called Unique Consignment Reference (UCR). By means of hand-held devices, customs officers can read the UCR off a TREC device and obtain access to the corresponding data in the EPCIS databases of supply chain partners. See Figure 3. Subject to periodic auditing, BeerCo will enjoy the status of Authorized Economic Operator (AEO). This means that Customs NL can put a higher level of trust in the information provided by BeerCo in the excise declarations, and in the EPCIS repository. It also means that Customs NL can delegate part of the control activities – the most burdensome part – to other parties. One such party

[4] For further details see `http://www.epcglobalinc.org`, last accessed on Oct 31, 2007.

Table 4. A-A-D Checklist applied to current and redesigned excise procedures

Principle	P1	P2	P3	P4	P5	P6	P7	P8
Current	5% verification	3rd party	upon request	yes	**no**	yes	yes	yes
Redesigned	yes, always	3rd party	yes, always	yes	yes	yes	yes	yes

will be the TREC service provider, a company who will install and maintain TREC devices. Given the sensitive nature of the technology, the TREC service provider will also have to be certified by the customs and tax office. Thanks to the GPS built into the TREC devices, one can automatically witness when a container has crossed a border, and create an alert for customs NL. The electronic seal will make sure that the container is secure en route. If anything goes wrong, alert information will be sent to Customs NL as an extraordinary event. The original administrative burden for Customs NL is lowered. The only remaining control activities are checking extraordinary events and periodic auditing. This means that resources become available to enhance the level of control.

3.3 Applying A-A-D to the Case Study

A pending issue for the redesign in this case is whether the redesigned control procedure can still preserve or even improve the original control. Our Actor-Activity-Document analysis, can serve as a guideline for such control validation. We first identify the A-A-D components. For the current control procedures, the components are listed in Table 3. Note that if some A-A-D components cannot be identified, this is already an indicator for potential control problems. After identifying the A-A-D components, the checklist (Table 2) is used to identify control problems. A brief summary of the results is listed in the first column of Table 4.

In general, when assessing the effectiveness of control measures, one has to verify the adequacy of the *design*, their *existence* in regulations, and their *operational effectiveness*. Table 4 shows that with regard to existence and design, the current procedures are sufficient, with a notable exception for P5, that a supporting document must be directly transferred to the control actor. The AAD is handled by all parties in the supply chain, and can therefore be tampered with. Regarding operational effectiveness, the current procedures are clearly lacking. In practice only about 5% of the excise declarations are checked against the AAD. This often happens months after the shipment was made. What is needed is reliable evidence, that the goods where indeed exported.

Does the redesigned procedure preserve or even improve on the original controls? In the new situation, the number of parties has decreased. BeerCo UK, Excise Warehouse and Customs UK have been removed. Customs UK can now concentrate on its own task, of collecting excise duties within the UK. The role of evidencing actor is played by the TREC service provider, who should be certified. The TREC alerts Customs NL when the container has left the Netherlands. Such a location alert can play the role of supporting document. Because the message is sent directly to Customs NL, manipulation by other parties has become impossible. The redesigned system supports an

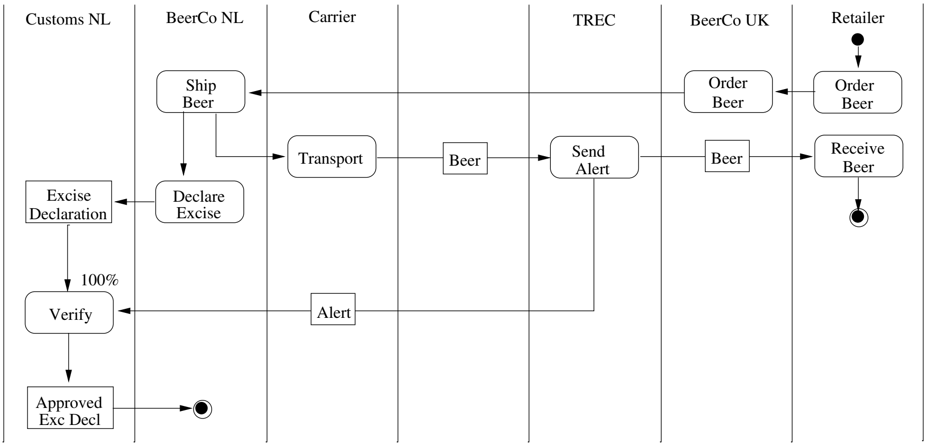


Fig. 4. Redesigned control in BeerLL

Table 5. Analysis of redesigned control in BeerLL; changes in bold

control objective:	be exempt from excise duties provided beer is sold abroad
responsible actor:	BeerCo NL
operational activity:	export beer, without excise duties
control actor:	Customs NL
control activities:	**physical inspection supported by TREC and EPSIS (random sample)** **verify excise declaration against EPSIS, TREC alerts** **regular audit AEO, TREC service provider**
evidence actor:	**supply chain partners (EPSIS), TREC service provider**
evidence activity:	testify arrival in UK, testify sale to end customers
to-be-verified document:	excise declaration
supporting document:	**TREC alert, EPSIS**
verified document:	approved excise declaration

automated 100% check of excise declarations. In essence, it becomes a preventive control: excise reimbursement is only granted when the excise declaration is verified. This is opposed to the current situation, in excise is reimbursed by default, and tax officers perform sample tests of paper AADs.

Table 5 summarizes these changes. A graphical depiction of the new process is given in Figure 4. Table 4 summarizes that the deficiencies of the current design, have been removed in the new situation.

It is interesting to note that in EPSIS, shipment data is used and provided by different supply chain partners. This increases the reliability of the shipment data, because different partners may have opposed commercial interests. For example, if BeerCo NL would overstate the amount of beer exported, in order to reclaim more excise duties than due, the buyer of the beer would protest, because no such amount was delivered, and because the buyer will have to pay excise duties in the UK for this amount.

3.4 Discussion on Research Limitations

In this section we briefly discuss the limitations of our research, and the relevance of formal specification and verification to this kind of research problem. Control procedures are often evaluated by means of a risk assessment. The risk that a misstatement or violation goes undetected is called audit risk [17]:

$$\text{audit risk} = \text{inherent risk} \times \text{control risk} \times \text{detection risk}.$$

The inherent risk, is the risk that errors are made or that anyone would want to violate excise laws in the first place. This is beyond our control. The control risk, is the risk that control procedures do not prevent, or detect and correct a violation. The detection risk is the risk that remaining cases are not detected during investigation by the auditor, e.g. by manually checking samples of transactions. Redesign of administrative procedures has the greatest impact on control risk. In our case, direct transfer of the TREC message prevents collusion and tampering. Also the changes of detection and correction are improved, since all excise declarations are now verified. When the control risk is reduced, there is less reason to deal with the residual risk by random inspections.

That means that we must demonstrate that the control procedures are 'watertight': that there is a reasonable assurance that potential incidents are prevented, or detected and corrected. For such a task formal specification of (relevant parts of) the control mechanism, and subsequent verification of desirable properties, is possible and useful. Formal techniques can be used in various ways. If we focus on the transfer of evidence, the processes in Figure 2 and 4 should be translated into a formal representation, such as a Petri net [13] for example. The crucial part for demonstrating P5, are assumptions about access to evidence documents and the possibility of tampering. If we focus on the motivation of actors for collusion or tampering with evidence, a much more elaborate model would be needed. One candidate would be the qualitative game theory, developed by Boella and Van der Torre [2]. In our opinion, deontic logic itself would not be a good candidate to verify these control procedures. The properties we want to show (no tampering; no collusion) are not themselves normative. However, together with other properties of the system, these properties help to assure that a norm – the excise law – is implemented in a effective and efficient way.

In this paper however, we have not used formal specification and verification techniques. The research is still in an initial stage. Detailed protocols for e.g. communication with TREC devices do not yet exist. For the 'proof of concept' needed at this stage of the project, an informal argument is more appropriate. In future research we plan to further analyze the reasoning of an auditor when they are auditing procedure redesigns such as was done in the BeerLL. For this we will apply formal modeling techniques. The aim is to discover the underlying rules and principles that guide the auditor's decision making process. These will also be used to develop an expert system that can support the auditor in its auditing tasks.

4 Conclusion

We have investigated the possibility of redesigning administrative procedures, by delegating part of the control activities.

Generally, we observe a shift in the design of controls from transaction-based control, where each individual transaction must be checked, to relation-based control, where part of the control activities are delegated to trusted actors. Such a shift has a huge potential for improving the operational efficiency and effectiveness of controls, but there are also risks involved. In particular, one must assess which actors can be trusted to delegate control to. Moreover, the redesigned administrative procedures must be shown to guarantee, or improve upon, the control objectives of the original procedures.

In this paper we have proposed an analysis approach called *actor-activity-document* (A-A-D), based on ideas taken from agency theory, trust, internal control and international trade procedures. For each scenario, we first identify the actors with their objectives. From an analysis of the dependencies between objectives, we can deduce which activities need to be controlled, and what control activities and evidencing activities are required. When the control actor (principal) is not in a position to control the activity, control must be delegated. Finally, we analyze which documents are involved in the control activity. Then we apply a checklist of principles about evidence and control.

We have applied the approach to a case study of the redesign of customs procedures related to excise. In the current situation, proof that goods have been exported and are therefore exempt from excise duties, is given by a paper document, called AAD. This document travels along the supply chain. The current procedure is inefficient, ineffective, and insecure. In practice only about 5% of excise declarations are verified, and parties in the supply chain can tamper with the AAD. The new situation makes use of TREC devices, which can monitor the state of a container, provide access control, and send alerts. In particular, TREC can send a message when the container crosses the border. Moreover, customs officers can use TREC to get access to the EPCIS system, which contains reliable shipment data shared by all parties in the supply chain. In the new situation, 100 % of the excise declarations can be automatically checked.

With the actor-activity-document method we demonstrate that the design and existence of the control measures in the new situation, can guarantee the control objectives of the original procedures, and indeed improve on them. However, it is too early to draw any conclusions about their operational effectiveness.

The redesign of the EU customs code faces a number of challenges. First, there are technical challenges, related to the safety and security of the TREC device, and access control to the EPCIS systems. To help address these challenges, formal methods can be useful. Second, there are organizational challenges related to standardization and inter-operability of commercial systems and the customs organizations. Third, the scenario assumes that parties in the market are willing to provide TREC-like services. A joint venture between a technology provider and an international shipping company or customs broker, might provide TREC services. Currently there is no evidence to support this view. Finally, the scenario requires huge investments by supply chain partners and AEOs. It is unclear, whether companies will have enough operational benefits from the simplified customs procedures, to warrant becoming an AEO. These challenges are exemplary for the challenges facing control delegation in general.

References

1. Baida, Z., Liu, J., Tan, Y.-H.: Towards a methodology for designing e-government control procedures. In: Wimmer, M., Scholl, H., Grönlund, A. (eds.) EGOV. LNCS, vol. 4656, pp. 56–67. Springer, Heidelberg (2007)
2. Boella, G., van der Torre, L.: A game theoretic approach to contracts in multiagent systems. IEEE Transactions on Systems, Man and CyBernetics - Part C 36(1), 68–79 (2006)
3. Bons, R.W.H., Lee, R.M., Wagenaar, R.W.: Designing trustworthy interorganizational trade procedures for open electronic commerce. International Journal of Electronic Commerce 2(3), 61–83 (1998)
4. Cardoso, H.L., Oliveira, E.: Electronic institutions for b2b: dynamic normative environments. Artificial Intelligence and Law 10.1007/s10506-007-9044-2 (2007)
5. Chen, K.: Schematic Evaluation of Internal Accounting Control Systems. PhD thesis, University of Texas at Austin (1992)
6. Davenport, T., Short, J.: The new industrial engineering: Information technology and business process redesign. Sloan Management Review, 11–27 (Summer 1990)
7. Eisenhardt, K.M.: Agency theory: An assessment and review. Academy of Management Review 14(1), 57–74 (1989)
8. European Commission. Regulation no 648/2005 of the European Parliament and of the Council, amending regulation no 2913/92 establishing the community customs code (2005)
9. Hammer, M.: Reengineering work: Dont automate, obliterate. Harvard Business Review, 104–112 (July/August 1990)
10. Hu, X., Lin, Z., Whinston, A., Zhang, H.: Hope or hype: On the viability of Escrow services as trusted third parties in online auction environments. Information Systems Research 15(3), 236–249 (2004)
11. Kartseva, V., Gordijn, J., Tan, Y.-H.: Towards a modelling tool for designing control mechanisms in network organisations. International Journal of Electronic Commerce 10(2), 57–84 (2005)
12. Kolp, M., Giorgini, P., Mylopoulos, J.: Multi-agent architectures as organizational structures. Auton Agent Multi-Agent Sys. 23, 3–25 (2006)
13. Lee, R.: Automated generation of electronic procedures: Procedure constraint grammars. ecision Support Systems 33, 291–308 (2002)
14. Liu, J., Baida, Z., Tan, Y.-H.: e-Customs control procedures redesign methodology: Model-based application. In: Österle, H., Schelp, J., Winter, R. (eds.) 15th European Conference of Information Systems (ECIS 2007), pp. 93–105 (2007)
15. Mayer, R., Davis, J., Schoorman, F.: An integrative model of organizational trust. Academy of Management Review 20(3), 709–734 (1995)
16. Moffett, J.D.: Control principles and role hierarchies. In: Proceedings of the third ACM workshop on Role-based access control, pp. 63–69. ACM, New York (1998)
17. Ronmey, M., Steinbart, P.: Accounting Information Systems, 10e. Prentice Hall, NJ (2006)
18. Schaad, A.: A Framework for Organisational Control Principles. PhD thesis, Department of Computer Science, University of York (2003)
19. Starreveld, R., de Mare, B., Joels, E.: Bestuurlijke Informatieverzorging (in Dutch), 4th edn., Samsom, Alphen aan den Rijn, vol. 1 (1994)
20. Tan, Y.-H., Thoen, W.: Formal aspects of a generic model of trust for electronic commerce. Decision Support Systems 33(3), 233–246 (2002)
21. TAXUD. Electronic customs multi-annual strategic plan (2006)
22. TAXUD. Authorised economic operators guidelines (2007)
23. Williamson, O.E.: Transaction cost economics: The governance of contractual relations. Journal of Law and Economics 22, 3–61 (1979)

Variations in Access Control Logic

Martín Abadi

University of California, Santa Cruz
Microsoft Research, Silicon Valley

Abstract. In this paper we investigate the design space of access control logics. Specifically, we consider several possible axioms for the common operator `says`. Some of the axioms come from modal logic and programming-language theory; others are suggested by ideas from security, such as delegation of authority and the Principle of Least Privilege. We compare these axioms and study their implications.

1 Introduction

While access control appears in various guises in many aspects of computer systems, it is attractive to reduce it, as much as possible, to few central concepts and rules [17]. The development and use of general logics for access control is an ongoing effort in this direction. In this paper, we examine and compare several logics for access control.

The logics that we consider all have the same operators and intended applications, but they differ in their axioms and rules. They all start from propositional logic with the `says` operator, which is central in several theories and systems for access control (e.g., [1, 4, 5, 6, 8, 9, 12, 14, 16, 19, 20]). Moreover, they all allow the definition of a "speaks for" relation [4, 16, 18] from `says` and quantification: A speaks for B if, for every X, if A `says` X then B `says` X. In a formula A `says` s, the symbol A represents a principal and s represents a statement (such as a request or a delegation of authority). Intuitively, A `says` s means that A supports s, whether or not A has uttered s explicitly.

Perhaps because intuitive explanations of `says` are invariably loose and open-ended, the exact properties that `says` should satisfy do not seem obvious. The goal of this paper is to investigate the space of options, exploring the formal consequences and the security interpretations of several possible axiomatizations, and thus to help in identifying logics that are sufficiently strong but not inconsistent, degenerate, or otherwise unreasonable.

Some of the axioms that we study come from modal logic [15], computational lambda calculus [21], and other standard formal systems. Other axioms stem from ideas in security, such as delegations of authority and the Principle of Least Privilege [22]. For instance, we consider the hand-off axiom, which says that if A says that B speaks for A, then B does speak for A [16]. We evaluate these axioms in both classical and intuitionistic contexts.

More specifically, we start with the basic axioms of standard modal logic, in particular that `says` is closed under consequence (if A says s_1 and A says that

R. van der Meyden and L. van der Torre (Eds.): DEON 2008, LNAI 5076, pp. 96–109, 2008.

s_1 implies s_2, then A says s_2), together with the necessitation rule (if s is valid then A says s). In addition, the axioms that we consider include the following:

1. The hand-off axiom, as described above, and a generalization: if A says that s_1 implies A says s_2, then s_1 does imply A says s_2. In the special case where s_1 is B says s_2, we obtain a hand-off from A to B for s_2.
2. A further axiom that if A can make itself speak for B, then A speaks for B in the first place. This axiom may be seen roughly as a dual to the hand-off axiom.
3. The axiom that s implies A `says` s. This axiom is similar to the necessitation rule but stronger, and has been considered in access control in the past. It is also suggested by the computational lambda calculus. We call it Unit.
4. The other main axiom from the computational lambda calculus, which we call Bind: if s_1 implies A says s_2, then A says s_1 implies A says s_2.
5. The axiom that if A `says` s then s or A `says false`. We call this axiom Escalation, because it means that whenever A `says` s, either s is true or A says anything—-possibly statements intuitively "much falser" than s.
6. An axiom suggested by the Principle of Least Privilege, roughly that if a principal is trusted on a statement then it is also trusted on weaker statements.

We obtain the following results:

- In classical logics, the addition of axioms beyond the basic ones from modal logic quickly leads to strong and surprising properties that may not be desired. Bind is equivalent to Escalation, while Unit implies Escalation.

  Pictorially, we have:

$$\begin{tikzcd}
\text{Unit} \arrow[d, Rightarrow] \\
\text{Escalation} \arrow[d, Leftrightarrow] \\
\text{Bind}
\end{tikzcd}$$

  There are systems intermediate between the basic modal logic and Escalation. For instance, one may require the standard axiom C4 from modal logic (if A says A says s then A says s) without obtaining Escalation. However, these intermediate systems appear quite limited in their support of delegation and related concepts.
- In intuitionistic logics, we have a little more freedom. In particular, a system that includes Unit and Bind, which we call CDD [2, Section 8], does not lead to Escalation.

  Pictorially, we have:

$$\begin{tikzcd}
\text{CDD} \arrow[d, Rightarrow] \arrow[dr, Rightarrow] & \text{Escalation} \arrow[d, Rightarrow] \\
\text{Unit} & \text{Bind}
\end{tikzcd}$$

  Many further refinements become possible, in particular because Escalation and Unit are independent intuitionistically.

- The general form of the hand-off axiom (1) is equivalent to Bind.
- Unit implies axiom (2). This axiom is equivalent to Unit if there is a truth-telling principal.
- Finally, Escalation implies axiom (6). Conversely, this axiom and C4 imply Escalation.

In addition to occasional trickiness in proofs, the main difficulties of this work are in identifying and formulating the results summarized above. While some previous work also explores various axiomatizations of access control logics (e.g., [4]), those explorations have focused on classical logics, dealing for instance with properties of compound principals. We previously knew that Unit implies Escalation [1], and that Bind implies the hand-off axiom [2]. All the other results appear to be new.

Section 2 reviews the basic intuitionistic and classical logics that serve as our starting point. Section 3 studies CDD, considering axioms (1), (2), (3), and (4). Section 4 focuses on Escalation (axiom (5)). Section 5 considers axiom (6). Section 6 concludes with a brief discussion.

2 Basic Logics

In this section we briefly review the basic logics on which we build.

2.1 Formulas

Formulas are given by the grammar:

$$s ::= \texttt{true} \mid (s \vee s) \mid (s \wedge s) \mid (s \rightarrow s) \mid A \texttt{ says } s \mid X \mid \forall X.\, s$$

where A ranges over elements of a set $\mathcal{P}$ (intuitively the principals), and X ranges over a set of variables. The variable X is bound in $\forall X.\, s$, and subject to renaming.

We write $\texttt{false}$ for $\forall X.\, X$. We write $s_1 \equiv s_2$ for $(s_1 \rightarrow s_2) \wedge (s_2 \rightarrow s_1)$. We write $A \Rightarrow B$ as an abbreviation for

$$\forall X.\, (A \texttt{ says } X \rightarrow B \texttt{ says } X)$$

This formula is our representation of "A speaks for B". We write $A \texttt{ controls } s$ as an abbreviation for $(A \texttt{ says } s) \rightarrow s$.

2.2 Basic Axioms and Rules

All of the logics that we consider are based on second-order propositional intuitionistic logic. We review this logic in Appendix A. In addition, we rely on a standard axiom (closure under consequence):

$$\forall X, Y. ((A \texttt{ says } (X \rightarrow Y)) \rightarrow (A \texttt{ says } X) \rightarrow (A \texttt{ says } Y))$$

and a standard rule (necessitation):

$$\frac{s}{A \texttt{ says } s}$$

Thus, we obtain a second-order, intuitionistic, multi-modal version of the standard logic K. It is the least system that we consider in this paper.

Sometimes we consider classical variants. In those, we use the following additional principle:

$$[\textit{Excluded-middle}] \quad \forall X. (X \vee (X \rightarrow \texttt{false}))$$

Throughout the paper, we also introduce other additional axioms, as explained in the introduction.

3 CDD

CDD arose as a simplified version of the Dependency Core Calculus (DCC) [3], but it is similarly adequate as a logic for access control [2, Section 8]. CDD is related to lax logic [10] and the computational lambda calculus [21]. It has been used for language-based authorization [11], and its central rules also appear in other systems for access control, such as Alpaca [19].

In comparison with DCC, CDD may be seen as straightforward and conservative. For instance, while DCC proves $(A \texttt{ says } B \texttt{ says } s) \rightarrow (B \texttt{ says } A \texttt{ says } s)$, CDD does not. Although we do not discuss DCC in detail, the results of this paper are relevant to DCC as well.

A self-contained definition of CDD is in Appendix B. In the context of the basic intuitionistic logic presented in Section 2.2, however, CDD amounts to adopting the following two additional axioms, Unit and Bind:

$$\begin{array}{ll} [\textit{Unit}] & \forall X. (X \rightarrow A \texttt{ says } X) \\ [\textit{Bind}] & \forall X, Y. ((X \rightarrow A \texttt{ says } Y) \rightarrow (A \texttt{ says } X) \rightarrow (A \texttt{ says } Y)) \end{array}$$

It is easy to show that neither of these axioms is derivable in the logic of Section 2.2, neither intuitionistically nor classically. We prove some stronger results below, in Section 3.1, also considering the axiom C4 mentioned in the introduction. In Sections 3.2 and 3.3, we relate Unit and Bind to formulas motivated by security considerations.

3.1 C4 in CDD

This section is devoted to some simple results on the relation between CDD and the axiom C4:

$$[\textit{C4}] \quad \forall X. (A \texttt{ says } A \texttt{ says } X \rightarrow A \texttt{ says } X)$$

We can replace Bind with the simpler C4 when we have Unit:

Proposition 1. *Starting from the basic logic (without Excluded-middle), we have:*

1. *Bind implies C4;*
2. *Unit and C4 (together) imply Bind;*
3. *C4 does not imply Bind;*
4. *Unit does not imply C4 (and a fortiori not Bind).*

Proof. 1. In Bind, take X to be $A\ \texttt{says}\ Y$.

2. For arbitrary X and Y, assume $(X \rightarrow A\ \texttt{says}\ Y)$ and $A\ \texttt{says}\ X$. We wish to show $A\ \texttt{says}\ Y$.
   By Unit, we have $(X \rightarrow A\ \texttt{says}\ Y) \rightarrow A\ \texttt{says}\ (X \rightarrow A\ \texttt{says}\ Y)$.
   By closure under consequence, $(X \rightarrow A\ \texttt{says}\ Y)$ yields $(A\ \texttt{says}\ X) \rightarrow (A\ \texttt{says}\ A\ \texttt{says}\ Y)$. By C4, we obtain $(A\ \texttt{says}\ X) \rightarrow (A\ \texttt{says}\ Y)$.
   It follows that $A\ \texttt{says}\ Y$.
3. We prove a stronger result in Proposition 2, with Excluded-middle.
4. Mapping the logic to its fragment without says (to System F [7, 13], essentially), we interpret $A\ \texttt{says}\ s$ as

$$(X_A \rightarrow s) \vee X_A$$

where X_A is a distinct type variable used only for this purpose for each principal A. This interpretation satisfies Unit. It does not satisfy C4, because

$$A\ \texttt{says}\ A\ \texttt{says}\ \texttt{false} \rightarrow A\ \texttt{says}\ \texttt{false}$$

translates to

$$((X_A \rightarrow ((X_A \rightarrow \texttt{false}) \vee X_A)) \vee X_A) \rightarrow ((X_A \rightarrow \texttt{false}) \vee X_A)$$

The left-hand side of this implication is intuitionistically provable, and the right-hand side is not, so the implication is not. ■

Bind does not imply Unit in the basic logic. We do not state it explicitly in the intuitionistic case (in Proposition 1, above) because it is a corollary from a stronger result in the classical case (Proposition 2, below). Conversely, Bind implies C4 in the classical case, but we do not state explicitly there because it follows from a stronger result in the intuitionistic case.

Proposition 2. *Starting from the basic logic plus Excluded-middle, we have:*

1. *C4 implies neither Bind nor Unit.;*
2. *Unit implies C4 (and therefore Bind);*
3. *Bind does not imply Unit.*

Proof. 1. We consider a Kripke model with two possible worlds w and w', with the accessibility relation $\{w, w'\} \times \{w'\}$ associated with A. This model satisfies C4. It does not satisfy the instance of Bind

$$(X \to A\ \texttt{says false}) \to (A\ \texttt{says}\ X) \to (A\ \texttt{says false})$$

for a proposition X that holds in w' but not in w (so A `says` X holds in w). It does not satisfy the instance of Unit $X \to A$ `says` X for a proposition X that holds in w but not in w'.

2. In classical logic, we assume A `says` A `says` X in order to prove A `says` X. We proceed by cases on A `says` X, using Excluded-middle. If A `says` X holds, we are done. On the other hand, if $(A\ \texttt{says}\ X) \to \texttt{false}$ holds, Unit yields $A\ \texttt{says}\ ((A\ \texttt{says}\ X) \to \texttt{false})$, and closure under consequence yields A `says false`, and then A `says` X.
   That Unit implies Bind follows from Proposition 1, which says that Unit and C4 together imply Bind.
3. This part follows from Theorem 4. ■

3.2 Hand-Off in CDD

In CDD, we obtain the hand-off axiom as a theorem:

$$[\textit{Hand-off}] \quad A\ \texttt{controls}\ (B \Rightarrow A)$$

A slight generalization of the hand-off axiom is also interesting and also a theorem:

$$[\textit{Generalized-hand-off}] \quad \forall X, Y.\ A\ \texttt{controls}\ (X \to A\ \texttt{says}\ Y)$$

Theorem 1. *Starting from the basic logic: Bind is equivalent to Generalized-hand-off.*

Proof. First we establish that Bind implies Generalized-hand-off. In order to prove that, for all X and Y, we have $A\ \texttt{controls}\ (X \to A\ \texttt{says}\ Y)$, we assume X and $A\ \texttt{says}\ (X \to A\ \texttt{says}\ Y)$ in order to prove A `says` Y. By Bind, we have:

$$\begin{array}{l}((X \to A\ \texttt{says}\ Y) \to A\ \texttt{says}\ Y) \\ \quad\to \\ A\ \texttt{says}\ (X \to A\ \texttt{says}\ Y) \to A\ \texttt{says}\ Y\end{array}$$

Since we have $A\ \texttt{says}\ (X \to A\ \texttt{says}\ Y)$, we obtain:

$$\begin{array}{l}((X \to A\ \texttt{says}\ Y) \to A\ \texttt{says}\ Y) \\ \quad\to \\ A\ \texttt{says}\ Y\end{array}$$

Since we also have X, and hence $(X \to A\ \texttt{says}\ Y) \to A\ \texttt{says}\ Y$, we conclude A `says` Y.

For the converse, let us assume that $A\ \texttt{controls}\ (X \to A\ \texttt{says}\ Y)$ in order to prove that $(X \to A\ \texttt{says}\ Y) \to (A\ \texttt{says}\ X) \to (A\ \texttt{says}\ Y)$. So let us assume that $X \to A\ \texttt{says}\ Y$ and A `says` X in order to prove A `says` Y. If A `says` X, by closure under consequence we have $A\ \texttt{says}\ ((X \to A\ \texttt{says}\ Y) \to A\ \texttt{says}\ Y)$ since $X \to ((X \to A\ \texttt{says}\ Y) \to A\ \texttt{says}\ Y)$ is valid. By Generalized-hand-off, we obtain $(X \to A\ \texttt{says}\ Y) \to A\ \texttt{says}\ Y$. Applying this to $X \to A\ \texttt{says}\ Y$, we conclude A `says` Y. ■

3.3 The Limits of Hand-Off in CDD

Suppose that a principal A is trusted on whether it speaks for another principal B on every statement. In CDD, it follows that A must speak for B in the first place, whether it says so or not. If A does not wish to speak for B, it should reduce its authority, for instance by adopting an appropriate role [16, Section 6.1]. This result might be seen as a reassuring characterization of who can attribute the right to speak for B; it may also be seen as a dual or a limitation of hand-off in the context of CDD.

More precisely, we define:

$$[\textit{Authority-shortcut}] \quad (\forall X.\, A \;\texttt{controls}\; (A \;\texttt{says}\; X \rightarrow B \;\texttt{says}\; X)) \rightarrow (A \Rightarrow B)$$

We obtain:

Theorem 2. *Unit implies Authority-shortcut.*

Proof. Suppose that, for all X, $A \;\texttt{controls}\; (A \;\texttt{says}\; X \rightarrow B \;\texttt{says}\; X)$ and suppose that, for some particular X, we have $A \;\texttt{says}\; X$. We wish to derive $B \;\texttt{says}\; X$.

Because $A \;\texttt{says}\; X$, Unit implies $A \;\texttt{says}\; B \;\texttt{says}\; X$. (Here we apply Unit under says.) Then by closure under consequence we have $A \;\texttt{says}\; (A \;\texttt{says}\; X \rightarrow B \;\texttt{says}\; X)$.

By our assumption that, for all X, $A \;\texttt{controls}\; (A \;\texttt{says}\; X \rightarrow B \;\texttt{says}\; X)$, we obtain $A \;\texttt{says}\; X \rightarrow B \;\texttt{says}\; X$.

Combining $A \;\texttt{says}\; X \rightarrow B \;\texttt{says}\; X$ with $A \;\texttt{says}\; X$, we obtain $B \;\texttt{says}\; X$, as desired.

The proof is peculiar, not least because the hypothesis $A \;\texttt{says}\; X$ is used twice in different roles. ■

A small variant of the proof of Theorem 2 shows that Unit also implies:

$$\forall X.\, ((A \;\texttt{controls}\; (A \;\texttt{says}\; X \rightarrow B \;\texttt{says}\; X)) \rightarrow (A \;\texttt{says}\; X \rightarrow B \;\texttt{says}\; X))$$

In other words, writing $A \Rightarrow_X B$ for $A \;\texttt{says}\; X \rightarrow B \;\texttt{says}\; X$ [18], we have that Unit implies:

$$\forall X.\, ((A \;\texttt{controls}\; (A \Rightarrow_X B)) \rightarrow (A \Rightarrow_X B))$$

The converse of Theorem 2 is almost true. Suppose that there is a truth-telling principal A, that is, a principal for which $\forall X.\, X \equiv (A \;\texttt{says}\; X)$. Applying Authority-shortcut to this principal, we can derive $s \rightarrow B \;\texttt{says}\; s$ by propositional reasoning, for every B and s. In other words, given such a truth-teller, we obtain Unit.

Nevertheless, the converse of Theorem 2 is not quite true. All basic axioms plus rules, plus Authority-shortcut, hold when we interpret $A \;\texttt{says}\; s$ as true, for every A and s. Unit does not hold under this interpretation.

In addition, we can prove that Authority-shortcut does not follow from other axioms (such as Bind), even in classical logic. In other words, Authority-shortcut appears to be very close to Unit, and can be avoided by dropping Unit.

4 Escalation

As indicated in the introduction, Escalation is the following axiom:

$$[\mathit{Escalation}] \quad \forall X, Y. \, ((A \; \texttt{says} \; X) \rightarrow (X \vee (A \; \texttt{says} \; Y)))$$

Equivalently, Escalation can be formulated as:

$$\forall X, Y. \, ((A \; \texttt{says} \; X) \rightarrow (X \vee (A \; \texttt{says} \; \texttt{false})))$$

Escalation embodies a rather degenerate interpretation of says. At the very least, great care is required when Escalation is assumed. For instance, suppose that two principals A and B are trusted on s, and that we express this as $(A \; \texttt{controls} \; s) \wedge (B \; \texttt{controls} \; s)$; with Escalation, if $A \; \texttt{says} \; B \; \texttt{says} \; s$ then s follows. Formally, we can derive:

$$(A \; \texttt{controls} \; s) \wedge (B \; \texttt{controls} \; s) \rightarrow ((A \; \texttt{says} \; B \; \texttt{says} \; s) \rightarrow s)$$

This theorem may be surprising. Its effects may however be avoided: A should not say that B says s unless A wishes to say s. As a result, though, the logic loses flexibility and expressiveness.

On the whole, we consider that Escalation is not a desirable property. Unfortunately, it can follow from the combination of properties that may appear desirable in isolation, as we show.

Theorem 3. *Starting from the basic logic (without Excluded-middle),*

1. *Unit and Bind (together) do not imply Escalation (in other words, Escalation is not a theorem of CDD);*
2. *Escalation implies Bind (and therefore C4).*

Proof. 1. Following Tse and Zdancewic [23], we can interpret CDD in System F [7, 13]. We map $A \; \texttt{says} \; s$ to $X_A \rightarrow s$, where X_A is a distinct type variable used only for this purpose. If s is provable in CDD, then its translation is provable in System F.
The translation of Escalation is:

$$\forall X, Y. \, ((X_A \rightarrow X) \rightarrow (X \vee (X_A \rightarrow Y)))$$

This formula is not provable in System F.

2. Suppose that $X \rightarrow A \; \texttt{says} \; Y$ and that $A \; \texttt{says} \; X$. We wish to prove $A \; \texttt{says} \; Y$.
By Escalation, $A \; \texttt{says} \; X$ implies $X \vee A \; \texttt{says} \; Y$. Combining this with $X \rightarrow A \; \texttt{says} \; Y$, we obtain $A \; \texttt{says} \; Y \vee A \; \texttt{says} \; Y$, that is, $A \; \texttt{says} \; Y$. ■

Theorem 4. *Starting from the basic logic plus Excluded-middle, we have:*

1. *Unit implies Escalation (and therefore Bind);*
2. *Escalation (and a fortiori Bind) does not imply Unit;*

3. *Bind implies Escalation;*
4. *C4 does not imply Escalation.*

Proof. 1. Suppose A $\texttt{says}$ X. If X is true, then we are done, as we obtain $X \vee (A\ \texttt{says}\ Y)$. If X is false, that is, $X \rightarrow \texttt{false}$ is true, then Unit yields $A\ \texttt{says}\ (X \rightarrow \texttt{false})$, and by closure under consequence we obtain $A\ \texttt{says}\ \texttt{false}$ and then $A\ \texttt{says}\ Y$ for any Y, and then $X \vee (A\ \texttt{says}\ Y)$.

2. Escalation (and a fortiori Bind) is true in a Kripke model with two possible worlds w and w', in which every principal is mapped to the universal accessibility relation $\{w, w'\} \times \{w, w'\}$. This Kripke model does not satisfy the instance of Unit $X \rightarrow A\ \texttt{says}\ X$ for a proposition X that holds in w but not in w'.
3. We prove Escalation by cases on whether X is true or not. If it is true, then Bind yields X vacuously, and hence Escalation. If it is false, then that means $X \rightarrow \texttt{false}$, which entails $X \rightarrow A\ \texttt{says}\ \texttt{false}$, and applying Bind with $\texttt{false}$ for Y we obtain $(A\ \texttt{says}\ X) \rightarrow (A\ \texttt{says}\ \texttt{false})$.
4. The Kripke model described in part 1 of Proposition 2 does not satisfy Escalation: $A\ \texttt{says}\ X$ at w means that X is true in w', while X may be false in w and $A\ \texttt{says}\ \texttt{false}$ is false in w. ■

Going further, in classical logic Unit implies that each principal A is either a perfect truth-teller or says $\texttt{false}$. In the former case, A speaks for any other principal; in the latter case, any other principal speaks for A. Formally, we can derive $(A \Rightarrow B) \vee (B \Rightarrow A)$. While this conclusion does not represent a logical contradiction, it severely limits the flexibility and expressiveness of the logic: policies can describe only black-and-white situations. This point is a further illustration of the fact that usefulness degrades even before a logic becomes inconsistent.

5 On the Monotonicity of Controls

The monotonicity of controls means that, if a principal controls a formula X, then it controls every weaker formula Y. Formally, we write:

$$[\textit{Control-monotonicity}] \quad \forall X, Y. \begin{pmatrix} (X \rightarrow Y) \\ \rightarrow \\ ((A\ \texttt{controls}\ X) \rightarrow (A\ \texttt{controls}\ Y)) \end{pmatrix}$$

This monotonicity property may seem attractive. In particular, it may make it easier to comply with the Principle of Least Privilege. This principle says [22]:

> Every program and every user of the system should operate using the least set of privileges necessary to complete the job.

The monotonicity of controls implies that, if A wants to convince a reference monitor of Y, and it can convince it of a stronger property X, then A should be able to state Y directly, rather than the stronger property X. For instance, suppose that Y is the statement that B may access a file f_1, and that X is the statement that B may access both f_1 and another file f_2. When A wishes to allow B to access f_1, it should not have to state also that B may access f_2. The monotonicity of controls allows A to say only that B may access f_1.

Nevertheless, the monotonicity of controls has questionable consequences.

Proposition 3. *Starting from the basic logic (without Excluded-middle), Control-monotonicity implies:*

$$A \texttt{ controls } s_1 \rightarrow A \texttt{ says } s_2 \rightarrow (s_1 \vee s_2)$$

Proof. We obtain $A \texttt{ controls } s_1 \rightarrow A \texttt{ says } s_2 \rightarrow (s_1 \vee s_2)$ from Control-monotonicity, as follows: Let X be s_1 and Y be $s_1 \vee s_2$. We have $X \rightarrow Y$. Suppose that $A \texttt{ controls } s_1$. Control-monotonicity yields $A \texttt{ controls } (s_1 \vee s_2)$. If $A \texttt{ says } s_2$, then we obtain $A \texttt{ says } (s_1 \vee s_2)$, and hence $s_1 \vee s_2$. ■

In Proposition 3, the formulas s_1 and s_2 may be completely unrelated. For instance, suppose that A controls whether B may access a file f_1, and A says that B may access another file f_2; curiously, we obtain that B may access f_1 or B may access f_2, by Proposition 3.

In fact, the monotonicity of `controls` is equivalent to Escalation in the presence of C4. (Intuitionistically, C4 is strictly required for this equivalence.)

Theorem 5. *Starting from the basic logic (without Excluded-middle), the following are equivalent:*

- *Escalation,*
- *C4 and Control-monotonicity.*

However, neither Control-monotonicity nor C4 implies the other, not even in combination with Unit.

Proof.
- Escalation implies C4, by Theorem 3.
- Escalation implies Control-monotonicity:
  Suppose $X \rightarrow Y$ and $A \texttt{ controls } X$. We wish to prove $A \texttt{ controls } Y$, so we assume $A \texttt{ says } Y$ in order to derive Y.
  By Escalation, we obtain $Y \vee A \texttt{ says } X$. Since $A \texttt{ controls } X$, it follows that $Y \vee X$. Since $X \rightarrow Y$, it follows that Y, as desired.
- C4 and Control-monotonicity together imply Escalation:
  We have $A \texttt{ controls } A \texttt{ says false}$ by C4, and $(A \texttt{ says false}) \rightarrow (Y \vee A \texttt{ says false})$ by propositional reasoning, so Control-monotonicity yields $A \texttt{ controls } (Y \vee A \texttt{ says false})$.
  Since $A \texttt{ says } Y$ implies $A \texttt{ says } (Y \vee A \texttt{ says false})$ by propositional reasoning and closure under consequence, we obtain that $A \texttt{ says } Y$ implies $Y \vee A \texttt{ says false}$.
- C4 does not imply Control-monotonicity, even in combination with Unit, by Proposition 1 (which says that Bind implies C4) and Theorem 3 (which says that Unit and Bind do not imply Escalation).

– Starting from the basic logic (without Excluded-middle), Control-monotonicity and Unit (together) do not imply C4, and therefore not Bind nor Escalation.

  We construct an interpretation of the logic that satisfies the basic axioms, Unit, and Control-monotonicity, but not C4.

  In this interpretation, each formula is mapped to an open set in the Sierpinski space, that is, to one of the sets $\emptyset$, $\{1\}$, and $\{0,1\}$. These open sets form a Heyting algebra with the usual inclusion ordering, so they provide a model for intuitionistic logic. In this model, $\emptyset$ corresponds to `false`. Importantly $\{1\} \rightarrow$ `false` is $\emptyset$ (and not $\{0\}$, since this is not an open set). Quantification works as a finite conjunction. For every A, we let the meaning of A `says` s be $\{1\}$ if the meaning of s is $\emptyset$, and $\{0,1\}$ otherwise.

  This interpretation satisfies Unit, since the meaning of s is always contained in the meaning of A `says` s. A fortiori, it also satisfies necessitation.

  Since `says` is monotonic, we have that A `says` $((X \rightarrow Y) \wedge X) \rightarrow A$ `says` Y. Moreover, since the inclusion ordering is linear, monotonicity implies that `says` distributes over conjunctions, so $((A$ `says` $(X \rightarrow Y)) \wedge (A$ `says` $X)) \rightarrow A$ `says` $((X \rightarrow Y) \wedge X)$. Closure under consequence follows.

  These definitions also imply that the meaning of A `controls` s is the same as the meaning of s:

  - for $s = \emptyset$, A `controls` s is $\{1\} \rightarrow \emptyset$, that is, $\emptyset$;
  - for $s = \{1\}$, A `controls` s is $\{0,1\} \rightarrow \{1\}$, that is, $\{1\}$;
  - for $s = \{0,1\}$, A `controls` s is $\{0,1\} \rightarrow \{0,1\}$, that is, $\{0,1\}$.

  Therefore, `controls` is monotonic.

  The meaning of A `says false` is $\{1\}$. The meaning of A `says` A `says false` is $\{0,1\}$. So we do not have C4. ■

Although Control-monotonicity does not imply C4 in intuitionistic logic, it does in classical logic, as the following theorem implies:

Theorem 6. *Starting from the basic logic plus Excluded-middle, the following are equivalent:*

– *Escalation,*
– *Control-monotonicity.*

Proof. By Theorem 5, Escalation implies Control-Monotonicity. Conversely, we instantiate Control-monotonicity in the case the stronger propositions (X) is `false`; we obtain a formula that is classically equivalent to Escalation. ■

The theorems of this section should not be construed as a criticism of the Principle of Least Privilege. Formulations weaker than Control-monotonicity might be viable and less problematic.

6 Discussion

Overall, the results of this paper indicate that, while in a classical setting we may want to stay close to basic modal logic, in an intuitionistic setting we may adopt CDD. This move may be attractive, in particular, because CDD supports hand-off. These results also suggest that a great deal of caution should be applied in selecting axioms, considering both their formal properties and their security implications.

We do not argue that the use of a particular set of axioms is required for writing good security policies. It is possible that reasonable security policies and other assertions can be formulated in many different systems, with different underlying logics. However, understanding the properties and consequences of these logics is essential for writing appropriate formulas reliably.

The literature contains models for some of these axioms (e.g., [4]), and we are currently developing others (in collaboration with Deepak Garg). Semantics can be helpful in providing a different perspective on axiomatizations. In this paper, we employ semantics as a tool in some of the proofs; more extensive uses of semantics remain attractive but a subject for further research.

Acknowledgments

This work benefited from conversations with Butler Lampson.

References

1. Abadi, M.: Logic in access control. In: Proceedings of the Eighteenth Annual IEEE Symposium on Logic in Computer Science, pp. 228–233 (2003)
2. Abadi, M.: Access control in a core calculus of dependency. Electronic Notes in Theoretical Computer Science 172, 5–31 (2007); Computation, Meaning, and Logic: Articles dedicated to Gordon Plotkin
3. Abadi, M., Banerjee, A., Heintze, N., Riecke, J.G.: A core calculus of dependency. In: Proceedings of the 26th ACM Symposium on Principles of Programming Languages, pp. 147–160 (January 1999)
4. Abadi, M., Burrows, M., Lampson, B., Plotkin, G.: A calculus for access control in distributed systems. ACM Transactions on Programming Languages and Systems 15(4), 706–734 (1993)
5. Bauer, L., Garriss, S., Reiter, M.K.: Distributed proving in access-control systems. In: Proceedings of the 2005 IEEE Symposium on Security and Privacy, pp. 81–95 (May 2005)
6. Becker, M.Y., Fournet, C., Gordon, A.D.: Design and semantics of a decentralized authorization language. In: 20th IEEE Computer Security Foundations Symposium, pp. 3–15 (2007)
7. Cardelli, L.: Type systems. In: Tucker, A.B. (ed.) The Computer Science and Engineering Handbook, ch.103, pp. 2208–2236. CRC Press, Boca Raton (1997)
8. Cirillo, A., Jagadeesan, R., Pitcher, C., Riely, J.: Do as I SaY! programmatic access control with explicit identities. In: 20th IEEE Computer Security Foundations Symposium, pp. 16–30 (July 2007)
9. DeTreville, J.: Binder, a logic-based security language. In: Proceedings of the 2002 IEEE Symposium on Security and Privacy, pp. 105–113 (May 2002)
10. Fairtlough, M., Mendler, M.: Propositional lax logic. Information and Computation 137(1), 1–33 (1997)

11. Fournet, C., Gordon, A.D., Maffeis, S.: A type discipline for authorization in distributed systems. In: 20th IEEE Computer Security Foundations Symposium, pp. 31–45 (2007)
12. Garg, D., Pfenning, F.: Non-interference in constructive authorization logic. In: 19th IEEE Computer Security Foundations Workshop, pp. 283–296 (2006)
13. Girard, J.-Y.: Interprétation Fonctionnelle et Elimination des Coupures de l'Arithmétique d'Ordre Supérieur. Thèse de doctorat d'état, Université Paris VII (June 1972)
14. Gurevich, Y., Neeman, I.: DKAL: Distributed-knowledge authorization language. Technical Report MSR-TR-2007-116, Microsoft Research (August 2007)
15. Hughes, G.E., Cresswell, M.J.: An Introduction to Modal Logic. Methuen Inc., New York (1968)
16. Lampson, B., Abadi, M., Burrows, M., Wobber, E.: Authentication in distributed systems: Theory and practice. ACM Transactions on Computer Systems 10(4), 265–310 (1992)
17. Lampson, B.W.: Protection. In: Proceedings of the 5th Princeton Conference on Information Sciences and Systems, pp. 437–443 (1971)
18. Lampson, B.W.: Computer security in the real world. IEEE Computer 37(6), 37–46 (2004)
19. Lesniewski-Laas, C., Ford, B., Strauss, J., Kaashoek, M.F., Morris, R.: Alpaca: extensible authorization for distributed services. In: 14th ACM Conference on Computer and Communications Security, pp. 432–444 (2007)
20. Li, N., Grosof, B.N., Feigenbaum, J.: Delegation logic: A logic-based approach to distributed authorization. ACM Transactions on Information and System Security 6(1), 128–171 (2003)
21. Moggi, E.: Notions of computation and monads. Information and Control 93(1), 55–92 (1991)
22. Saltzer, J.H., Schroeder, M.D.: The protection of information in computer system. Proceedings of the IEEE 63(9), 1278–1308 (1975)
23. Tse, S., Zdancewic, S.: Translating dependency into parametricity. Journal of Functional Programming (to appear)

Appendix

A Second-Order Propositional Intuitionistic Logic

The axioms are:

- $\mathtt{true}$
- $s_1 \rightarrow (s_2 \rightarrow s_1)$
- $(s_1 \rightarrow (s_2 \rightarrow s_3)) \rightarrow ((s_1 \rightarrow s_2) \rightarrow (s_1 \rightarrow s_3))$
- $(s_1 \wedge s_2) \rightarrow s_1$
- $(s_1 \wedge s_2) \rightarrow s_2$
- $s_1 \rightarrow s_2 \rightarrow (s_1 \wedge s_2)$
- $s_1 \rightarrow (s_1 \vee s_2)$
- $s_2 \rightarrow (s_1 \vee s_2)$
- $(s_1 \rightarrow s_3) \rightarrow ((s_2 \rightarrow s_3) \rightarrow ((s_1 \vee s_2) \rightarrow s_3))$
- $(\forall X.\, s) \rightarrow s[t/X]$
- $(\forall X.\, (s_1 \rightarrow s_2)) \rightarrow (s_1 \rightarrow \forall X.\, s_2)$ where X is not free in s_1

The rules are modus ponens and universal generalization:

$$\frac{s_1 \qquad s_1 \rightarrow s_2}{s_2} \qquad\qquad \frac{s}{\forall X.\, s}$$

B CDD

The rules of CDD are:

$[\mathit{Var}] \quad \Gamma, s, \Gamma' \vdash s$

$[\mathit{Unit}] \quad \Gamma \vdash \texttt{true}$

$$[\mathit{Lam}] \quad \frac{\Gamma, s_1 \vdash s_2}{\Gamma \vdash (s_1 \rightarrow s_2)}$$

$$[\mathit{App}] \quad \frac{\Gamma \vdash (s_1 \rightarrow s_2) \qquad \Gamma \vdash s_1}{\Gamma \vdash s_2}$$

$$[\mathit{Pair}] \quad \frac{\Gamma \vdash s_1 \qquad \Gamma \vdash s_2}{\Gamma \vdash (s_1 \wedge s_2)}$$

$$[\mathit{Proj\ 1}] \quad \frac{\Gamma \vdash (s_1 \wedge s_2)}{\Gamma \vdash s_1}$$

$$[\mathit{Proj\ 2}] \quad \frac{\Gamma \vdash (s_1 \wedge s_2)}{\Gamma \vdash s_2}$$

$$[\mathit{Inj\ 1}] \quad \frac{\Gamma \vdash s_1}{\Gamma \vdash (s_1 \vee s_2)}$$

$$[\mathit{Inj\ 2}] \quad \frac{\Gamma \vdash s_2}{\Gamma \vdash (s_1 \vee s_2)}$$

$$[\mathit{Case}] \quad \frac{\Gamma \vdash (s_1 \vee s_2) \qquad \Gamma, s_1 \vdash s \qquad \Gamma, s_2 \vdash s}{\Gamma \vdash s}$$

$$[\mathit{TLam}] \quad \frac{\Gamma \vdash s}{\Gamma \vdash \forall X.\, s} \quad (X \text{ not free in } \Gamma)$$

$$[\mathit{TApp}] \quad \frac{\Gamma \vdash \forall X.\, s}{\Gamma \vdash s[t/X]}$$

$$[\mathit{UnitM}] \quad \frac{\Gamma \vdash s}{\Gamma \vdash A \ \mathsf{says}\ s}$$

$$[\mathit{BindM}] \quad \frac{\Gamma \vdash A \ \mathsf{says}\ s \qquad \Gamma, s \vdash A \ \mathsf{says}\ t}{\Gamma \vdash A \ \mathsf{says}\ t}$$

As is typical for type systems, the rules are presented in a sequent-calculus format, rather than as a Hilbert system. In this definition, though, we simply omit all the terms, as well as declarations for variables. An environment Γ denotes a list of formulas. In the case where Γ is empty, we write $\vdash s$, and say that s is a theorem, when $\vdash s$ is derivable by these rules.

Reasoning about Conditions and Exceptions to Laws in Regulatory Conformance Checking[*]

Nikhil Dinesh, Aravind Joshi, Insup Lee, and Oleg Sokolsky

Department of Computer Science
University of Pennsylvania
Philadelphia, PA 19104-6389, USA
{nikhild,joshi,lee,sokolsky}@seas.upenn.edu

Abstract. This paper considers the problem of checking whether an organization conforms to a body of regulation. Conformance is cast as a trace checking question – the regulation is represented in a logic that is evaluated against an abstract trace or run representing the operations of an organization. We focus on a problem in designing a logic to represent regulation.

A common phenomenon in regulatory texts is for sentences to refer to others for conditions or exceptions. We motivate the need for a formal representation of regulation to accomodate such references between statements. We then extend linear temporal logic to allow statements to refer to others. The semantics of the resulting logic is defined via a combination of techniques from Reiter's default logic and Kripke's theory of truth.

1 Introduction

Regulations, laws, and policies that affect many aspects of our lives are represented predominantly as documents in natural language. For example, the Food and Drug Administration's Code of Federal Regulations [1] (FDA CFR) governs the operations of American bloodbanks. The CFR is framed by experts in the field of medicine, and regulates the tests that need to be performed on donations of blood before they are used. In such safety-critical scenarios, it is desirable to assess formally whether an organization (bloodbank) conforms to the regulation (CFR).

There is a growing interest in using formal methods to assist organizations in complying with regulation [2,3,4]. Assisting an organization in compliance involves a number of tasks related to the notion of a violation. For example, it is of interest to detect or prevent violations, assign blame, and if possible, recover from violations. In this paper, we focus on *conformance checking* which involves detecting the presence of violations.

We cast conformance checking as a trace-checking question. The regulation is translated to statements in a logic which are evaluated against a trace or run representing the operations of an organization. The result of evaluation is either an affirmative answer to conformance, or a counterexample representing a subset of the operations of the organization and the specific law that is violated.

[*] This research was supported in part by NSF CCF-0429948, NSF-CNS-0610297, ARO W911NF-05-1-0158, and ONR MURI N00014-07-1-0907.

R. van der Meyden and L. van der Torre (Eds.): DEON 2008, LNAI 5076, pp. 110–124, 2008.

There are two important features of regulatory texts that need to be accomodated by a representation in logic. First, regulations convey constraints on an organization's operations, and these constraints can be obligatory (required) or permitted (optional). Second, statements in regulation refer to others for conditions or exceptions. An organization conforms to a body of regulation iff it satisfies all the obligations. However, permissions provide exceptions to obligations, indirectly affecting conformance. Our formulation of obligations and permissions follows the theory of Ross [5], and we will discuss the relationship to other theories (cf. [6]) in Section 3.1.

The central focus of this work is the function of regulatory sentences as conditions or exceptions to others. This function of sentences makes them dependent on others for their interpretation, and makes the translation to logic difficult. We call this the problem of *references to other laws*. In Section 2, we argue that a logic to represent regulation should provide mechanisms for statements to refer to others. We provide motivation using examples from the FDA CFR. We discuss how these sentences can be represented in a logic without references, and conclude that this would make the translation difficult.

We then turn to the task of defining a logic that lets statements refer to and reason about others. In Section 3.1, we define a trace or run-based representation for the operations of an organization, and a predicate-based linear temporal logic (PredLTL) to make assertions about runs. PredLTL is extended to express two kinds of normative statements (obligations and permissions), leading to a formal definition of conformance.

In Sections 3.2 and 3.3, we extend PredLTL to allow references between laws thereby making permissions relevant to conformance. Specifically, we introduce *an inference predicate*, whose interpretation is determined by inferences from laws. The justifications in default logic [7] can be cast as an instance of this predicate. Default logic has been used in computing extensions to a theory, in the manner of logic programs [8,9]. In conformance checking, we need to separate two uses of statements: (a) extending a theory (the regulation), and (b) determining facts about an organization. This separation is achieved using the inference predicate. Statements are evaluated using the fixed points of an appropriate function, based on a technique used in Kripke's theory of truth [10]. An axiomatization is discussed in Section 3.4.

Section 4 concludes with a discussion of related and future work.

2 Motivation

In this section, we argue that a logic to represent regulation should provide a mechanism for sentences to refer to others. We discuss shortened versions of sentences from the CFR Section 610.40, which we will use as a running example throughout the paper. Consider the following sentences:

(1) Except as specified in (2), every donation of blood or blood component must be tested for evidence of infection due to Hepatitis B.

(2) You are not required to test donations of source plasma for evidence of infection due to Hepatitis B.

Statement (1) conveys an obligation to test donations of blood or blood component for Hepatitis B, and (2) conveys a permission not to test a donation of source plasma (a blood component) for Hepatitis B. To assess an organization's conformance to (1) and (2), it suffices to check whether "All non-source plasma donations are tested for Hepatitis B". In other words, (1) and (2) imply the following obligation:

(3) Every non-source plasma donation must be tested for evidence of infection due to Hepatitis B.

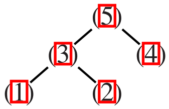
There are a variety of logics in which one can capture the interpretation of (3), as needed for conformance. Now suppose we have a sentence that refers to (1):

(4) To test for Hepatitis B, you must use a screening test kit.

The reference is more indirect here, but the interpretation is: "If (1) requires a test, then the test must be performed using a screening test kit". A bloodbank is not prevented from using a different kind of test for source plasma donations. (4) can be represented by first producing (3), and then inferring that (3) and (4) imply the following:

(5) Every non-source plasma donation must be tested for evidence of infection due to Hepatitis B using a screening test kit.

It is easy to represent the interpretation of (5) directly in a logic. However, (5) has a complex relationship to the sentences from which it was derived, i.e., (1), (2) and (4). The derivation takes the form of a tree:



To summarize, if one wishes to use a logic with no support for referring to other sentences, derived obligations must be created manually. We argue that the manual creation of derived obligations is impractical in terms of the amount of effort involved. We give two (pragmatic) reasons. First, the derived obligation can become very complex. The full version of statement (1) in the CFR contains six exceptions, and these exceptions in turn have statements that qualify them further. It is difficult to inspect a derived obligation, and determine if it captures the intended interpretation of the sentences from which it came. Second, references between laws are frequent, amplifying the effort in creating a logic representation. In [11], we discuss lexical statistics which suggest that references are a common way of establishing relationships between sentences in the CFR, and [12,3] point out their frequency in other bodies of regulation.

We advocate an approach that allows us to introduce references into the syntax of the logic, and resolve references during evaluation.

3 Representing Regulatory Documents in Logic

In this section, we extend linear temporal logic (LTL) to distinguish between obligations and permissions, and allow references between statements. We begin, in Section 3.1, by representing a bloodbank as a run or trace. LTL is extended to distinguish between obligations and permissions, leading to definitions of conformance. We then extend the logic to allow sentences to refer to others. Section 3.2 gives an informal example-driven account, and Section 3.3 provides a formal account. In Section 3.4, we discuss an axiomatization.

Sections 3.1 is intended as background, in which we discuss several underlying assumptions. Our goal is to focus on the problem of references, and to treat the representation of obligations and permissions as an important but orthogonal issue.

3.1 Predicate-Based Linear Temporal Logic (PredLTL)

Representing regulated operations: Given the need to demonstrate conformance to the regulation in case of an audit, regulated organizations such as bloodbanks keep track of their operations in a database, for example, donor information and the tests they perform. Such a system can be thought of abstractly as a relational structure evolving over time. At each point in time (state), there are a set of objects (such as donations and donors) and relations between the objects (such as an association between a donor and her donations). The state changes by the creation, removal or modification of objects. We represent this as a run.

Definition 1 (A Run of a System). *Given a set O (of objects) and countable sets $\Phi_1, ..., \Phi_n$ (where Φ_j is a set of predicate names of arity j), a run of a system $R(O, \Phi_1, ..., \Phi_n)$, abbreviated as R, is a tuple $(r, \pi_1, ..., \pi_n)$ where:*

- *$r : N \rightarrow S$ is a sequence of states. N is the set of natural numbers, and S is a set of states.*
- *$\pi_j : \Phi_j \times S \rightarrow 2^{O^j}$ is a truth assignment to predicates of arity j. Given $p \in \Phi_j$, we will say that $p(o_1, ..., o_j)$ is true at state s iff $(o_1, ..., o_j) \in \pi_j(p, s)$.*

Given a run R and a time $i \in N$, the pair (R, i) is called a point (statements in linear temporal logic are evaluated at points). Given the predicate names $(\Phi_1, ..., \Phi_n)$, the corresponding space of runs is denoted by $\mathcal{R}(\Phi_1, ..., \Phi_n)$, abbreviated as $\mathcal{R}$.

Conceivably, we could construct a state-transition diagram representing all possible behaviors of the system and explore conformance from the model checking perspective (e.g., [13]). We chose to restrict our attention to traces for two reasons. First, checking of traces is easier to explain, and all interesting theoretical and algorithmic aspects that we explore in this paper manifest themselves in trace checking. Second, many parts of the operations of an organization, such as a bloodbank, do not involve computers. A complete model of operations has to include a model of human users, which is a research problem in its own right that is well beyond the scope of this paper. However, if a finite-state model of an organization can be created, the propositional version of the logic developed here can be adapted to work with available model-checkers.

Representing the regulation: The logic that we define in this section is a restricted fragment of first-order modal logic. The restriction is that we allow formulas with free variables, but no quantification over objects. Formulas will be interpreted using the universal generalization rule, i.e., over all assignments to free variables. The restrictions are similar in spirit to the logic programing approaches to regulation [8,9]. PredLTL is less expressive than the variants of first-order logic used by [2,4]. However, when references are added, the logic becomes more expressive than first-order logic (Section 3.3).

Definition 2 (Syntax). *Given sets $\Phi_1, ..., \Phi_n$ (of predicate names) and a set of variables X, the language $L(\Phi_1, ..., \Phi_n, X)$, abbreviated as L, is the smallest set such that:*

- *$p(y_1, ..., y_j) \in L$ where $p \in \Phi_j$ and $(y_1, ..., y_j) \in X^j$.*
- *If $\varphi \in L$, then $\neg\varphi \in L$ and $\Box\varphi \in L$. If $\varphi, \psi \in L$, then $\varphi \wedge \psi \in L$.*

Disjunction $\varphi \vee \psi = \neg(\neg\varphi \wedge \neg\psi)$ and implication $\varphi \Rightarrow \psi = \neg\varphi \vee \psi$ are derived connectives. The temporal operator is understood in the usual way: $\Box\varphi$ (φ holds and will always hold (globally)). $\Diamond\varphi$ (φ will eventually hold) is defined as $\neg\Box\neg\varphi$.

We now extend the syntax to express normative statements in a body of regulation, by distinguishing between obligations and permissions.

Definition 3 (Syntax of Regulation). *Given a finite set of identifiers ID, a body of regulation Reg is a set of statements such that for each $id \in ID$, there exist $\varphi, \psi \in L$ such that either: id.***o***: $\varphi \rightsquigarrow \psi \in Reg$, or id.***p***: $\varphi \rightsquigarrow \psi \in Reg$*

id.**o**: $\varphi \rightsquigarrow \psi$ (id.**p**: $\varphi \rightsquigarrow \psi$) is read as: "it is obligated (permitted) that the precondition φ leads to the postcondition ψ". The distinction between preconditions and postconditions corresponds to the distinction between input and output in input-output logic [14].

Definition 4 (Semantics). *Given a run $R = (r, \pi_1, ..., \pi_n)$, $i \in N$, $\varphi \in L$, and an assignment $v : X \to O$, the relation $(R, i, v) \models \varphi$ is defined inductively as follows:*

- *$(R, i, v) \models p(y_1, ..., y_j)$ iff $(o_1, ..., o_j) \in \pi_j(p, r(i))$ where $o_k = v(y_k)$ if $y_k \in O$.*
- *The semantics of conjunction and negation is defined in the usual way.*
- *$(R, i, v) \models \Box\varphi$ iff for all $k \geq i : (R, k, v) \models \varphi$*

We extend the semantic relation to regulatory statements. We take $\models$ to stand for "conforms to":

- *$(R, i, v) \models$ id.***o***: $\varphi \rightsquigarrow \psi$ iff $(R, i, v) \models \varphi \Rightarrow \psi$ ($\Rightarrow$ is implication)*
- *$(R, i, v) \models$ id.***p***: $\varphi \rightsquigarrow \psi$. Runs vacuously conform to permissions. Permissions will become relevant when references from obligations are present (Section 3.2).*

Consider again our example from Section 2. We use three predicates defined as follows. $d(x)$ is true iff x is a donation. $sp(x)$ is true iff x consists of source plama. $test(x)$ is true iff x is tested for Hepatitis B. Statement (3) is represented as:

3.**o**: $d(x) \wedge \neg sp(x) \rightsquigarrow \Diamond test(x)$

Statement (2) is be represented as: 2.**p**: $d(y) \wedge sp(y) \rightsquigarrow \neg\Diamond test(y)$. However, statement (1) cannot be represented directly.

We will now define conformance, and then discuss the various definitions in the context of related work. Given a run R, let $V(R)$ denote the set of variable assignments. Conformance is defined using the notion of validity. A formula φ is valid at the point (R, i), denoted $(R, i) \models \varphi$, iff for all $v \in V(R)$: $(R, i, v) \models \varphi$. A formula φ is valid on R iff it is valid at all points, that is, $R \models \varphi$ iff for all $i \in N : (R, i) \models \varphi$.

Definition 5 (Run Conformance). *Given a body of regulation* Reg *and a run* R *representing the operations of an organization, we say that* R *conforms to the regulation iff for all obligations* $id.\mathbf{o}\colon \varphi \rightsquigarrow \psi \in Reg$*, we have* $R \models id.\mathbf{o}\colon \varphi \rightsquigarrow \psi$*.*

Discussion: The deontic concepts of obligation and permission are treated as properties of sentences. Only obligations matter for conformance. If a non-source plasma donation is not tested, there is a problem. On the other hand, a bloodbank may choose to test a donation of source plasma or not. In assessing conformance, the function of a permission is to serve as an exception to an obligation, and in this indirect manner it becomes relevant. We will give a semantics to this function of permissions in Section 3.2. Such a treatment of permissions has its basis in the legal theory of Ross [5].

Ross' approach to permission is by no means the only one. Theories have distinguished between various kinds of permission (cf. [6]), the most common distinction being that of positive and negative permission. We discuss the analysis by Makinson and van der Torre [15]. φ is said to positively permitted iff it is explictly permitted by the laws, and φ is negatively permitted iff it is not forbidden. The key issue is whether positive permissions can give rise to violations. In regulations phrased exclusively in terms of permissions, it is desirable to say that *if* φ *denotes a "relevant" condition which is not explicitly permitted, then it should not hold in conforming implementations*. While this has been analysed as a property of permission, following Ross, we take such violations as arising from an implicit obligation, i.e., the italicized clause. This implicit obligation can be represented using the techniques we discuss in Section 3.2, provided that the relevance of the condition is known.

In the formulation here, obligations and permissions are top-level operators and cannot be negated. This restriction can be removed by treating obligation and permission as KD modalities (c.f. [16]), and using a many-valued interpretation to decide if a run belongs to the set of ideal runs. However, we avoid this to simplify presentation. A more crucial restriction is that iterated deontic constructs cannot be expressed directly, i.e., sentences of the form "required to allow x" or "allowed to require x.". One has to decide what top-level obligations or permissions are implied by these constructs. To our knowledge, handling iterated constructs is an open problem in deontic logic [17].

3.2 References to Other Laws – An Informal Description

In this section, we give an informal account of *reference logic* (RefL), which is used to handle references. We extend the syntax of PredLTL with *an inference predicate* $\mathrm{by}_{\mathrm{Id}}(\varphi)$, where Id is a set of identifiers. $\mathrm{by}_{\mathrm{Id}}(\varphi)$ is read as "by the laws in Id, φ holds". There are two restrictions: (a) φ is a statement in PredLTL (Definition 2) and (b) the predicate $\mathrm{by}_{\mathrm{Id}}(\varphi)$ can appear only in preconditions of laws. These restrictions are similar to those that apply to justifications in default logic [7].

Consider again our example statements (1) and (2), which are represented in RefL as follows:

- 1.**o**: $d(x) \wedge \neg\mathrm{by}_{\{2\}}(\varphi(x)) \rightsquigarrow \Diamond test(x)$, and
- 2.**p**: $d(y) \wedge sp(y) \rightsquigarrow \neg\Diamond test(y)$

In the obligation above, the subformula $\mathrm{by}_{\{2\}}(\varphi(x))$ is understood as "by the law (2) the formula $\varphi(x)$ holds". It remains to define the formula $\varphi(x)$. Intuitively, this should

Table 1. A run and its annotations

Time	Objects	Predicates	Annotations
1	o_1	$d(o_1)$, $sp(o_1)$, $\neg test(o_1)$	2: $\neg\Diamond test(o_1)$
2	o_1	$d(o_1)$, $sp(o_1)$, $\neg test(o_1)$	2: $\neg\Diamond test(o_1)$
	o_2	$d(o_2)$, $\neg sp(o_2)$, $\neg test(o_2)$	1: $\Diamond test(o_2)$
3	o_1	$d(o_1)$, $sp(o_1)$, $test(o_1)$	2: $\neg\Diamond test(o_1)$
	o_2	$d(o_2)$, $\neg sp(o_2)$, $\neg test(o_2)$	1: $\Diamond test(o_2)$

be the negation of the postcondition of (1). In other words, if $\neg\Diamond test(x)$ follows from (2), then the postcondition of (1) need not hold. This gives us:

1.**o**: $d(x) \wedge \neg \mathrm{by}_{\{2\}}(\neg\Diamond test(x)) \rightsquigarrow \Diamond test(x)$

We interpret the predicate $\mathrm{by}_{\{2\}}(\neg\Diamond test(x))$, by letting formulas have output. In other words, when the precondition of an obligation or permission is true at a point, the point is *annotated* with the postcondition.

Table 3.2 shows a run of a bloodbank augmented with annotations. First, an object o_1 is entered into the system. o_1 is a donation of source plasma ($d(o_1)$ and $sp(o_1)$ are true). When a donation is added, its test predicate is initially false. Then, an object o_2 is added, which is a donation but not of source plasma. In the third step, the object o_1 is tested. At this point, unless the run is extended to test o_2 as well, it does not conform with the regulation. We now discuss how the annotations are arrived at and used to assess the regulation.

We begin by defining an annotation. Given a run R, an assignment $v \in V(R)$, and $\varphi \in L$, $v(\varphi)$ is the formula obtained by replacing all variables x by the unique name for the object $v(x)$. We assume that all variables are free. Note that $v(\varphi)$ is equivalent to a propositional LTL formula, as the variables have been replaced by constant symbols. An annotation, id: $v(\varphi)$, is a propositional LTL formula associated with an identifier.

Given a point (R, i) and an assignment $v \in V(R)$, first we consider the permission 2.**p**: $d(y) \wedge sp(y) \rightsquigarrow \neg\Diamond test(y)$. If $(R, i, v) \models d(y) \wedge sp(y)$, then (R, i) is *annotated* with 2: $v(\neg\Diamond test(y))$. Otherwise, there is no annotation.

Since the precondition of statement (2) is true for the assignment of y to o_1, we have the annotation 2: $\neg\Diamond test(o_1)$ at all points. However, since o_2 is not a donation of source plasma, there is no correponding annotation.

Now consider the formula $\mathrm{by}_{\{2\}}(\neg\Diamond test(x))$. This is evaluated as follows. We evaluate 2.**p**: $d(y) \wedge sp(y) \rightsquigarrow \neg\Diamond test(y)$ at (R, i) w.r.t. all variable assignments. Let ψ_2 be the conjunction of the annotations produced by the formula for (2).

$(R, i, v) \models \mathrm{by}_{\{2\}}(\neg\Diamond test(x))$ iff $\models \psi_2 \Rightarrow v(\neg\Diamond test(x))$

Notice that this requires a validity check in propositional LTL, which can be decided in space polynomial in the size of the formula [18].

Returning to the run in Table 3.2, the states are annotated with 2: $\neg\Diamond test(o_1)$ and $\models \neg\Diamond test(o_1) \Rightarrow \neg\Diamond test(o_1)$, since $\varphi \Rightarrow \varphi$ is a propositional tautology. So $(R, i, v) \models \mathrm{by}_{\{2\}}(\neg\Diamond test(x))$ when $v(x) = o_1$.

We can evaluate 1.**o**: $d(x) \wedge \neg \mathrm{by}_{\{2\}}(\neg\Diamond test(x)) \rightsquigarrow \Diamond test(x)$ similarly by annotating states with $\Diamond test(x)$ if the precondition holds. In Table 3.2, this results in an annotation of 1: $\Diamond test(o_2)$ on the appropriate states. If o_2 is never tested, the run will be declared non-conforming (by Definition 5), but the annotation will remain. This lets a law which depends on (1) draw the correct inference.

3.3 Reference Logic (RefL)

The semantic evaluation outlined in Section 3.2 works only when the references are acyclic, since an order of evaluation needs to be defined. To handle cycles, we adopt a fixed-point technique from Kripke's theory of truth [10]. The idea is to move to a three-valued logic where the third (middle) value stands for *ungrounded*. Initially, all statements are ungrounded and there are no annotations. Using an inflationary function, we add annotations until a fixed point in reached. In this section, we define this inflationary function and show that it has least and maximal fixed points. We begin by extending the syntax described in Section 3.1:

Definition 6 (Syntax of Preconditions). *Given sets $\Phi_1, ..., \Phi_n$ (of predicate names), a set of variables X, and a finite set of identifiers ID, the language $L'(\Phi_1, ..., \Phi_n, X, ID)$, abbreviated as L', is the smallest set such that:*

- *$p(y_1, ..., y_j) \in L'$ where $p \in \Phi_j$ and $(y_1, ..., y_j) \in X^j$.*
- *If $\varphi \in L'$, then $\neg\varphi \in L'$ and $\Box\varphi \in L'$. If $\varphi, \psi \in L'$, then $\varphi \wedge \psi \in L'$*
- *If $Id \subseteq ID$ and $\varphi \in L(\Phi_1, ..., \Phi_n, X)$ (Definition 2), then $\mathrm{by}_{\mathrm{Id}}(\varphi) \in L'$*

The syntax of regulatory statements (Definition 3) is modified so that the preconditions of laws are statements from L'. We use $id.\mathbf{x} : \varphi \rightsquigarrow \psi$ to stand for a normative statement (either obligation or permission). We now define an annotation:

Definition 7 (Annotation). *Given a run R, a set of identifiers ID, a body of regulation Reg and $v \in V(R)$, an annotation is a statement id: $v(\psi)$ such that id $\in ID$ and $id.\mathbf{x} : \varphi \rightsquigarrow \psi \in Reg$. The set of annotations is denoted by $A(R, ID, Reg)$, abbreviated A.*

Definition 8 (Annotation Function). *Given a run R, an annotation function $\alpha : N \rightarrow 2^A$ assigns a set of annotations to each point. We use $\alpha.Id(i)$ to denote the set of annotations id: $\psi \in \alpha(i)$ such that id $\in Id$.*

We will formalize the semantics using the fixed point technique outlined in [10]. Before we turn to the formal definitions, we sketch some of the key ideas involved.

Let us assume as given a run R. Statements in L' and Reg are divided into three classes corresponding to true ($\mathbf{T}(i, v)$), false ($\mathbf{F}(i, v)$) and ungrounded ($\mathbf{U}(i, v)$) for all times $i \in N$ and assignments $v \in V(R)$. Intuitively, $\mathbf{U}(i, v)$ is the set of statements that are waiting for the evaluation of another statement.

As we discussed in Section 3.2, to determine whether $\mathrm{by}_{\mathrm{Id}}(\varphi) \in \mathbf{T}(i, v)$, we need to check if there is a set of annotations which imply $v(\varphi)$. We construct the annotation function α such that for all assignments v, we have id: $v(\psi) \in \alpha(i)$ iff $\varphi \in \mathbf{T}(i, v)$ for some $id.\mathbf{x} : \varphi \rightsquigarrow \psi \in Reg$ and $id \in Id$. We will say that $\mathrm{by}_{\mathrm{Id}}(\varphi) \in \mathbf{T}(i, v)$ only if $\alpha.Id(i) \cup \{v(\neg\varphi)\}$ is not satisfiable.

To determine whether $\mathrm{by}_{\mathrm{Id}}(\varphi) \in \mathbf{F}(i, v)$, we need to ensure that there is no ungrounded statement that could make it true. To check this condition, we construct the annotation function α' such that id: $v(\psi) \in \alpha'(i)$ iff $\varphi \in \mathbf{T}(i,v) \cup \mathbf{U}(i,v)$ for some $id.\mathbf{x}: \ \varphi \rightsquigarrow \psi \in Reg$ and $id \in Id$. The condition for falsity w.r.t. α' is simply the negation of the condition for truth w.r.t. α. More formally, $\text{by}_{\text{Id}}(\varphi) \in \mathbf{F}(i,v)$ only if $\alpha'.Id(i) \cup \{v(\neg\varphi)\}$ is satisfiable.

When there are circular references, one cannot always evaluate a statement to be true or false. The Nixon-diamond problem (introduced in [7]) is a well-known example. We rephrase it in "legalese":

(6) Except as otherwise specified, Quakers must be pacifists.

(7) Except as otherwise specified, Republicans must not be pacifists.

These statements can be represented in RefL as follows:
6.**o**: $q(x) \wedge \neg\text{by}_{\{6,7\}}(\neg p(x)) \rightsquigarrow p(x)$, and
7.**o**: $r(x) \wedge \neg\text{by}_{\{6,7\}}(p(x)) \rightsquigarrow \neg p(x)$
Suppose we are given a state with an individual n (for Nixon), who is both quaker and republican, i.e., $q(n)$ and $r(n)$ hold. How should we evaluate the statements above? [10] suggests two answers to this question: (A) The statements are neither true or false (they are ungrounded). This corresponds to skeptical reasoning in non-monotonic logic. (B) Exactly one of $\text{by}_{\{6,7\}}(p(n))$ and $\text{by}_{\{6,7\}}(\neg p(n))$ is true, which leads us to conclude $p(n)$ (by (6)) or $\neg p(n)$ (by (7)) resply. This corresponds to credulous reasoning in non-monotonic logic.

In the semantics we give below, different answers correspond to different fixed points. We refer the reader to [10] for examples and discussion of the various possibilities with regard to fixed points. The choice of what to do when there are multiple fixed points depends on the application, and we discuss this issue further at the end of this section.

Definition 9 (Evaluation). *Given a run R and a body of regulation Reg, an evaluation is a tuple $E = (\mathbf{T}, \mathbf{F}, \mathbf{U})$, where $\mathbf{T}$, $\mathbf{F}$ and $\mathbf{U}$ are functions of the form $N \times V(R) \rightarrow 2^{L^+}$, where $L^+ = Reg \cup L'$. Furthermore, for all $i \in N$ and $v \in V(R)$, we have $\mathbf{T}(i,v) \cap \mathbf{F}(i,v) = \emptyset$ and $\mathbf{U}(i,v) = 2^{L^+} - (\mathbf{T}(i,v) \cup \mathbf{F}(i,v))$.*

Given an evaluation E, α_E is the annotation such that for all $i \in N$ and $id \in ID$, we have id: $v(\psi) \in \alpha_E(i)$ iff $\varphi \in \mathbf{T}(i,v)$, where $id.\mathbf{x}: \ \varphi \rightsquigarrow \psi \in Reg$. Similarly, α'_E is the annotation such that id: $v(\psi) \in \alpha'_E(i)$ iff $\varphi \in \mathbf{T}(i,v) \cup \mathbf{U}(i,v)$.

Definition 10 (Consistent Evaluation). *An evaluation E is consistent iff for all $i \in N$ and $v \in V(R)$, $\mathbf{T}(i,v) = \mathbf{F}(i,v) = \emptyset$, or $\mathbf{T}(i,v)$ and $\mathbf{F}(i,v)$ are sets such that:*

1. *$p(x_1, ..., x_j) \in \mathbf{T}(i,v)$ iff $(v(x_1), ..., v(x_j)) \in \pi_j(p, r(i))$*
   $p(x_1, ..., x_j) \in \mathbf{F}(i,v)$ iff $(v(x_1), ..., v(x_j)) \notin \pi_j(p, r(i))$
2. *If $\phi \in \mathbf{T}(i,v)$ and $\psi \in \mathbf{T}(i,v)$, then $\phi \wedge \psi \in \mathbf{T}(i,v)$*
   If $\phi \in \mathbf{F}(i,v)$ or $\psi \in \mathbf{F}(i,v)$, then $\phi \wedge \psi \in \mathbf{F}(i,v)$
   and similarly for negation and temporal operators
3. *If $\varphi \Rightarrow \psi \in \mathbf{T}(i,v)$, then id.**o**: $\varphi \rightsquigarrow \psi \in \mathbf{T}(i,v)$*
   *If $\varphi \Rightarrow \psi \in \mathbf{F}(i,v)$, then id.**o**: $\varphi \rightsquigarrow \psi \in \mathbf{F}(i,v)$*
   *id.**p**: $\varphi \rightsquigarrow \psi \in \mathbf{T}(i,v)$. Runs vacuously conform to permissions.*
4. *If $\text{by}_{\text{Id}}(\varphi) \in \mathbf{T}(i,v)$, then $\alpha_E.Id(i) \cup \{v(\neg\varphi)\}$ is not satisfiable.*
   If $\text{by}_{\text{Id}}(\varphi) \in \mathbf{F}(i,v)$, then $\alpha'_E.Id(i) \cup \{v(\neg\varphi)\}$ is satisfiable.

The set of all consistent evaluations for a run R and regulation Reg is denoted by $\mathcal{E}(R, Reg)$, abbreviated $\mathcal{E}$.

Observe that in consistent evaluations, if $\mathrm{by}_{\mathrm{Id}}(\varphi) \in \mathbf{T}(i, v)$, then $\alpha_E.Id(i) \cup \{v(\neg\varphi)\}$ is not satisfiable (Clause 4 in Definition 10). The converse need not be true.

Definition 11 (Partial Order). *Given evaluations $E_1 = (\mathbf{T}_1, \mathbf{F}_1, \mathbf{U}_1)$ and $E_2 = (\mathbf{T}_2, \mathbf{F}_2, \mathbf{U}_2, \alpha_2)$, we say that $E_1 \leq E_2$ iff for all $i \in N$ and $v \in V(R)$, $\mathbf{T}_1(i, v) \subseteq \mathbf{T}_2(i, v)$ and $\mathbf{F}_1(i, v) \subseteq \mathbf{F}_2(i, v)$.*

The pair $(\mathcal{E}, \leq)$, where $\mathcal{E}$ is the set of consistent evaluations is a partially ordered set (poset).

We now define the inflationary function whose fixed points we will be interested in.

Definition 12 (Inflationary function). *Given $(\mathcal{E}, \leq)$, the function $\mathcal{I} : \mathcal{E} \to \mathcal{E}$ is defined as follows. Given a consistent evaluation $E_1 = (\mathbf{T}_1, \mathbf{F}_1, \mathbf{U}_1)$, $\mathcal{I}(E_1)$ is the smallest consistent evaluation $E_2 = (\mathbf{T}_2, \mathbf{F}_2, \mathbf{U}_2)$ such that $E_1 \leq E_2$, for all $i \in N$ and $v \in V(R)$, $\mathbf{T}_2(i, v) \neq \emptyset$, $\mathbf{F}_2(i, v) \neq \emptyset$, and E_2 extends E_1.*

We say that E_2 extends E_1 iff for all $i \in N$ and assignments $v \in V(R)$:
If $\alpha_{E_1}(i) \cup \{v(\neg\varphi)\}$ is not satisfiable, then $\mathrm{by}_{\mathrm{Id}}(\varphi) \in \mathbf{T}_2(i, v)$
If $\alpha'_{E_1}(i) \cup \{v(\neg\varphi)\}$ is satisfiable, then $\mathrm{by}_{\mathrm{Id}}(\varphi) \in \mathbf{F}_2(i, v)$

It remains to show that $\mathcal{I}$ is well-defined, has maximal fixed points and a unique least fixed point. We give a brief sketch here, and refer the reader to [19] for detailed proofs.

Proposition 1. *Given $(\mathcal{E}, \leq)$ and $E_1 \in \mathcal{E}$, let $\mathcal{E}_2 \subseteq \mathcal{E}$ be the set of consistent evaluations such that $E_2 \in \mathcal{E}_2$ iff $E_1 \leq E_2$, for all i and v, $\mathbf{T}_2(i, v) \neq \emptyset$, $\mathbf{F}_2(i, v) \neq \emptyset$, and E_2 extends E_1. Then, $\mathcal{E}_2$ has a smallest element.*

The existence of fixed points is established using Zorn's lemma, which applies to chain-complete posets. Given the poset $(\mathcal{E}, \leq)$, a set $\mathcal{E}' \subseteq \mathcal{E}$ is called a chain (totally ordered set) iff for all $E_1, E_2 \in \mathcal{E}'$, we have $E_1 \leq E_2$ or $E_2 \leq E_1$. A poset is chain complete iff every chain has a supremum. The following can be shown:

Proposition 2. *$(\mathcal{E}, \leq)$ is a chain-complete poset.*

Lemma 1 (Zorn (c.f. [20])). *Every chain complete poset has a maximal element*

The existence of maximal fixed points is immediate from Zorn's lemma and the fact that $\mathcal{I}$ is inflationary, i.e., $E \leq \mathcal{I}(E)$. Let E^* be a maximal element in $\mathcal{E}$, since E^* is maximal and $E^* \leq \mathcal{I}(E^*)$ it follows that $E^* = \mathcal{I}(E^*)$.

To show the existence of a least fixed point, as [10] notes, we will need the observation that $\mathcal{I}$ *is monotonic*, i.e., if $E_1 \leq E_2$ then $\mathcal{I}(E_1) \leq \mathcal{I}(E_2)$. This can be shown by an argument similar to the proof of Proposition 1. With monotonicity, we obtain the following corollary to Zorn's lemma:

Corollary 1. *Given $E_1 \in \mathcal{E}$, let $\sigma(E_1)$ be the smallest set such that: (a) E_1 in $\mathcal{E}$, (b) if $E \in \sigma(E_1)$ then $\mathcal{I}(E) \in \sigma(E_1)$, and (c) if $C \subseteq \sigma(E_1)$ is a non-empty chain, then $E_{sc} \in \sigma(E_1)$, where E_{sc} is the supremum of C w.r.t. $\mathcal{E}$. Then:*

1. *$\sigma(E_1)$ is a chain whose supremum is a fixed point of $\mathcal{I}$*
2. *$\sigma(E_1)$ contains a unique fixed point*
3. *If $E_1 \leq E_2$, then $E_{s1} \leq E_{s2}$, where E_{s1} and E_{s2} are the suprema of $\sigma(E_1)$ and $\sigma(E_2)$ resply., and*
4. *$\mathcal{I}$ has a unique least fixed point.*

The first claim follows from a technique to prove Zorn's lemma [20]. The second and third claims follow from the first using monotonicity. And, for the last claim, consider the evaluation $E_0 = (\mathbf{T}_0, \mathbf{F}_0, \mathbf{U}_0)$, where for all $i \in N$, $v \in V(R)$, $\mathbf{T}_0(i, v) = \mathbf{F}_0(i, v) = \emptyset$, and $U_0(i, v) = 2^{L^+}$. Since $E_0 \leq E$ for all $E \in \mathcal{E}$, it follows from the third claim that $\sigma(E_0)$ is the least fixed point. The results are summarized in the following theorem, which provides a base for extending RefL with other inference predicates. We discuss the need for other predicates at the end of this section, and in Section 4.

Theorem 1. *Given the poset of consistent evaluations $(\mathcal{E}, \leq)$ and a function $\mathcal{I} : \mathcal{E} \rightarrow \mathcal{E}$ which is inflationary and monotonic, $\mathcal{I}$ has a least fixed point and a maximal fixed point.*

We mention the upper and lower bounds for the complexity of conformance checking w.r.t. the least fixed point. Given a run R and regulation Reg, we say that $R \models Reg$ iff all obligations are valid in R at the least fixed point. R is assumed to be finite in two ways: (a) The set of objects O is finite, and (b) There exists n, such that for all $j \geq n$, $r(n) = r(j)$, i.e., R eventually reaches a stable state.

Lemma 2 (Upper Bound). *Given a finite run R and regulation Reg, $R \models Reg$ can decided in EXPSPACE (space exponential in the size of Reg)*

The upper bound is obtained by turning Corollary 1 into a decision procedure. We start with the evaluation E_0, and apply $\mathcal{I}$ until a fixed point is reached. The worst-case size of the satisfiability tests are exponential in the size of the regulation. Since testing satisifiablity for propositional LTL is PSPACE-complete [18], applying $\mathcal{I}$ requires EXPSPACE. For the fragment of LTL discussed in this paper (using only $\Box$) satisfiability is NP-complete [18], and $R \models Reg$ can be decided in EXPTIME.

Lemma 3 (Lower Bound). *Given a finite run R and regulation Reg, $R \models Reg$ is hard for EXPTIME (time exponential in the size of Reg)*

The lower bound is shown by a reduction from first order logic enriched with a least fixed point predicate (the system YF in [21]). With circular references, we can encode reachability computations that cannot be expressed in first order logic.

Discussion: We now discuss some options in defining conformance, depending on the needs of the application. The sections of the FDA CFR that we have examined can be formalized so that there is a unique fixed point, and conformance is simply the satisfaction of obligations at this fixed point.

However, examples discussed in the literature suggest that it may not be desirable to always have a unique fixed point. A well-known example is that of contrary-to-duty (CTD) obligations (c.f. [16]). CTD obligations are those that arise when other obligations have been violated. Prakken and Sergot [16] point out an inflexibility in casting

CTD structures as an instance of non-monotonic reasoning. We outline how this inflexibility can be avoided, using alternate definitions of conformance. Consider the following example from [14] (similar to one in [16]): *The cottage must not have a fence or a dog. If it has a dog, then it must have both a fence and a warning sign*. The question is what are the obligations when the cottage has a dog. We discuss two possible solutions.

The first solution is to treat the CTD norm as an exception to the first:

1.**o**: $\neg\mathrm{by}_{\{2\}}(f \vee d) \rightsquigarrow \neg(f \vee d)$ and 2.**o**: $d \rightsquigarrow f \wedge w$

The propositions f, d and w correpond to the cottage having a fence, dog and warning sign resply. Since there is a dog, the precondition of the second law is true, and this leads to the precondition of the first law being false. So if $f \wedge w$ holds, there is no violation. However, as [16] points out, it may be useful to detect that the situation is worse than the one in which there is no dog. In the second solution, we represent the laws as excluding each other, i.e., we conjoin $\neg\mathrm{by}_{\{1\}}(\neg(f \wedge w))$ to the precondition of the second law. At the least fixed point, both obligations are ungrounded, and we get two maximal fixed points – one in which $\neg(f \vee d)$ is obligated, and one in which $f \wedge w$ is obligated. Since d holds, there is a violation w.r.t. the former fixed point. In a scenario where there is no dog, a unique fixed point is obtained.

Our analysis of CTD structures achieves the same effect as the analyses in [16,14]. However, [16,14] characterize the CTD norm as presupposing the violation of the other, and then revising the situation. In future work, we plan to investigate predicates that capture this presuppositional analysis more directly.

3.4 Axiomatization

As we discussed in the context of Lemma 3, RefL contains first order logic enriched with a least fixed point predicate. It can be shown that the validity problem is Π^1_1-hard, and as a result, it cannot be recursively axiomatized. In this section, we briefly discuss an axiomatization of the propositional fragment of L' (the language of preconditions).

We assume as given a fixed finite domain of quantification, and replace variables by identifiers for domain elements. Given a set of identifiers ID, a propositionalized body of regulation has one or more statements of the form $id.\mathbf{x} : \varphi \rightsquigarrow \psi$ for each $id \in ID$. For example, the presence of $id.\mathbf{x} : \varphi_1 \rightsquigarrow \psi_1$ and $id.\mathbf{x} : \varphi_2 \rightsquigarrow \psi_2$ corresponds to different assignments to the variables.

To simplify presentation, we will assume that the references in the regulation are acyclic. This lets us obtain a unique fixed point and restrict attention to a two-valued logic. We discuss the general case at the end of this section.

A1 All substitution instances of propositional tautologies
A2 $\Box(\varphi \Rightarrow \psi) \Rightarrow (\Box\varphi \Rightarrow \Box\psi)$
A3 $\Box\varphi \Rightarrow \varphi \wedge \Box\Box\varphi$
R1 From $\vdash \varphi \Rightarrow \psi$ and $\vdash \varphi$, infer $\vdash \psi$
R2 From $\vdash \varphi$ infer $\vdash \Box\varphi$

We characterize the inference predicate by the laws it refers to. To axiomatize $\mathrm{by}_{\mathrm{Id}}(\varphi)$, we need to reason about provability in the language L (propositional LTL). We say that $\varphi \in L$ is is provable (denoted $\vdash_L \varphi$) iff it is an instance of the axioms A1-A3, or follows from the axioms using the rules R1 and R2. Crucially, we will use the negation of provability in the premise of a rule. Similar mechanisms have been used to axiomatize default logic, e.g., in [22], satisfiability is used in the premise of a rule, and in [23], a modal language is augmented with an operator for satisfiability.

We begin by developing some notation. Given a set of regulatory statements $F = \{id_1.\mathbf{x} : \ \varphi_1 \rightsquigarrow \psi_1, ..., id_n.\mathbf{x} : \ \varphi_n \rightsquigarrow \psi_n\}$, let $F_{pre} = \{\varphi_1, ..., \varphi_n\}$ be the set of preconditions, $F_{post} = \{\psi_1, ..., \psi_n\}$ be the set of postconditions, and $F_{id} = \{id_1, ..., id_n\}$ be the set of identifiers. Given a finite set of formulas Γ, we denote the conjunction by $\bigwedge \Gamma$. The conjunction of the empty set is identified with $\top$ (a tautology). We use two rules for the inference predicate:

R3 For all $F \subseteq Reg$ with $F_{id} \subseteq Id$, from $\vdash_L \bigwedge F_{post} \Rightarrow \phi$, infer $\vdash \bigwedge F_{pre} \Rightarrow \mathrm{by}_{\mathrm{Id}}(\phi)$

R4 For all $\psi \in L'$, if for all $F \subseteq Reg$ with $F_{id} \subseteq Id$, either $\not\vdash_L \bigwedge F_{post} \Rightarrow \phi$, or $\vdash \psi \Rightarrow \neg \bigwedge F_{pre}$, then infer $\vdash \psi \Rightarrow \neg\mathrm{by}_{\mathrm{Id}}(\phi)$.

Informally, R3 says that $\mathrm{by}_{\mathrm{Id}}(\phi)$ is true, if there exists a set of laws whose postconditions imply ϕ, and whose preconditions are true. R4 says that $\mathrm{by}_{\mathrm{Id}}(\phi)$ is false, if one of the preconditions is false for all sets of laws whose postconditions imply ϕ. In particular, if $\not\vdash_L \bigwedge F_{post} \Rightarrow \phi$ for all appropriate subsets, then $\vdash \top \Rightarrow \neg\mathrm{by}_{\mathrm{Id}}(\phi)$, and using R1, $\vdash \neg\mathrm{by}_{\mathrm{id}}(\phi)$.

The rules have an equivalent axiomatic characterization, which is important in establishing completeness. Given $\phi \in L$, let $\mathcal{F}_{(Id,\phi)}$ be the set of subsets ($F \subseteq Reg$ with $F_{id} \subseteq Id$) such that $F \in \mathcal{F}$ iff $\vdash_L \bigwedge F_{post} \Rightarrow \phi$. Let $\Gamma_{(Id,\phi)}$ be the set such that $\neg \bigwedge F_{pre} \in \Gamma_{(Id,\phi)}$ iff $F \in \mathcal{F}_{(Id,\phi)}$. Finally, let $\Delta_{(Id,\phi)}$ be the set such that $\bigwedge F_{pre} \in \Delta_{(Id,\phi)}$ iff $F \in \mathcal{F}_{(Id,\phi)}$.

Proposition 3. *The following are provable:*

1. $\vdash \bigwedge \Gamma_{(Id,\phi)} \Rightarrow \neg\mathrm{by}_{\mathrm{Id}}(\phi)$
2. $\vdash \mathrm{by}_{\mathrm{Id}}(\phi) \Rightarrow \bigvee \Delta_{(Id,\phi)}$

The first claim is an immediate consequence of R4. And, the second claim follows from the first by propositional reasoning. It is easy to show that the axioms A1-A3, together with Proposition 3, and the rules R1 and R2 imply the rules R3 and R4. The inference predicate behaves like a modality:

Proposition 4. $\vdash \mathrm{by}_{\mathrm{Id}}(\varphi \Rightarrow \psi) \Rightarrow (\mathrm{by}_{\mathrm{Id}}(\varphi) \Rightarrow \mathrm{by}_{\mathrm{Id}}(\psi))$

Completeness is be established by a standard pre-model construction (see [19] for details). We now discuss the general case, i.e., when there are circular references and multiple fixed points. In the presence of multiple fixed points, we can define validity w.r.t. all fixed points, the least fixed point, or maximal fixed points. The axioms and rules discussed here can be adapted to characterize valdity w.r.t. all fixed points [19]. However, we have not obtained a direct characterization of validity w.r.t. the least or maximal fixed points. [22] provides an axiomatization of these three notions of validity for default logic, by translating the default rules into an autoepistemic logic. A question of interest is whether the the translation procedure in [22] can be adapted for RefL.

4 Conclusions and Future Work

We have motivated and described a logic (RefL) that accomodates references between laws. RefL separates two uses of statements – drawing inferences from regulation, and determining facts about an organization. We believe that this separation is crucial to the application of conformance checking.

The inference predicate blends two ideas from logic programming. First, the Kripke-Kleene-Fitting semantics [24], which uses three values for negation in logic programs. In RefL, we place the burden on a predicate, rather than on negation. The advantage is that connectives can behave as they do in a many valued logic. Second, contextual logic programs [25] use operations to restrict the context from which inferences are derived. Referring to specific laws (via identifiers) gives us a fine-grained control of context.

RefL provides a staring point in bringing the advantages of non-monotonic reasoning to systems such as [2,4]. [2] represents business contracts as SQL queries, and [4] uses first-order logic augmented with real time operators. The inference predicate can be added to these systems, provided that the existential quantification is relativized to either the preconditions or the postconditions. However, restrictions are needed to ensure that the satisfiability tests remain decidable. [3] discusses the importance of anlayzing references, but do not provide a formalization.

In this work, we have considered references to laws that appear in preconditions. There is also the need for references in postconditions. An obvious case is for laws that cancel obligations and permissions given by another, e.g., *if a donation is not used for transfusion, exemption (3) no longer applies*. A more speculative case can be made for iterated deontic constructs [17], e.g., “required to allow x”. We suggest that the semantics will involve representing agents who introduce laws that reason about each other, e.g., *You are required to (introduce laws that) allow a patient to see his records*.

On the computational side, our goal is to be able to scale up to runs with a large number of objects, and incorporate RefL into a runtime checking framework for LTL. In a companion paper [26], we identify a fragment of RefL motivated by a case study of the FDA CFR. The fragment assumes that $\mathrm{by}_{\mathrm{Id}}(\varphi)$ can be evaluated by using at most one of the laws referred to. This assumption allows us to replace satisfiability tests with tests of lower complexity, and lets us scale up to runs with a large number of objects. In this paper, we have focussed on formally characterizing the semantics and complexity of RefL, and in [26], we focus on optimizations that are needed in practice.

References

1. U.S. Food and Drug Administration: Code of Federal Regulations, http://www.gpoaccess.gov/cfr/index.html
2. Abrahams, A.: Developing and Executing Electronic Commerce Applications with Occurrences. PhD thesis, Univeristy of Cambridge (2002)
3. Breaux, T.D., Vail, M.W., Anton, A.I.: Towards regulatory compliance: Extracting rights and obligations to align requirements with regulations. In: Proceedings of the 14th IEEE International Requirements Engineering Conference (2006)
4. Giblin, C., Liu, A., Muller, S., Pfitzmann, B., Zhou, X.: Regulations Expressed as Logical Models (REALM). In: Moens, M.F., Spyns, P. (eds.) Legal Knowledge and Information Systems (2005)

5. Ross, A.: Directives and Norms. Routlege and Kegan Paul (1968)
6. Boella, G., van der Torre, L.: Permissions and obligations in hierarchical normative systems. In: Proceedings of the 9th international conference on AI and law (2003)
7. Reiter, R.: A logic for default reasoning. In: Readings in nonmonotonic reasoning, pp. 68–93. Morgan Kaufmann Publishers Inc., San Francisco (1987)
8. McCarty, L.T.: A language for legal discourse - i. basic features. In: Proceedings of ICAIL (1989)
9. Sergot, M., Sadri, F., Kowalski, R., Kriwaczek, F., Hammond, P., Cory, H.: The british nationality act as a logic program. Communications of the ACM 29(5), 370–386 (1986)
10. Kripke, S.: Outline of a theory of truth. Journal of Philosophy 72, 690–716 (1975)
11. Dinesh, N., Joshi, A., Lee, I., Sokolsky, O.: Logic-based regulatory conformance checking. In: Proceedings of the 14th Monterey Workshop (2007)
12. Bench-Capon, T., Robinson, G., Routen, T., Sergot, M.: Logic programming for large scale applications in law: A formalisation of supplementary benefit legislation. In: Proceedings of the 1st International Conference on AI and Law (1987)
13. Holzmann, G.: The Spin model checker. IEEE Trans. on Software Engineering 23(5), 279–295 (1997)
14. Makinson, D., van der Torre, L.: Input/output logics. Journal of Philosophical Logic 29, 383–408 (2000)
15. Makinson, D., van der Torre, L.: Permissions from an input/output perspective. Journal of Philosophical Logic 32(4) (2003)
16. Prakken, H., Sergot, M.: Contrary-to-duty obligations. Studia Logica 57(1), 91–115 (1996)
17. Marcus, R.B.: Iterated deontic modalities. Mind 75(300) (1966)
18. Sistla, A.P., Clarke, E.M.: The complexity of propositional linear temporal logic. ACM 32, 733–749 (1985)
19. Dinesh, N., Joshi, A., Lee, I., Sokolsky, O.: A default temporal logic for regulatory conformance checking. Technical Report MS-CIS-08-07, University of Pennsylvania (2008)
20. Rudin, W.: Real and Complex Analysis. McGraw-Hill Book Company (1987)
21. Vardi, M.: The complexity of relational query languages. In: STOC (1982)
22. Lakemeyer, G., Levesque, H.: Towards an axiom system for default logic. In: Proceedings of the AAAI Conference (2006)
23. Halpern, J., Lakemeyer, G.: Multi-agent only knowing. Journal of Logic and Compuation 11(1) (2001)
24. Fitting, M.: A Kripke/Kleene Semantics for logic programs. Journal of Logic Programming 2 (1985)
25. Monteiro, L., Porto, A.: A language for contextual logic programming. In: Apt, K., de Bakker, J., Rutten, J. (eds.) Logic Programming Languages: Constraints, Functions and Objects (1993)
26. Dinesh, N., Joshi, A., Lee, I., Sokolsky, O.: Checking traces for regulatory conformance. In: Proceedings of the Workshop on Runtime Verification (2008)

Need to Know: Questions and the Paradox of Epistemic Obligation

Joris Hulstijn

Faculty of Economics and Business Administration
Vrije Universiteit, Amsterdam
jhulstijn@feweb.vu.nl

Abstract. Åqvist's paradox of epistemic obligation can be solved, if we use knowledge-wh instead of knowledge-that in specifications of the 'need to know': the knowledge which an agent in a certain organisational role is required to have. Knowledge-wh is knowledge of an answer to a question, which depends on the context. We show how knowledge-wh can be formalised in a logic of questions, which is combined with standard deontic logic to represent epistemic obligations. We demonstrate that under the new interpretation, the paradox can no longer be derived. The resulting logic is useful for representation of access control policies.

1 Introduction

Many types of computer systems require a way to specify the knowledge which agents in an organisational role need or need not have. Consider an access control policy for a distributed database. Access is regulated by rules, which indicate who may access which documents for what purpose. In police records, for example, only those officers who are working on a specific case, and who therefore have the 'need to know', are authorised to access the case files. Managing access control policies is becoming increasingly complex. One must demonstrate that a set of rules satisfies certain properties, such as consistency, least privilege, or segregation of duties. A declarative representation of access control rules can be managed by a so called trust management system, and be used to assist in formal specification and verification [3]. To help develop a conceptual model for such specifications, a logic could be useful. Which logics are appropriate?

Traditionally, norms are studied in *deontic logic*, see e.g. [18]. Permissions and obligations are expressed by formulas of the form $\mathsf{P}\varphi$ and $\mathsf{O}\varphi$, respectively. Standard Deontic Logic (SDL) has axioms, which make it a normal modal logic. This means among other things that obligations can be distributed over implication, and that necessitation can be used as a derivation rule. The logic of knowledge and information is called *epistemic logic* [14, 6]. Knowledge of an agent i is expressed by formulas of the form $\mathsf{K}_i\varphi$. Knowledge is characterised by modal logic $S5$, which means among other things, that knowledge is supposed to be true. To specify a 'need to know', the obvious thing to do is to combine deontic and epistemic logic, and study the resulting *epistemic obligations*, expressed as $\mathsf{OK}_i\varphi$. However, the combination is not as straightforward as it may seem. There are various problems and paradoxes related to the combination of deontic and epistemic logic [17]. In this paper, we discuss one such problem: Åqvist's paradox of epistemic obligation [2].

R. van der Meyden and L. van der Torre (Eds.): DEON 2008, LNAI 5076, pp. 125–139, 2008.

The problem is the following: by combining the truth axiom of knowledge with the distribution axiom and the necessitation derivation rule of standard deontic logic, we can derive from the fact that someone needs to know some facts, that these facts are true. This is counterintuitive, as illustrated by the following example.

(1) The bank is being robbed.
It ought to be the case that Jones (the guard) knows that the bank is being robbed.
So, it ought to be the case that the bank is being robbed.

In this paper, we want to demonstrate that the paradox can be avoided, if only we would specify epistemic obligations in a different way. Instead of specifying that agents in some organisational role need to know *that* some information is true, the designer of a system should specify *what* an agent needs to know, or *whether* some information is true or not. Such knowledge-wh corresponds to knowing the answer to some question. Note that the answer to a question crucially depends on the context, but that at design time, the true answer is often not yet known. In the example above, Jones, the guard, ought to know *whether* the bank is being robbed or not. In other words, *if* the bank is being robbed, Jones should know that this is the case, and if not, not.

The approach can be made precise using the logic of questions and answers developed by Groenendijk and Stokhof [12]. This logic can express an entailment relation between questions, as well as relationships between questions and propositions, such as the answerhood relation. The answerhood relation states under what circumstances a proposition is said to give a complete answer to a question. Because the semantics is expressed in terms of sets of possible worlds, the logic of questions can be combined relatively easily with epistemic logic. We use formulas of the form $\mathsf{K}_i?\varphi$ to express that agent i has knowledge of an answer to the question $?\varphi$, i.e., whether φ is true or not.

Epistemic obligations with embedded questions have a great potential beyond the paradox. Modern information systems require specification of the knowledge which agents enacting a specific role in an organisation must possess, or are not allowed to possess. Examples of such specifications are found in work-flow management systems, information systems security, electronic institutions, or web-service applications. Because the future is unknown, the legislator or designer of a system is generally not in a position to specify the actual knowledge of an agent for all situations. Instead, a legislator should specify to which (kinds of) questions an agent in a certain organisational role ought to know – or not know – the answer. Consider examples (2) and (3).

(2) Passwords should only be known by their owner.

(3) Jones (the guard) should know which key fits on which door.

The remainder of the paper is structured as follows. In section 2 we present deontic logic and epistemic logic, and show how the paradox arises. In section 3 we present a logic of questions, and show how it can be used to express knowledge-wh. We also show how to combine this logic with epistemic logic and standard deontic logic. In section 4 we demonstrate that, if we would use the knowledge-wh representation of Jones' job description, the paradox no longer arises. Finally, in section 5 we consider some applications of the logic.

2 Åqvist's Paradox

To show how the paradox can be derived, we first need to characterise the base logics: Standard Deontic Logic and epistemic logic.

We want a logical language to express knowledge $\mathsf{K}_i\varphi$, obligations $\mathsf{O}\varphi$ and permissions $\mathsf{P}\varphi$ as well as epistemic obligations of the form $\mathsf{OK}_i\varphi$. We also want to say something about other relations between knowledge and obligations, for instance, the property that obligations are known: $\vdash \mathsf{O}\varphi \rightarrow \mathsf{K}_i\mathsf{O}\varphi$. So the language should allow knowledge and obligation to be embedded both ways.

Definition 1 (Syntax). Let $P = \{p, q, ...\}$ be a set of proposition variables, and let $A = \{i, j, ...\}$ be a set of agents, then language L is characterised by:
$\varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \psi) \mid (\varphi \rightarrow \psi) \mid \mathsf{K}_i\varphi \mid \mathsf{O}\varphi \mid \mathsf{P}\varphi$.

As usual we define $(\varphi \rightarrow \psi) \equiv \neg(\neg\varphi \wedge \psi)$ and $\mathsf{P}\varphi \equiv \neg\mathsf{O}\neg\varphi$.

Regulations and norms are studied in deontic logic [18]. Standard Deontic Logic is characterised by the following axioms and derivation rules [17].

P All tautologies of proposition logic.
K $\vdash \mathsf{O}(\varphi \rightarrow \psi) \rightarrow (\mathsf{O}\varphi \rightarrow \mathsf{O}\psi)$
D $\vdash \mathsf{O}\varphi \rightarrow \mathsf{P}\varphi$
MP If $\vdash \varphi$ and $\vdash \varphi \rightarrow \psi$, then $\vdash \psi$.
Nec If $\vdash \varphi$, then $\vdash \mathsf{O}\psi$.

The **K** axiom expresses the usual distribution requirement for a modal logic; the **D** axiom guarantees consistency. Together with **P** this means SDL is a normal modal logic. **MP** is the rule of Modus Ponens, and **Nec** is the necessitation rule, applied to obligation.

The logic of knowledge and information is called epistemic logic [6]. Epistemic logic originates with Hintikka [14]. Usually, the logic of knowledge is taken to be S5. It is characterised by the following axioms and derivation rules.

P All tautologies of proposition logic.
K $\vdash \mathsf{K}_i(\varphi \rightarrow \psi) \rightarrow (\mathsf{K}_i\varphi \rightarrow \mathsf{K}_i\psi)$
T $\vdash \mathsf{K}_i\varphi \rightarrow \varphi$
4 $\vdash \mathsf{K}_i\varphi \rightarrow \mathsf{K}_i\mathsf{K}_i\varphi$
5 $\vdash \neg\mathsf{K}_i\varphi \rightarrow \mathsf{K}_i\neg\mathsf{K}_i\varphi$
MP If $\vdash \varphi$ and $\vdash \varphi \rightarrow \psi$, then $\vdash \psi$.
Nec If $\vdash \varphi$, then $\vdash \mathsf{K}_i\psi$.

In addition to **P**, **K**, **MP** and **Nec**, for knowledge we also have **T**, the truth axiom, which expresses that knowledge is veridical: when we say that someone knows something, this knowledge is presupposed to be true. This feature distinguishes knowledge from belief, among other things. Axioms **4** and **5** represent positive and negative introspection respectively: when someone knows or doesn't know something, he is supposed to know that he does or doesn't know. Although the S5 axioms are usually not true for humans, who have bounded rationality, they are true for idealised situations, like databases or the statements collected in a dispute.

Regarding the interaction between knowledge and obligation, we consider the following positive and negative introspection axioms for obligation.

O4 $\vdash \mathsf{O}\varphi \rightarrow \mathsf{K}_i\mathsf{O}\varphi$
O5 $\vdash \neg\mathsf{O}\varphi \rightarrow \mathsf{K}_i\neg\mathsf{O}\varphi$

These are strong assumptions. Although it is strong, we believe that axiom **O4** is not more idealised than positive and negative introspection of knowledge itself, or the fact that knowledge is closed under deduction. Axiom **O5** however, is too strong. This can be seen more easily, by taking the contraposition: $\vdash \mathsf{M}_i\mathsf{O}\varphi \rightarrow \mathsf{O}\varphi$, where $\mathsf{M}_i\varphi \equiv \neg\mathsf{K}_i\neg\varphi$ is the dual of knowledge. Now it reads: "when i considers it possible that φ is obliged, φ is indeed obliged", which is clearly wrong.

In general **O4** is a nice property: one should know the rules. But getting to know the rules takes time and effort. An agent caught not knowing the rules, would be a violation; not a failure. So for resource bounded applications this principle is better represented by a specific norm, like $M \models \mathsf{O}(\mathsf{O}\varphi \rightarrow \mathsf{K}_i\mathsf{O}\varphi)$, rather than an axiom.

Regarding the converse interaction, about being obliged or forbidden to know something, we do not want to impose any axioms. To the contrary, we might consider a 'freedom of thought' axiom, $\vdash \mathsf{K}_i\varphi \rightarrow \mathsf{P}\mathsf{K}_i\varphi$, to exclude the definition of 'thought crimes', as in Orwell's 1984. However, such an axiom goes against the purpose of access control policies. In some cases there is good reason to specify or restrict what may be known.

Finally, we could consider an interaction axiom comparable to feasibility in BDI logic. This would come out as $\vdash \mathsf{O}\varphi \rightarrow \mathsf{K}_i\varphi$, which is wrong, or else as $\vdash \mathsf{O}\varphi \rightarrow \mathsf{M}_i\varphi$. The contraposition, $\vdash \mathsf{K}_i\varphi \rightarrow \mathsf{P}\varphi$, indicates that this suggestion is wrong too.

2.1 Semantics

The semantics of epistemic logic is traditionally given in terms of Kripke structures, with accessibility relations $\mathcal{R}_i$, for agents i. The semantics is sound and complete with respect to S5, when the accessibility relation $\mathcal{R}_i$ is an equivalence relation [6]. Relation $\mathcal{R}_i$ is interpreted as indistinguishability: two worlds are related when they are *indistinguishable* with respect to the knowledge of agent i. The set of worlds accessible from w, denoted by $\mathcal{R}_i(w)$, characterises the information state of agent i: $\mathcal{R}_i(w) = \{v \mid \mathcal{R}_i(w, v)\}$.

Standard Deontic Logic has a possible worlds semantics too. Here the accessibility relation $\mathcal{I}$ represents an ideal situation with respect to compliance. The set of ideal worlds is denoted by $\mathcal{I}(w) = \{v \mid \mathcal{I}(w, v)\}$. We require that $\mathcal{I}$ is serial: for all w there is at least one ideal world v. This will ensure the consistency axiom **D**. It is well known that this semantics is sound and complete with respect to SDL [17].

What about the interaction between deontic and epistemic logic? To ensure **O4**, positive introspection for obligation, we require the following frame property, which looks a bit like transitivity. A similar property is known from KB-logic, which studies the interaction of knowledge and belief [15].

If $\mathcal{R}_i(w, v)$ and $\mathcal{I}(v, u)$, then $\mathcal{I}(w, u)$. ($\mathcal{RI}$-transitive)

Incidently, negative introspection for obligations **O5** would correspond to frames being $\mathcal{RI}$-Euclidean: if $\mathcal{R}_i(w, v)$ and $\mathcal{I}(w, u)$, then $\mathcal{I}(v, u)$. However, we already decided that **O5** is not a desirable property.

Definition 2 (Semantics L). Let $M = \langle W, A, \{\mathcal{R}_i\}_{i \in A}, \mathcal{I}, V\rangle$ be a model consisting of a set of possible worlds W, a set of agents A, a set of reflexive, symmetric and transitive accessibility relations $\mathcal{R}_i \subseteq W \times W$ for each agent i, a serial accessibility relation $\mathcal{I} \subseteq W \times W$, which together are $\mathcal{RI}$-transitive, and a valuation function V. Define the satisfaction relation '$\models'$ as follows:

$$\begin{array}{lll}
M, w \models p & \text{iff} & V(w)(p) = 1 \\
M, w \models \neg\varphi & \text{iff} & M, w \not\models \varphi \\
M, w \models \varphi \wedge \psi & \text{iff} & M, w \models \varphi \text{ and } M, w \models \psi \\
M, w \models \mathsf{K}_i\varphi & \text{iff} & M, v \models \varphi, \text{ for all } v \text{ such that } \mathcal{R}_i(w, v). \\
M, w \models \mathsf{O}\varphi & \text{iff} & M, v \models \varphi, \text{ for all } v \text{ such that } \mathcal{I}(w, v).
\end{array}$$

Now we should prove soundness and completeness. For the two base logics, this is already well known [6, 17]. What remains to be shown is the correspondence between **O4** and frames being $\mathcal{RI}$-transitive.

Proposition 1 (Correspondence)
A frame $\langle W, A, \{\mathcal{R}_i\}_{i\in A}, \mathcal{I}\rangle$ is $\mathcal{RI}$-transitive iff $\vdash \mathsf{O}\varphi \rightarrow \mathsf{K}_i\mathsf{O}\varphi$ is valid. Proof.

($\Rightarrow$) Suppose the frame is $\mathcal{RI}$-transitive. We prove the contraposition: $\vdash \mathsf{M}_i\mathsf{P}\psi \rightarrow \mathsf{P}\psi$. Take any valuation V such that $M, w \models \mathsf{M}_i\mathsf{P}\psi$, for some w. By definition 2 there is some v such that $\mathcal{R}_i(w, v)$ and $M, v \models \mathsf{P}\psi$. Again by definition 2, there is some u such that $\mathcal{I}(v, u)$ and $M, u \models \psi$. Because the frame is $\mathcal{RI}$-transitive, we have $\mathcal{I}(w, u)$. So by Def 2 we have $M, w \models \mathsf{P}\psi$.

($\Leftarrow$) Suppose $\vdash \mathsf{O}\varphi \rightarrow \mathsf{K}_i\mathsf{O}\varphi$ is valid. Let $\mathcal{R}_i(w, v)$ and $\mathcal{I}(v, u)$ for arbitrary w, v, u. To be shown: $\mathcal{I}(w, u)$. This is the case when $M, w \models \mathsf{O}\varphi$ implies $M, u \models \varphi$ for all φ. Take a valuation V such that $M, w \models \mathsf{O}\varphi$. By **O4** $M, w \models \mathsf{K}_i\mathsf{O}\varphi$. So by definition 2 $M, v \models \mathsf{O}\varphi$, and $M, u \models \varphi$.

That completes our presentation of a framework to express epistemic obligations.

2.2 Paradox of Epistemic Obligation

What about the paradox? Åqvist's [2] Paradox derives from the combination of two assumptions, which are plausible in isolation [17]. The first is the assumption that knowledge is veridical, expressed by the truth axiom. The second assumption is known as **RM** and follows from necessitation (**Nec**) and distribution for obligations (**K**).

RM If $\vdash \varphi \rightarrow \psi$, then $\mathsf{O}\varphi \rightarrow \mathsf{O}\psi$

Let us reconsider the example from the introduction. The following account of the paradox is based on [17].

(1)	1.	The bank is being robbed.	$\texttt{robbed}$	
	2.	It ought to be the case that Jones knows that the bank is being robbed.	$\mathsf{OK}_j\,\texttt{robbed}$	
	3.		$\mathsf{K}_j\,\texttt{robbed} \rightarrow \texttt{robbed}$	(T)
	4.		$\mathsf{OK}_j\,\texttt{robbed} \rightarrow \mathsf{O}\,\texttt{robbed}$	(3,RM)
	5.	It ought to be the case that the bank is being robbed.	$\mathsf{O}\,\texttt{robbed}$	(2,4,MP)

Suppose that a bank is being robbed. Jones is the guard of the building. It is part of his job description to know that the bank is being robbed. We can formalise these assumptions in premise 1 and 2. Premise 3 is an instantiation of the truth axiom. In premise 4 and 5 we derive the counterintuitive result that it ought to be the case that the bank is being robbed. Note that we do not really need premise 1 in the derivation. In the general discussion it is used to support premise 2, but it does re-enter in the formal version of the paradox. We will come back to this observation later.

Usually, when discussing the paradox, people blame Standard Deontic Logic, and there are many different ways of changing the logic to avoid this paradox [17]. Most of these attack the RM principle, because it also leads to other paradoxes, in particular to Ross' Paradox and The Good Samaritan paradox. In this paper however, we will attack the truth axiom for knowledge, or to be more precise, the way the job description of Jones, the guard, is represented in the logic. We believe that Jones' job description – or epistemic obligations in general – should take the form of an embedded question. It is his job to know *whether* the bank is being robbed or not, not *that* the bank is being robbed. In other words, it is Jones' job to know the answer to a question. This answer crucially depends on the context. The context dependency in our approach is compatible with other solutions of deontic paradoxes, which stress the conditional nature of obligations: if the bank is being robbed, then Jones ought to know that. See Tomberlin [21] for a critical evaluation of such approaches.

3 Questions and Answers

There are different kinds of knowledge. Traditionally, knowledge-how is contrasted with knowledge-that. But there is yet another type: knowledge-wh. In school, for example, you learn *what* the capital of France is. Should we express such knowledge as a proposition? Should we require for all pupils i, that $\mathsf{OK}_i\, \texttt{capital}(\texttt{Paris}, \texttt{France})$? But what about the capital of Guatemala? Or knowing who is the current president? In other words, what if those who specify the required knowledge, do not know these things themselves? In such cases, it seems more natural to require that pupils know the answer to the question, regardless of the answer at the time of specification.

In linguistics, Groenendijk and Stokhof [11, 12] have developed a logic of interrogative expressions, in particular questions. Like propositions, questions stand in entailment relations, and there is a logical relation between propositions and questions, called answerhood. Crucially, questions can express the embedded content of verbs like 'wonder', 'doubt' and 'know'. In this paper we use the 'update' version of the logic of interrogatives [10]. This version has been given a sound and complete axiomatisation [5]. Initially, we only present a propositional logic version. In Section 5 we discuss a predicate logic version, to cover examples like (3) and (2).

First, we define a logical language, which only contains questions.

Definition 3 (Syntax). Let L be as in Definition 1. Then $L' = \{?\varphi \mid \varphi \in L\}$.

What about the interpretation? An answer is a proposition: a set of possible worlds. A question is a specification of the possible answers to the question, one for each possible world. Therefore it is a set of sets of worlds. It turns out that this set of propositions forms a partition: the answer sets are disjoint and together cover the whole space of possibilities [11]. Technically, a partition corresponds to an equivalence relation. This relation can also be interpreted as an indistinguishability relation: two worlds are connected whenever they agree on their specific answer to the question.

Definition 4 (Interpretation). Let M be a model as in Definition 2. For any $\varphi \in L$ and $?\psi \in L'$ define interpretation $[\![.]\!]$ as follows:
$[\![\varphi]\!] = \{w \mid M, w \models \varphi\}$
$[\![?\psi]\!] = \{\langle w, v\rangle \mid M, v \models \psi \text{ iff } M, w \models \psi\}$

Groenendijk [10] interprets an information exchange as a dynamic process. Questions raise distinctions between different possible alternatives. Answers provide information to eliminate some alternatives. The semantics of questions and answers is therefore most naturally given in terms of the changes they bring to an information state. This is called *update semantics* [24]. The idea is illustrated by the following exchange.

(4) A: Should I bring my umbrella?
B: Well, it is raining.

Here, B answers A's question with what is strictly speaking the answer to a different question. This makes sense, when the two questions are related: when A and B both know that one should bring an umbrella whenever it is raining.

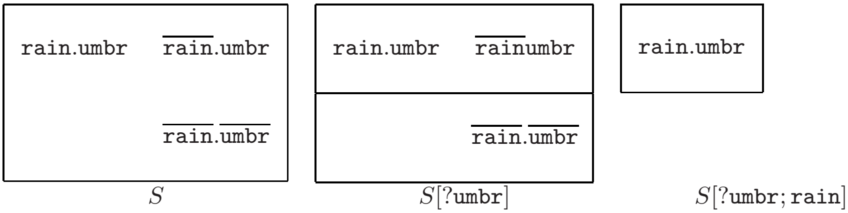

$\mathtt{rain.umbr}$
$\overline{\mathtt{rain}}\mathtt{.umbr}$
$\overline{\mathtt{rain}}\mathtt{.}\overline{\mathtt{umbr}}$
S
$\mathtt{rain.umbr}$
$\overline{\mathtt{rain}}\mathtt{umbr}$
$\overline{\mathtt{rain}}\mathtt{.}\overline{\mathtt{umbr}}$
$S[?\mathtt{umbr}]$
$\mathtt{rain.umbr}$
$S[?\mathtt{umbr};\mathtt{rain}]$

Fig. 1. Information exchange as a process of asking questions and giving answers

An information state S consists of a set of possible worlds F, structured by an equivalence relation Q. Worlds in F are compatible with the information of an agent. Adding information, will eliminate all worlds that are incompatible with it. The equivalence relation Q models which distinctions are considered relevant. Adding a question makes the distinctions more fine grained. All pairs of worlds that cut across a distinction, are eliminated. To illustrate this idea, the dialogue of example (4) is shown in Figure 1. Suppose S represents the common ground, which already contains the fact that $\mathtt{rain} \rightarrow \mathtt{umbr}$, so world $\mathtt{rain.}\overline{\mathtt{umbr}}$ has been eliminated. When we ask $?\mathtt{umbr}$ we divide the set into two parts: umbrella and no umbrella. Adding the information $\mathtt{rain}$ removes the not-rain worlds, and therefore also answers the original question.

Definition 5 (Auxiliary Notation). Let X, Y be any set and let $R \subseteq X \times X$ be an equivalence relation. Define the following notation:

$R \downarrow Y = \{\langle x, y\rangle \in R \mid x \in Y, y \in Y\}$	restriction of R to Y
$(x)_R = \{y \in X \mid R(x, y)\}$	equivalence class induced by R
$Y/R = \{(x)_R \mid x \in Y\}$	partition induced by R

Definition 6 (Update Semantics). Let $S = \langle F, Q\rangle$ be an information state, $F \subseteq W$ and $Q \subseteq F \times F$ an equivalence relation. Define an update function $[.]$ as follows:
$\langle F, Q\rangle[\varphi] = \langle F \cap [\![\varphi]\!], Q \downarrow [\![\varphi]\!]\rangle$
$\langle F, Q\rangle[?\varphi] = \langle F, Q \cap [\![?\varphi]\!]\rangle$

Update semantics employs a content dependent notion of entailment [24]. The idea is that an information states *supports* some information χ, written $S \Vdash \chi$, when an update with it does not add any information. This holds for both propositions and questions.

Definition 7. For all $\chi_1, ...\chi_n, \psi \in L \cup L'$ and information states S, define:
$S \Vdash \psi$ iff $S[\psi] = S$.
$\chi_1, ...\chi_n \Vdash_S \psi$ iff $S[\chi_1], ...[\chi_n] \Vdash \psi$.
$\chi_1, ...\chi_n \Vdash \psi$ iff $\chi_1, ...\chi_n \Vdash_S \psi$, for any S.

For propositions, one can prove that this notion of entailment corresponds to the regular one: $\varphi_1, ..., \varphi_n \Vdash \psi$ iff $[\![\varphi_1]\!] \cap ... \cap [\![\varphi_n]\!] \subseteq [\![\psi]\!]$. For questions, this entailment can be explained by reducing a question to its answerhood conditions. So $?\varphi$ implies $?\psi$ when all answers to $?\varphi$ also qualify as answers to $?\psi$. This is the case when the distinctions made by $?\varphi$ are more fine grained than those made by $?\psi$. So one can prove that $?\varphi_1, ..., ?\varphi_n \Vdash ?\psi$ iff $[\![?\varphi_1]\!] \cap ... \cap [\![?\varphi_n]\!] \subseteq [\![?\psi]\!]$ So question “Does John come and does Mary come too?” implies the question “Does John come?”, but not vice versa.

A crucial notion for the logic of questions is answerhood. A proposition qualifies as an answer to a question , when the corresponding set of worlds is subsumed in one of the equivalence classes of the partition, induced by the question [12]. It turns out that this is a special case of the notion of support, where the consequent is a question [10].

Definition 8 (Answerhood). For any information state S, $\varphi \in L$ and $?\psi \in L'$, we say that φ provides an an *answer* to $?\psi$ in S, written $\varphi \Vdash_S ?\psi$, whenever $S[\varphi] \Vdash ?\psi$. Moreover, $\varphi \Vdash ?\psi$ in general, whenever $\varphi \Vdash_S ?\psi$, for any S.

This definition means, for instance, that $\neg p \Vdash ?p$ or that $(p \wedge q) \Vdash ?p$, but $p \not\Vdash ?(p \wedge q)$. In Figure 1 we have $\texttt{rain} \Vdash_S ?\texttt{umbr}$ for example, but $\neg\texttt{rain} \not\Vdash_S ?\texttt{umbr}$.

We show that this version of answerhood corresponds to the original intuition.

Proposition 2. $\varphi \Vdash ?\psi$ iff there is an $X \in W/[\![?\psi]\!]$ and $[\![\varphi]\!] \subseteq X$.
Proof sketch ($\Rightarrow$) Let $\varphi \Vdash ?\psi$, so $S[\varphi][?\psi] = S[\varphi]$, for any S. Let $S_0 = \langle W, W \times W\rangle$. By Def 6 $S_0[\varphi] = \langle [\![\varphi]\!], [\![\varphi]\!] \times [\![\varphi]\!]\rangle$. Adding $[\![?\psi]\!] = \{\langle w, v\rangle \mid M, w \models \psi \text{ iff } M, v \models \psi\}$ does not make any changes. So for all $w, v \in [\![\varphi]\!]$ either both $w, v \in [\![\psi]\!]$ or $w, v \notin [\![\psi]\!]$. In the first case $X = [\![\psi]\!]$, in the second $X = [\![\neg\psi]\!]$.
($\Leftarrow$) Let $[\![\varphi]\!] \subseteq X$, for some $X \in W/[\![?\psi]\!]$. Because ψ is propositional, there are two candidates: $X = [\![\psi]\!]$ or $X = [\![\neg\psi]\!]$. In the first case, $\varphi \models \psi$ and in the second case $\varphi \models \neg\psi$. In both cases, by Def 8 $\varphi \Vdash ?\psi$.

Defined in this way, answerhood is a rather weak constraint on responses. Inconsistencies and tautologies also qualify as answers. Therefore Groenendijk [10] defines some pragmatic constraints on answerhood, inspired by Grice. In particular, answers should be consistent, informative, and relevant to some question, and should not be over-informative. This means that answers may only eliminate complete equivalence classes from the partition induced by the question.

This logic of questions, and in particular the notion of answerhood, has been given an axiomatisation [5]. Intuitively, this works by a syntactic characterisation of the answerhood relation, based on Beth's definability theorem. We return to this in Section 5.

3.1 Questions and Knowledge

Issues can be embedded under the verb 'to know'. In linguistics, it turns out that 'to know' is one of a class of verbs that express epistemic states that answer some implicit question or resolve some open issue: discover, tell, guess, etc. These verbs are called *resolutive* by Ginzburg [9], in contrast to other verbs, like wonder, doubt or consider, which can also embed a question, but do not require resolution. Formula $\mathsf{K}_i?\varphi$ is a proposition, in L, which expresses that agent i knows an answer to question $?\varphi$.

Definition 9 (Syntax). Extend L in Definition 1 with $\{K_i?\varphi \mid ?\varphi \in L'\}$.

The definition of answerhood can be adapted quite straightforwardly to epistemic logic. See [8] for a different suggestion of how to do this.

Definition 10 (Semantics). Extend Definition 2 with the following clause:
$M, w \models \mathsf{K}_i?\psi$ iff there is φ such that $\varphi \Vdash ?\psi$ and $M, w \models \mathsf{K}_i\varphi$.

This definition is a bit unsatisfactory. It would be nicer to have a characterisation of $\mathsf{K}_i?\varphi$ directly in terms of $\mathcal{R}_i$. And indeed, we can prove the following.

Proposition 3 (Knowledge-wh). $M, w \models \mathsf{K}_i?\psi$ iff $R_i(w) \subseteq X$ for an $X \in W/[\![?\psi]\!]$.
Proof. Suppose $M, w \models \mathsf{K}_i?\psi$. By Def 10 $M, w \models \mathsf{K}_i\varphi$ for some φ. So by Def 2 $M, v \models \varphi$ for all $v \in \mathcal{R}_i(w)$, so by Def 4 $R_i(w) \subseteq [\![\varphi]\!]$. Moreover $\varphi \Vdash ?\psi$, so $[\![\varphi]\!] \subseteq X$ for some $X \in W/[\![?\psi]\!]$ (Prop 8). By transitivity of $\subseteq$ we have $R_i(w) \subseteq X$.

We can derive a number of properties to characterise $\mathsf{K}_i?\varphi$. Proposition 4 states that knowledge-that implies knowledge-wh. The converse does not hold, but if the answer happens to be true, we can get the converse (Proposition 6). Moreover, we can reduce knowledge-wh to a disjunction of knowledge-that statements (Proposition 5).

Proposition 4. $\models \mathsf{K}_i\varphi \rightarrow \mathsf{K}_i?\varphi, \quad \models \mathsf{K}_i\neg\varphi \rightarrow \mathsf{K}_i?\varphi$
Proof. Suppose $M, w \models \mathsf{K}_i\neg\varphi$. Then by Def 4 $R_i(w) \subseteq [\![\neg\varphi]\!]$. Take $X = [\![\neg\varphi]\!]$. Verify that $X \in W/[\![?\varphi]\!]$ and apply Prop 3. Similar for $M, w \models \mathsf{K}_i\varphi$.

Proposition 5. $\models \mathsf{K}_i?\varphi \leftrightarrow (\mathsf{K}_i\varphi \vee \mathsf{K}_i\neg\varphi)$.
Proof. ($\Rightarrow$) For question $?\varphi$, there are two possible answers: φ and $\neg\varphi$. By proposition 3 either $\mathcal{R}_i(w) \subseteq [\![\varphi]\!]$ or $\mathcal{R}_i(w) \subseteq [\![\neg\varphi]\!]$. Apply Def 2.
($\Leftarrow$) Suppose $M, w \models (\mathsf{K}_i\varphi \vee \mathsf{K}_i\neg\varphi)$. So either $M, w \models \mathsf{K}_i\varphi$ or $M, w \models \mathsf{K}_i\neg\varphi$. In both cases by Prop 4 $M, w \models \mathsf{K}_i?\varphi$.

Proposition 6. $\models \varphi \wedge \mathsf{K}_i?\varphi \leftrightarrow \mathsf{K}_i\varphi, \quad \models \neg\varphi \wedge \mathsf{K}_i?\varphi \leftrightarrow \mathsf{K}_i\neg\varphi$.
Proof. ($\Rightarrow$) Let $\models \varphi \wedge \mathsf{K}_i?\varphi$. By Prop 5 $\models \varphi \wedge (\mathsf{K}_i\varphi \vee \neg\mathsf{K}_i\varphi)$, so by **T** and propositional logic $M, w \models \mathsf{K}_i\varphi$. ($\Leftarrow$) By Prop 4 and **T**.

Predicate logic versions of these properties also exist. In those cases, proposition φ and $\neg\varphi$ above, are replaced by some substitution expression $\varphi\{c/x\}$, to characterise one of the appropriate answers to the question $?x\varphi$. See also Section 5.

4 Analysis of the Paradox

To get back to the Åqvist case, we reconsider (1), but now we represent Jones' job description by $\mathsf{OK}_j?\texttt{robbed}$: Jones must know whether the bank is being robbed.

(5) 1. The bank is being robbed. $\texttt{robbed}$
2. It ought to be the case that Jones knows whether the bank is being robbed. $\mathsf{OK}_j?\texttt{robbed}$

One of the causes of the paradox is the truth axiom. But also to know-wh is veridical: to know-wh something is to know a true answer to the question. By proposition 6 we have that *if* the bank is being robbed, Jones knows this, and *if* the bank is not being robbed, Jones knows that. We can put this into premise 3.

3. If the bank is robbed Jones knows that the bank is robbed. $\texttt{robbed} \wedge \mathsf{K}_j?\texttt{robbed} \rightarrow \mathsf{K}_j\texttt{robbed}$
If the bank is not robbed Jones knows that it is not robbed. $\neg\texttt{robbed} \wedge \mathsf{K}_j?\texttt{robbed} \rightarrow \mathsf{K}_j\neg\texttt{robbed}$

Is this enough to re-generate the paradox? Note that this conditional veridicality is much weaker than the truth axiom. We really need premise 1. Lets try to find an analogue of premise 4. We recognise an implication. So by RM, we derive the following.

4. $\mathsf{O}(\texttt{robbed} \wedge \mathsf{K}_j?\texttt{robbed}) \rightarrow \mathsf{OK}_j\texttt{robbed}$
$\mathsf{O}(\neg\texttt{robbed} \wedge \mathsf{K}_j?\texttt{robbed}) \rightarrow \mathsf{O}(\mathsf{K}_j\neg\texttt{robbed})$

If we now try an analogous step to 5, it fails. We require $\mathsf{O}\texttt{robbed} \wedge \mathsf{OK}_j?\texttt{robbed}$, which begs the question. This shows that a similar derivation of the paradox is blocked. To get a proper proof that the paradox is blocked, we need a counter example.

Proposition 7. $\not\models \mathsf{O}?\mathsf{K}_i\varphi \rightarrow \mathsf{O}\varphi$.
To be shown for some M and w: $M, w \models \mathsf{OK}_j?\texttt{robbed}$, but $M, w \not\models \mathsf{O}\texttt{robbed}$.
Proof. The model in Figure 2 has worlds w, u, and v such that $V(w)(\texttt{robbed}) = 1$, $V(v)(\texttt{robbed}) = 1$ and $V(u)(\texttt{robbed}) = 0$, $\mathcal{R}_j(v, v)$, $\mathcal{R}_j(u, u)$, $\mathcal{I}(w, v)$, $\mathcal{I}(w, u)$, $\mathcal{I}(v, v)$, $\mathcal{I}(u, u)$. We verify that $\mathcal{R}_i$ is an equivalence relation, and that $\mathcal{I}$ is serial. We observe that $M, v \models \mathsf{K}_i\texttt{robbed}$ and $M, u \models \mathsf{K}_i\neg\texttt{robbed}$, so by Proposition 4 $M, v \models \mathsf{K}_i?\texttt{robbed}$ and $M, u \models \mathsf{K}_i?\texttt{robbed}$ and by Definition 2 we get indeed $M, w \models \mathsf{OK}_i?\texttt{robbed}$. We verify that, although $M, w \models \texttt{robbed}$, $M, w \not\models \mathsf{O}\texttt{robbed}$, because there is a world u such that $\mathcal{I}(w, u)$ but $M, u \not\models \texttt{robbed}$.

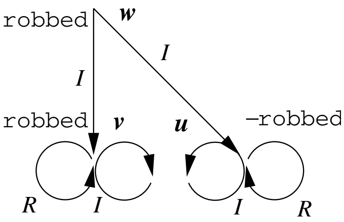


Fig. 2. Counter Example

Like many other deontic paradoxes, also this paradox has to do with contextual obligations, and how they should be represented. Many obligations are implicitly contextual. The general rule expressed in premise 3 can be rephrased as: in the context of $\texttt{robbed}$, Jones must know that $\texttt{robbed}$. Suppose we would have conditional obligation, like proposed for instance in [22], we could represent such a rule as follows:

(6) $\mathsf{O}(\mathsf{K}_j\texttt{robbed}|\texttt{robbed}) \wedge \mathsf{O}(\mathsf{K}_j\neg\texttt{robbed}|\neg\texttt{robbed})$

However, such a solution would mean that we would have to specify all possible answers that must be known, in advance. By contrast, if we use questions embedded under know, we can delegate finding the answers to the person who must know them.

5 Applications of Epistemic Obligations

The main point of this story, is to show the relevance of the following empirical observation: almost all epistemic obligations are most naturally formulated in terms of knowledge-wh. The reason is simple: the legislator can specify what kind of knowledge a person must have, but since this knowledge may concern the future, the legislator can not know what this knowledge is at specification time. In some cases, the legislator is himself not even allowed to know what the knowledge is, such as in the case of passwords. We have already seen some examples of that in the introduction.

To deal with such examples, the language can be extended to a version of epistemic predicate logic. For questions that involve words like who, what, where or when, we use formulas of the form $?x\varphi$, where variable x occurs free in formula φ. We also allow questions with multiple variables. "What is the capital of which country?" is represented by $?xy\texttt{capital}(x,y)$. When we ask about a closed formula, with no free variables, we get a 'whether-question' as before. So $?\texttt{rain}$ expresses the question whether it is raining or not, and "Does France have a capital?" is represented by $?\exists x\texttt{capital}(x,\texttt{France})$.

We use notation $\bar{x}$ for a sequence of zero or more variables, so $\bar{x} = x_1, ..., x_n$, where $n \geq 0$. When $n = 0$ the vector $\bar{x}$ is the empty sequence, and we get a whether-question.

Definition 11 (Syntax). Let P be a predicate symbol, $x, x_1, ..., x_n$ variables, and c a constant, then define terms T, propositions L_2 and questions L_2' as follows

$$\begin{array}{lll} T & t ::= & x \mid c \\ L_2 & \varphi ::= & t_1 = t_2 \mid P(x_1,...,x_n) \mid \neg\varphi \mid \varphi \wedge \psi \mid \forall x\varphi \mid K_t\varphi \mid K_t?\bar{x}\varphi \mid \mathsf{O}\varphi \\ L_2' & \chi ::= & ?\bar{x}\varphi \qquad\qquad\qquad\qquad\qquad\qquad\qquad (\varphi \in L_2; ?\bar{x}\varphi \in L_2') \end{array}$$

With this language we can express the motivational examples from the introduction. Example (9) is added to illustrate epistemic obligations for roles.

(7) Passwords ought to be known only by the owner.
$\mathsf{O}(\forall xy\mathsf{K}_x?w.\texttt{password\_of}(y,w) \rightarrow y = x))$

(8) Jones (the guard) ought to know which key fits on which door.
$\texttt{role}(\texttt{j},\texttt{guard}) \wedge \mathsf{OK}_\mathtt{j}(?kd.\texttt{key}(k) \wedge \texttt{door}(d) \rightarrow \texttt{fit}(k,d))$

(9) Generally, guards ought to know who is allowed to enter.
$\forall x(\texttt{role}(x,\texttt{guard}) \rightarrow \mathsf{OK}_x?y.\mathsf{Penter}(y))$

Epistemic operators and quantification make for a complex combination. To express the examples we need to be able to quantify over agents. So we require $A \subseteq D$ and allow variables as subscripts to modal operators. This gives rise to technical issues about epistemic predicate logic, which are outside the scope of this paper. For an introduction to modal predicate logic and some related philosophical problems, we refer to [7] and [6, Ch 3.7]. We assume one domain is shared by all worlds and constants are rigid designators: they denote the same individual in all possible worlds.

Definition 12 (Semantics L_2). Let M be a model as in Definition 2, and let D be a domain, shared by all possible worlds. Let $G = \{g, h, ...\}$ be a set of assignments from variables to entities in D. Let $g[x/d]$ the assignment that is just like g, except that it assigns d to x. V is re-interpreted as a function from constants and predicates to sets of tuples of entities from D, such that for all w, v we have $V(w)(c) = V(v)(c)$.
For all w and g the interpretation of terms is defined as follows:
$V_g(w)(t) = V(w)(t)$, if t is a constant
$V_g(w)(t) = g(t)$, if t is a variable.
The satisfaction relation '$\models$' is re-interpreted as follows:

$$\begin{array}{lll}
M, w, g \models P(t_1, ..., t_n) & \text{iff} & \langle V_g(w)(t_1), ..., V_g(w)(t_n)\rangle \in V_g(w)(P) \\
M, w, g \models x = t & \text{iff} & V_g(w)(x) = V_g(w)(t_1) \\
M, w, g \models \neg\varphi & \text{iff} & M, w, g \not\models \varphi \\
M, w, g \models \varphi \wedge \psi & \text{iff} & M, w, g \models \varphi \text{ and } M, w \models \psi \\
M, w, g \models \forall x\varphi & \text{iff} & M, w, g[x/d] \models \varphi, \text{ for all } d \in D \\
M, w, g \models \mathsf{K}_t\varphi & \text{iff} & M, v, g \models \varphi, \text{ for all } v \text{ such that } \mathcal{R}_{V_g(w)(t)}(w, v). \\
M, w, g \models \mathsf{K}_t?\bar{x}\varphi & \text{iff} & M, w, g \models \mathsf{K}_t\psi \text{ and } \psi \Vdash ?\bar{x}\varphi, \text{ for some formula } \psi \\
M, w, g \models \mathsf{O}\varphi & \text{iff} & M, v, g \models \varphi, \text{ for all } v \text{ such that } \mathcal{I}(w, v).
\end{array}$$

Also the interpretation of questions needs to be adapted.

Definition 13 (Interpretation). Let $\varphi \in L_2'$ be any formula with free variables $x_1, ..., x_n$. Now define the extension of a formula as follows:
$V_g(w)(\varphi) = \{\langle d_1, ..., d_n\rangle \mid M, w, g[x_1/d_1, ..., x_n/d_n] \models \varphi\}$
$[\![?\bar{x}\varphi]\!]_g = \{\langle w, v\rangle \models V_g(w)(\varphi) = V_g(v)(\varphi)\}$

Again, the meaning of a question is seen as an indistinguishability relation. Two worlds are connected, whenever they agree on their answer to the question, which is modelled here by an extension, a set of tuples of entities. For example, the question "Who will come to the party?", is represented by a formula $?x\texttt{come\_to\_party}(x)$. Given that John and Mary are the only feasible party goers in the context, the question would induce a partition consisting of the following answers: "No one will come", "John will come", "Mary will come" and "John and Mary will come".

The way the model has been set up, it is difficult to express information about 'values of variables', such as questions like "Who is he?", represented by $?x(x = y)$. After all, such information is stored in assignments g, rather than possible worlds. In dynamic logic, in particular [10], information states are defined over indices: pairs of worlds and assignments. Adding the information that 'he' is actually Jones, eliminates all assignments incompatible with that information.

The dynamic version of the logic of questions [10] has been given a sound and complete axiomatisation [5]. It is based on a syntactic characterisation of the answerhood relation. The main theorem is repeated here in simplified form, without proof.

Proposition 8 (Theorem 3.2 [5]). Entailment $\varphi_1, ..., \varphi_n \Vdash ?\psi$ holds iff the set of propositions $\{\varphi_1, .., \varphi_n\}$ implicitly defines an answer ψ in terms of $\{x = c \mid c \text{ a constant}\}$.

The theorem uses the notion 'implicitly defines', which has to do with Beth's definability theorem. Beth's Definability is the proposition that a set of formulas Σ *implicitly defines* ψ in terms of Γ iff there is a development χ of Γ with the same free variables $\bar{x}$ as ψ has, so that $\Sigma \models_{fol} \forall\bar{x}(\psi \leftrightarrow \chi)$. Here '$\models_{fol}$' stands for first order logic entailment. A development is essentially a syntactic reformulation of an answer. Formally, a *development* of a set of formulas Σ is defined as a formula that is built up from elements of Σ and formulas of the form $(x = y)$ using $\neg, \wedge, \rightarrow, \forall$.

Given a syntactic characterisation of answers, it is likely that we can prove predicate logic versions of Proposition 4, 5 and 6.

6 Related Research

To our knowledge, a combination of questions and knowledge with deontic logic is new, although the notion of 'knowledge-whether' has been around for some time.

Regarding the combination of epistemic logic and questions, we would like to mention Gerbrandy and Groeneveld [8]. They too take answerhood as the core. Balder ten Caste and Chung-Chieh Shan [5] have developed an axiomatisation of the answerhood relation, which also works for the predicate logic version. In a different tradition, Hart, Heifetz and Samet [13] demonstrate that it is much easier to trace dependencies between the knowledge of different agents, when you use a notion of knowledge-whether.

It is interesting that Åqvist himself also published about the semantics of questions [1]. In his theory, a question is interpreted as a request for knowledge: "Bring it about that I know whether...". This instrumental view would not solve the paradox.

A combination of knowledge and obligation is addressed in a recent paper by Pacuit, Parikh and Cogan [19]. They focus on the way in which the obligations to which an agent is subjected, depend on what the agent knows. Lomuscio and Sergot [16] also combine epistemic and deontic notions. They use interpreted systems to give a computationally feasible semantics of the knowledge that an agent is permitted to have, or the knowledge that an agent ideally must have, given that the system behaves according to some protocol. Among logics for expressing more practical problems regarding confidentiality and security, one of the best known ones is BAN logic [4].

How does our work compare to solutions of other paradoxes? The central point is that knowing an answer to a question depends on the context. This context dependency makes our solution similar to approaches which have also stressed the conditional nature of the obligation: "If the bank is being robbed, Jones ought to know that". Such a conditional solution also works for another paradox that is a victim of the RM rule: The

Good Samaritan paradox[1]. Interestingly, both paradoxes involve a presupposition. The verb 'to know' presupposes, rather than entails, its complement. In a similar way, a definite description like 'the man who is robbed' presupposes that there is a robbed man. Linguistic theory suggests that presupposition inferences do not 'escape' from a conditional [20]. "If a man is robbed, the robbed man ought to be assisted" does not presuppose that there is a robbed man. In those cases, the presupposition is said to be bound by the antecedent of the conditional.

7 Conclusions

In this paper we have discussed Åqvist's paradox of epistemic obligation, which results from a combination of Standard Deontic Logic axioms, and the truth axiom of knowledge. From the fact that someone is obliged to know some facts, we can derive that those facts are also obliged. The paradox can be solved, when epistemic obligations, expressions of some 'need to know', are represented by obligations over knowledge of the answer to a question. Such knowledge is crucially dependent on the context, so the paradox can no longer be derived. In general, descriptions of what someone ought to know, should be specified by an embedded question, because the system designer may not have the required knowledge available at design time.

Why bother with the O operator at all? Why not simply specify the knowledge required by agents in epistemic logic, if needed extended with questions? The usual reply to such remarks is that deontic logic allows one to reason explicitly about violations. For some applications, especially those dealing with the human aspect of confidentiality, we prefer to see a broken secret as a violation, rather than a system failure.

In future research we plan to further investigate the 'interaction axioms', like **O4** and **O5**. We also need to look more at the details of quantification in epistemic predicate logic. In particular, the use of quantifiers to express properties of agents, is useful. The high expressivity of our logic comes at a price however. The nice computational properties of modal logic are lost, when the complexity of quantification is added. Remember however that our language is primarily used as a way to develop conceptual models of access control; not necessarily to prove formal properties.

Before you start thinking that our logic is rather farfetched, and far removed from practical applications, please consider the similarity between the logic of questions and database query languages like SQL. Moreover, there is a similarity between the semantics of relational databases, and possible worlds semantics. A set of possible worlds – a proposition – can be compared to a set of tuples in the Carthesian product of all tables (relations)[23]. A query specifies which sets of tuples would count as an appropriate answer. Which answers are true, depends on the current state of the database. So a query specifies a set of sets of tuples: a partition. So the main message of the paper is that one should use query languages inside the specification of access control policies. Further practical research will have to point out wether this idea is indeed being used in practice, and how such query-based access control rules would look like.

[1] "The man who is robbed ought to be assisted" implies "There ought to be a robbed man", because of RM and the fact that "the man who is robbed" implies that a man is robbed.

References

[1] Åqvist, L.: A new approach to the logical theory of interrogatives. Almqvist and Wiksell, Uppsala (1965)

[2] Åqvist, L.: Good samaritans, contrary-to-duty imperatives, and epistemic obligations. Noûs 1(4), 361–379 (1967)

[3] Blaze, M., Feigenbaum, J., Lacy, J.: Decentralized trust management. In: IEEE Symposium on Security and Privacy, pp. 164–173 (1996)

[4] Burrows, M., Abadi, M., Needham, R.: A logic of authentication. In: Practical Cryptography for Data Internetworks, pp. 1871–1989. IEEE, Los Alamitos (1996)

[5] ten Cate, B., chieh Shan, C.: Axiomatizing groenendijk's logic of interrogation. In: Questions in dynamic semantics, pp. 63–82. Elsevier, Amsterdam (2007)

[6] Fagin, R., Halpern, J.Y., Moses, Y., Vardi, M.: Reasoning about Knowledge. MIT Press, Cambridge (1995)

[7] Gamut, L.: Logic, Language and Meaning. Intensional Logic and Logical Grammar, vol. II. University of Chicago Presss, Chicago (1991)

[8] Gerbrandy, J., Groeneveld, W.: Reasoning about information change. Journal of logic, language and information 6(2), 147–170 (1997)

[9] Ginzburg, J.: Resolving questionsI. Linguistics and Philosophy 18, 459–527 (1995)

[10] Groenendijk, J.: The logic of interrogation: classical version. In: Matthews, T., Strolovitch, D. (eds.) Proceedings of SALT-9. CLC Publications (1999)

[11] Groenendijk, J., Stokhof, M.: Studies on the Semantics of Questions and the Pragmatics of Answers. PhD thesis, University of Amsterdam (1984)

[12] Groenendijk, J., Stokhof, M.: Questions. In: van Benthem, J., ter Meulen, A. (eds.) Handbook of Logic and Language, pp. 1055–1124. Elsevier, Amsterdam (1996)

[13] Hart, S., Heifetz, A., Samet, D.: knowing whether, knowing that and the cardinality of state spaces. Journal of Economic Theory 70(1), 249–256 (1996)

[14] Hintikka, J.: Knowledge and Belief. Cornell University Press, Ithaca (1962)

[15] van der Hoek, W.: Systems for knowledge and belief. Journal of Logic and Computation 3(1), 173–193 (1993)

[16] Lomuscio, A., Sergot, M.: Deontic interpreted systems. Studia Logica 75, 63–92 (2003)

[17] McNamara, P.: Deontic logic. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy, CSLI , Stanford University (Spring 2006)

[18] Meyer, J.-J.C., Wieringa, R.J. (eds.): Deontic Logic in Computer Science: Normative System Specification. Wiley and Sons, Chichester (1993)

[19] Pacuit, E., Parikh, R., Cogan, E.: The logic of knowledge based obligation. Synthese 149(2), 311–341 (2006)

[20] van der Sandt, R.: Presupposition projection as anaphora resolution. Journal of Semantics 9, 333–377 (1992)

[21] Tomberlin, J.E.: Contrary-to-duty imperatives and conditional obligation. Noûs 15(3), 357–375 (1981)

[22] van der Torre, L., Tan, Y.-H.: Contrary-to-duty reasoning with preference-based dyadic obligations. Annals of Math. and Artificial Intelligence 27, 49–78 (1999)

[23] Ullman, J.: Principles of Database Systems. Computer Science Press (1982)

[24] Veltman, F.: Defaults in update semantics. Journal of Philosophical Logic 25(3), 221–262 (1996)

A Logical Analysis of the Interaction between 'Obligation-to-do' and 'Knowingly Doing'

Jan Broersen

Department of Information and Computing Sciences, Utrecht University,
The Netherlands

Abstract. Within a STIT framework, this paper presents a logical study of the interaction between 'ought-to-do', and an epistemic notion of 'knowingly doing'. We start out with some motivating examples concerning the interaction between action, obligation and knowledge. Then we present a complete temporal STIT logic including operators for action, obligation and knowledge. We use the logic to analyze the examples and discuss open problems.

1 Introduction

This paper investigates some new problems that are introduced by considering epistemic modalities within a deontic STIT logic. In particular we will focus on epistemic modalities that are directly applied to (STIT) actions. The resulting concept is one of 'knowingly doing', or maybe even 'consciously doing'. Several new and fascinating question arise by introducing this notion. For instance, the concept presupposes that things can also be done 'unknowingly' or 'unawarely'. And what are the implications for the conditions under which obligations are violated? For instance, are violations also avoided if agents comply with an obligation unknowingly? And, are violations resulting from unknowingly performed actions excusable? We further elaborate on some of these questions by first considering some motivating examples.

2 Examples and Questions

The first example is one designed by John Horty [1]. Consider the following three scenarios:

> The first scenario is as follows. There are two agents, α and β. β puts money in box 1 and leaves box 2 empty. α, not having seen in which box the money has been put, has to make a choice between the two boxes. It knows β has put the money in one of the boxes, but it does not know that it is box 1. The right thing to do for the agent is then to pick box 1. So, α knows that one of its choices is the right thing to do. α might take a gamble, but it cannot do the right thing *knowingly*, since it does not know in which box the money is.

R. van der Meyden and L. van der Torre (Eds.): DEON 2008, LNAI 5076, pp. 140–154, 2008.

> In the second scenario, α first chooses between one of the two boxes. After α has chosen, β, who does not know which box is selected by α, puts the money in one of the two boxes. Now both choices for α seem equally good, since we assume there is no communication between the agents, which means that α cannot know beforehand in which box β is going to put the money.
>
> The third scenario is where both agents choose simultaneously. From the perspective of agent α this is equivalent to the second scenario. It cannot know what β chooses, so by no means there is a right thing to do for the agent.

Question 1: The first question is whether or not, in the first scenario, the agent has an obligation to pick the right box or not. On the one hand indeed there 'exists' a right thing to do. But on the other hand, the agent will not be able to (knowingly) guarantee that its action will comply with the obligation.

Question 2: For the second scenario, we are much less inclined to say that the agent is obliged to put the money in the right box. But why exactly is that? After all the only difference between both scenario's is the order of choosing.

A second example is about the issue of violation. The question is whether or not an agent violates an obligation when it performs the obliged action *unknowingly*.

> Imagine a (doctor) agent is obliged to see to it that a patient is cured. The agent injects the patient with a drug the agent believes cures the patient. However, the agent is mistaken, because the drug it believes to inject, in fact does not cure the patient. But then, coincidentally, without the doctor agent knowing it, it turns out that another drug is in the hypodermic, and this drug is indeed the right drug to cure the patient. So, the doctor *unknowingly* cures the patient.

Question 3: Now, does the (doctor) agent violate his obligation to cure the patient or not?

Question 4: What concepts do we need to formalize, to express the differences between the three scenario's in the boxes example and to explain the hesitation in the doctor's example? Can we give such a formalization?

In the next section we start by defining a logic that we think answers this final question affirmatively. Then, in section 6 we will come back to the other questions and examples of the present section, using the insights obtained from defining the formal logic.

3 A Temporal Epistemic STIT Logic

In this section we define a complete STIT logic with operators for knowledge. The present STIT logic is a little different from the one in [2] since it encompasses an axiom for uniformity of strategies. With respect to the predecessor STIT logics in [3] and [4], the differences are bigger. In particular, the present logic drops some of the axioms in [4], and adds new ones, most prominently some for the interactions of action and knowledge. Also we define a two dimensional semantics, closer to the STIT semantics in the philosophical literature. Furthermore, we obtain completeness for several intuitive properties concerning the interaction of time, action and knowledge.

One of the distinguishing features of the present STIT logic is that actions only take effect in 'next' states, where 'next' refers to immediate successors of the present state. This distinguishes the present STIT logic not only from the STIT variants in the above mentioned papers, but also from any STIT-logic in the (philosophical) literature. However, there are very good reasons for taking this approach. The first reason is that it can be shown (see [5]) that the logics of the multi-agent versions of, what we might call, the standard 'instantaneous' STIT, are undecidable. The second reason is that the view that actions only take effect in some immediate next state, is the standard view in formal models of computation in computer science. And finally, also from a philosophical perspective, the choice can be advocated. Given that acting always seems associated with some effort or process, and given that these take time, we may conclude that actions take place 'in' time.

For the extension of this framework with an operator for 'ought-to-do', we adapt the approach taken by Bartha [6] who introduces Anderson style ([7]) violation constants in STIT theory. The approach with violation constants is very well suited for theories of ought-to-do, witness the many logics based on adding violation constants to dynamic logic [8,9]. However, we believe that the STIT setting is even more amenable to this approach. Some evidence for this is found in Bartha's article ([6]), which shows that many deontic logic puzzles (paradoxes) are representable in an intuitive way. And a clear advantage in the present approach is that since our base STIT logic is complete, defining obligation as a reduction using violation constants guarantees that completeness is preserved under addition of the obligation operator.

Besides the usual propositional connectives, the syntax of the logic comprises an operator $K_a\varphi$ for knowledge of individual agents a, an operator $\Box\varphi$ for historical necessity, which plays the same role as the well-known path quantifiers in logics such as CTL and CTL^* [10], an operator $[A \text{ xstit}]\varphi$ for 'agents A jointly see to it that φ in the next state', and finally, a violation constant V denoting that a violation occurs. Given a countable set of propositions P, the violation constant V, and a finite set Ags of agent names, formally the language can be described as:

Definition 1. *Given a propositional constant V and a countable set of propositions P and $p \in P$, and given a finite set Ags of agent names, and $a \in Ags$ and $A \subseteq Ags$, formally the language can be described as:*

$$\varphi, \psi, \ldots := p \mid V \mid \neg\varphi \mid \varphi \wedge \psi \mid K_a\varphi \mid \Box\varphi \mid [A \; \mathit{xstit}]\varphi$$

We define operators for 'next' $X\varphi$, and several operators for obligation as abbreviations in the language. In this section we only give the definition for the 'next' because the explanation of the definition of the obligation operators can better be done after the formal semantics of the base operators is given. We define the ‘next’ operator as the current action performed by the complete set of agents Ags:

Definition 2

$$X\varphi \equiv_{def} [Ags\ \textsf{\textit{xstit}}]\varphi$$

The view that the complete set of agents uniquely determines the next state is a common one. Not only it can be found in the multi-agent STIT logics in the philosophical literature [11], but also in related computer science formalisms such as ATL [12,13]. For the relation between STIT formalisms and ATL and Coalition Logic [14], see [15,16]. In the description of the models below, we will actually use terminology from Coalition Logic, and call the relations interpreting the STIT operator ‘effectivity’ relations.

Definition 3. *A frame is a tuple $\mathcal{F} = \langle H, S, R_\Box, \{R_A \mid A \subseteq Ags\}, \{\sim_a \mid a \in Ags\}\rangle$ such that:*

- *H is a non-empty set of histories. Elements of H are denoted h, h', etc.*
- *S is a non-empty set of states. Elements of S are denoted s, s', etc.*
- *$R_\Box$ is a ‘historical necessity’ relation over the elements of $H \times S$ such that $\langle h, s\rangle R_\Box \langle h', s'\rangle$ if and only if $s = s'$*
- *The R_A are ‘effectivity’ relations over the elements of $H \times S$ such that:*
  - *R_{Ags} is a ‘next time’ relation such that if $\langle h, s\rangle R_{Ags} \langle h', s'\rangle$ then $h = h'$, and R_{Ags} is serial and deterministic (the next state is completely determined by the choice made by the complete set of agents). So, histories ‘contain’ linearly ordered sets of states.*
  - *$R_\Box \circ R_{Ags} \subseteq R_\emptyset$ (the empty set of agents is ineffective)*
  - *$R_A \subseteq R_\Box \circ R_{Ags}$ for any A (an action undertaken by A in the present state ensures the next state is element of a specific subset of all possible next states)*
  - *$R_{Ags} \circ R_\Box \subseteq R_A$ for any A (no actions constitute a choice between histories that are undivided in next states)*
  - *$R_A \subseteq R_B$ for $B \subset A$ (super-groups are at least as effective)*
  - *if $\langle h, s\rangle (R_\Box \circ R_A) \langle h', s'\rangle$ and $\langle h, s\rangle (R_\Box \circ R_B) \langle h'', s''\rangle$ and $A \cap B = \emptyset$ then there is a $\langle h, s\rangle R_\Box \langle h''', s\rangle$ such that both $\langle h''', s\rangle R_A \langle h', s'\rangle$ and $\langle h''', s\rangle R_B$ $\langle h'', s''\rangle$ (independence of agency)*
- *The $\sim_a$ are epistemic equivalence relations over the elements of $H \times S$ such that:*
  - *$\sim_a \circ R_a \subseteq \sim_a \circ R_{Ags}$ (agents cannot know what choices other agents perform concurrently)*
  - *$R_{Ags} \circ \sim_a \subseteq \sim_a \circ R_a$ (agents recall the effects of the actions they knowingly perform themselves)*
  - *if $\langle h, s\rangle R_\Box \langle h', s'\rangle$ and $\langle h, s\rangle \sim_a \langle h'', s''\rangle$ then there is a $\langle h''', s'''\rangle$ for which $\langle h', s'\rangle R_\Box \langle h''', s'''\rangle$ and if $\langle h''', s'''\rangle R_a \langle h'''', s''''\rangle$ then $\langle h', s'\rangle R_a$ $\langle h'''', s''''\rangle$ (uniformity of conformant action)*

Definition 4. *A frame* $\mathcal{F} = \langle H, S, R_{\Box}, \{R_A \mid A \subseteq Ags\}, \{\sim_a \mid a \in Ags\}\rangle$ *is extended to a model* $\mathcal{M} = \langle H, S, R_{\Box}, \{R_A \mid A \subseteq Ags\}, \{\sim_a \mid a \in Ags\}, \pi\rangle$ *by adding a valuation* π *of atomic propositions:*

- π *is a valuation function* $\pi : P \longrightarrow 2^{H \times S}$ *assigning to each atomic proposition the set of history/state pairs in which they are true.*

The truth conditions for the semantics of the operators on these models is standard for a two-dimensional modal logic [17].

Definition 5. *Validity* $\mathcal{M}, \langle h, s\rangle \models \varphi$*, of a formula* φ *in a history/state pair* $\langle h, s\rangle$ *of a model* $\mathcal{M} = \langle H, S, R_{\Box}, \{R_A \mid A \subseteq Ags\}, \{\sim_a \mid a \in Ags\}, \pi\rangle$ *is defined as:*

$$\begin{array}{ll}
\mathcal{M}, \langle h, s\rangle \models p & \Leftrightarrow \langle h, s\rangle \in \pi(p) \\
\mathcal{M}, \langle h, s\rangle \models \neg\varphi & \Leftrightarrow \textit{not } \mathcal{M}, \langle h, s\rangle \models \varphi \\
\mathcal{M}, \langle h, s\rangle \models \varphi \wedge \psi & \Leftrightarrow \mathcal{M}, \langle h, s\rangle \models \varphi \textit{ and } \mathcal{M}, \langle h, s\rangle \models \psi \\
\mathcal{M}, \langle h, s\rangle \models K_a\varphi & \Leftrightarrow \langle h, s\rangle \sim_a \langle h', s'\rangle \textit{ implies that } \mathcal{M}, \langle h', s'\rangle \models \varphi \\
\mathcal{M}, \langle h, s\rangle \models \Box\varphi & \Leftrightarrow \langle h, s\rangle R_{\Box} \langle h', s'\rangle \textit{ implies that } \mathcal{M}, \langle h', s'\rangle \models \varphi \\
\mathcal{M}, \langle h, s\rangle \models [A\ \mathsf{xstit}]\varphi & \Leftrightarrow \langle h, s\rangle R_A \langle h', s'\rangle \textit{ implies that } \mathcal{M}, \langle h', s'\rangle \models \varphi
\end{array}$$

Satisfiability, validity on a frame and general validity are defined as usual.

While the semantics is very standard from a (two-dimensional) modal logic perspective, the relation with standard STIT semantics deserves some explanation. In the conditions on the frames we recognize standard STIT properties like 'no choice between undivided histories' and properties that are specific for the present STIT version, like 'actions take effect in successor states'. Actually, the frames can easily be pictured as trees where histories branch in states, like in standard STIT theory. The main difference is that *states* are not partitioned into choice sets. The choice sets appear here (implicitly) as sets of possible *next* states (like in Coalition Logic). From a given 'actual' history/state pair, we reach these choice sets by first jumping (along $R_{\Box}$) to another history through the same state, and then looking (along R_A) what next states are reachable through the choice made by agents on that history.

One aspect of the present semantics needs extra clarification. Like in standard STIT semantics, history/state pairs for the same state can have different valuations of atomic propositions. In standard STIT formalisms this is actually needed to give semantics to the instantaneous effects of actions. But here, as said, the effects are not instantaneous. Therefore, in the present logic, the fact that different histories through the same state can have different valuations of non-temporal propositions, does not carry much meaning. Of course, in the logic we can talk about atomic propositions being true or not in other histories through the same state. For instance, the formula "$\Box p$" expresses that all the histories through the present state have in common that the atomic proposition p holds on them. But the point is that one might think that actually we should *impose* on the models that all histories through a state come with identical valuations of atomic propositions. That would induce the property $\varphi \rightarrow \Box\varphi$ for

φ any 'STIT-operator-free' formula (in [4] a system involving such an axiom is given). However, this would complicate establishing a completeness result, and does not strengthen the logic in any essential or interesting way. We think there is no need at all to impose such a condition. Since actions only take effect in next states, alternative valuations for *atomic* propositions on other histories through the same state are just not relevant for the semantics of the STIT fragment of our logic.

Now we go on to the axiomatization of the logic. Actually, axiomatization is fairly easy. The approach we have taken for constructing this logic is to build up the semantic conditions on frames and the corresponding axiom schemes simultaneously, while staying within the Sahlqvist class. This ensures that the semantics cannot give rise to more logical principles than can be proven from the axiomatization.

Definition 6. *The following axiom schemas, in combination with a standard axiomatization for propositional logic, and the standard rules (like necessitation) for the normal modal operators, define a Hilbert system:*

	S5 for $\Box$
	KD for each $[A\ \mathsf{xstit}]$
(C-Mon)	$[A\ \mathsf{xstit}]\varphi \rightarrow [A \cup B\ \mathsf{xstit}]\varphi$
(Indep)	$\Diamond[A\ \mathsf{xstit}]\varphi \wedge \Diamond[B\ \mathsf{xstit}]\psi \rightarrow \Diamond([A\ \mathsf{xstit}]\varphi \wedge [B\ \mathsf{xstit}]\psi)$ *for* $A \cap B = \emptyset$
(Det)	$\neg X \neg\varphi \rightarrow X\varphi$
(Ineff-$\emptyset$)	$[\emptyset\ \mathsf{xstit}]\varphi \rightarrow \Box X\varphi$
(X-Eff)	$\Box X\varphi \rightarrow [A\ \mathsf{xstit}]\varphi$
(n-c-u-h)	$[A\ \mathsf{xstit}]\varphi \rightarrow X\Box\varphi$
	S5 for each K_a
(Know-X)	$K_a X\varphi \rightarrow K_a[a\ \mathsf{xstit}]\varphi$
(Rec-Eff)	$K_a[a\ \mathsf{xstit}]\varphi \rightarrow XK_a\varphi$
(Unif-Str)	$\Diamond K_a[a\ \mathsf{xstit}]\varphi \rightarrow K_a\Diamond[a\ \mathsf{xstit}]\varphi$

Theorem 1. *The Hilbert system of definition 6 is complete with respect to the semantics of definition 5.*

Proof. The axioms correspond one-to-one to the semantic conditions defined on the frames (proofs omitted[1]). Also the axioms are all within the Sahlqvist class. This means that the axioms are all expressible as first-order conditions on frames and that they are complete with respect to the defined frame classes, cf. [18, Th. 2.42].

As part of the above axiomatization, we recognize Ming Xu's axiomatization for multi-agent STIT logics (see the article in [19]). Xu's axiomatization is for the standard, instantaneous STIT variant. But, it should not come as a surprise that the same axioms apply to the present logic. The central property in Xu's

[1] Thanks to Heleen Booy for finding the correspondence of the uniform strategy axiom using the Sahlqvist algorithm.

axiomatization is the 'independence of agency' property. But the issue of independence of choices of different agents does not depend on the condition that effects are instantaneous or occur in next states.

As a proposition we list some theorems. Derivation of these is just a little exercise in normal modal logic. The last theorem in the list below is the well known 'perfect recall' or 'no forgetting' axiom, known from the literature on the interaction between epistemic and temporal modalities.

Proposition 1. *The following are derivable:*

$$\begin{array}{l}
[A \textit{ xstit}]\varphi \land [B \textit{ xstit}]\psi \to [A \cup B \textit{ xstit}](\varphi \land \psi) \\
\Box X\varphi \to X\Box\varphi \\
[A \textit{ xstit}]\varphi \to X\varphi \\
X\neg\varphi \to \neg X\varphi \\
\Box X\varphi \leftrightarrow [\emptyset \textit{ xstit}]\varphi \\
K_a X\varphi \leftrightarrow K_a[a \textit{ xstit}]\varphi \\
K_a X\varphi \to X K_a\varphi
\end{array}$$

Pauly's Coalition logic [14] is a logic of ability that is very closely related to STIT formalisms. In particular, in [15] it is shown that Coalition Logic can be embedded in STIT logic. Since in Coalition Logic actions also take effect in next states, restricting the STIT formalism by only allowing effects in next state, as in the logic of this paper, does not inhibit definability of Coalition Logic. See [2] for some more details.

Finally a word on the 'deliberative' STIT. The kind of STIT operator we defined above has often been criticized for properties like $[A \text{ xstit}]\top$. The idea is that agents should not be able to bring about things that are true inevitably, but only things that without their intervention might not become true. If we want an operator that takes this into account we can easily define a deliberative version of the STIT operator, as follows:

$$[A \text{ dxstit}]\varphi \equiv_{def} [A \text{ xstit}]\varphi \land \neg\Box X\varphi$$

An interesting question is how deliberateness of actions relates to the concept of 'knowingly doing'. We leave this aside as an opportunity for future research.

4 The Concept of 'Knowingly Doing'

Because the notion is central to the present paper, in this section we elaborate on the notion of 'knowingly doing'. We explain what it means to do something *(un)knowingly*. In the previous section we gave a semantics in terms of models with epistemic equivalence sets (information sets) containing history/state pairs. An agent knowingly does something if its action 'holds' for all the history/state pairs in the epistemic equivalence set containing the *actual* history/state pair.

Several closure conditions apply. The first one says that epistemic equivalence sets are closed under choices[2]. The corresponding axiom, is $K_a X\varphi \to$

[2] An extreme case is where the information sets are exactly the choices in each state. In that case an agent knows all the consequences of his actions.

$K_a[a\ \mathsf{xstit}]\varphi$ (this property does not hold if the STIT operator is replaced by a *deliberative* STIT oparator). This property ensures that an agent cannot know that two histories belonging to the same choice are different, or, in other words, for any agent the histories within its own choices are indistinguishable. This means that agents cannot knowingly do *more* then what is affected by the choices they have. In particular, the property $K_aX\varphi \rightarrow K_a[a\ \mathsf{xstit}]\varphi$ says that agents can only know things about the (immediate) future if they are the result of an action they themselves knowingly perform. Then, an agent *unknowingly* does everything that is (1) true for all the history/state pairs belonging to the actual *choice* it makes in the actual *state*, but (2) not true for all the history/state pairs it considers possible. In general the things an agent does unknowingly vastly outnumber the things an agent *knows* it does. For instance, by sending an email, we may enforce many, many things we are not aware of, which are nevertheless the result of me sending the email. All these things we do *unknowingly* by knowingly sending the email.

Another, equivalent way of interpreting the property $K_aX\varphi \rightarrow K_a[a\ \mathsf{xstit}]\varphi$ is to say that it expresses that agents cannot know what actions other agents perform concurrently. This is because the independence property (Indep) guarantees that choices of other agents always refine the choices of the agent we consider. Then, knowing the choice of the other would mean that the agent would be able to know more about the future state of affairs then is guaranteed by his own action.

A second constraint on the interaction between knowledge and action is the one expressed by the axiom $K_a[a\ \mathsf{xstit}]\varphi \rightarrow XK_a\varphi$. The issue here is that if agents knowingly see to it that a condition holds in the next state, in that same next state they will recall that the condition holds.

Finally, we discuss the interaction property $\Diamond K_a[a\ \mathsf{xstit}]\varphi \rightarrow K_a\Diamond[a\ \mathsf{xstit}]\varphi$. It says that if an agent can knowingly see to it that φ, then it knows that among its repertoire of choices there is one ensuring φ. This property is the STIT version of the constraint concerning ‘uniform strategies’ game theorists talk about. In game theory, *uniform* strategies require that agents have the same choices in all states within information sets. Since in game theory the choices are given names, a constraint is formulated saying that each state within the information set should have choices of the same type (that is, choices with the same name). In the present STIT setting, we do not have names. But the intuition that the same choices should be possible in different states of an information set, still applies. The property $\Diamond K_a[a\ \mathsf{xstit}]\varphi \rightarrow K_a\Diamond[a\ \mathsf{xstit}]\varphi$ exactly captures this intuition. It says that if an agent can knowingly see to it that φ, then at least one of its choices in the states it considers possible actually ensures φ (that is, a φ-action is possible in all states of the information set). The axiom can thus be said to express that ‘true ability’ obeys the property of uniformity of strategies.

In section 3 we already mentioned that recent computer science formalisms like Alternating Time Temporal Logic (ATL) [12], its epistemic extension ATEL [20], and Coalition Logic (CL) [14] are closely related to STIT formalisms. That STIT is the right formalism to solve the many conceptual problems raised for AT(E)L is clearly demonstrated by the properties on 'knowingly doing' we discuss in this section. Actually, the STIT operators enable us to axiomatize properties that also play a prominent role in the discussions surrounding ATEL and its derivatives, but that hitherto have not been characterized in these logics. The first example is the property that agents cannot know what actions are concurrently performed by other agents. This is also a basic assumption in ATL, ATEL, etc. However, in these logics, this assumption does not correspond directly with a property of the logic. In the present logic, it corresponds with the axiom $K_aX\varphi \rightarrow K_a[a\ \mathsf{xstit}]\varphi$. The second witness is the property for uniformity of strategies. In ATEL it is impossible to give an axiom that corresponds with this constraint on the models. This is clear right away, since in ATEL models the uniformity property is defined in terms of the names given to actions, while there are no explicit actions in the object language. This means that dropping the uniformity constraints for ATEL models does not change the *logic* ATEL (which prompts the question why in the semantics of this logic the uniformity assumption is made in the first place). The same holds for all extensions and adaptations of ATEL meant to solve the problem without making the actions explicit in the object language (e.g. [21,22]). Finally, also in formalisms that do have the actions symbolized in the object language under an assumption of uniformity of strategies, like [23], logical properties resulting from this uniformity, let alone an axiom *characterizing* it, are lacking.

5 Defining Deontic Modalities

To define an operator for 'obligation to do', we adapt the approach of Bartha [6] to the present situation where actions only take effect in next states. The intuition behind the definition is straightforward: an agent is obliged to do something if and only if by not performing the obliged action, it performs a violation. As said, the difference with Bartha's definition is that the effect of the obliged action can only be felt in next states, which is why also violations have to be properties of next states. Formally, our definition is given by:

$$O[a\ \mathsf{xstit}]\varphi \equiv_{def} \Box(\neg[a\ \mathsf{xstit}]\varphi \rightarrow [a\ \mathsf{xstit}]V)$$

First note that we slightly abuse notation by denoting $[\{a\}\ \mathsf{xstit}]\varphi$ as $[a\ \mathsf{xstit}]\varphi$. Also note that $\neg[a\ \mathsf{xstit}]\varphi$ expresses that A do not see to it that φ, which is the same as saying that A 'allow' a choice for which $\neg\varphi$ is a possible outcome. The definition then says that all such choices *do* guarantee that a violation occurs.

The $\Box$ operator in the definition ensures that obligations are 'moment determinate'. This means that their validity only depends on the state, and not on the history (see [11] for a further explanation of this concept). We think that this is correct. But see [24] for an opposite opinion.

The above defined obligation is a 'personal' one. If, by 'coincidence', φ occurs, apparently due the action of other agents, while the agent bearing the obligation did not make a choice that *ensured* that φ would occur, a violation is guaranteed. So agents do not escape an obligation by having other agents do the work for them. But, although the definition states the agent itself should perform the action to avoid violation, it does not state that the agent should *knowingly* perform the action to guarantee that the violation does not occur. It can be that the agent is not even aware that actually it performs an action that ensures the obligation is complied to. However, there is good reason to say that that should also lead to a violation. We can view this as making the obligation even more 'personal': the agent should perform the action 'knowingly' to avoid violation. It simply does not count if the agent only complies 'coincidentally'. The corresponding definition is:

$$OK[a \ \mathsf{xstit}]\varphi \equiv_{def} \Box(\neg K_a[a \ \mathsf{xstit}]\varphi \rightarrow [a \ \mathsf{xstit}]V)$$

Note that this obligation is *stronger* than the previous one, because unknowingly complying to the obligation now also counts as a violation.

Finally, we discuss a third variant for the obligation operator. For the above two variants, nothing is said about whether or not the agent actually knows whether or not it has the obligation. We associate awareness of an obligation directly with the awareness of the act of bringing about a violation in case the agent does not comply. We incorporate this by adapting the previous definition as follows:

$$KOK[a \ \mathsf{xstit}]\varphi \equiv_{def} \Box(\neg K_a[a \ \mathsf{xstit}]\varphi \rightarrow K_a[a \ \mathsf{xstit}]V)$$

In this definition also violations are knowingly brought about. This expresses that the agent bearing the obligation actually knows about the obligation, that is, the agent will knowingly bring about a violation if it does not comply with the obligation.

Of course, looking at the formal structure of the above definitions, a fourth definition suggests itself: one where it is not necessary to perform the obliged action knowingly, while at the same time, in case of non-compliance, the violation *is* brought about knowingly. But it seems clear right away that this combination is absurd. We cannot knowingly bring about a violation by unknowingly failing to comply with an obligation.

6 Back to the Examples

We now go back the the examples and question of section 2. First we answer the questions raised for the example with the boxes. We take the perspective of agent α and analyze its position using the logic defined in section 3.

In the first scenario, agent α faces the situation where agent β has already put money in one of the boxes (which in fact is in box 1), and where it has to choose the right box to collect the money. Its problem is that it does not know in which of the two boxes β has put the money. Now does it hold that $O[\alpha \ \mathsf{xstit}]CollectFrom1$? Yes. There is a right thing for the agent to do, and not doing that right thing, will lead to a violation, independent of what the agent knows it is doing and not doing. It seems interesting to investigate whether or not it would make a difference for this analysis whether or not we adopt the 'ought implies can' principle. 'Ought implies can' for the weak notion of obligation simply means that the agent should be able to perform the action unknowingly, which indeed is the case. In the example, agent α can choose the right box unknowingly: $\Diamond[\alpha \ \mathsf{xstit}]CollectFrom1$, but it cannot do it knowingly: $\neg\Diamond K_\alpha[\alpha \ \mathsf{xstit}]CollectFrom1$.

Now, in the first scenario do we also have the *stronger* obligation saying that $OK[\alpha \ \mathsf{xstit}]CollectFrom1$? The first part of our answer is: not if to this form of obligation, where the agent can only escape a violation if it knowingly sees to it that it collects the money, we apply the 'ought implies can' principle. The problem is of course, that it cannot knowingly do that, because it misses the information saying in which box the money is. So if for this type of obligation, where the agent should knowingly collect the money from the right box, we require 'ought implies can', and thus that the agent should be *able* to knowingly collect the money from the right box, the agent does *not* have an obligation. The second part of the answer is: if we do *not* adopt the 'ought implies can' principle for this type of obligation, then it depends on the intention behind the obligation, which is something that is not made explicit in the description of the example. In particular, it depends on whether or not the agent issuing this obligation accepts that the agent being subject to the obligation may coincidentally pick the right box or not.

Finally, for the first scenario, we ask whether or not the agent is aware of its obligation. And if so, what type of obligation is it aware of? This question is difficult to answer, again due to under-specification of the example. A possible interpretation is that the agent has the obligation in the stronger sense where it has to comply knowingly, and that it also knows this: $KOK[\alpha \ \mathsf{xstit}]CollectFrom1$. Then, if the agent takes the gamble of choosing a box, it knows it is violating its obligation. Another possible interpretation is that the obligation is of the weaker form, where the agent is allowed to comply unknowingly, and that the agent does not know whether it has the obligation $O[\alpha \ \mathsf{xstit}]CollectFrom1$ or the obligation $O[\alpha \ \mathsf{xstit}]CollectFrom2$, while it knows that it has one of the two: $K_\alpha(O[\alpha \ \mathsf{xstit}]CollectFrom1 \vee O[\alpha \ \mathsf{xstit}]CollectFrom2)$. More interpretations are possible. Of course, one of the good things of having a formalization is that we can use it to actually talk and reason about these different interpretations.

We now turn to the question why in scenario 2 we are not inclined to say agent α has an obligation at all. In the first scenario, the obligation can be said to be conditional on agent β's previous action. And in the example, in fact β has put the money in box 1. This could thus in principle be known to agent α. But agent α does not know. However, in the second scenario the obligation is conditional on a future choice of agent β. So, the obligation is conditional on something that is a priori not knowable (agents only can know the future if that future is knowingly brought about by these same agents presently: axiom $K_aX\varphi \rightarrow K_a[a \ \mathsf{xstit}]\varphi$). That is a possible explanation for why we are inclined to say that in the second case, indeed, the agent does not have an obligation.

We view the third scenario as analogous to the second one. Also here the agent cannot be said to be obliged to pick the right box, because what is the right box, has not been settled yet: it depends on an action of the other agent that takes place simultaneously.

Finally a brief word on the doctor's example. The problem is here that we hesitate between saying that the doctor fulfilled his obligation and that it did not. The difference is expressed by the first two notions of obligation. Either we have that $O[\alpha\ \mathsf{xstit}]CureThePatient$ or we have that $OK[\alpha\ \mathsf{xstit}]CureThePatient$. Which one is the case is under-specified in the example.

7 Related Work

In [25] a logic is presented whose semantics shares several features with ours. In particular, the logic has epistemic indistinguishability relations ranging over history/state pairs. However, actions are omitted. In [23] actions are added to this framework by using action names in the models and the object language. So, the authors take a, what we might call 'dynamic logic view' on action. The work focusses on so called 'knowledge based obligations'. The central idea is that when agents get to know more, there are less histories they consider possible, which in turn may induce that the subset of deontically optimal histories, may give rise to new obligations. So the phenomenon being studied is that new knowledge may induce new obligations.

In our setting the phenomenon of getting more obligations by an increase in knowledge can occur in different ways. One way is simply by becoming aware of an obligation, that is, getting to know that one knowingly performs a violation by not performing some obliged action. Another route to enabling that obligations arise as the result of new knowledge, is by adopting the 'ought implies can' principle for the stronger variants of our obligation operator. If agents get to know how to do something knowingly, they might incur an obligation that previously did not apply due to 'ought implies can'. This demonstrates that there seems to be more sides to the problem of 'knowledge based obligation'.

Another well-known interaction between epistemic and deontic modalities is Åqvist's puzzle of 'the knower' [26]. If knowledge is modeled using S5 and obligation using KD (SDL [27]), from $OK\varphi$ we derive $O\varphi$, which is clearly undesirable in an ought-to-be reading. However, this problem does not arise in the present logic, because obligation is strictly limited to apply to *actions*. In particular, if in Åqvist's example, for φ we substitute a STIT action $[\alpha\ \mathsf{xstit}]\varphi$, then we can read the derivation as 'the obligation to knowingly see to something implies the obligation to see to that same something'. This actually comes down to our second, stronger notion of obligation implying the first, weaker one. In section 5 we mentioned this not as a problem, but as a desirable property.

8 Future Research

The logic we presented in section 3 asks for extension in several ways. Note first that while the operators for agency are group operators, the operators for knowledge and obligation only refer to single agents. Actually, there are many open questions about how to generalize these operators to group operators. As is well-known, there are several notions of group-knowledge, such as 'shared knowledge', 'common knowledge' and 'distributed knowledge'. Which ones combine with which interaction properties for knowledge and group-action is yet unclear. Likewise we can consider generalizing the obligation operator to a group operator. Given the definitions of section 5 this actually hinges on providing group operators for the knowledge modalities.

Another issue concerns the violation constants. According to the present definitions, they are not relativized to agents or sets of agents. This corresponds to a 'consequentialist's' view on obligation, as in [11], where deontic optimality is determined according to an ordering of all possible histories. We could also take the view, like in [28], that deontic optimality orderings should be relative to agents or groups of agents. For our setting, using violation constants, that would mean that we introduce a violate constant for each agent or each group.

Related to this we want to discuss one final issue. We can also analyze the boxes example as a coordination problem. That is, we no longer take the viewpoint of agent α alone, but see the task as one where α and β have to coordinate their actions such that the money is transferred from β to α. This is a 'cooperative' view on the problem. The agents have to bring about the right thing, together. In particular, both agents should choose the same box: one puts the money in one of the boxes, the other collects it from that same box. But for coordination, we need communication. And that is absent from the example, since the agents do not know each other's choices.

Let us, for the moment, assume that we have a notion of strategic STIT, as in [29]. In strategic STIT, actions possibly involve *series* of choices. We denote the associated operator by $[C\ \mathsf{strat}]\varphi$, with C a group of agents, and φ the action result. Let us also assume that we have a notion of strategic 'ought-to-do'. We denote the associated operator by $O[C\ \mathsf{strat}]\varphi$. There are many options for the semantics of such an operator. Here we assume we have fixed one. Finally, we assume that there is also a 'some time in the future' operator, denoted by $F\varphi$, with the standard interpretation. Now, *all three* scenarios satisfy (in the initial state) the following formulas:

$$\Box(K_\alpha\neg\text{AlphaHasTheMoney} \wedge K_\beta\neg\text{AlphaHasTheMoney})$$
$$O_{\{\alpha,\beta\}}[\{\alpha,\beta\}\ \mathsf{strat}]F(\text{AlphaHasTheMoney})$$

$$K_\alpha\Diamond[\alpha\ \mathsf{strat}]F(\text{AlphaHasTheMoney})\ ,\ \neg\Diamond K_\alpha[\alpha\ \mathsf{strat}]F(\text{AlphaHasTheMoney})$$
$$K_\beta\Diamond[\beta\ \mathsf{strat}]F(\text{AlphaHasTheMoney})\ ,\ \neg\Diamond K_\beta[\beta\ \mathsf{strat}]F(\text{AlphaHasTheMoney})$$

Of course the issues surrounding the transfer of money could be modeled in much more detail, using propositions like $MoneyInBox1$, etc. But that is not important. What is important is that the key properties concerning knowledge, ability, obligation and time that are common to all three scenarios are captured by these formulas. In particular: the formulas express that there is a way in which α and β might succeed in transferring possession of the money from β to α, thereby obeying their joint obligation, but there is no way in which they can knowingly do that because they are not allowed to communicate or observe each other's behavior. This shows that from a more abstract perspective, where we use a strategic version of the STIT operator, we can also view the models associated to the three scenario's as semantically equivalent.

9 Conclusions

This paper presents an epistemic temporal STIT formalism that is complete with respect to a standard two-dimensional Kripke semantics. It introduces the new notion of 'knowingly doing' and discusses some of its possible properties. Using this notions, new 'epistemic' variants of operators for 'ought-to-do' are defined. The logic encompassing these new concepts is used to analyze some intriguing examples concerning the interaction of knowledge and obligation. Clearly this paper is not the last word on the issue of introducing both epistemic and deontic modalities in STIT logic. But it is a good start, and it contains some promising directions for further research.

References

1. Horty, J.: Personal communication (2007)
2. Broersen, J.M.: A complete STIT logic for knowledge and action, and some of its applications. In: Baldoni, M., Son, T.C., Riemsdijk, M.B.v., Winikoff, M. (eds.) Proceedings Workshop on Declarative Action Languages and Technologies (DALT) 2008. LNCS. Springer, Heidelberg (to appear, 2008)
3. Herzig, A., Troquard, N.: Knowing How to Play: Uniform Choices in Logics of Agency. In: Weiss, G., Stone, P. (eds.) 5th International Joint Conference on Autonomous Agents & Multi Agent Systems (AAMAS 2006), Hakodate, Japan, May 8-12, pp. 209–216. ACM Press, New York (2006)
4. Broersen, J., Herzig, A., Troquard, N.: A normal simulation of coalition logic and an epistemic extension. In: Proceedings Theoretical Aspects Rationality and Knowledge (TARK XI), Brussels
5. Balbiani, P., Gasquet, O., Herzig, A., Schwarzentruber, F., Troquard, N.: Coalition games over Kripke semantics: expressiveness and complexity. In: Dègremont, C., Keiff, L., Rückert, H. (eds.) Festschrift in Honour of Shahid Rahman. College Publications (to appear, 2008)
6. Bartha, P.: Conditional obligation, deontic paradoxes, and the logic of agency. Annals of Mathematics and Artificial Intelligence 9(1-2), 1–23 (1993)
7. Anderson, A.: A reduction of deontic logic to alethic modal logic. Mind 67, 100–103 (1958)
8. Meyer, J.J.: A different approach to deontic logic: Deontic logic viewed as a variant of dynamic logic. Notre Dame Journal of Formal Logic 29, 109–136 (1988)
9. Broersen, J.: Modal Action Logics for Reasoning about Reactive Systems. PhD thesis, Faculteit der Exacte Wetenschappen, Vrije Universiteit Amsterdam (February 2003)
10. Emerson, E.: Temporal and modal logic. In: Leeuwen, J.v. (ed.) Handbook of Theoretical Computer Science. Formal Models and Semantics, vol. B, pp. 996–1072. Elsevier Science, Amsterdam (1990)

11. Horty, J.: Agency and Deontic Logic. Oxford University Press, Oxford (2001)
12. Alur, R., Henzinger, T., Kupferman, O.: Alternating-time temporal logic. In: Proceedings of the 38th IEEE Symposium on Foundations of Computer Science, Florida (October 1997)
13. Alur, R., Henzinger, T., Kupferman, O.: Alternating-time temporal logic. Journal of the ACM 49(5), 672–713 (2002)
14. Pauly, M.: A modal logic for coalitional power in games. Journal of Logic and Computation 12(1), 149–166 (2002)
15. Broersen, J., Herzig, A., Troquard, N.: From coalition logic to STIT. In: Proceedings LCMAS 2005. Electronic Notes in Theoretical Computer Science, vol. 157, pp. 23–35. Elsevier, Amsterdam (2005)
16. Broersen, J., Herzig, A., Troquard, N.: Embedding Alternating-time Temporal Logic in strategic STIT logic of agency. Journal of Logic and Computation 16(5), 559–578 (2006)
17. Gabbay, D., Kurucz, A., Wolter, F., Zakharyachev, M.: Many-Dimensional Modal Logics: Theory and Applications. Elsevier, Amsterdam (2003)
18. Blackburn, P., Rijke, M.d., Venema, Y.: Modal Logic. Cambridge Tracts in Theoretical Computer Science, vol. 53. Cambridge University Press, Cambridge (2001)
19. Belnap, N., Perloff, M., Xu, M.: Facing the future: agents and choices in our indeterminist world, Oxford (2001)
20. Hoek, W.v.d., Wooldridge, M.: Cooperation, knowledge, and time: Alternating-time temporal epistemic logic and its applications. Studia Logica 75(1), 125–157 (2003)
21. Jamroga, W., Hoek, W.v.d.: Agents that know how to play 63(2) (2004)
22. Jamroga, W., Ågotnes, T.: Constructive knowledge: what agents can achieve under incomplete information. Technical Report IfI-05-10, Institute of Computer Science, Clausthal University of Technology, Clausthal-Zellerfeld (2005)
23. Pacuit, E., Parikh, R., Cogan, E.: The logic of knowledge based obligation. Knowledge, Rationality and Action a subjournal of Synthese 149(2), 311–341 (2006)
24. Wansing, H.: Obligations, authorities, and history dependence. In: Wansing, H. (ed.) Essays on Non-classical Logic, pp. 247–258. World Scientific, Singapore (2001)
25. Parikh, R., Ramanujam, R.: A knowledge based semantics of messages. Journal of Logic, Language and Information 12(4), 453–467 (2003)
26. Åqvist, L.: Good samaritans contarary-to-duty imperatives and epistemic obligations. NOUS 1, 361–379 (1967)
27. Wright, G.v.: Deontic logic. Mind 60, 1–15 (1951)
28. Kooi, B.P., Tamminga, A.M.: Conflicting obligations in multi-agent deontic logic. In: Goble, L., Meyer, J.-J.C. (eds.) DEON 2006. LNCS (LNAI), vol. 4048, pp. 175–186. Springer, Heidelberg (2006)
29. Broersen, J., Herzig, A., Troquard, N.: A STIT-extension of ATL. In: Fisher, M., van der Hoek, W., Konev, B., Lisitsa, A. (eds.) JELIA 2006. LNCS (LNAI), vol. 4160, pp. 69–81. Springer, Heidelberg (2006)

Reactive Kripke Models and Contrary to Duty Obligations

Dov M. Gabbay

King's College London

Version 1: 31 March 08

This is an intuitive description of our approach to modelling contrary to duty obligations. We shall describe our ideas through the analysis of typical problematic examples taken from Carmo and Jones [6], L. van der Torre [14] and Prakken and Sergot [5].

1 Preliminary Discussion

Contrary to duties (CTD) are dealt with in the framework of standard deontic logic (SDL), and ordinary Kripke possible world models. Given a world t, one associates statically a non-empty set $I(t)$ of ideal worlds for t and $t \vDash Oq$ (q is obligatory for t) if q holds in all the worlds of $I(t)$.

This is a static perception of obligation. If we have to list as Δ_t the set of all obligations for the world t then $I(t)$ would be the set of all models of Δ_t. The contrary to duty examples have some implicit dynamics in them. It is therefore not surprising that there are problems with the formalisation of various CTD examples within SDL. There are currently in the literature various proposals for solutions, however all are still largely within the STL possible world semantics approach or its extensions, with additional operators or preferential ordering. See footnote 2 below and references [18], [15] and [13].

Reactive Kripke models is a stronger version of possible world semantics, affording the semantic characterisation of more modal systems (this is a theorem in [1]. They have a dynamic dimension to them. Therefore using this new semantics might simplify existing solutions to CTD problems as well as offer new sharper solutions.

Note that this new approach does not necessarily abandon or challenge any of the existing solutions, since ordinary Kripke models are a special case of reactive Kripke models. This is an important point to bear in mind. We can proceed on two fronts.

1. Take an existing solution, say the Carmo and Jones model of [6] and view it in the context of the richer reactive Kripke semantics and maybe simplify the models or sharpen the semantics, etc.
2. Offer a new solution of our own, maybe disagree with existing proposals but make our case using the stronger tool of reactive Kripke models.

Either way all benefit and we are in a win–win situation.

Our plan is to give some examples in detail to familiarise the reader with our ideas, leaving the formal machinery and the extensive discussion to the full version of the paper.[1]

[1] We illustrate in this outline how CTD problems can be solved but we do not commit that our examples are the final solution. In the full paper we will offer some final solutions after looking at the problems more thoroughly. The spirit of this paper is correct but the details may change.

R. van der Meyden and L. van der Torre (Eds.): DEON 2008, LNAI 5076, pp. 155–173, 2008.

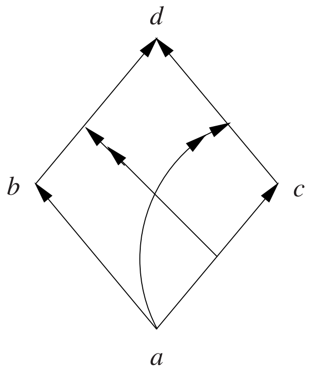


Fig. 1.

2 Reactive Kripke Models

Example 1 (A Reactive Kripke Model). Figure 1 shows such a model.

The single arrows show accessibility relation R. So in this figure we have aRc, aRb, bRd and cRd.

The double arrows are connections which can deactivate accessibility (we can use triple arrows to activate). We have

$aR(c,d)$ double arrow from a to the connection (c,d).
$(a,c)R(b,d)$ a double arrow from the connection (a,c) to the connection (b,d).

The best way to explain how evaluation works in such a model is by actually doing it.

So assume our model is (S,R,h), where

$$\begin{aligned} S &= \{a,b,c,d\} \\ R &= \{(a,b),(a,c),(b,d),(c,d),(a,(c,d)),((a,c),(b,d))\} \end{aligned}$$

If Q is the set of atomic sentences, then h is the assignment. For each $q \in Q$ and $s \in S$, $h(s,q)$ is a truth value.

The language contains the usual classical connectives and $\Diamond$.

Let us evaluate

$$a \vDash \Diamond\Diamond q, q \text{ atomic}$$

We need to start at point a and move two steps through the accessibility relation and land at a point $x \vDash q$. We can either make our first step to b or to c.

First observe that the minute we leave point a the double arrow from a to (c,d) will cancel the connection cRd. So if we leave a to go to c, then when we get to c the point d will no longer be accessible to c. Furthermore, to go to c we pass along the arc (a,c). The minute we pass through (a,c) the double arrow from (a,c) to (b,d) will cancel the connection bRd.

So when we get to c the model will have changed. Figure 2 shows the model as it is when we go to c from a.

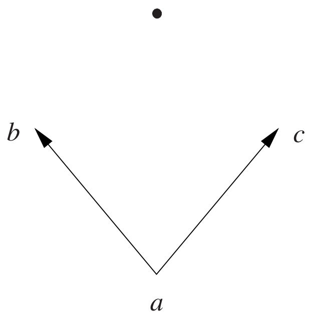


Fig. 2.

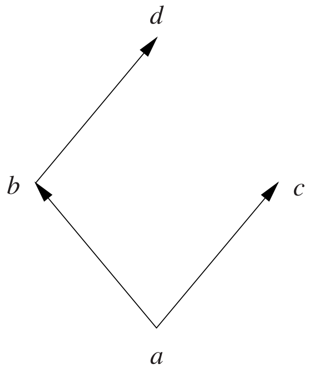


Fig. 3.

On the other hand, if we go from a to b, the connection (c,d) will be cancelled, but the connection bRd is still on and so we can continue to point d.

Figure 3 shows the model as it is when we get to d through the path a,b,d.

If indeed $d \vDash q$, then $a \vDash \Diamond\Diamond q$.

By the way, we have also shown that $a \vDash \Diamond\Box\bot$, because starting from a going to c we get $c \vDash \Box\bot$ as in Figure 2. Note that we need to know how we get to c in order to evaluate at c.

So the correct evaluation metapredicate should be

$$(x,y) \vDash A$$

where there is a unique path from x to y and we are evaluating A at y.

So y is our current evaluation point and x is our starting point. (We assume there is a unique path from x to y, otherwise we have to specify the path.) So we should write:

$$\begin{gathered}(a,a) \vDash \Diamond\Diamond q\\ (a,b) \vDash \Box\bot\\ (a,b,d) \vDash q.\end{gathered}$$

Note that (a,c,d) is not a path. so really (a,b,d) is a unique path to d.

Before we go on to contrary to duties, let us highlight the "take a walk" point of view of the evaluation. We imagine ourselves as agents standing at point a of Figure 1 and given a formula to evaluate or trying to reach a world (say get to d). We move along the arcs towards our goal worlds and evaluate along the way. Connections change as we walk up the arcs.

The logic is determined by the nature of the connections we allow and by the algorithm which tells us how to walk and evaluate. This point of view is dynamic, not static, and is very compatible with the semantic view of ideal worlds as objects to aspire for and contrary to duties. If we want to satisfy an obligation we must move towards an ideal world. If we deviate we might go in a direction which strays away from the ideal in which case some double arrows will change the connections and steer us in the direction of other subideal worlds. This view is very intuitive. It has implicit dynamics in it even though the model itself is static.

So given a world t and the obligations Δ_t for t, we do not just use semantics to describe Δ_t, i.e. use the set of ideal world $I(t)$ to characterise Δ_t. We actually think of $I(t)$ as worlds spread inside the possible world model and expect our agent to move along the accessibility relation to one of the worlds $I(t)$.

This is an action possible world model. The people at world t take action to move towards an ideal world. In this context contrary to duties become natural.

We must warn the reader that in any model there may be three types of movement.

1. Virtual movement towards the ideal world which is not temporal at all[2]
2. Temporal movement. See, e.g. [12]
3. A combination of (1) and (2).

There are CTD examples of all the above types. In fact some papers solved CTD puzzles of type (2) exploiting temporal operators. See [16] and the thesis of J. Broersen [17] (his use of 'reactive' is not the same as ours).

The model of Figure 1 is by no means the most general reactive model. We can allow double arrows with specific tasks, either to switch on a connection or to switch off a connection. We can also annotate connections and arrows as on or off.

Figure 4 is an example. Double arrows switch connections off. Triple arrows switch connections on.

The annotations 'on', 'off' say which arrows, double arrows and triple arrows are active at the start position before we move out of a. So, for example, (b,d) is off and $((a,b),(b,d))$ is on, and (a,b) is on.

We start at a. Moving out of a to b the double arrow from a to (c,d) switches (c,d) off. The triple arrow from (a,b) to (b,d) switches (b,d) on.

If we carry on from b to d then the double arrow from (b,d) to the double arrow $(a,(c,d))$ switches it off. We also see that triple or double arrows can go to other triple or double arrows, etc.

[2] Some CTD examples are completely static and do not involve time. Still the CTD aspect of the problem does involve movement towards the ideal worlds. How can this be? How can we give a static (non-temporal) model which still involves virtual movement? Well, we have such examples in classical mechanics. We solve a static distribution of forces in a structure by imagining a slight movement. This is called the principle of virtual work. For reference look up "virtual work" in Wikipedia.

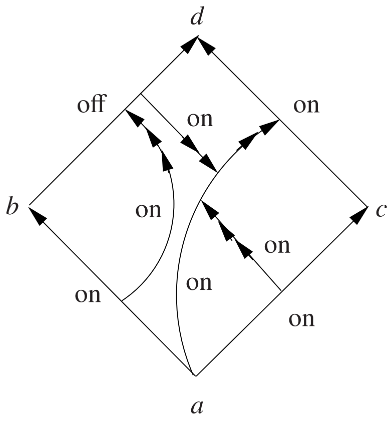


Fig. 4.

We can have suitable modal operators for walking along the arcs in any suitable way.

The next section gives a specific reactive semantics suitable for some analysis of CTD.

3 Contrary to Duty Models

We consider a reactive model of the form (S, R, I, h), where R is a relation say as in Figure 1 (no triple arrows) and I is the ideal world function giving for each $s \in S$ a non-empty set of ideal worlds $I(s) \subseteq S$. So actually $(S, I(s), h)$ is an SDL model for O, R gives us the reactivity. We assume two additional modalities. $\Diamond$ evaluated as above, taking into account the effect of the double arrows and $\diamondsuit\!\!\!\!\mid$ which is an ordinary modality which ignores the double arrows. So for $\diamondsuit\!\!\!\!\mid$, Figure 1 becomes Figure 5.

Consider now the model in Figure 6. Its points are

S =$\{a, b, c, w^+, w^-, f^-\}$is as in Figure 6, which also indicates the meaning of the worlds
$R = \{(a, b), (a, c), (c, f^-), (b, w^+), (b, w^-),(a, (c, f^-)), ((a, b), (b, w^-))\}$

The function I satisfies $I(a) = \{f^-\}$. We don't care about the other values of I.

The minute we leave point a the connection with f^- is severed. This means the agent beginning at a is not able, according to this model, to follow a path to the ideal world f^-. Note that double arrows emanating from points are local properties of the model (which we can interpret as having to do with agent's circumstances). Double arrows emanating from connections are systems contrary to duties.

Thus the agent is not able to comply to his obligation and has to go for fence. He can go to point c and get stuck there or go to point b to continue to a world with a fence. As he passes the connection (a, b) the double arrow from (a, b) disconnects his way to the non-white fence world and he has to go to w^+.

Note that we need a starting point an an evaluation point. So the model has the form (S, R, I, h, a, x) (we fix a as the starting point for our example of Figure 6).

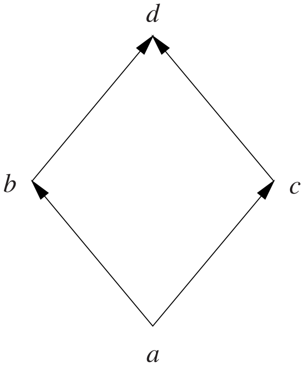


Fig. 5.

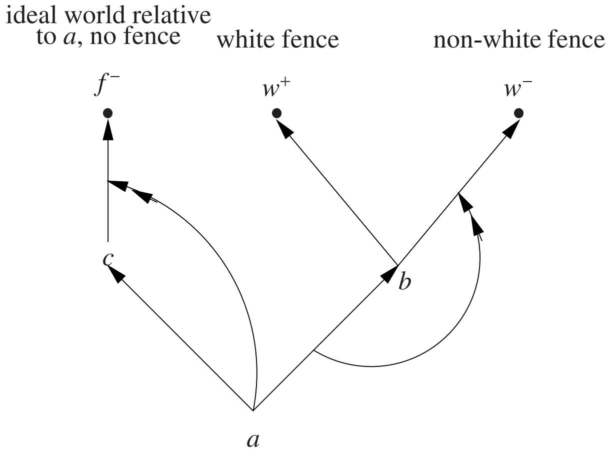


Fig. 6.

So by looking at a, w^+ we know the agent is not able to comply to his obligation and has to go for fence and he has a white fence because of CTD.

So a set of CTD sentences determines a class of reactive models. Given a model we can read from it the CTD sentences it suggests. We are saying ‘suggests’ rather than ‘holds’ because existence of double arrows suggests STD sentences, see section 5.

Let us look at the analysis of the scholarly work of Carmo and Jones [6, p. 305].

Statement of the Problem

(d1) *There must be no fence*
In the model this is implemented by f^- being ideal world relative to a, $f^- \in I(a)$ with a as the starting point.

(d2) *But if there is a fence it must be white*
This is implemented by the fact that any arc in the model where there is no continuation to an ideal world has double arrows emanating from it cutting access

to any non-white fence world ahead, i.e. in Figure 6 this is the double arrow $((a,b),(b,w^-))$.

We can add

(d3) *There is a fence*

This is implemented by saying we are at point $x \in \{w^+, w^-\}$. So, for example, the model of Figure 6 with starting point a and evaluation point x, $x \in \{w^+, w^-\}$ does model (d1)–(d3).

Carmo and Jones [6] did not have reactive models. They added additional modalities of actually possible and potentially possible and used them to make case analysis. We now quote their case analysis and show how to express the cases in our reactive models. The reader should note that we did not construct our model to simulate Carmo and Jones. Had we done so systematically we probably would have come up with a slightly different model, which implements additional modalities using reactivity.

Case 3.1

(f1) *There is no fence and it is still actually possible not to erect a fence and actually possible to erect a fence, white or not.*

In the model we look at point a without the double arrows emanating from it, i.e. to model Case 3.1, we take a model without double arrows at all. The starting point is a (which allows for all the possibilities) and the evaluation point is f^- (which allows for the fact that there is no fence). Alternatively, we can say $a \vDash$ no fence[3]

Case 3.2

(f1) *There is a white fence and it is actually fixed that there will be a fence, possibly white or another colour.*

To model this take a as the starting point and w^+ as the evaluation point. The double arrow from a makes it actually not possible to have no fence and the double arrow $((a,b),(b,w^-))$ blocks w^-.

(f2) *It is potentially possible to have no fence.*

This is clear since there is a connection path to f^-, if we ignore the double arrows (i.e. use $\Diamond\!\!\!\!\mid\;$ for potentially and $\Diamond$ for actually).

Let us do another example:

Example 2 (Chisholm paradox). The following is from Carmo and Jones [6, p. 299]

(d1) *It ought to be that a certain man go to help his neighbours*
(d2) *It ought to be that if he goes he tell them he is coming*
(d3) *If he does not go, he ought not to tell them he is coming*
(d4) *he does not go.*

Consider Figure 7. (d1) is modelled by $d \in I(a)$, i.e. d is an ideal world for a. (d4) is modelled by taking e as an evaluation point.

[3] Our purpose here is not necessarily to model and simplify Carmo and Jones [6] but to show we have the power to offer our own models or to model other approaches. See Remark 11. In fact, we do not need to make such a case analysis. In the full paper, we shall study in detail the case analysis of [6] as well as the works in [12], [18] and [19].

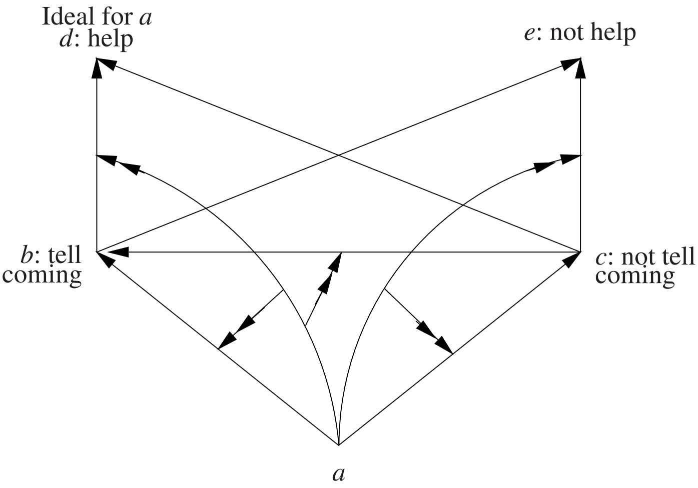


Fig. 7.

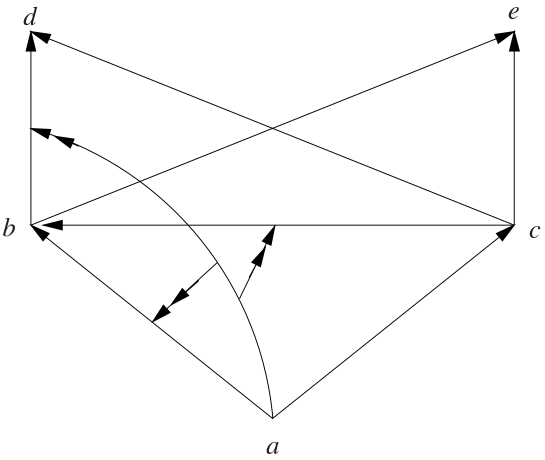


Fig. 8.

Modelling (d2) and (d3) is a bit challenging, because both help and tell are in the future and tell comes before help. See Example 6 below for a discussion. To model (d2) and (d3) first note that the double arrow from a to (b,d), triggers the system to send a double arrow from $(a,(b,d))$ to (a,b) and to (c,b). This models (d3). Second note that to model (d2) we have the double arrows $(a,(c,e))$ and $((a,(c,e)),(a,c))$. However, putting them both in the same model means that the man decides to block his path from going anywhere.

The way to solve it is to split Figure 7 to Figure 8 one with the single arrows of Figure 7 and one with only the double arrows $\{(a,(b,d))((a,(b,d)),(c,b)),((a,(b,d),(a,b))\}$ modelling (d3) and Figure 9 with all the single arrows plus only the double arrows $\{(a,(,e)),((a,(c,e)),(a,c))\}$, modelling (d2).

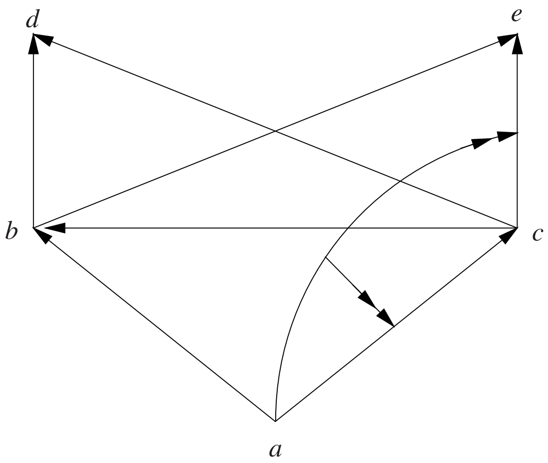


Fig. 9.

The actual modelling of the Chisholm paradox is the pair of reactive models done in parallel disjunctively. See Definition 8 and also [19].

We now follow the case analysis of Carmo and Jones [6, pp. 300].

Case 1.1

(f1) *The man decides not to go to help.*
This is modelled by Figure 8.

(f2) *It is potentially possible for the man to help and to tell and potentially possible for the man to help and not to tell*
This is modelled in Figure 8 by choosing the evaluation point as a. The beginning point is always a, so we have (a, a) as our pair. The decision in (f1) is the choice of Figure 8. The potentiality comes from the fact that the man has not started yet (evaluation point a) and he potentially can change his mind and choose the model of Figure 9.

(f3) *The man has not in fact told that he is coming to help although it is still actually possible that he does tell and actually possible that he does not tell.*
This is modelled by taking the evaluation point as point c. The man can actually move either from c to e or from c to b and then to c.

The other cases of Carmo and Jones, namely case 1.2 and case 1.3, [6, pp. 300 and 301] can be done similarly.

Case 1.4 contains the fact that the man helped but that it was potentially impossible for the man to tell his neighbour. For this we need a variation of Figure 9 where there is no connection from a to b. The man took the path a to c to d. See Figure 10.

Case 1.5 is where it is not potentially possible to help but the man does tell he is coming but he might have not told.

This is covered by Figure 11.

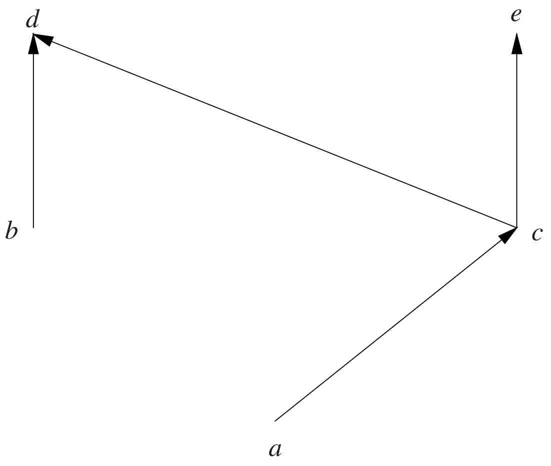


Fig. 10.

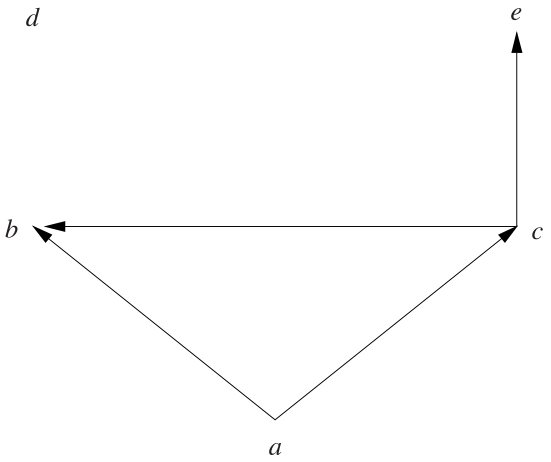


Fig. 11.

In the full version of this paper we shall analyse the Carmo and Jones examples in detail. Note that since we are using reactive models we do not need any case analysis. Our two diagrams 8 and 9 actually solve the paradox in principle.

4 Concluding Remarks

Remark 1 (The Idea of Reactivity). The idea of reactivity is a general one applicable across research areas. Whenever we have a system with states and algorithms involving these states we can turn it reactive by allowing signals from states or transitions which are being used to some other states causing change in the other states.

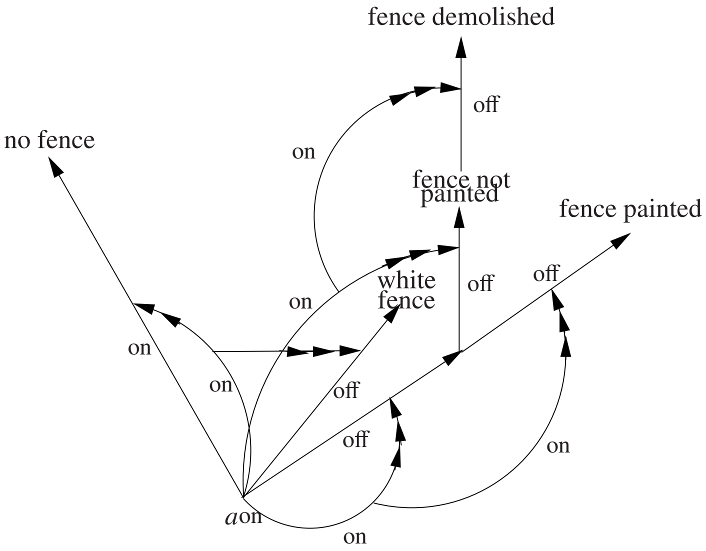


Fig. 12.

This change can be due to faults and overuse of the system, or feedback in the system or object level implementation of norms regimentation in the system or just efficiency shortcuts embedded in the system by design. We are now systematically studying reactive automata, reactive grammars, reactive conditionals, reactive proof theory and more.

Remark 2 (Proof Theory for Reactive Semantics). Proof theory can be provided for reactive Kripke models in the methodology of LDS (Labelled Deductive Systems). This means we can propose models for CTD systems as well as proof theory. This also means that we can provide LDS proof theory for existing CTD systems such as, for example, the Carmo and Jones proposals [6]. More on this in the full version of the paper.

Remark 3 (Multiple Level CTD). We have no inherent difficulties with multiple level contrary to duty.

For example:

1. It is obligatory to have no fence.
2. If there is a fence it should be white.
3. If it is not white it should be painted white.
4. If it cannot be painted (some plastics cannot take paint, I have some in my office at King's) then it should be demolished.

Figure 12 illustrates a possible model.

The beginning position is that all arrows are off except the one leading to no fence. The arrows emanating from *a* block the path to no fence and clear a path to fence not painted. The arrows from the agent's arrows force him to demolish.

Remark 4 (Conflicting Norms). We can cope more easily with conflicting norms. The modern world is full of them. Think of

1. There should be no fence
2. There should be no dog
3. If there is a dog there should be a fence
4. If there is a fence it should be demolished
5. There is a dog

In the reactive model we can loop, repeatedly erecting and demolishing a fence and thus fulfill our obligations, that is assuming we insist on a dog.

Remark 5 (Expressing Ideal Worlds using Double Arrows). The additional power of the double arrows and triple arrows can be used to eliminate the ideal world function $\lambda t I(t)$. In a model with connections capable of being on or off, we can characterise $I(t)$ as all those worlds accessible to t by an active direct connection and the non-ideal worlds as those not accessible. The minute we make our first move we can activate and deactivate connections to bring us back to whatever connections we want. So in fact the ideal worlds are recognised by the way we do our on and off switches. Figure 7 becomes the new Figure 13 below.

Here we use arrows, double arrows and triple arrows.

The starting position is that this access (on) only to the ideal world f^-. The minute we move from a, we cancel access to it and activate access in the $\{b, w^+, w^-\}$ direction. We know f^- is ideal because disconnecting access to it makes the difference.

There may be better ways to do the coding. We just want to illustrate the principle involved.

Remark 6 (Solving CTD Problems using Reactive Proof Theory). Reactive proof theory can be used directly to solve problems of CTD. The idea of reactive proof theory is that using a rule can activate or de-activate other rules. So, for example, we may have

1. $O\neg f$. There must be no fence.
2. $f \to Ow^+$ If there is a fence then it must be a white fence.
3. $\vdash w^+ \to f$. White fence is a fence

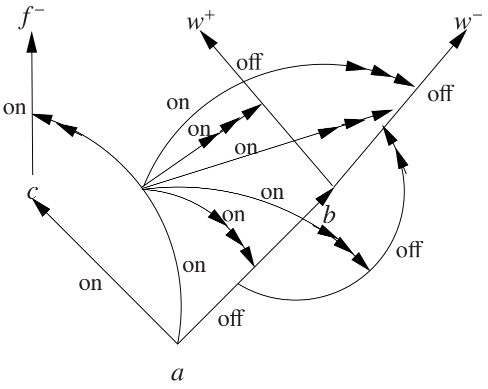


Fig. 13.

4. f. There is a fence.
5. Reactivity rules: if we use 6.2 we cannot use item 6.1.

This is crude but effective.

Before the time when we understood the reactive proof theory, such rules might have looked (to the traditional logicians) as a hack, a trick without semantical meaning. However, with reactive semantics, a competently crafted system with such rules may actually be characterised by a class of reactive models.

The above example may be too crude and some balanced refinement may be needed, but it does illustrate the idea. Note that the use of 6(2) as a ticket corresponds the move from point a to point b in Figure 8. The ideal world f^- is no longer accessible so 6(1) cannot be used. Ow^+ is derived. This corresponds to the CTD rule.

I have studied rule cancellations and deletion in logic extensively. I assure you this is workable. See my papers [8,9] and [10].

Remark 7 (Comparison with Dyadic Obligations $O_A B$). We now comment on the use of the binary operator $O_A B$ to express CTD B if A (also denoted $O(B/A)$. Semantically this is a powerful operator on the set of possible worlds corresponding to a binary relation $xR_A y$ indexed by subsets (the truth set of A). If you look at Figure 8 again, you see that the double arrow depends on a pair of points. So, for example, we formally can write $\neg bR_{(a,b)}w^-$, i.e. $R_{(a,b)}$ represents the double arrow. So formally we need relations indexed by pairs of points at the most, not all subsets. $O_A B$ is too powerful. I also think in addition to the above technical points that $O_A B$ as a different concept and should not be used as a hacking coding ground for solving the CTD problems. We leave it at that, more in the full paper. Note that Carmo and Jones [6] whose a dyadic connective $O_A B$ use for it a semantical function from subsets to families of subsets of the form

$$ob : w^S \mapsto 2^{2^S}$$

which is a very high level function.

Remark 8 (Input Output Logic). There is affinity between our reactive Kripke models and the work of Makinson and Torre on input output logics. I refer to [11]. Given a CTD of the form If A then obligatory that B, we can feed A as input to an input output node and get B as output. This idea can be incorporated into reactive Kripke models if we allow the worlds in the model to have input output facilities.

We shall elaborate on this in the full paper.

5 Technical Definitions

This section supplies the technical definitions supporting the intuitive ideas presented in this paper.

Definition 1 (Language). *Our propositional language contains a set of atomic propositions Q, the classical connectives and the modal connectives OA, $O_A B$, $\Diamond A$, $\Diamond\!\!\!\!\mid\; A$, and possibly more.*

Definition 2 (Reactive Arcs)

1. *Let S be a non-empty set. The set $\mathcal{A}$ of arcs on S is defined as the smallest set $\mathcal{A}$ containing $S \times S$ and closed under the following operation.*
   - *If $x \in S \cup \mathcal{A}$ and $y \in \mathcal{A}$ then $(x, y) \in \mathcal{A}$.*

   The above condition says that we can have for example $(t, (a, b)) \in \mathcal{A}$ but not $((a, b), t) \in \mathcal{A}$. We do not allow $(x, y) \in \mathcal{A}$ with $y \in S$.
2. *A set $\mathcal{A}_0 \subseteq \mathcal{A}$ of arcs is said to be well founded if whenever $(x, y) \in \mathcal{A}_0$, then $x \in S \cup \mathcal{A}_0$ and $y \in \mathcal{A}_0$.*

Definition 3 (Reactive Accessibility)

1. *An element $(x, y) \in S \times S$ is called a single arrow. We can also write $x \rightarrow y$.*
2. *Let $(x, y) \in \mathcal{A}$. (x, y) can be used either as a negative switch or as a positive switch. We regard (x, y) as a negative switch by writing it as a double arrow $x \twoheadrightarrow y$. We regard it as a positive switch by writing it as a triple arrow $x \rightarrow\mkern-14mu\twoheadrightarrow y$.*
3. *A single arrow, double arrow or triple arrow can be either on (we put $+$ in front of it) or it can be off (we put $-$ in front of it).*
4. *An accessibility relation R is obtained from a well founded base of arcs $\mathcal{A}_R$ as follows:*
   - *(a) If $(x, y) \in \mathcal{A}_R$ and $x, y \in S$ then either $+(x, y) \in R$ or $-(x, y) \in R$, but not both.*
   - *(b) If $(x, y) \in \mathcal{A}$ with $y \notin S$ then exactly one of the following must be in R:*
     either $+(x \twoheadrightarrow y)$
     or $-(x \twoheadrightarrow y)$
     or $+(x \rightarrow\mkern-14mu\twoheadrightarrow y)$
     or $-(x \rightarrow\mkern-14mu\twoheadrightarrow y)$
5. *A reactive Kripke model has the form (S, I, R, a, e, h), where $S \neq \varnothing$ is the set of possible worlds, $I : S \mapsto (w^S - \varnothing)$ is the ideal world function, R is a reactive accessibility relation and $a \in S$ is the beginning world and $e \in S$ is the evaluation world. h is the assignment to the atoms. For each $q \in Q, h(q) \subseteq S$.*

Remark 9 1. The meaning of $x \twoheadrightarrow y$ is that as we pass through x the arc y is put in an off position if it is on. The meaning of $x \rightarrow\mkern-14mu\twoheadrightarrow y$ is that as we pass through x the arc y is put in an on position if it is off.

2. An arc y is in off position in R if $-y \in R$. It is in an on position in R if $+y \in R$. The formal definition is Definition 4 below.

Example 3 Let us describe the model introduced in Figure 4.

$$
\begin{aligned}
S &= \{a, b, c, d\} \\
R &= \{+(a, c), +(a, b), -(b, c), +(c, d)\} \cup \\
&\quad \{+(a \twoheadrightarrow (c, d)), +((a, c) \rightarrow\mkern-14mu\twoheadrightarrow (a, (c, d))), +((a, b) \rightarrow\mkern-14mu\twoheadrightarrow (b, c)), +((b, d)) \rightarrow\mkern-14mu\twoheadrightarrow (a, (c, d))\}
\end{aligned}
$$

I is not specified in the figure, neither is h or e.

Definition 4 (Movement in a Reactive Model). *Let (S, R, e) be part of a model. Assume $+(e, e') \in R$. We explain what it means to move along the arc (e, e') from (S, R, e) to $(S, R_{(e,e')}, e')$.*

$R_{(e,e')}$ is obtained from R by executing the following actions:

1. *For every $+(e \twoheadrightarrow y) \in R$ such that $+y \in R$, replace $+y$ by $-y$.*
2. *For every $+(e \mathrel{\rightarrow\!\!\!\!\twoheadrightarrow} y) \in R$ such that $-y \in R$, replace $-y$ by $+y$.*
3. *For every $+((e, e') \twoheadrightarrow y) \in R$ such that $+y \in R$ replace $+y$ by $-y$.*
4. *For every $+((e, e') \mathrel{\rightarrow\!\!\!\!\twoheadrightarrow} y) \in R$ such that $-y \in R$ replace $-y$ by $+y$.*
5. *$R_{(e,e')}$ is the set obtained from R by doing exactly the above actions.*

Define R_a as the set obtained from R by executing actions (1) and (2) only.

Definition 5 (Reachability). *Let (S, R, a) be a part of a model. Let $x \in S$. We define the notion of 1x is reachable from a in (s, R, a)' by induction:*

1. *If $+(a, x) \in R$ then x is reachable in one step from a in (S, R, a).*
2. *x is reachable in $n + 1$ steps in (S, R, a) if for some $a' \in S$ $+(a, a') \in R$ and x is reachable in n steps from a' in $(S, R_{(a,a')}, a')$.*
3. *x is reachable from a in (S, R, a) if for some n, x is reachable from a in (S, R, a) in n steps.*

Definition 6 (Contrary to Duties). *Let (S, I, R, a, e) be part of a model. We now define the contrary to duties suggested by this model relative to a.*

The ideal worlds are $I(a)$, and assume that none of $I(a)$ is reachable from a in (S, R_a, a)

Let a' be such that$+(a, a') \in R$. Let $CTD_{(a,a')}$ be the points reachable from a' in $(S, R_{(a,a')}, a')$. Let $\mathrm{CTD}_a = \bigcap_{+(a,a')\in R} \mathrm{CTD}_{(a,a')}$. Then CTD_a are the contrary to duty worlds relative to a. In words:

> *It is obligatory to go to $I(a)$ but if not go to CTD_a.*

Remark 10. The double and triple arrows emanating from a (i.e. $+(a \twoheadrightarrow y)) \in R$ or $+(a \mathrel{\rightarrow\!\!\!\!\twoheadrightarrow} y) \in R$) can be seen to indicate the intention of the agent, or the restrictions on the user imposed by the model (if we do not want to ascribe intentions to users).

If activated the agent will move to a model with R_a. If $I(a)$ world are no longer accessible then our agent is not able to execute his obligations at a. The contrary to duties are the adjustments (activation and cancellation of arcs) firing as the agent passes through to any a' such that $+(a, a') \in R$. The CTD_a are the worlds which are always accessible from any a' the agent goes to. These are the contrary to duty worlds. Note that $I(a')$ are the ideal worlds of a'. This is not the same as the contrary to duties at a as imposed from node a onto node a'.

The reader may ask what if at a' some $x \in I(a)$ is still reachable? The definition of CTD_a still works. What is the meaning of it? The answer is in the next example. We call these *preventive* CTD, PCTD.

Example 4. At a world where you have a fence, it is obligatory to point the fence white within seven days. Say after five days nothing has been done. There are two options. Do nothing and the fence will not be painted. Hire extra hands and the fence will be painted. A preventive contrary to duty is to put pressure on the agent by blocking some of his moves. See Figure 14.

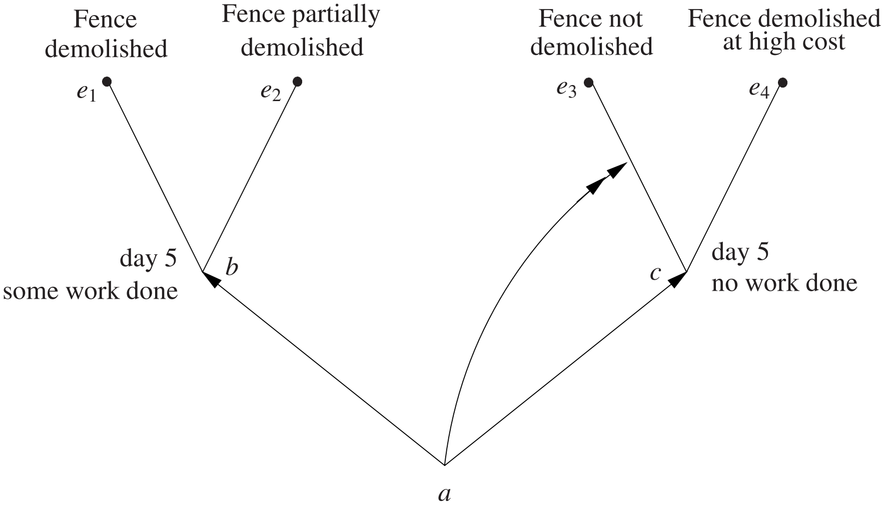


Fig. 14.

It is not correct to say that a CTD is that if after five days no work is done then the agent has a duty to bring extra workers to do the job. His duty remains to demolish the fence within seven days and he can still do it.

So the double arrow $(+((a,c),(e,e_3))$ is a preventive measure, a PCTD.

Definition 7 (Evaluation of Modalities in a Model). *Consider a model* (S,R,I,a,e,h). *We define the notion of* $e \vDash A$ *for a wff* A.

1. $e \vDash q$, *for* q *atomic, if* $e \in h(q)$.
2. *We adopt the usual definition for the classical connectives.*
3. $e \vDash OA$ *iff for all* $x \in I(e)$, $x \vDash A$ *in the model* (S,R,I,a,x,h).
4. $e \vDash \Diamond\!\!\!\!\mid\; A$ *iff for some* x *such that* $+(e,x) \in R$, $x \vDash A$ *in* (S,R,I,a,x,h).
5. $e \vDash \Diamond A$ *iff for some* x *such that* $+(e,x) \in R$, $x \vDash A$ *in* $(S,R_{(e,x)},I,a,x,h)$.

Definition 8 (Suggested CTDs). *A CTD multimodel is a family* $\mathcal{M}$ *of of models of the form* $\mathcal{M}_i = (S,I,R_i,a,e,h)$ *where all* (S,I,a,e,h) *are the same for all models and only* R_i *change. The CTD rules suggested by the family are all the* CTD^i_a *suggested by each model* M_i.

Remark 11. Note that for formulas without modalities, we have $x \vDash_i A$ iff $x \vDash_j A$ for any i, j, since they all agree on (S,I,h). Write $x \vDash A$ if A holds in any i. So for such formulas we can extract a syntactical CTD. Let $\Delta_a = \{A \mid x \vDash A \text{ for all } x \in I(a)\}$.

Let $\Theta^i_a = \{B \mid y \vDash B \text{ for all } y \in \mathrm{CTD}^i_a\}$.

Then our syntactical CTDs suggested by M_i are Δ_a and if not then Θ^i_a.

Example 5 (The Reykjavic Paradox). We show how to handle this paradox, see [14]. we have

1. X should not tell the secret to Reagan.
2. X should not tell the secret to Gorbachev.

3. If X tells Reagan, then X should tell Gorbachev.
4. If X tells Gorbachev, then X should tell Reagan.
5. X tells Reagan and Gorbachev.

(1)–(5) is the paradox. It is easy to model it in our system.

So, just to add to the problem, let me add (6) as a challenge

6. If X insists on telling exactly one of them, then it should be Reagan.

Figure 15 is a model for the above.

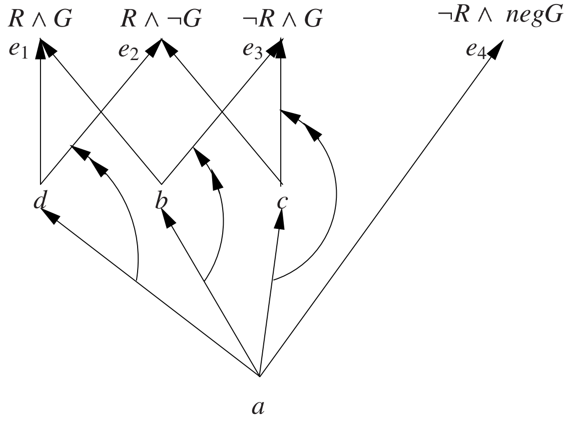


Fig. 15.

Point e_4 is the ideal world for a, $e_4 \in I(a)$. This models (1) and (2). Point d forces telling Reagan. Point b forces telling Gorbachev. Point c forces telling exactly one of them.

The double arrows are the contrary to duties. They model (3), (4) and (6). The arc to point d for example, means X is going to point d which forces telling Reagan. We want him to be under CTD to tell Gorbachev. So we disconnect the arc (d, e_2). So we need the double arrow $+((a, d) \twoheadrightarrow (d, e_2))$. Similarly, we need $+((a, b) \twoheadrightarrow (b, e_3))$ and also $+((a, c) \twoheadrightarrow (c, e_3))$.

The beginning point is a, the evaluation point is e_1 which models (5).

The reader may ask how did we construct the model? Well, there are some heuristics.

Remark 12 (Heuristics for Building a Reactive Model in Cases where there is no Temporal Element Involved). Let $A_1, \ldots, A_n$ be obligations. Let If $\neg A_i$ then $B_i, i = 1, \ldots, n$ be CTDs.

First let $\Theta = \{X_1, \ldots, x_{2^{2n}}\}$ be the set of all Boolean combinations of $\{A_1, \ldots, A_n, B_1, \ldots, B_n\}$.

The contrary to duties say that if $\neg A_i$ then B_i. When we move along the arc $+(a, t_i)$, we are getting to a point where $\neg A_i$ is committed. So we must force B_i. So any t_j such that $X_j \vdash \neg A_i \wedge \neg B_i$ must be disconnected. So we include in R all double arrows of the form $+((a, t_i) \twoheadrightarrow (t_i, e_j))$ such that $X_j \vdash \neg A_i \wedge \neg B_i$.

The ideal worlds for a are all e_j such that $X_j \vdash \bigwedge_{i=1}^{n} A_i$.

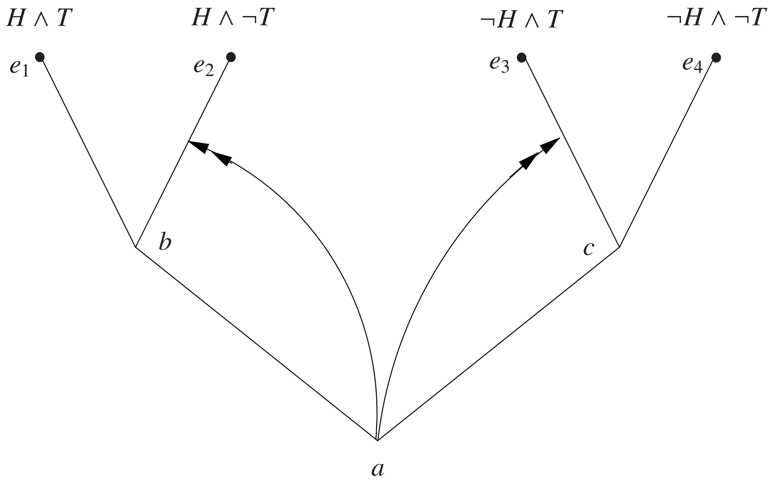


Fig. 16.

The above construction is a Henkin-like type of construction. It works when there is no temporal element.

Example 6 (Chisholm Paradox Revisted). Let us try and model the Chisholm paradox, using the Henkin-like method as in the previous example 5. We get Figure 16

Figure 16 does the job but it does not take into account the temporal aspect of the problem. See also [10] and [19].

Evaluation point is $x \in \{e_3, e_4\}$, $x = e_4$ if the agent complies with the CTD and $x = e_3$ if he does not.

This solution is more appropriate if instead of 'Tell' we have 'Wear his overalls'. So

(d1) It ought to be that a certain man go to help his neighbour.
(d2) It ought to be that if he goes he wear his overalls.
(d3) If he does not go he ought not wear his overalls.
(d4) He does not go.

We need to develop special methods to deal with the temporal aspects of CTD.

We can improve the situation in the Chisholm example by reading 'T' as 'T''

T' = having told in the past.

This would help. However, we do not get the best model. We do need to develop a general theory for time dependence. Again I ask the reader to wait for the full paper.

Let us stop here. Full analysis in the expanded full version of the paper possibly with more co-authors.

Acknowledgements

I am grateful to L. van der Torre and O. Rodrigues for valuable comments.

References

1. Gabbay, D.M.: Reactive Kripke Semantics and Arc Accessibility. In: Avron, A., Dershowitz, N., Rabinovich, A. (eds.) Pillars of Computer Science: Essays Dedicated to Boris (Boaz) Trakhtenbrot on the Occasion of His 85th Birthday. LNCS, vol. 4800, pp. 292–341. Springer, Berlin (2008); Earlier version published. In: Carnielli, W., Dionesio, F. M., Mateus, P. (eds.) Proceeding of CombLog04, Centre of Logic and Computation University of Lisbon, pp. 7–20 (2004), http://www.cs.math.ist.utl.pt/comblog04/ ftp://logica.cle.unicamp.br/pub/e-prints/comblog04/gabbay.pdf
2. Gabbay, D.M., Barringer, H., Rydeheard, D.: Reactive Grammars. Draft
3. Gabbay, D.M., Crochemore, M.: Reactive Automata. Draft
4. Gabbay, D.M., D'Agostino, M.: Reactive Conditionals. Draft
5. Prakken, H., Sergot, M.J.: Contrary-to-duty obligations. Studia Logica 57(1), 91–115 (1996)
6. Carmo, J., Jones, A.J.I.: Deontic Logic and Contrary-to-Duties. In: Gabbay, D.M., Guenthner, F. (eds.) Handbook of Philosophical Logic, vol. 8, pp. 265–343. Springer, Heidelberg (2002)
7. Gabbay, D.M.: Reactive Proof Theory. Draft
8. Gabbay, D.M., Rodrigues, O., Woods, J.: Belief Contraction, Anti-formulas, and Resource Overdraft: Part I. Logic Journal of the IGPL 10, 601–652 (2002)
9. Gabbay, D.M., Rodrigues, O., Woods, J.: Belief Contraction, Anti-formulae and Resource Overdraft: Part II. In: Gabbay, D.M., Rahman, S., Symons, J., van Bendegem, J.-P. (eds.) Logic, Epistemology and the Unity of Science, pp. 291–326. Kluwer, Dordrecht (2004)
10. Gabbay, D.M., Reyle, U.: N-Prolog: An Extension of Prolog with Hypothetical Implications I. Journal of Logic Programming 1, 319–355 (1984)
11. Makinson, D., van der Torre, L.: Constraints for input-output logics. Journal of Philosophical Logic 30(2), 155–185 (2001)
12. van der Torre, L.W.N., Tan, Y.-H.: The Temporal Analysis of Chisholm's Paradox. In: Proceedings of the Fourteenth National Conference on Artificial Intelligence and the Ninth Innovative Applications of Artificial Intelligence Conference (1998)
13. Hansson, B.: Standard Dyadic Denotic Logic. Noûs 3, 373–398 (1969)
14. van der Torre, L.: Violated obligations in a defeasible deontic logic. In: Proceedings of ECAI 1994, Amsterdam, pp. 371–375 (1994)
15. Makinson, D.: Five faces of minimality. Studia Logica 52, 339–379 (1993)
16. Broersen, J., van der Torre, L.: Reasoning About Norms, Obligations, Time and Agents. In: Proceedings of PRIMA 2007. LNCS. Springer, Heidelberg (to appear)
17. Broersen, J.: Modal Action Logics for Reasoning about Reactive Systems, Jan Broersen, PhD-thesis Vrije Universiteit Amsterdam (January 2003)
18. Hansen, J., Pigozzi, G., van der Torre, L.: Ten Philosophical Problems in Deontic Logic. In: Boella, G., van der Torre, L., Verhagen, H. (eds.) Normative Multi-agent Systems, Dagstuhl Seminar Proceedings, Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany (2007)
19. Broersen, J., van der Torre, L.W.N.: Semantic Analysis of Chisholm's Paradox. In: BNAIC 2005, pp. 28–34 (2005)

Normative Consequence: The Problem of Keeping It Whilst Giving It up

Audun Stolpe

Department of Philsosophy, University of Bergen, Norway
humas@uib.no

Abstract. The problem of deriving implicit norms from explicitly given ones is at the heart of normative reasoning. *In abstracto* the problem is that of formalizing a plausible consequence relation taking norms to norms. I argue that any such relation should allow norms to be chained, even when the consequent of one is strictly stronger than the antecedent of another—i. e. even if logical inference is required to complete the chain. However, since it is commonly agreed that the set of items classically entailed by an obligatory proposition are not in general obligatory, we are left with the following problem: How do reserve the right to reason classically for the purpose of chaining, whilst not committing to the view that all items entailed by a norm are obligatory in the same sense. I shall argue that the problem can be given a natural solution with reference to different *uses* of a norm in a normative system.

Keywords: Normative systems, input/output logic, dynamics.

1 Introduction

A central problem of normative reasoning is to clarify the inference from explicitly given norms to other norms that are, in some sense, implied. For instance a prohibition against processing personal information under certain circumstances intuitively entails a prohibition against processing *sensitive* personal information under the same conditions. Posed in general terms the problem is that of capturing, in a principled and illuminating way, the notion of *normative consequence.* If there is to be a logic of norms, this is certainly one of its basic problems.

It is generally agreed, however, that the *normative* consequences of an obligation or duty cannot be identified with the *logical* consequences of *fulfilling* it. Ross is usually given credit for pointing this out—his well known counterexample runs as follows: "If it is obligatory to mail the letter, then it is obligatory to mail it or to burn it". If the two occurrences of 'obligatory' are given the same meaning then the logical consequences of mailing the letter are construed as obligatory in the same sense as mailing the letter is. It follows that burning the letter really does discharge an obligation, since it entails mailing or burning it. The unintuitiveness of this conclusion is evident. It is more plausible therefore, to see the second occurrence of 'obligatory' simply as a marker for a consequence of fulfilling the duty to mail the letter—in other words, something is obligatory

R. van der Meyden and L. van der Torre (Eds.): DEON 2008, LNAI 5076, pp. 174–188, 2008.

in the second sense given that it is *true* upon fulfillment of a duty proper. On this reading Ross' example does not assert anything about the deontic status of burning the letter. What it *does* say is that burning the letter will *make true* something that is also *made true* if the letter is posted. I. e. if I ever fulfill my obligation to mail the letter, it will at that point be *true* that I either mail it or burn it [1]. One may therefore be tempted to say that logically entailed items are not *genuine* duties and may therefore safely be ignored.

Consider now another candidate notion of a derived norm; namely one obtained by chaining. Chains of norms are frequent in Law. It is invariably the case, for instance, that criminal law is linked with other parts of law, say administrative law, in the sense that legal offenders are prevented from seeking higher office, and/or are barred from certain learning institutions educating and licensing government officials. According to the Norwegian Criminal Code, for instance, if a person kills another man then he or she ought to be sentenced to imprisonment. The Norwegian Police Act in turn states that only applicants who do not have a prior sentence ought to be admitted to the police academy. The question, then, is; should we construe Norwegian law as containing a norm barring anyone who is guilty of manslaughter from the police academy?

The way we have posed this question, we are not assuming that the applicant has *in fact* been sentenced, only that he *ought* to be. The intuitiveness of the proposed chaining principle, therefore, turns on the plausibility of construing the *obligatoriness* of one state of affairs as a sufficient condition for the *obligatoriness* of another—in this case; the *obligatoriness* of criminal punishment and the *obligatoriness* of rejecting the applicant on which the punishment ought to be inflicted. The validity of this form of inference has been denied by several writers; Sven Ove Hansson, for instance, argues as follows:[1] Suppose that for some reason you are morally required to come to a conference. You are also required not to come unannounced. Let p denote that you stay away from the conference and q that you give notice you will come. Then $\mathcal{O}(\neg p \rightarrow q)$ and $\mathcal{O}\neg p$ both hold, but since you should not notify unless you come, $\mathcal{O}q$ does not hold [2, p. 155]. Although this argument is not crystal clear, what Hansson wants to say, I think is that, if you are required to go to a conference, *and you don't*, then you are under no obligation to give notice that you're coming. Hansson's argument straightforwardly extends to our motivating example: If a person ought to be sentenced for manslaughter, but *in fact* he is not, then the police academy is under no obligation to reject his application. McLaughlin, to which Hansson refers, makes essentially the same point [9, p. 400]. What makes both example tick, though, is that something is pictured as going amiss. It seems odd to say that one is under an obligation to give notice of coming to a conference one isn't going to go to—agreed. However, not going to the conference is already in breach of an obligation, and it is this transgression that is responsible for the sense of oddity. Thus, the example makes an *assumption of fact* not implicit in the norm itself. Take away such extra assumptions—i.e. picture a world as good as can be—and a case against chaining becomes difficult to establish. Indeed, associating eligibility of police academy applicants with the absence

[1] Thanks to one of the reviewers for bringing this example to my attention.

of a prior sentence would not have much of a point were it not to say that *if all goes according to plan* evil-doers in the eyes of the law will not be policemen.

What I am advocating, then, is the *prima facie* plausibility of chaining. In the absence of evidence to the contrary—i.e. in the absence of defeating circumstances—chaining should be accepted. What would otherwise be the point of making the fulfilment of one obligation a condition for the application of another (which is, after all, a standard technology of legal drafting)? Generally speaking, a legal order, or normative system may be said to envisage an *ideal* state of things where no obligation is ever neglected or unfulfillable. Such a description is almost certainly false, in the sense that it does not agree completely with reality [10]. There are limits on what it is possible for people to do and not to do, and it is perfectly common for people to ignore a standard of correct behaviour. Such factors, will certainly have consequences for which norms can plausibly be said to be *operative* or *in force*—one cannot return a book by March 18 on the 19th of March the same year. It is folly, therefore, to deny that norms interact with facts. Nevertheless, the optimum described by the norms themselves *sans* extra assumptions of fact is the standard that conduct should be measured against. The first question of normative reasoning, then, is, in my opinion: What are the properties of this optimum? How does the world look according to the description implicit in the norms themselves? Answering this question, I think, requires that we treat the initial obligations, as well as all subsequently inferred ones *as if* they were true, in a sequence of steps that fleshes out the theory of ideality. Of course, the theory of ideality needs, at some later stage, to be supplemented by a theory of what turns a *prima facie* obligation into an operative duty, true, but that is secondary. The very concept of failure or deviation refers essentially to the standard of which something falls short. That standard, then, is the primary object of study.

Returning now to the Ross problem and the status of items entailed by obligatory states of affairs, it is not too difficult to see that we were to hasty when we dismissed these items as irrelevant to the notion of normative consequence. The reason is that chaining is often facilitated by information from the fulfillment context (by which I shall mean the logical closure of the obligatory propositions in question). To bring this forth more clearly, let's regiment the motivating example. Put;

- $m = i$ is convicted of manslaughter
- $s = i$ is sentenced to imprisonment, and
- $r = i$ is inadmissible to the police academy

If, for the time being, we represent the norms simply as pairs and construe the problem as one of moving from (m, s) and (s, r) to (m, r), then all we need is the transitivity of the relation to which the pairs belong. However, the situation is usually less clear cut than that. The punishments of criminal law are usually specific—at least within a reasonable range— about the form, duration and severity of the sanction, whereas the rules for admission to an institution such as a police academy are not. Say the relevant provision in the criminal code states that 'he who kills another man shall be sentenced for manslaughter to imprisonment from 5 years and into lifetime'. Clearly, being sentenced to imprisonment from 5 years into lifetime implies being sentenced to imprisonment, so intuitively the provision should still make any murderer inadmissible for the police academy. Something more is involved in this latter inference than plain old transitivity though. Extracting the two respective patterns, we have;

Transitivity: From (m, s) and (s, r) infer (m, r), and

Mediated transitivity: From (m, s') and (s, r) infer (m, r) whenever $s' \vdash s$

The two forms differ insofar as the second but not the first utilizes logical entailment as a stepping-stone to complete the inference. Is there any reason we should accept the first, but not the second as valid? I think not. If the obligation to see to it that s' is *fulfilled* then s is *true*. But, if s is true, then the norm (s, r) is triggered. The point, then, is that upon fulfillment of one obligation the condition for the applicability of another may often be inferred. If that is the case then we should be able to move from the former *via the consequences of its fulfillment* to the latter. Stated differently, the context of fulfillment of one obligation may, and frequently does, provide information about which other obligations are applicable. We should be able to utilize this information in an iterative detachment of applicable duties. Fleshing out the theory of ideality requires that we treat the initial obligation and all subsequently inferred ones *as if* they were fulfilled. Whenever we do, we are relating hypothetically to a context of *truth*, and should therefore be allowed to avail ourselves of the full power of classical logic. There is a tangible tension, then, between the need to avoid the Ross problem, on the one hand, whilst reserving the right to reason iteratively on the other: As norms are frequently chainable (mediately), we ought to be able to reason classically about fulfillment contexts—so logical closure is a feature. On the other hand, as the Ross example shows, logical closure is bound to introduce an infinity of elements that intuitively should not be given the status of obligations—so it is a bug. Seen in this light, the closure problem is more of a problem than is usually conceded—it is the problem of keeping what one has to give up. This metaphor is perhaps not quite apt, though, since the key issue, described with a bit more precision, is really *when*, or for what purposes or uses, to retain vs. *when*, or for what purposes or uses, to discard information about the fulfillment context. These 'whens' should not be run together lest the problem become a dilemma.

We should forestall one important objection to this analysis: One may argue that what is essential to a logic of norms is not so much how we identify the set of operative obligations, but rather how we devise a criterion to tell when an obligation is met. At the end of the day the central question for deontic logic is 'are things as they should be?'. If we can answer that question, then it may seem that nothing remains to be said. Hence, closure under logical consequence may appear to be only a harmless limiting case, for when an obligation is fulfilled then so are all of its logical consequences. Perhaps, therefore, we need not worry about whether these consequences are *really* duties - i. e. perhaps the problem of keeping what one gives up is just a pseudo-problem and the tension only apparent. An argument of Carmo and Jones [3] suffices to refute this objection, I think: Obligations are essentially *violatable* entities, they say. It is precisely when the possibility of norm violation is kept open that deontic logic has a potentially useful role to play [3, p. 261]. Hence, we should be able to say not only when an obligation is fulfilled but also when it is transgressed. The closure problem, Carmo and Jones argue, should therefore also be seen from the point of view of violation. But then its air of harmlessness vanishes, for if I am under an obligation to see to it that p, and I don't, then, if we accept that the logical consequences of p are obligatory in the same sense as p is, we must conclude that I have *violated* $p \vee q$ whenever q happens to be false (whatever the reason may be), but *complied* with $p \vee q$ whenever q happens to be true (whatever the reason may be). In other words if the concept of obligatoriness is construed as closed under logical consequence, then, upon *violation* of a primary duty, everything is partitioned into observed and transgressed in a philosophically random way. Ergo, the Ross problem is not harmless, and the problem of *when* to keep it vs. *when* to give it up is a live issue.

2 A Solution in an Input/Output-ish Idiom

Recapitulating briefly, the task we have set for ourselves is essentially that of formalizing a plausible notion of normative consequence, which in turn is the problem of deriving implicit norms from explicitly given ones. Any such relation should, *prima facie*, be closed under mediated chaining, reflecting the fact that fulfillment contexts in general contain information that partly determines which norms are applicable. The relation should not, however, link an obligation proper to all classically entailed items, since it would then multiply, in an intuitively random way, the number of items the code renders violated. There is no *a priori* reason why this problem could not be solved in modal logic, say, by adopting a modal operator closed under logical equivalence only, and adding the principle of mediated transitivity as an axiom or rule of inference.[2] Nevertheless, my idiom of choice in this study is input/output logic (as set out in a series of papers by Makinson and van der Torre [5], [6], [7]). With a new idiom often comes an opportunity to view old questions in a new light. I do not believe that there is any overriding reason why we should feel compelled to encode everything we wish to say about a system of norms in terms of truth-sets and relations between them. On the contrary, if such a strategy is pursued too rigorously, one runs the risk, I think, of reducing the dynamics of a *system* to the static properties of *sets*, thereby distorting an otherwise reasonably clear picture. The case in hand constitutes one example (a modest one, but nevertheless): The problem of *when* to retain vs. *when* to discard information, is most naturally solved, I think, in terms describing how information is animated in different ways by different components of a system. Instead of talking about what's true where, therefore, we view ourselves more as talking about the *uses* to which information may be

[2] This was pointed out to me by one of the reviewers.

put and *how*. Input/output logic seems well-suited to this task, and an idiom worth exploring.

A bit of context first: In input/output logic a norm is just a pair of boolean formulae. In other words a norm (a, x) is a *logically arbitrary* stipulation connecting an *input* a with an *output* x—it is logically arbitrary in the sense that a pair is not a formula, so there is nothing to the norm (a, x) above and beyond the fact that some authority requires that x be done given a. Notably, since the pair as such has no logic, the contrapositive of a norm is not necessarily a norm, nor is any pair of the form (a, a). A normative code G is seen simply a set of such pairs, from whence it follows that the explicitly declared mandates, in any given situation a according to G, can be obtained by taking the image of G under a. The basic notion of a normative *system* allows implicit norms to be derived from the explicit ones—i.e. from the ones contained in G —by accepting logical consequences of inputs as inputs, and logical consequences of outputs as outputs. In other words, where L is a boolean language, the basic model of a normative system in input/output logic, is an operation $out_1 : 2^{L^2} \times L \mapsto 2^L$ defined as follows:

Definition 1. $out_1(G, a) = Cn(G(Cn(a)))$.

As Makinson and van der Torre observe, putting $(a, x) \in out(G)$ iff $x \in out(G, a)$, one may also construe the operator out_1 as a closure operator on a set of norms. As they say, the two formulations give a rather different gestalt, and one is sometimes more convenient than the other. Since all the operators I shall introduce are of the same type, I will switch freely between these two modes of expression. A basic result of [5] is that out_1 is characterized by the following system of axioms and rules

Definition 2. *$(a, x) \in deriv_1(G)$ iff (a, x) is derivable from axioms $(t, t) \cup G$ by the rules of inference,*

$$SI \; \frac{(b,x)}{(a,x)} \; if \; a \vdash b \quad AND \; \frac{(a,z),(a,y)}{(a,z \wedge y)} \quad WO \; \frac{(a,z)}{(a,x)} \; if \; z \vdash x$$

There are various ways to modify the definition of the out_1-operation, yielding other systems satisfying more rules. One of particular interest, given our concerns, is the version that allows outputs to be recycled as inputs. Adhering to Makinson and van der Torres' typology, this is the operation out_3:

Definition 3. *$x \in out_3(G, a)$ iff $x \in \bigcap\{Cn(G(B)) : a \in B = Cn(B) \supseteq G(B)\}$.*

Recycling is thus effected by taking the intersection of the outputs of out_1 for all the respective sets that include the input as well as the image of the set under G. This yields the least operation closed under chaining. In other words out_3 is characterized by the following system:

Definition 4. *$(a, x) \in deriv_3(G)$ iff (a, x) is derivable in the system containing all the axioms and rules of $deriv_1$, plus the rule of cumulative transitivity;*

$$CT\ \frac{(a,x),\ (a \wedge x, y)}{(a,y)}$$

Outputs from out_3 cannot straightforwardly be interpreted as obligations, since out_3 validates output weakening.[3] Nevertheless, weakening of the output is often involved in the chaining process, I have argued, for instance in derivations such as (where we assume that $x' \vdash x$):

$$CT\ \dfrac{WO\ \dfrac{(a,x')}{(a,x)} \qquad (a \wedge x, y)}{(a,y)}$$

The task that confronts us then, is essentially how to preserve this kind of inference—as it instantiates the pattern of mediated transitivity—whilst not allowing weakening of outputs in general. Stated differently, we want to disallow weakening of outputs in the *last* step of a derivation, whilst reserving the right to connect the output of one norm to the input of another via the logical consequences of the former. It seems natural therefore, to replace the rule of cumulative transitivity with one that, for want of a better name, I have chosen to dub mediated cumulative transitivity:

$$MCT\ \frac{(a,x') \qquad (a \wedge x, y)}{(a,y)}\ \ \textit{whenever } x' \vdash x$$

Mediated cumulative transitivity *simulates* weakening of the output in the sense that it itself makes logical entailment sufficient for chaining. Our question now becomes: Can we find an intuitively plausible semantics that yields a system which has MCT but not WO? It may be helpful to make a detour through the out_1-idiom to get a hint as to what is required. Consider the following definition

Definition 5. *$x \in PN_1(G, a)$ iff x is equivalent to a subset of $G(Cn(a))$*

This is just like out_1 except that the output is no longer closed under logical consequence. Instead we require that an output be equivalent to the set of heads of a subset of the *explicitly* given norms, since these norms can all safely be assumed to be *genuine*. Hence, we think of the operation PN_1 as picking out the *proper norms* of G—intuitively the set of norms that stipulate *genuine, violatable* duties. It is not difficult to give this notion a syntactical characterization:

Definition 6. *$(a, x) \in IN_1(G)$ iff (a, x) is derivable from axioms $(t, t) \cup G$ by the rules of inference SI, AND and*

$$Eq\quad \frac{(a,x')}{(a,x)}\quad \textit{whenever } x' \equiv x$$

[3] The same goes for out_1 of course.

The letters IN are meant to stand for *implicit norms*—a notion which we think of as the syntactical counterpart of the more semantically tainted notion of a proper norm. Note that the rule Eq is new to the system. It allows us to pass from any derived rule to a new rule with the same body and an equivalent head. Due to weakening of the output Eq is a derived rule of $deriv_1$. Since weakening of the output is not a rule of IN_1, however, Eq must be added. Now, the following theorem—which is just a suitably modified version of observation 1 from [5]—shows that the set of proper norms and the set of implicit norms, as the two notions have so far been defined, coincide.

Theorem 1. *$PN_1(G) = IN_1(G)$.*

Proof (Sketch). The right-in-left inclusion is straightforward, so we prove the converse only: Suppose x is equivalent to a subset of $G(Cn(a))$. Then by compactness and monotony for classical consequence there is a finite set $x_1, \ldots, x_n \in G(Cn(a))$ with $x \equiv x_1 \wedge \ldots \wedge x_n$. If $n = 0$ then x is a tautology so $x \in IN_1(G, a)$ by axiom (t, t) and SI. If $n \neq 0$ then for each x_i there is a $b_i \in Cn(a)$ with $(b_i, x_i) \in G$. Since $a \vdash b_i$ for each i we may apply SI to each (b_i, x_i) to obtain (a, x_i). We may further collect all the latter rules by AND obtaining $(a, \bigwedge_{i=1}^{n} x_i)$, before deriving (a, x) by one application of Eq, as desired.

The case for recycling (*sans* output weakening) is considerably more complex, however, as we need to find a way of pumping outputs back as inputs. Intersecting images in the manner of definition 3 may work—if we close the images under logical equivalence—but the following inductive definition gives a more illuminating picture, I think:

Definition 7. *$x \in PN_3(G, a)$ iff x is equivalent to a subset of $\bigcup_{i=0}^{\omega} A_i$ where*

- *$A_0 = G(Cn(a))$, and*
- *$A_{n+1} = A_n \cup G(Cn(A_n \cup \{a\}))$.*

This semantics gives a concrete example of turning away from truth-talk towards the *uses* to which information may be put: When a norm is used to produce an output, then its consequent—i.e. what the norm decrees to be ideal or obligatory—is dissociated from logically weaker items so that its normative force, so to speak, does not extend to items that are merely true upon fulfillment. Hence all obligations generated are *genuine* in the sense that they correspond to accumulations of *explicitly* given duties pertaining to the circumstances. On the other hand, when an obligation is pumped back into the input, we reserve the right to consider the context of fulfillment as such, in order to determine which other norms are applicable. Thus, the obligation in question is used in different ways depending on its current coordinates in the wider geography of the system. The overall behaviour that results is captured by the system defined below:

Definition 8. *$(a, x) \in IN_3(G)$ iff (a, x) is derivable from axioms $(t, t) \cup G$ by the rules of inference SI, AND, Eq and MCT.*

The proof requires a few lemmata:

Lemma 1. *Let $\{A_i : i < \omega\}$ be any sequence defined according to 7. Then $A_n \subseteq A_{n+1}$ for all $n < \omega$.*

Proof. $A_{n+1} = A_n \cup G(Cn(A_n \cup \{a\})) \supseteq A_n$, so $A_n \subseteq A_{n+1}$ for any n.

When $\{A_i : i < \omega\}$ is the chain defining $PN_3(G, a)$ I shall say that a is an input to the chain, and that the chain is generated by G. We have:

Lemma 2. *Let $\{A_i : i < \omega\}$ and $\{B_i : i < \omega\}$ be sequences generated by G, where a and b are the respective inputs. Then, if $a \in Cn(B_k \cup \{b\})$ for some k, then $A_i \subseteq B_n$ for all i and some n such that $k \leq n$.*

Proof. Proof proceeds by induction on the sequence $\{A_i : i < \omega\}$. For the base case we reason as follows: By assumption $a \in Cn(B_k \cup \{b\})$, so $Cn(a) \subseteq Cn(B_k \cup \{b\})$, by monotony and idempotence for classical consequence, whence $G(Cn(a)) \subseteq G(Cn(B_k \cup \{b\}))$, by the monotony of image-formation. It follows, by general set-theory that $A_0 = G(Cn(a)) \subseteq B_k \cup G(Cn(B_k \cup \{b\})) = B_{k+1}$. For the induction step, suppose that $A_n \subseteq B_p$ for some p such that $k \leq p$. Now, $A_{n+1} = A_n \cup G(Cn(A_n \cup \{a\}))$. Since $B_k \subseteq B_p$, by lemma 1, we have $a \in Cn(B_p \cup \{b\})$. By the induction hypothesis it follows that $A_n \cup \{a\} \subseteq Cn(B_p \cup \{b\})$. Hence $A_n \cup G(Cn(A_n \cup \{a\})) \subseteq B_p \cup G(Cn(B_p \cup \{b\}))$, by the same steps as for the base case. Hence $A_{n+1} \subseteq B_{p+1}$, so the proof is complete.

Note that lemma 2 is monotony in the input in the particular case where $i = k = 0$.

Lemma 3 (Cumulativity in the input). *$PN_3(G, a \wedge b) \subseteq PN_3(G, a)$ whenever $b' \in out(G, a)$ for $b' \vdash b$.*

Proof. Let $PN_3(G, a \wedge b)$ and $PN_3(G, a)$ be defined by $\{A_i : i < \omega\}$ and $\{B_i : i < \omega\}$ respectively. It suffices to show that $a \wedge b \in Cn(B_k \cup \{a\})$ for some k, because then we have $\bigcup_{i=0}^{\omega} A_i \subseteq \bigcup_{i=k}^{\omega} B_i$, by lemma 2, whence $\bigcup_{i=0}^{\omega} A_i \subseteq \bigcup_{i=0}^{\omega} B_i$ since $\bigcup_{i=k}^{\omega} B_i \subseteq \bigcup_{i=0}^{\omega} B_i$. By assumption $b' \in PN_3(G, a)$, i. e. b' is equivalent to a subset of $\bigcup_{i=0}^{\omega} B_i$. By compactness and monotony for classical consequence there is thus a finite set $b_1, \ldots, b_i \subseteq \bigcup_{i=0}^{\omega} B_i$ such that $b_1 \wedge \ldots \wedge b_i \equiv b'$. Let B_k be the set such that $b_1, \ldots, b_i \subseteq B_k$. Since $b_1, \ldots, b_i$ is finite k exists. Now, $b' \vdash b$ by assumption, so $b \in Cn(B_k) \subseteq Cn(B_k \cup \{a\})$, whence $a \wedge b \in Cn(B_k \cup \{a\})$ as desired.

These lemmata will suffice to establish soundness:

Theorem 2. *$IN_3(G) \subseteq PN_3(G)$, i. e. all implicit norms are proper norms.*

Proof. Suppose that $PN_3(G, a)$ is determined by the chain $\{A_i : i < \omega\}$. We prove, by induction on the length of the derivation, that $x \in PN_3(G, a)$ whenever $(a, x) \in IN_3(G, a)$. In the base case (a, x) is an axiom, i. e. $(a, x) \in G$ or $a \equiv x \equiv t$. If $(a, x) \in G$ then $x \in G(Cn(a)) = A_0$. If, on the other hand,

$a \equiv x \equiv t$, then $Cn(x) = Cn(\emptyset)$ and, clearly, $\emptyset \subseteq G(Cn(a)) = A_0$. In both cases, therefore, we have that x is equivalent to a subset of $\bigcup_{i=0}^{\omega} A_i$, so we are done. For the induction step, suppose the theorem holds for shorter derivations, then: For EQ, suppose (a, x) is derived from (a, x') by EQ. Then $x \equiv x'$. By the induction hypothesis $x' \in PN_3(G, a)$. Hence, x' is equivalent to a subset of $\bigcup_{i=0}^{n} A_i$, whence, since logical equivalence is transitive, $x \in PN_3(G, a)$ as desired. For SI, suppose (a, x) is derived from (b, x) by SI. Then $a \vdash b$. By the induction hypothesis $x \in PN_3(G, b) \subseteq PN_3(G, a)$, by monotony in the input, so we are done. For AND, suppose (a, x) is derived from (a, z) and (a, y) by AND. Then, by the induction hypothesis, we have $y, z \in PN_3(G, a)$. Hence $x \wedge y \in PN_3(G, a)$ since $x \wedge y$ is equivalent to $\{x, y\}$. For MCT, Suppose (a, x) is derived from (a, y') and $(a \wedge y, x)$ by MCT. By the induction hypothesis we have $x \in PN_3(G, a \wedge y)$ and $y' \in PN_3(G, a)$. Since $y' \vdash y$, we may apply cumulativity in the input to obtain $PN_3(G, a \wedge b) \subseteq PN_3(G, a)$, so $x \in PN_3(G, a)$ as desired. This completes the proof.

Theorem 3. *$PN_3(G) \subseteq IN_3(G)$, i. e. all proper norms are implicit norms.*

Proof. Suppose $PN_3(G, a)$ is defined by the sequence $\{A_i : i < \omega\}$. First we prove that $\bigcup_{i=0}^{\omega} A_i \subseteq IN_3(G, a)$. Proof proceeds by induction on A_n. For the base case, if $x \in G(Cn(a))$ then there is a rule $(b, x) \in G$ such that $a \vdash b$. Since (b, x) is an axiom in $IN_3(G)$ it follows that x in $IN_3(G, a)$ by SI. For the induction step suppose that $A_{k-1} \subseteq IN_3(G, a)$. We need to show that $A_k = A_{k-1} \cup G(Cn(A_{k-1} \cup \{a\})) \subseteq IN_3(G, a)$. Suppose therefore that $x \in A_k$. Then $x \in A_{k-1}$ or $x \in G(Cn(A_{k-1} \cup \{a\}))$. The first case is covered by the induction hypothesis. Hence we may assume that $x \in G(Cn(A_{k-1} \cup \{a\}))$. If x is a tautology then we immediately have $x \in IN_3(G, a)$ by axiom (t, t). If not, then there is a rule $(b, x) \in G$ such that $b \in Cn(A_{k-1} \cup \{a\})$. By compactness for classical consequence there is thus a finite subset $b_1, ..., b_i$ of A_{k-1} with $b \in Cn(\{b_1, \ldots, b_i\} \cup \{a\})$, whence $\bigwedge_{j=1}^{i} b_j \vdash a \to b$. By the induction hypothesis, each $b_j \in A_{k-1}$ is such that $b_j \in IN_3(G, a)$. Hence, we have the following derivation:

$$MCT\ \dfrac{AND\ \dfrac{(a, b_1), \ldots, (a, b_i)}{(a, \bigwedge_{j=1}^{i} b_j)} \qquad SI\ \dfrac{SI\ \dfrac{(b, x)}{(a \wedge b, x)}}{(a \wedge a \to b, x)}}{(a, x)}$$

Hence, $\bigcup_{i=0}^{\omega} A_i \subseteq IN_3(G, a)$. It remains to show that if x is equivalent to a subset of $\bigcup_{i=0}^{\omega} A_i$, say $x \equiv \bigwedge_{i=1}^{n} x_i$, then $x \in IN_3(G, a)$. By the induction $x_j \in IN_3(G, a)$ for any $1 \leq j \leq n$, so $(a, x_1), \ldots (a, x_n) \in IN_3(G)$. Hence,

$$AND\ \dfrac{(a, x_1), \ldots, (a, x_n)}{Eq\ \dfrac{(a, \bigwedge_{i=1}^{n} x_i)}{(a, x)}}$$

Summing up, I have considered two different notions of a derived norm; norms obtained by chaining and norms obtained by weakening of the obligatory proposition in question. Only the former of these, I have argued, should be recognized as giving us a genuine notion of normative implicature. The need to retain the former whilst discarding the latter initially pull in opposite directions, but a balance can be struck by introducing the rule of mediated cumulative transitivity which simulates output weakening for the purposes of chaining. The problem of when to keep vs. when to discard information is thereby given a solution—surely not the only one conceivable—which has the additional virtue, in my opinion, of giving the notion of consequence an *operational* semantics that reconstructs the reasoning process in terms of a sequence of discrete computational steps.

3 Further Prospects: Amplification of Output

Throwing a third concept into the pot, consider now norms derived in conjunction with *material dependencies*.[4] Say, for the sake of argument, that doctor-assisted deaths are not approved of by a given system—i. e. killing another person is never permitted. In other words we assume that $(t, \neg kill)$ is an explicitly given norm in the system in question. Assume further that turning off a respirator will kill a certain patient i. Needless to say, this is just a contingent fact about i's physical constitution, and not a law of logic. Turning the respirator off would not be fatal to the patient if he or she were, say, rolled over to a sunny spot and could sustain his vital functions by photosynthesis. Since the system in question is categorical about not killing it seems plausible to say that it also forbids turning off the respirator. Let $o \rightarrow kill$ stand for the fact that turning off the respirator kills i. The question now becomes; how do we bring that information to bear on the norms of the system? Where should that information go? Obviously, the conditional $o \rightarrow kill$ is not a norm. It is not something the system prescribes, or decrees to be ideal. Thus, putting $(t, o \rightarrow kill)$ or $(o, kill)$ in the code itself—even though the former *would* give us the desired inference—is out, as it distorts the intuitive picture. A more principled, I think, solution is not far to seek though. Consider again the rule of output weakening; from (a, x) to (a, y) $\underline{\textit{whenever } x \vdash y}$. It is natural to view the underlined side-constraint, or auxiliary hypotheses, as an *environment* in which the system of norms in question is currently deployed. The rule-of-inference format has the virtue of factoring out this environment, making the rule plug-and-play compatible with different ones. In our motivating example, the conditional $o \rightarrow kill$ belongs to the environment—it is true when the norms of the system are applied. Hence it should be regarded as an *amplifier* for the side-constraint—something we should be allowed to take into consideration when computing what should or should not be done. In other words, a solution that seems principled enough, is to replace classical consequence with some kind of supraclassical consequence in the auxiliary hypothesis of the rule; from (a, x) to (a, y) $\underline{\textit{whenever } x \vdash_k y}$, let's call the latter rule WO^K. The standard rule of output weakening WO becomes, then,

[4] In [3] Henry Prakken is given credit for identifying this problem.

just a limiting case of the amplified rule WO^K—the case where K is empty, i. e. where nothing is known about the context of deployment. Assume for the moment—since these are simple to work with, and will do well enough for our purposes—that the supraclassical operator in question is a *pivotal-assumption* operator, in the terminology of [8]. In other words, $x \vdash_K y$ iff $x \cup K \vdash y$, or in terms of closure operators, $Cn_K(x) = Cn(K \cup x)$. It seems intuitively clear that amplifying the rule of output weakening corresponds to taking the closure under Cn_K of the output of a given operation, viz.:

Definition 9. *Put* $x \in out_3^K(G,a)$ *iff* $x \in \bigcap\{Cn_K(G(B)) : a \in B = Cn(B) \supseteq G(B)\}$.

Theorem 4. *Let* $deriv_3^K(G)$ *be exactly like* $deriv_3(G)$ *except that* WO *is replaced with* WO^K. *Then* $deriv_3^K(G) = out_3^K(G)$.

The proof is a straightforward modification of the completeness theorem for out_3 wrt. $deriv_3$. We leave it out of the text to keep the paper at a reasonable length.

It is worth pausing to note that $deriv_3^K$ can also be defined by leaving the rule of output weakening unaltered whilst incorporating an axiom (t,y) for each $y \in K$. For suppose $y \in K$ then (t,y) is derivable from (t,t) by one application of WO^K. Hence any system that has WO^K has (t,y). Conversely, suppose that $x \cup K \vdash z$. By compactness for logical consequence, then, there is a finite set $k_1,\dots,k_n \in K$ such that $x \wedge \bigwedge_{i=1}^n k_i \vdash z$. For each k_i we have assumed that there is a corresponding axiom (t,k_i) so we have:

$$WO\,\dfrac{AND\,\dfrac{SI\,\dfrac{AND\,\dfrac{(t,k_1),\dots,(t,k_n)}{(t,\bigwedge_{i=1}^n k_i)}}{(a,\bigwedge_{i=1}^n k_i)} \qquad (a,x)}{(a,x \wedge \bigwedge_{i=1}^n k_i)}}{(a,z)}$$

Hence any system that has (t,y) for each $y \in K$ has WO^K as a derived rule.

Amplification is indeed one way of utilizing background information, for suppose $o \rightarrow kill$ is true in the context described by K—i. e. suppose that $o \rightarrow kill \in K$. Then;

$$WO^K\,\dfrac{(t,\neg kill)}{(t,\neg o)}$$

Since $\neg kill \vdash_K \neg o$. Alternatively, by the equivalence of contextual axioms:

$$WO\,\dfrac{AND\,\dfrac{(t,\neg kill) \qquad WO\,\dfrac{(t,o \rightarrow kill)}{(t,\neg kill \rightarrow \neg o)}}{(t,\neg kill \wedge (\neg kill \rightarrow \neg o))}}{(t,\neg o)}$$

Alas, what was pushed out the door now comes back in through the window. Amplifying output *in this way* is clearly not coherent with our earlier efforts to shake the Ross problem, since pivotal-assumption consequence is supraclassical. This adds a new and interesting twist to the problem of keeping what one gives up. It can now be seen to comprise the additional problem of how to facilitate material inferences in spite of the fact that logical entailment is not available. A natural way to compromise between the two, is to amplify the system of *proper* norms. As before, then, we substitute logical equivalence for logical consequence, thereby avoiding the Ross problem, but this time we define equivalence *modulo the environment*, thereby absorbing information about the context of deployment:

Definition 10. *Put $x \in PN_3^K(G,a)$ iff x is equivalent, modulo K, to a subset of $\bigcup_{i=0}^{\omega} A_i$ where the chain $\{A_i : i \leq \omega\}$ is defined exactly as in definition 7.*

Theorem 5. *Let $IN_3^K(G)$ be exactly like $IN_3(G)$ except that Eq is replaced with, Eq^K; from (a,x) to (a,y)* whenever *$y \equiv_K x$, i. e. whenever x and y are equivalent modulo K. Then $IN_3^K(G) = PN_3^K(G)$.*

Again, the proof is a simple re-run and is left out. Obviously, the concept of material inference captured by this system is rather weak. Nevertheless, it solves the motivating example fairly well: Since $\neg kill \wedge (o \rightarrow kill) \equiv \neg kill \wedge \neg o$ we have;

$$Eq^K \, \frac{(t, \neg kill)}{(t, \neg kill \wedge \neg o)}$$

or, since Eq^K too may be traded for axioms describing the context:

$$\begin{array}{rc} AND & \dfrac{(t, \neg kill) \qquad (t, o \rightarrow kill)}{(t, \neg kill \wedge (o \rightarrow kill))} \\ Eq & \dfrac{}{(t, \neg kill \wedge (\neg kill \rightarrow \neg o))} \\ Eq & \dfrac{}{(t, \neg kill \wedge \neg o)} \end{array}$$

The Ross problem does not arise, since the system does not have a rule of output weakening. A word of criticism could perhaps be directed at the fact that the norm we are able to infer is $(t, \neg kill \wedge \neg o)$, *not* $(t, \neg o)$, and that the latter cannot be derived from the former. In other words, leaving the respirator on is not mandatory unless one simultaneously refrains from killing the patient. Nevertheless, if we recognize the existence of a proper norm $(t, \neg kill \wedge \neg o)$ then we do have sufficient warrant to say that the system in question prohibits turning off the respirator. Consider the following definition of *norm violation*:

Definition 11. *b violates (a,x) iff $b \vdash a \wedge \neg x$ and (a,x) is a proper norm.*

The definition seems intuitive enough, once the norm acting as the standard of conduct is a *proper* norm. It follows immediately from the definition that o violates $(t, \neg kill \wedge \neg o)$ and is therefore prohibited by the code. Another potentially problematic feature, is that contextual information becomes encoded in norms. If

$o \to kill$ is true in the context of deployment, then the amplified system contains $(t, o \to kill)$. Doesn't this go against our initial determination not to represent $o \to kill$ as something decreed by the system? Note that if it is decreed by the system, then, presumably, it is also violatable. But should we say that a norm has been violated if the respirator is turned off and the patient survives? The proper response to these worries, I think, is to point to the difference between incorporating a norm $(t, o \to kill)$ in the code itself vs. having it emerge from context. In the former case, the norm becomes an invariant feature of the system itself, whereas in the latter case it is part of a plug-and-playable module that may vary from one application of the system to the next. Intuitively the 'norms' belonging to this module are not violatable since if, say, $o \wedge \neg kill$ is true, then one is no longer in a context where $o \to kill$ holds, so the system no longer yields $(t, o \to kill)$. In other words it makes no sense to suppose that the system contains norms describing the context, which are nevertheless not fulfilled.

4 Summary

There are at least three candidate notions of a derived norm that deserve serious study; norms derived by logical consequence, norms obtained by chaining, and norms generated by material dependencies. I have argued that we should accept the latter two as presenting us with genuine notions of normative implicature, but reject the first. This is not easy to accomplish, since these needs pull in different directions so to speak. Chaining and material inference clearly require that *some* notion of logical inference be available, whilst avoiding the Ross problem depends on logical inference being ignored. The rather delicate balancing of needs required to solve the problem is conveniently expressed, and illuminatingly represented, I think, by turning towards the dynamics of systems. Progress can be made if we decompose a system into a code and a context, and look at how these sources of information interact.

Acknowledgements

The author wishes to thank Dag Elgesem, Andrew Jones and Michal Walicki for helpful comments. Thanks also to the anonymous reviewers for their observations and criticism.

References

1. Brown, M.A.: Conditional obligation and positive permission for agents in time. Nordic Journal of Philsophical Logic 5(2), 83–111 (2000)
2. Hansson, S.O.: The Structure of Values and Norms. Cambridge University Press, Cambridge (2001)
3. Carmo, J., Jones, A.J.I.: Deontic logic and contrary-to-duties. In: Gabbay, D., Guenthner, F. (eds.) Handbook of Philsophical Logic, vol. 8, pp. 263–265. Kluwer Academic Publishers, Dordrecht (2002)

4. Makinson, D.: Bridges Between Classical and Nonmonotonic Logic. Texts in Computing, vol. 5. King's College London Publications (2005)
5. Makinson, D., van der Torre, L.: Input-output logics. Journal of Philosophical Logic 30, 155–185 (2001)
6. Makinson, D., van der Torre, L.: Constraints for input-output logics. Journal of Philosophical Logic 32(4), 391–416 (2003)
7. Makinson, D., van der Torre, L.: What is input/output logic? In: Foundations of the Formal Sciences II: Applications of Mathematical Logic in Philosophy and Linguistics. Trend in Logic, vol. 17. Kluwer, Dordrecht (2003)
8. Makinson, D.: Bridges Between Classical and Nonmonotonic Reasoning. Texts in Computing, vol. 5. Kings' College London Publications (2005)
9. McLaughlin, R.N.: Further problems of derived obligation. Mind, new series 4(64), 400–402 (1955)
10. von Wright, G.H.: Is and Ought. In: Paulson, Paulson (eds.) Normativity and Norms, pp. 365–383. Clarendon Press, Oxford (1998)

On the Strong Completeness of Åqvist's Dyadic Deontic Logic G

Xavier Parent

54 avenue de l'Elisa, 83100 Toulon, France
Xavier.Parent@univ-provence.fr

Abstract. Åqvist's dyadic deontic logic **G**, which aims at providing an axiomatic characterization of Hansson's seminal system DSDL3 for conditional obligation, is shown to be strongly complete with respect to its intended modelling.

Keywords: Conditional obligation, preference-based semantics, strong completeness, DSDL3.

1 Introduction

The present study is mainly concerned with so-called preferential semantics for conditional obligation. These rely on a binary relation, which ranks all possible worlds in terms of comparative goodness or betterness. Structures of this sort seem to have made their first explicit appearance in print with the paper of Hansson [1]. There they are used to give a semantic analysis of contrary-to-duty (or secondary) obligations, which tell us what comes into force when some other (primary) obligations are violated. A number of researchers have followed Hansson's suggestion, providing a more comprehensive investigation of the treatment of contrary-to-duty obligations within a preference-based approach. It is not the purpose of this paper to evaluate such a treatment. The interested reader should consult the relevant literature (see, e.g., [2,3,4,5,6,7,8]).

In what follows, I shall focus on another long-standing problem, that of axiomatizing the logic of conditional obligation as outlined by Hansson in the aforementioned pioneering paper.[1] An important step towards resolving such an issue has been taken by Spohn [9]. There the focus is on the class of models corresponding to the system known as DSDL3, which Hansson wished to be regarded as his 'official' one. An axiomatic characterization of the logic is given, and proved semantically complete with respect to the model theory, in the sense that every formula of this calculus is shown to be provable if and only if it is valid. Metatheorems of this sort are frequently called *weak completeness theorems*— the object of the present paper is to extend Spohn's result to obtain a *strong* completeness theorem for dyadic deontic logic; i.e., I will show that a formula A of

[1] The systems proposed by Hansson (he confidently calls them 'dyadic standard systems of deontic' - DSDL) are purely semantical.

R. van der Meyden and L. van der Torre (Eds.): DEON 2008, LNAI 5076, pp. 189–202, 2008.

this calculus can be deduced from a (possibly infinite) set Γ of formulae if and only if Γ entails A. Reference will be made to Åqvist's axiomatic system **G** (see, e.g., [10,11]). It is essentially a reformulation of the Hansson-Spohn calculus in terms of modal logic. Unless I am mistaken, the strong completeness problem for **G** has not been settled yet. It will here be answered in the affirmative. Moreover, it will be shown that (as conjectured by Åqvist himself) **G** remains complete if the assumption of a linear or total ordering among possible worlds is dropped. It is the assumption that any pair of possible worlds are mutually comparable under the betterness relation: either one is better than the other, or they are of equal value. The fact that such an assumption does no work was already known by Spohn, at least for his reconstruction of DSDL3.

The plan of this paper is as follows. In section 2, I present Åqvist's dyadic deontic system **G**, and its associated semantics. Two classes of models will be discussed, one of them corresponding to Hansson's system DSDL3. In section 3, I introduce the notion of a canonical structure, and prove a number of lemmata, which in section 4 will suffice to establish the desired completeness of the system with respect to the two classes of models.

2 Syntax, Semantics and Proof Theory

The language of **G** has, in addition to a set Prop of propositional variables and the usual Boolean sentential connectives, the following characteristic primitive logical connectives : the alethic modal operators $\Box$ (for necessity) and $\Diamond$ (for possibility) ; and the two dyadic deontic operators $\bigcirc(-/-)$ and $P(-/-)$, which may be read as *'It ought to be that ..., given that ...'* and *'It is permitted that ..., given that ...'*, respectively. The set $\mathcal{L}$ of well-formed formulae (wffs) is defined in the usual way. There are no restrictions as to iterations of dyadic deontic operators and modal ones.

The system comes with a possible worlds semantics *à la* Kripke. I begin with the idea of an H-model ('H' is mnemonic for Hansson), by which I understand a structure

$$\mathcal{M} = (W, \succeq, V)$$

in which

(i) $W \neq \emptyset$ (W is a set of 'possible worlds')
(ii) $\succeq \subseteq W \times W$ (Intuitively, $\succeq$ is a betterness or comparative goodness relation; '$x \succeq y$' can be read as 'world x is at least as good as world y'.)
(iii) V : Prop $\rightarrow \mathcal{P}(W)$ (V is an assignment, which associates a set of possible worlds to each propositional letter p).

I write $\mathcal{M} \models_x A$ to mean that *sentence A is true at world x in $\mathcal{M}$*. Such a notion is defined in the usual way except that, for $x, y \in W$,

$$\begin{aligned}
\mathcal{M} \models_x \Box A &\text{ iff } \forall y \ \ (\mathcal{M} \models_y A) \\
\mathcal{M} \models_x \Diamond A &\text{ iff } \exists y \ \ (\mathcal{M} \models_y A) \\
\mathcal{M} \models_x \bigcirc(B/A) &\text{ iff } \forall y\,(\,(\,\mathcal{M} \models_y A \ \& \ \forall z(\mathcal{M} \models_z A \Rightarrow y \succeq z)\,)\Rightarrow \mathcal{M} \models_y B\,) \\
\mathcal{M} \models_x P(B/A) &\text{ iff } \exists y\,(\,(\,\mathcal{M} \models_y A \ \& \ \forall z(\mathcal{M} \models_z A \Rightarrow y \succeq z)\,)\ \& \ \mathcal{M} \models_y B\,)
\end{aligned}$$

The clauses for $\Box$ and $\diamond$ are self-explanatory. These modalities are interpreted by the relation $W \times W$, and thus correspond to the universal modalities further studied by Goranko and Passy [12] among others. In fact, these modalities are not part of Hansson's account. Informally speaking, the evaluation rule for the obligation operator says that $\bigcirc(B/A)$ is true at a world x in $\mathcal{M}$ just in case B is true at all among the *best* (according to $\succeq$) worlds satisfying A. The evaluation rule for the permission operator is obtained by replacing the universal quantifier (ranging over the set of best A-worlds) with the existential one. It is worth noticing that both evaluation rules are formulated in terms of what is sometimes called *optimal* or *last* elements. These are members of S that are at least as good as any other element of S. Formally:

$$y \in \text{opt}_{\succeq}(S) \quad \Leftrightarrow \quad y \in S \;\;\&\;\; y \succeq z \text{ for all } z \in S$$

A last or optimal element of S is, thus, an upper bound of S that is contained in S.[2]

The comparative goodness relation $\succeq$ may be constrained by suitable conditions as desired. The following two classes of models will be discussed further throughout this paper. One is the class of (Åqvist's terminology) H$_3$-models. In such models, the relation $\succeq$ satisfies the following restrictions:

- reflexivity:

  For all $x \in W, x \succeq x$ (δ_1)

- limitedness:

  If $[\![A]\!]^{\mathcal{M}} \neq \emptyset$ then $\{x \in [\![A]\!]^{\mathcal{M}} : (\forall y \in [\![A]\!]^{\mathcal{M}})\, x \succeq y\} \neq \emptyset$, (δ_2)

  where $[\![A]\!]^{\mathcal{M}}$ is $\{x \in W : \mathcal{M} \models_x A\}$, the 'truth-set' of A in $\mathcal{M}$

- transitivity:

  For all $x, y, z \in W, x \succeq y$ and $y \succeq z$ entail $x \succeq z$ (δ_3)

The class of H$_3$-models will henceforth be denoted $\mathcal{H}_3$.

The other class of structures studied in this paper is the class of (Åqvist's terminology) strong H$_3$-models. This class of models corresponds to Hansson's

[2] This is a non-trivial alteration of the account initially proposed by Hansson [1, pp. 143-6]. He works with so-called *maximality* under the strict order induced by $\succeq$. For a given y in S to qualify for the set of maximal elements of S, no other z in S must be strictly better than y. Formally: $y \in \max_{\succ}(S) \Leftrightarrow (y \in S \;\;\&\;\; \nexists\, z \in S\; (z \succ y))$. Here $\succ$ denotes the 'strengthened converse complement' of $\succeq$, defined by $z \succ y$ iff $z \succeq y$ and $y \not\succeq z$. So the previous definition can be rephrased as:

$$y \in \max_{\succeq}(S) \quad \Leftrightarrow \quad y \in S \;\;\&\;\; \forall z \in S\, (\, z \succeq y \Rightarrow y \succeq z\,)$$

Clearly, $\text{opt}_{\succeq}(S) \subseteq \max_{\succeq}(S)$, but not generally the converse. In particular, the maximal set will not necessarily match the optimal set if S is only partially ordered by $\succeq$. The notions of 'optimality' and 'maximality' are more fully discussed by Sen [13].

official system DSDL3. In such models, the following additional constraint is placed on $\succeq$:

- strong connectedness (totalness, or linearity) :

$$\text{For all } x, y \in W, \text{ either } x \succeq y \text{ or } y \succeq x \qquad (\delta_4)$$

There is, then, no more need to explicitly require $\succeq$ to be reflexive. For (δ_1) follows from (δ_4). The class of strong H$_3$-models will be denoted by $\mathcal{H}_3{}^s$.

Care should be taken with the limitedness condition (δ_2). Its main purpose is to forbid infinite (ascending) sequences of ever more perfect worlds. (δ_2) should not be confused with the following condition, of which (δ_4) is just a special case:

- well-orderedness:

$$\text{For all } X \subseteq W \text{ if } X \neq \emptyset \text{ then } \{x \in X : (\forall y \in X)\, x \succeq y\} \neq \emptyset \qquad (\delta_2')$$

(δ_2') entails (δ_2). The converse does not hold generally, but only in special cases. One of them is worth mentioning. It is the case where the language is generated from a finite set of atomic propositions. Notoriously, any subset X of the set of all valuations is, then, definable, in the following sense: for all $X \subseteq W$ there exists a formula $A \in \mathcal{L}$ such that $X = [\![A]\!]^{\mathcal{M}}$.[3] Using this further assumption, (δ_2') − and, by the same way, (δ_4) − can easily be derived from (δ_2). The distinction between the class of H$_3$-models and the class of strong H$_3$-models vanishes.

The notion of semantic consequence is used in its 'local' sense. A set Γ of formulae is said to be true at a state x in $\mathcal{M}$ (notation: $\mathcal{M} \models_x \Gamma$) if all members of Γ are true at x. A formula A is said to be a (local) semantic consequence of Γ over some class $\mathcal{C}$ of models (notation: $\Gamma \models_{\mathcal{C}} A$) if for all models $\mathcal{M}$ from $\mathcal{C}$, and all points x in $\mathcal{M}$, if $\mathcal{M} \models_x \Gamma$ then $\mathcal{M} \models_x A$. Finally, Γ is said to be satisfiable in $\mathcal{C}$ if there is a model $\mathcal{M}$ from $\mathcal{C}$, and a point x in $\mathcal{M}$, such that $\mathcal{M} \models_x \Gamma$. Brackets will be omitted when Γ is a singleton, i.e. a wff A will be said to be satisfiable in $\mathcal{C}$, if the set $\{A\}$ is satisfiable in $\mathcal{C}$.[4]

In Åqvist [10,11] the proof theory for **G** is defined as shown below:

$$\text{All truth functional tautologies} \qquad \text{(PL)}$$

$$\text{S5-schemata for } \Box \text{ and } \Diamond \qquad \text{(S5)}$$

$$P(B/A) \leftrightarrow \neg \bigcirc (\neg B/A) \qquad \text{(DfP)}$$

$$\bigcirc (B \to C/A) \to (\bigcirc(B/A) \to \bigcirc(C/A)) \qquad \text{(COK)}$$

$$\bigcirc (B/A) \to \Box \bigcirc (B/A) \qquad \text{(Abs)}$$

$$\Box A \to \bigcirc(A/B) \qquad \text{(CON)}$$

[3] Cf. Makinson [14, p. 62]. For an example showing that (in the infinite case) a set may not be definable by any formula, see Schlechta [15, p. 29].

[4] As mentioned, the universal modality $\Box$ is not used by Hansson. It is natural to ask whether such a modal operator can be dispensed with, by switching to the so-called global semantic consequence. Perhaps the job done by one can equally be done by the other. This is a topic for future research.

$$\Box(A \leftrightarrow B) \rightarrow (\bigcirc(C/A) \leftrightarrow \bigcirc(C/B)) \tag{Ext}$$

$$\bigcirc(A/A) \tag{Id}$$

$$\bigcirc(C/A \wedge B) \rightarrow \bigcirc(B \rightarrow C/A) \tag{C}$$

$$\Diamond A \rightarrow (\bigcirc(B/A) \rightarrow P(B/A)) \tag{D*}$$

$$(P(B/A) \wedge \bigcirc(B \rightarrow C/A)) \rightarrow \bigcirc(C/A \wedge B) \tag{S}$$

$$\text{If } \vdash A \text{ and } \vdash A \rightarrow B \text{ then } \vdash B \tag{MP}$$

$$\text{If } \vdash A \text{ then } \vdash \Box A \tag{N}$$

A few comments on the axioms involving the deontic modalities might be in order. (DfP) introduces 'P' as the dual of '$\bigcirc$' in the usual way. (COK) is the conditional analogue of the familiar distribution axiom K. (Abs) is the absoluteness axiom of Lewis [16], and reflects my deliberate choice not to make the ranking world-relative. (CON) is the deontic counterpart of the familiar necessitation rule. (Ext) permits the replacement of equivalent sentences in the antecedent of deontic conditionals. (Id), (C) and (S) are familiar from the literature on non-monotonic logic. (Id) is the deontic analogue of the identity principle. The question of whether this is a reasonable law for deontic conditionals has been much debated. A defence of (Id) can be found in Hansson [1] and Prakken and Sergot [5] – this line of defence is discussed in Parent [17, ch. 3]. (C) corresponds to the so-called 'conditionalization' principle (also referred to as 'the hard half of the deduction theorem'), which is part of Kraus and colleagues' system C for cumulative inference relations (see [18]). Axiom (S) has been introduced into the literature by Spohn [9]. The latter axiom is very reminiscent of the restricted principle of strengthening of the antecedent known as 'rational monotony', which is part of so-called system R (see [19]). This other principle says the following:

$$(P(B/A) \wedge \bigcirc(C/A)) \rightarrow \bigcirc(C/A \wedge B) \tag{RM}$$

It is straightforward to show that (S) and (RM) are deductively equivalent given the rest of the system. The deontic version of (RM) is discussed in Goble [20]. (D*) is the conditional analogue of the familiar modal axiom D.

Now the usual notions of theoremhood, deducibility and consistency become available. First, a wff A is said to be a theorem of **G** (written $\vdash_{\mathbf{G}} A$) if A belongs to the smallest subset of wffs that contains every instance of (PL)-(S), and is closed under (MP) and (N). Next, a wff A is said to be deducible in **G** from assumptions Γ (written $\Gamma \vdash_{\mathbf{G}} A$) if there are sentences B_1,..., $B_k \in \Gamma (k \geq 0)$ such that $\vdash_{\mathbf{G}} (B_1 \wedge ... \wedge B_k) \rightarrow A$. Finally, a set Γ of sentences is said to be consistent in **G** if $\bot$ is not deducible in **G** from Γ, and inconsistent otherwise. Again, I will omit brackets when Γ is a singleton.

It may be noted that deducibility is compact, in the sense that deducibility from a set of sentences always implies deducibility from a finite portion of that set. This follows at once from the fact that the number of conjuncts in the antecedent of the requisite conditional $(B_1 \wedge ... \wedge B_k) \rightarrow A$ is always finite. There is an alternative way of expressing compactness, using consistency: a set Γ of sentences is consistent iff every finite subset of Γ is consistent. The compactness property in these two (equivalent) forms will be used in the completeness proof below.

The soundness result, i.e. that

$$\Gamma \vdash_{\mathbf{G}} A \quad \Rightarrow \quad \Gamma \models_{\mathcal{C}} A \text{ (where } \mathcal{C} \in \{\mathcal{H}_3, \mathcal{H}_3^s\})$$

follows immediately from the definitions involved. Observe that the semantic validity of the Spohn sentence (S) – alias (RM) – depends on (δ_3) alone. This is in contrast to the situation in non-monotonic logics, where the validity of the principle of rational monotony is tightly connected to the assumption that the preference relation is a total order.

The adequacy result, i.e. the converse implication

$$\Gamma \models_{\mathcal{C}} A \quad \Rightarrow \quad \Gamma \vdash_{\mathbf{G}} A$$

takes a little bit of work. It can be established by adapting the standard modal technique of constructing a canonical model (see, for instance, Chellas [21] or Blackburn et al. [22]). The points of the canonical model are maximal consistent sets of sentences. In the present semantical context, the main difficulty is to define the comparative goodness relation in such a way that the semantic truth-conditions for formulae starting with a deontic operator coincide with the set-membership relation between formulae and maximal consistent sets. Åqvist [10,23,11] has developed the technique of so-called *systematic frame constants* as a solution to the latter difficulty. Such a technique provides a means of encoding the betterness relation into the syntax, whereby enabling us to talk and reason about the goodness of the maximal consistent sets in the canonical model. The idea behind the proposed construction (which has roots in Lewis [16]) involves extending the language with a family of propositional constants, $\{Q_i\}_{1\leq i<\omega}$ (the so-called "systematic frame constants"), which are indexed by the set of positive integers. These are used to attach a "rank" (or "level of perfection") to every maximal consistent sets. Intuitively, Q_1 refers to an ideal situation, Q_2 refers to a sub-ideal one, Q_3 refers to a sub-sub-ideal one, and so forth. The completeness of **G** is established indirectly, by taking a detour through the system $\mathbf{G}_q^\star$ that results from the addition of suitable axiom schemata. Some govern the behavior of all the Q_i, and others their interplay with the normative modalities. Further detail about how the latter system is used to establish the completeness of **G** can be found in Åqvist [10, p. 184-91]. The basic idea is to define a canonical model for **G** using maximal consistency in $\mathbf{G}_q^\star$ as the criterion for worldhood.

The following two observations have motivated my attempt to prove the completeness of **G** by other means. First, on Åqvist's own admission, the desired completeness remains conjectural, because the proposed argument rests on an unestablished lemma. Next, it has been argued by Hansen [24, p. 130] that Åqvist's conjectured proof fails with respect to strong completeness. To make his point, Hansen considers the case of an 'infinitely bad' set, call it Γ_0. It is made up of

- countably many propositional letters p_i $(1 \leq i < \omega)$
- the primary obligation $\bigcirc\neg p_1$, taken as a shorthand for $\bigcirc(\neg p_1/\top)$ (where $\top$ is any tautology), and
- the sequence of ever more specific contrary-to-duty obligations

$$\bigcirc(\neg p_{i+1}/p_1 \wedge p_2 \wedge ... \wedge p_i) \text{ for all } i \text{ such that } 1 \leq i < \omega$$

Hansen points out that Γ_0 is syntactically inconsistent in $\mathbf{G}_q^\star$. The systematic frame constants are indexed by the set of positive integers. Therefore, no systematic frame constants can consistently be added to Γ_0, and thus no rank (or level of ideality) can be assigned to such a set. The reason why should be obvious to the reader. Γ_0 cannot be ideal (i.e. $\Gamma_0 \cup \{Q_1\} \vdash_{\mathbf{G}_q^\star} \bot$), because Γ_0 violates the primary norm $\bigcirc\neg p_1$. Neither can Γ_0 be sub-ideal (i.e. $\Gamma_0 \cup \{Q_2\} \vdash_{\mathbf{G}_q^\star} \bot$), since Γ_0 violates the contrary-to-duty obligation $\bigcirc(\neg p_2/p_1)$. Neither can Γ_0 be sub-sub-ideal (i.e. $\Gamma_0 \cup \{Q_3\} \vdash_{\mathbf{G}_q^\star} \bot$), since it also violates the contrary-to-contrary-to-duty obligation $\bigcirc(\neg p_3/p_1 \wedge p_2)$. And so on indefinitely.

The purpose of the next section is to define a canonical model for **G** directly, without making reference to $\mathbf{G}_q^\star$ or to any other such system. The only notion of consistency I shall use is consistency in **G**. The worlds will be ordered from the standpoint of a given world, by just comparing the extent to which they comply with the obligations contained there.

3 A Canonical Model for G

The following derived rule and theorems are listed for future reference:

$$\text{If } \vdash B \rightarrow C \text{ then } \vdash \bigcirc(B/A) \rightarrow \bigcirc(C/A) \qquad \text{(RCOM)}$$

$$\bigcirc(B_1/A) \wedge ... \wedge \bigcirc(B_n/A) \rightarrow \bigcirc(B_1 \wedge ... \wedge B_n/A) (n \geq 2) \qquad \text{(AND)}$$

$$\Diamond A \rightarrow \neg \bigcirc (\bot/A) \qquad \text{(COD)}$$

$$\bigcirc (C/A \vee B) \rightarrow (\bigcirc(C/A) \vee \bigcirc(C/B)) \qquad \text{(DR)}$$

The abbreviations RCOM and COD are taken from Chellas [21]. The proofs of (RCOM), (AND) and (COD) are straightforward, and are omitted. (DR) is the deontic version of the principle usually referred to as 'disjunctive rationality' in the non-monotonic literature. The proof of (DR) requires a little more work. For the details, the reader is asked to consult, e.g., Makinson [25, p. 94]. The derivation presented there appeals to the following additional law, known as 'cautious monotony':

$$(\bigcirc(B/A) \wedge \bigcirc(C/A)) \rightarrow \bigcirc(C/A \wedge B) \qquad \text{(CM)}$$

It is perhaps easier to verify that the logic contains (CM) by breaking the argument into cases. If we have $\Diamond A$, then (CM) follows from (RM), since (D*) allows us to weaken $\bigcirc(B/A)$ into $P(B/A)$. If we do not have $\Diamond A$, then (CM) follows from (Ext), because $\neg\Diamond A$ implies $\Box(A \leftrightarrow (A \wedge B))$.

Definition 1. *Let $W^\star$ be the set of all maximal consistent sets of sentences (MCSs). Let w be a fixed element of $W^\star$. The canonical model generated by w can be defined as the triplet*

$$\mathcal{M}^w = (W, \succeq, V)$$

where:

(i) $W = \{x \in W^\star : \textit{ for each } A, \textit{ if } \Box A \in w \textit{ then } A \in x\}$
(ii) $x \succeq y$ *if and only if either*
(a) there is no consistent A such that $\{B : \bigcirc(B/A) \in w\} \subseteq y$ (the vacuous case) or
(b) there is a sentence $A \in x \cap y$ such that $\{B : \bigcirc(B/A) \in w\} \subseteq x$
(iii) V *= the valuation function such that for all p in Prop:*

$$V(p) = \{x \in W : p \in x\}$$

Condition (i) says that W is just the restriction of $W^\star$ to the set of MCSs containing all the wffs A for which $\Box A$ is in the 'generating' world w. This is needed to deal with the alethic modalities. Lemma 1 below clarifies the import of (ii). Intuitively, such a lemma says that the *best* (according to $\succeq$) MCSs among those containing A are precisely those containing all the wffs B for which $\bigcirc(B/A)$ is in the 'generating' world w.

Lemma 1. *If $\succeq$ is defined as in clause (ii)* supra*, then the following two conditions are equivalent (for any x and y in W):*

(I) $A \in x$ *and* $x \succeq y$ *for all y that contains the sentence A*
(II) $\{B : \bigcirc(B/A) \in w\} \subseteq x$

Proof. From the definition of $\succeq$ one sees that (II) entails (I) (given axiom Id). For the converse direction suppose (I) holds, and let B be such that $\bigcirc(B/A) \in w$. We need to show that $B \in x$. Consider the set $\Gamma = \{C : \bigcirc(C/A) \in w\}$. We make the following claims:

Claim 1. Γ is consistent, and can be extended to a maximal consistent set, call it Γ^+.

Verification. The second claim follows from the first (modulo Lindenbaum's lemma). To prove the first claim, suppose Γ is not consistent. By compactness, this means that there is some finite subset $\{C_1, ..., C_n\}$ of Γ such that $\vdash_{\mathbf{G}} (C_1 \wedge ... \wedge C_n) \rightarrow \bot$. By (AND) and (RCOM), $\bigcirc(\bot/A) \in w$. By (COD) and (S5), $\Box\neg A \in w$ so that $\neg A \in x$. Since x is consistent, $A \notin x$, contrary to assumption. We, thus, conclude that Γ is consistent after all.

Claim 2. Γ^+ belongs to W.

Verification. This follows from the fact that, in the presence of (CON), we have

$$\{C : \Box C \in w\} \subseteq \{C : \bigcirc(C/A) \in w\} \subseteq \Gamma^+$$

Claim 3. Γ^+ contains the sentence A.

Verification. Follows from (Id).

We can now apply hypothesis (I) to conclude that $x \succeq \Gamma^+$. By construction, $\{C : \bigcirc(C/A) \in w\} \subseteq \Gamma^+$. Therefore, $x \succeq \Gamma^+$ means that there exists a sentence $D \in x \cap \Gamma^+$ such that

$$\{E : \bigcirc(E/D) \in w\} \subseteq x \tag{1}$$

But we can see that $P(D/A) \in w$. If not, then (DfP) would yield $\neg D \in \Gamma^+$, and thus Γ^+ would be inconsistent. On the other hand, $\bigcirc(B/A) \in w$ entails $\bigcirc(D \to B/A) \in w$. By (S), (Ext) and (C) we conclude $\bigcirc(A \to B/D) \in w$. We can, then, apply (1) to get $A \to B \in x$ and, then, conclude. □

With this established, the rest is easy. First, we lift the 'truth = membership' equation to arbitrary formulae:

Theorem 1 (Truth Lemma). *Let w be a fixed maximal consistent set of sentences, and let $\mathcal{M}^w$ be the canonical model generated by w. Then, for any formula A and x in W,*

$$\mathcal{M}^w \models_x A \textit{ iff } A \in x$$

Proof. The proof is by induction on the complexity of A, as measured by the number of logical operators occurring in it. The base case follows from the definition of V in the canonical model. The boolean cases are handled in the usual way, and so are the modal cases. In the modal cases, it might be helpful first to show that (by virtue of the S5 schemata) the relation $R \subseteq W^\star \times W^\star$ defined by putting xRy whenever $\{C : \Box C \in x\} \subseteq y$ is an equivalence relation on the set $W^\star$ of all maximal consistent sets. The fact that R is symmetric, i.e.

$$\{C : \Box C \in x\} \subseteq y \Rightarrow \{D : \Box D \in y\} \subseteq x \tag{2}$$

will be used in the proof of the deontic cases, to which I now turn. I shall focus on the case where A is $\bigcirc(C/B)$. The following is to be established:

$$\mathcal{M}^w \models_x \bigcirc(C/B) \text{ iff } \bigcirc(C/B) \in x$$

For the right-to-left direction, assume $\bigcirc(C/B) \in x$ and let $y \in [\![B]\!]^{\mathcal{M}^w}$ be such that $y \succeq z$ for all z in $[\![B]\!]^{\mathcal{M}^w}$. By the inductive hypothesis, $B \in y$, and $y \succeq z$ for any $z \in W$ such that $B \in z$. Using lemma 1, we get

$$\{B' : \bigcirc(B'/B) \in w\} \subseteq y \tag{3}$$

Now, in the presence of (CON), $\Box \bigcirc (C/B) \in x$ can validly be inferred from $\bigcirc(C/B) \in x$. Using (2), we then get $\bigcirc(C/B) \in w$. From this together with (3), we obtain $C \in y$. By the inductive hypothesis, C is true at y. This shows that $\bigcirc(C/B)$ is true at x as wished.

For the left-to-right direction, assume that $\bigcirc(C/B)$ is true at x. Using the truth-clause for $\bigcirc$, the inductive hypothesis and lemma 1, and invoking the definition of W, we first get

$$\forall y \in W^\star : (\{E : \Box E \in w\} \subseteq y \;\&\; \{D : \bigcirc(D/B) \in w\} \subseteq y) \Rightarrow C \in y$$

This itself simplifies into (see claim 2 in the proof of lemma 1 above)

$$\forall y \in W^\star : \{D : \bigcirc(D/B) \in w\} \subseteq y \Rightarrow C \in y \tag{4}$$

(4) says that C belongs to every maximal consistent extension of

$$\{D : \bigcirc(D/B) \in w\}$$

By the second corollary to Lindenbaum's lemma, C is derivable from that set, i.e.,

$$\vdash_{\mathbf{G}} (D_1 \wedge ... \wedge D_n) \to C$$

for sentences $D_1, ..., D_n (n \geq 0)$ such that

$$\bigcirc(D_1/B), ..., \bigcirc(D_n/B) \in w$$

Without loss of generality, we can assume that the number of D_i is finite, given compactness. So, using (AND), we first obtain

$$\bigcirc(D_1 \wedge ... \wedge D_n/B) \in w$$

Using (RCOM), we get

$$\bigcirc(C/B) \in w$$

By (CON),

$$\Box \bigcirc (C/B) \in w$$

The definition of W, then, yields the desired conclusion $\bigcirc(C/B) \in x$.

The proof that the theorem holds when A is $P(C/B)$ is similar in structure. Details are omitted. □

We can now check that the comparative goodness relation $\succeq$ of the canonical model has the required properties:

Lemma 2 (Verification Lemma). *If $\succeq$ is taken as in definition 1, then $\succeq$ is limited (δ_2), transitive (δ_3) and strongly connected (δ_4).*

Proof. Limitedness is easily checked. Assume A is true at some x in W. By theorem 1, $A \in x$. Re-running the proof for the '(II) $\Rightarrow$ (I)' direction of lemma 1, claims 1 to 3, we get that W contains at least one y such that $\{A\} \subseteq \{B : \bigcirc(B/A) \in w\} \subseteq y$. Again, by theorem 1, A is true at y. Consider any z at which A is true. By theorem 1, A is in z, and hence in $y \cap z$. By definition 1 (ii), $y \succeq z$ as expected.

Strong connectedness can be proved by *reductio ad absurdum*. Assume $x \not\succeq y$ and $y \not\succeq x$. The former entails that there is a consistent A such that $\{B : \bigcirc(B/A) \in w\} \subseteq y$, whilst the latter implies that there is a consistent C such that $\{B : \bigcirc(B/C) \in w\} \subseteq x$. In virtue of (Id), $A \in y$ and $C \in x$ so that $A \vee C \in y \cap x$. Using (DR), one might then conclude that either

$$\{B : \bigcirc(B/A \vee C) \in w\} \subseteq \{B : \bigcirc(B/A) \in w\} \subseteq y$$

or

$$\{B : \bigcirc(B/A \vee C) \in w\} \subseteq \{B : \bigcirc(B/C) \in w\} \subseteq x$$

Either way we are done.

The proof that $\succeq$ is transitive is a bit tricky. Suppose $x \succeq y$ and $y \succeq z$. Assume $y \succeq z$ means that there exists a $B \in y \cap z$ such that

$$\{B' : \bigcirc(B'/B) \in w\} \subseteq y \tag{*}$$

(Otherwise, $x \succeq z$ holds trivially.) Given this, $x \succeq y$ entails that there is $C \in x \cap y$ such that

$$\{C' : \bigcirc(C'/C) \in w\} \subseteq x \tag{**}$$

Clearly, $B \vee C \in x \cap z$. The following is to be established:

$$\{D : \bigcirc(D/B \vee C) \in w\} \subseteq x \tag{5}$$

Note that $P(C/B \vee C) \in w$. For otherwise, using (DfP), the maximality of w and (DR), we would have either $\bigcirc(\neg C/C) \in w$ or $\bigcirc(\neg C/B) \in w$. None can occur, because a direct application of (**) and (*) would yield the result that C does not belong to the union of x and y − contradicting the assumption made above that C belongs to their intersection. The proof of (5) is then as follows. Assume $\bigcirc(D/B \vee C) \in w$. By (RCOM), $\bigcirc(C \to D/B \vee C) \in w$. By (S), $\bigcirc(D/(B \vee C) \wedge C) \in w$. By (Ext), $\bigcirc(D/C) \in w$. Using (**), we then get $D \in x$ as wished. □

The above results are similar to Boutilier's [26] theorem 3.36. There the focus is on belief revision theory. Due to this shift of emphasis, my proofs are different from those presented there.

4 Completeness

I first deal with the totally ordered case. The completeness of **G** with respect to the class $\mathcal{H}_3^s$ of strong H$_3$-models follows easily from the following:

Theorem 2. *Every consistent set of sentences is satisfiable in $\mathcal{H}_3^s$.*

Proof. Let Γ be any consistent set of sentences. By Lindenbaum's lemma, Γ has a maximal extension, call it Γ_ω. Form the canonical structure generated by Γ_ω, i.e., the structure $\mathcal{M}^{\Gamma_\omega}$ as defined *supra*. By lemma 2, $\mathcal{M}^{\Gamma_\omega}$ belongs to $\mathcal{H}_3^s$. By theorem 1 above, we obtain in particular that for each sentence A

$$\mathcal{M}^{\Gamma_\omega} \models_{\Gamma_\omega} A \text{ iff } A \in \Gamma_\omega$$

Since $\Gamma \subseteq \Gamma_\omega$, we thus have

$$\mathcal{M}^{\Gamma_\omega} \models_{\Gamma_\omega} A \text{ for any } A \text{ in } \Gamma$$

as required. □

Theorem 3 (Completeness, total order case). *For each set of formulae Γ and formula A, the equivalence*

$$\Gamma \vdash_{\mathbf{G}} A \quad \Leftrightarrow \quad \Gamma \models_{\mathcal{H}_3^s} A$$

holds.

Proof. The left-to-right implication is just soundness, so it suffices to check out the right-to-left implication. The argument is standard. Suppose $\Gamma \models_{\mathcal{H}_3^s} A$. Then $\Gamma \cup \{\neg A\}$ is not satisfiable in $\mathcal{H}_3^s$, and hence theorem 2 gives $\Gamma \cup \{\neg A\} \vdash_{\mathbf{G}} \bot$. By simple propositional manipulations, we get $\Gamma \vdash_{\mathbf{G}} A$ as required. □

Theorem 3 is a strong completeness result. As already emphasized, the argument uses compactness many times, and thus no restrictions are placed on the cardinality of the premisse set Γ, at least in principle. I say 'in principle', because the question of whether the above result is immune from a similar objection as the one raised against the systematic frame constants account remains an open problem. As mentioned, a drawback of the latter account is that it fails to assign a rank to the 'indefinitely' bad set as described on p. 194–195. Take the canonical structure generated from the maximal consistent extension of such a set. It is natural to ask if the definition of $\succeq$ in the canonical model does a better job. That issue calls for further exploration, which will not be attempted in this paper.

I now turn to the partially ordered case. In Åqvist [10, p. 182] and Åqvist [11, p. 249], the question is raised whether **G** is also strongly complete with respect to the class $\mathcal{H}_3$ of models. Based on our previous results we can give a positive answer to this last question.

Corollary 1 (Completeness, partial order case). *For each set of formulae Γ and formula A, the equivalence*

$$\Gamma \vdash_{\mathbf{G}} A \quad \Leftrightarrow \quad \Gamma \models_{\mathcal{H}_3} A$$

holds.

Proof. We already have the soundness part. The proof of the adequacy claim requires only two lines:

$$\begin{aligned} \Gamma \models_{\mathcal{H}_3} A \quad &\Rightarrow \quad \Gamma \models_{\mathcal{H}_3^s} A \\ &\Rightarrow \quad \Gamma \vdash_{\mathbf{G}} A \quad \text{(by th. 3)} \end{aligned}$$

□

The above result shows that within the present set-up the strong connectedness assumption (δ_4) has no import, in the sense that the logic is unaffected by imposing this requirement or not. At first, this may seem surprising. In a way, this is not, given the following:

- H_3-models include the requirement of limitedness (δ_2), and (as mentioned) in the finite case limitedness entails connectedness (δ_4),
- As can easily be verified, the schema $A \geq B \vee B \geq A$ is always valid as long as limitedness is assumed. Here the relation $\geq$ between formulas is defined in the usual fashion, i.e. by the rule: $A \geq B$ iff $\Diamond(A \vee B) \rightarrow P(A/A \vee B)$.

There is more to connectedness than meets the eye. Such a notion is central to, e.g., questions about the possibility of deontic dilemmas, which are key questions within deontic logic today (see, e.g., [27,20]). An in-depth discussion of the role of comparability for the logic of obligation falls outside the immediate scope of this paper, and must be postponed to another opportunity.

One reviewer suggested we start with a form of the limit assumption that does not have the effects described above. Suppose one such form is available. Suppose also we redefine H_3-models by requiring they satisfy this new version of the limit assumption, rather than the old one; the evaluation rules for the deontic modalities are then re-phrased in terms of maximality rather than optimality. We would certainly get a better understanding of the role of comparability within an Hansson-type semantics, by first axiomatizing this class of models, and then investigating the effects of adding the linearity requirement. The question of whether there are in fact alternative forms which the limit assumption might take, is the main focus of my current investigations. The notion of stopperedness from [18] and [28] would not do, but perhaps there are alternative forms available. Expressed in terms of maximality, the stopperedness condition says that whenever $x \in [\![A]\!]^{\mathcal{M}}$ there is a maximal $y \in [\![A]\!]^{\mathcal{M}}$ with $y \succeq x$. Connectedness would still be involved in the framework being used, because stopperedness validates the formula $A \geq B \vee B \geq A$.

Acknowledgments. I would like to thank the anonymous referees for their helpful comments.

References

1. Hansson, B.: An analysis of some deontic logics. Noûs 4, 373–398 (1969)
2. Mott, P.: On Chisholm's paradox. Journal of Philosophical Logic 2, 197–211 (1973)
3. Tomberlin, J.: Contrary-to-duty imperatives and conditional obligation. Noûs 16, 357–375 (1981)
4. Loewer, B., Belzer, M.: Dyadic deontic detachment. Synthese 54, 295–318 (1983)
5. Prakken, H., Sergot, M.: Dyadic deontic logic and contrary-to-duty obligation. In: [29], pp. 223–262
6. van der Torre, L., Tan, Y.H.: The many faces of defeasibility in defeasible deontic logic. In: [29], pp. 79–121

7. Carmo, J., Jones, A.: Deontic logic and contrary-to-duties. In: Gabbay, D., Guenthner, F. (eds.) Handbook of Philosophical Logic, 2nd edn., vol. 8, pp. 265–343. Kluwer Academic Publishers, Dordrecht (2002)
8. Parent, X.: Remedial interchange, contrary-to-duty obligation and commutation. Journal of Applied Non-Classical Logics 3/4, 345–375 (2003)
9. Spohn, W.: An analysis of Hansson's dyadic deontic logic. Journal of Philosophical Logic 4, 237–252 (1975)
10. Åqvist, L.: An Introduction to Deontic logic and the Theory of Normative Systems, Bibliopolis, Naples (1987)
11. Åqvist, L.: Deontic logic. In: Gabbay, D., Guenthner, F. (eds.) Handbook of Philosophical Logic, 2nd edn., vol. 8, pp. 147–264. Kluwer Academic Publishers, Dordrecht (2002)
12. Goranko, V., Passy, S.: Using the universal modality: Gains and questions. Journal of Logic and Computation 2, 5–30 (1992)
13. Sen, A.: Maximization and the act of choice. Econometrica 65, 745–779 (1997)
14. Makinson, D.: Bridges from Classical to Nonmonotonic logic. King's College Publications, London (2005)
15. Schlechta, K.: Coherent Systems. Elsevier, Amsterdam (2004)
16. Lewis, D.: Counterfactuals. Blackwell, Oxford (1973)
17. Parent, X.: Non-monotonic Logics and Modes of Argumentation − The Case of Conditional Obligation. PhD thesis, University of Aix-Marseille I, France (2002)
18. Kraus, S., Lehmann, D., Magidor, M.: Nonmonotonic reasoning, preferential models and cumulative logics. Artificial Intelligence 44, 167–207 (1990)
19. Lehmann, D., Magidor, M.: What does a conditional knowledge base entail? Artificial Intelligence 55, 1–60 (1992)
20. Goble, L.: A proposal for dealing with deontic dilemmas. In: Lomuscio, A., Nute, D. (eds.) DEON 2004. LNCS (LNAI), vol. 3065, pp. 74–113. Springer, Heidelberg (2004)
21. Chellas, B.: Modal Logic. Cambridge University Press, Cambridge (1980)
22. Blackburn, P., de Rijke, M., de Venema, Y.: Modal Logic, vol. 53. Cambrigde University Press, Cambridge (2001)
23. Åqvist, L.: A completeness theorem in deontic logic with systematic frame constants. Logique & Analyse 36, 177–192 (1993)
24. Hansen, J.: On relations between Åqvist's deontic system G and van Eck's deontic temporal logic. In: Mc Namara, P., Prakken, H. (eds.) Norms, Logics and Information Systems. Frontiers in Artificial Intelligence and Applications, pp. 127–144. IOS Press, Amsterdam (1999)
25. Makinson, D.: General patterns in nonmonotonic reasoning. In: Gabbay, D., Hogger, C., Robinson, J. (eds.) Handbook of Logic in Artificial Intelligence and Logic Programming, pp. 35–110. Clarendon Press, Oxford (1994)
26. Boutilier, G.: Unifying default reasoning and belief revision in a modal framework. Artificial Intelligence 68, 33–85 (1994)
27. van der Torre, L., Tan, Y.H.: Contrary-to-duty reasoning with preference-based dyadic obligations. Annals of Mathematics and Artificial Intelligence 27(1-4), 49–78 (1999)
28. Makinson, D.: General theory of cumulative inference. In: Reinfrank, M., Ginsberg, M.L., de Kleer, J., Sandewall, E. (eds.) Non-Monotonic Reasoning 1988. LNCS, vol. 346, pp. 1–18. Springer, Heidelberg (1989)
29. Nute, D. (ed.): Defeasible Deontic Logic. Kluwer Academic Publishers, Dordrecht (1997)

Strata of Intervenient Concepts in Normative Systems

Lars Lindahl[1] and Jan Odelstad[2]

[1] Faculty of Law, University of Lund, Sweden
lars.lindahl@jur.lu.se
[2] 1) Department of Mathematics, Natural and Computer Sciences, University of Gävle, Sweden, 2) DSV, KTH, Sweden
jod@hig.se

Abstract. Writing a contract of a specific content is a ground for purchase, purchase is a ground for ownership, ownership is a ground for power to dispose. Also power to dispose is a consequence of ownership, ownership is a consequence of purchase. etc. The paper presents a continuation of the authors' previous algebraic representation on ground - consequence chains in normative systems.The paper analyzes different kinds of "implicative closeness" between grounds and consequences in chains of legal concepts, in particular combinations of "weakest ground", "strongest consequence" and "minimal joining". The idea of a concept's being intermediate between concepts of two different sorts is captured by the technical notion of "intervenient", defined in terms of weakest ground and strongest consequence. A legal example concerning grounds and consequences of "ownership" and "trust" is used to illustrate the application of the formal theory.

Keywords: Normative system, Legal concept, Intermediate concept, Intervenient, Weakest ground, Strongest consequence, Intervenient minimality, Ownership.

1 Introduction: Intermediate Concepts in a Normative System

Janus, the Roman god of beginnings and endings, had two faces. Likewise, legal concepts have two faces, one turned towards facts and description, the other turned towards legal consequences. The ultimate grounds for there being a valid contract are described in an essentially empirical way, as a matter of actions, beliefs, intentions, absence of certain kinds of influence, such as violence or deceit etc. The ultimate consequences of there being a valid contract are described essentially in deontic terms as a matter of rights and duties between the parties. One set of rules relate to the factual requirements for a valid contract, another set of rules relate to the deontic consequences of a valid contract. Similarly

R. van der Meyden and L. van der Torre (Eds.): DEON 2008, LNAI 5076, pp. 203–217, 2008.

for other legal terms such as citizenship, guardianship, ownership, possession etc. We might say that ownership, valid contract, citizenship etc. are "intermediate" between certain facts (grounds) and certain deontic positions (legal consequences).

In a series of papers, the present authors have aimed at developing an algebraic framework for elucidating the role of intermediate concepts in normative systems. See [5], [6], [9], [10], [11]. Cf. [8], [12].[1] In these papers, emphasis is put on distinguishing various relations of "closeness" and "minimality". As a first hint, we consider the following three rules 1-3:

1. Legal rule linking descriptive concept a_1 (*ground*) to an intermediate concept a_2: For all x, y: If $a_1(x, y)$ then $a_2(x, y)$.
2. Legal rule linking an intermediate concept a_2 to deontic concept a_3 (*consequence*): For all x, y: If $a_2(x, y)$ then $a_3(x, y)$.
3. Legal rule directly linking descriptive concept a_1 (*ground*) to deontic concept a_3 (*consequence*): For all x, y: If $a_1(x, y)$ then $a_3(x, y)$.

The rules 1-3 are represented by three ordered pairs $\langle a_1, a_2\rangle$, $\langle a_2, a_3\rangle$, $\langle a_1, a_3\rangle$ where a_1, a_2, a_3 are considered to be of "different sorts". For any such pair $\langle u, v\rangle$ we consider the features of "weakest ground", "strongest consequence" and "minimality". Thus, if any ground for $v(x, y)$ implies $u(x, y)$, then u is a "weakest" ground for v, abbreviated $WG(u, v)$. Also, if $v(x, y)$ implies any consequence of $u(x, y)$, then v is a "strongest" consequence of u, abbreviated $SC(v, u)$. If both $WG(u, v)$ and $SC(v, u)$, then the implication from $u(x, y)$ to $v(x, y)$ is minimal and the pair $\langle u, v\rangle$ represents what we call a "minimal joining".

In our formal analysis of intermediate concepts, we make use of a technical notion "intervenient" intended to capture essential features of what, intuitively, can be regarded intermediate concepts in the law.[2] If a_1, a_2, a_3 are as in rules (1)-(3) and it holds both that $WG(a_1, a_2)$ and that $SC(a_3, a_2)$, then a_2 is called an intervenient "corresponding" to the pair $\langle a_1, a_3\rangle$.

In a full-fledged formal theory, the following features of intermediate concepts are important.

- *WG, SC relationships and effective expressiveness of intermediate concepts*

[1] Since the framework of our analysis in these papers is algebraic, there is a need for an algebraic framework as well for the deontic positions that are seen as ultimate legal consequences. An algebraic framework for deontic positions developed is introduced in [8].

[2] Our basic formal framework is abstract in the sense that the main algebraic results have other areas of applications than intermediate concepts in the law. We always endeavour to make the algebraic results independent of any specific interpretation. Thus, the so-called *cis* model (*cis* for "condition implication structure") of the abstract theory, envisaged in the present paper and intended as a tool e.g., for analysis of intermediaries in legal systems, only plays the part of one of several models for the theory.

An important issue in the classical debate on intermediate concepts was how these concepts (for example "ownership") served to reduce the number of legal rules needed for expressing the contents of the legal system. This feature can be called "economy of expression".[3] An essential element in the analysis performed by the present authors is that relations WG, SC, and minimality, as outlined above, are decisive for how economy of expression is accomplished and for how changes of a system can be effectively achieved. (See [5], [6], [8], [9], [11] and cf. [12].) Other notions defined by us for this purpose are those of "base of a system" and "base of intervenients" of a system. (See [9], [11].)

In the present paper we systematize different minimality relations, aiming at a typology.

– *Networks of Boolean structures (strata) of cognate intermediate concepts*

Within the classical debate, the analysis of legal concepts as intermediate only dealt with intermediate concepts (like "ownership") taken singly.

In a comprehensive system of legal concepts, however, sets of intermediate concepts constitute subsystems where the consequence-structure in one system can be the ground-structure in another.[4] Therefore, the pattern of a comprehensive system of legal concepts is usually that of a network of structures of intermediate concepts. (See the middle part of Fig.1.) Legal theory and concept formation essentially deals with the "box" between input and output. In the box in the middle of Fig.1, each of the nodes represents a structure of several interconnected concepts rather than a single concept.

A structure of concepts thus represented by a node is conceived by us as a Boolean algebra of cognate concepts. Within this framework, the status of Boolean combinations of intermediate concepts is an issue to be clarified. If, according to some criteria, m_1, m_2 are appropriately seen as intermediate within a specific system, it should be clarified whether negations $not - m_1$, $not - m_2$, conjunctions $m_1 \wedge m_2$ and disjunctions $m_1 \vee m_2$ are intermediate concepts as well according to these criteria. (On negation, see [10], [11]. On minimality see the theorems on "connections" in [8].)

In the present paper, we take a first step towards analyzing networks of structures of intermediate concepts.

– *Openness of intermediate concepts*

As is well-known, there are numerous cases where legal concepts are vague or "open textured", and power to interpret the concepts is conferred on judges and other persons who apply the law. Obvious examples are such concepts as "negligent" or "reasonable" but considerable openness also is a feature of such

[3] For references to Wedberg-Ross and the early Scandinavian debate, see [5], [6], [8], [11] For a recent contribution, cf. as well [13].

[4] In some recent versions of the so-called "Counts-as" theory, the concepts dealt with can be thought of as constituting a chain. See [11] cf. [2].

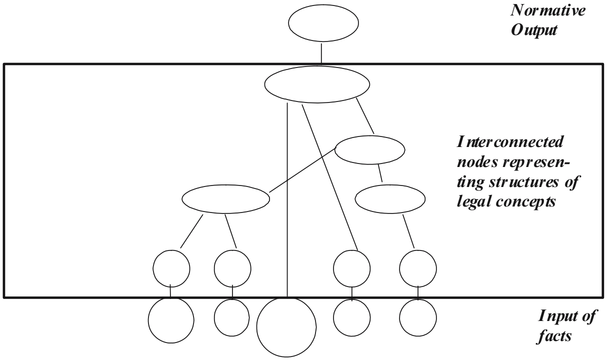

Normative Output
Interconnected nodes representing structures of legal concepts
Input of facts

Fig. 1.

concepts as "public interest", "contract" and "ownership".[5] Often the vague concepts occur within a chain or network, and, to arrive at deontic consequences,

[5] We recall that, at the end of the 18th century, Jeremy Bentham launched an attack on the traditional legal conceptual apparatus as used by the legal profession. According to Bentham, a large part of the legal terms in use were "impostor words". He dreamt of a "complete legal code" and envisaged a reformulation of legal rules in a language where clear commands are stated by the legislator or can be derived from such commands. In this way the influence of the inclinations and biases of lawyers applying the law should be minimized. The different deontic ("imperational") modes should be structured by "imperational logic". See, for example [3].

However, as clearly understood already by Aristotle, it is not possible to create a complete legal code without incurring into error by oversimplifying matters:

> ... all law is universal but about some things it is not possible to make a universal statement which shall be correct. In those cases, then, in which it is necessary to speak universally, but not possible to do so correctly, the law takes the usual case, though it is not ignorant of the possibility of error. And it is none the less correct; for the error is in the law nor in the legislator but in the nature of the thing, since the matter of practical affairs is of this kind from the start. When the law speaks universally, then, and a case arises on it which is not covered by the universal statement, then it is right, where the legislator fails us and has erred by oversimplicity, to correct the omission-to say what the legislator himself would have said had he been present, and would have put into his law if he had known. Hence the equitable is just, and better than one kind of justice-not better than absolute justice but better than the error that arises from the absoluteness of the statement. And this is the nature of the equitable, a correction of law where it is defective owing to its universality. In fact this is the reason why all things are not determined by law, that about some things it is impossible to lay down a law, so that a decree is needed. For when the thing is indefinite the rule also is indefinite, like the leaden rule used in making the Lesbian moulding; the rule adapts itself to the shape of the stone and is not rigid, and so too the decree is adapted to the facts. Aristotle, Nicomachean Ethics, EN 1137b.

deduction must be combined with step by step interpretative decisions for the concepts in the chain or network. The occurrence of "open" legal concepts is a strong argument against any reductionist idea that legal reasoning might in general proceed directly from facts to deontic consequences so as to dispense with intermediate concepts. (Cf. [11].) In previous papers [5], [9], [10], [11], the present authors have dealt algebraically with the problem of "open" legal intermediaries. If there is a chain or network of open concepts, the algebraic analysis will be very complex. In the present paper, we do not deal specifically with "open" intermediaries.

The paper is organized as follows. In Section 2, the formal framework is introduced, with an overview and some explicit definitions of basic theoretical tools. Section 3, which is the main part of the paper, starts with a legal example (ownership and trust), a small network intended to illustrate the formal results. Subsequently in this section, a number of results on weakest grounds, strongest consequences, minimality and intervenients are presented. Moreover, the case of a chain of more than three structures is analyzed. The section ends with a systematization of different kinds of minimality, summed up in a rudimentary typology. In Section 4, which is the conclusion, some suggestions are made with a view to future work.

2 The Basic Framework

The basic framework of our analysis is purely algebraic (abstract) but is developed with a preferred model in view.[6] This model, called "condition implication structure" (*cis*), has limitations. In our view, however, the model provides means for seeing and formulating distinctions and features that elucidate the different character of various kinds of concepts in actual normative systems. In this section we describe the formal framework in terms of the *cis*-model.

The *cis*-model consists of different strata $\mathcal{B}_1, \mathcal{B}_2, ...$, called "Boolean quasi-orderings" (*Bqo*'s) where, for $\mathcal{B}_i = \langle B_i, \wedge, ', R_i \rangle$, it is assumed that $\langle B_i, \wedge, ' \rangle$ is a Boolean algebra and R_i is a quasi-ordering on B_i.[7] The indifference part of R_i is denoted Q_i and the strict part is denoted S_i. The elements of the (domains of the) *Bqo*'s are called "conditions" and are assumed to be concepts such that elements a_i and a_j of two *Bqo*'s $\mathcal{B}_i$ and $\mathcal{B}_j$ are of "different kinds".[8]

Within one and the same *Bqo* $\mathcal{B}_i$, conditions $a_i, b_i, c_i, ...$ are connected by an implicative relation relation R_i. The relation R_i between conditions $a_i, b_i, c_i, ...$ within a *Bqo* $\mathcal{B}_i$, is thought of as representing logical or otherwise highly stable relationships, immune to changes. Two conditions a_i, a_j from two different *Bqo*'s

[6] For an extensive presentation of the formal theory, the reader is referred to [11] with further references.

[7] The formal definition of a *Bqo* requires, that (i) aR_ib and aR_ic implies $aR_i(b \wedge c)$, (ii) aR_ib implies $b'R_ia'$, (iii) $(a \wedge b)R_ia$, (iv) not $\top R_i \bot$, where $\bot$ is the zero element and $\top$ is the unit element.

[8] Note our notation where $\mathcal{B}_i$ is a *Bqo* and B_i the domain of $\mathcal{B}_i$. Sometimes we speak of "element of a *Bqo* $\mathcal{B}_i$" in the sense of element of the domain of $\mathcal{B}_i$.

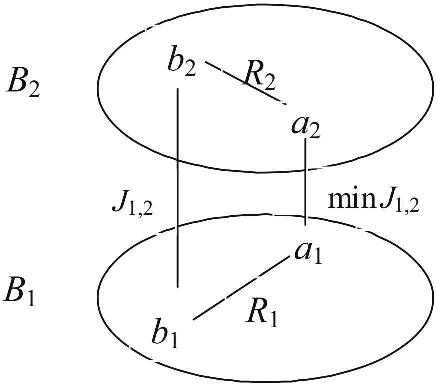


Fig. 2.

$\mathcal{B}_i$ and $\mathcal{B}_j$ can be connected by an implicative relation $J_{i,j}$. These implications (called "joinings") between conditions in two *Bqo*'s $\mathcal{B}_i$ and $\mathcal{B}_j$ are thought of as the prescriptions (the normative content) of the system. Changes of a system are thought of as changes of the joinings between elements of two *Bqo*'s rather than of the relations R_i within a *Bqo* $\mathcal{B}_i$. A triple $\langle \mathcal{B}_i, \mathcal{B}_j, J_{i,j} \rangle$ consisting of two *Bqo*'s $\mathcal{B}_i, \mathcal{B}_j$ interconnected by a set $J_{i,j}$ of joinings (implications) between their elements, is called a "Boolean joining system" (*Bjs*).[9]

If $J_{i,j}$ is the set of joinings from $\mathcal{B}_i$ to $\mathcal{B}_j$, then $\min J_{i,j}$, i.e., the subset of *minimal* joinings from $\mathcal{B}_i$ to $\mathcal{B}_j$, is of special importance for characterizing the interrelation between $\mathcal{B}_i$ and $\mathcal{B}_j$.[10] A pair $\langle a_i, a_j \rangle$ belongs to $\min J_{i,j}$ if, in B_i there is no weaker ground for a_j than a_i and, in $\mathcal{B}_j$, there is no stronger consequence of a_i than a_j. If for any pair $\langle b_i, b_j \rangle \in J_{i,j}$, there is a pair $\langle a_i, a_j \rangle \in \min J_{i,j}$ such that $\langle b_i, b_j \rangle$ "encloses" $\langle a_i, a_j \rangle$, then we say that $J_{i,j}$ satisfies connectivity.[11] Thus, in the figure below, if $\langle a_1, a_2 \rangle \in \min J_{1,2}$, the pair $\langle b_1, b_2 \rangle$, belonging to $J_{1,2}$ encloses $\langle a_1, a_2 \rangle$ in the sense that $b_1 R_1 a_1$ and $a_2 R_2 b_2$.[12]

The various *Bqo*'s of a normative system and the various implicative relations are seen against a background of the general framework of a Boolean algebra $\mathcal{B}$

[9] The formal definition of a *Bjs* presupposes the following definition: The *narrowness-relation determined by* the quasi-orderings $\langle B_1, R_1 \rangle$ and $\langle B_2, R_2 \rangle$ is the binary relation $\trianglelefteq$ on $B_1 \times B_2$ such that $\langle a_1, a_2 \rangle \trianglelefteq \langle b_1, b_2 \rangle$ if and only if $b_1 R_1 a_1$ and $a_2 R_2 b_2$. $\langle \mathcal{B}_1, \mathcal{B}_2, J \rangle$ is a *Bjs* if (i) for all $b_1, c_1 \in B_1$ and $b_2, c_2 \in B_2$, $\langle b_1, b_2 \rangle \in J$ and $\langle b_1, b_2 \rangle \trianglelefteq \langle c_1, c_2 \rangle$ implies $\langle c_1, c_2 \rangle \in J$, (ii) for any $C_1 \subseteq B_1$ and $b_2 \in B_2$, if $\langle c_1, b_2 \rangle \in J$ for all $c_1 \in C_1$, then $\langle a_1, b_2 \rangle \in J$ for all $a_1 \in lub_{R_1} C_1$, (iii) for any $C_2 \subseteq B_2$ and $b_1 \in B_1$, if $\langle b_1, c_2 \rangle \in J$ for all $c_2 \in C_2$, then $\langle b_1, a_2 \rangle \in J$ for all $a_2 \in glb_{R_2} C_2$. (Note that the definitions of least upper bound (*lub*) and greatest lower bound (*glb*) for partial orderings are easily extended to quasi-orderings, but the *lub* or *glb* of a subset of a quasi-ordering is not necessarily unique but can consist of a set of elements.)

[10] A minimal joining is minimal with respect to the narrowness-relation.

[11] See [11], Section 2.2 for conditions on a *Bjs* that imply connectivity.

[12] By Fig. 2, it is visualized that $\langle b_1, a_2 \rangle$ belongs to the relative product $R_1 | \min J_{1,2}$, and $\langle a_1, b_2 \rangle$ to the relative product $\min J_{1,2} | R_2$.

providing the language of the whole system. The various implicative relations (R_1,R_2,... $J_{1,2}$,$J_{2,3}$, ...and their combinations) of the system are seen as regimentations of a general implicative binary relation ρ in $\mathcal{B}$. Thus the most general structure of the system is expressed by what we call a supplemented Boolean algebra or *sBa* $\langle B, \wedge, ', \rho\rangle$.[13] This *sBa* is all-embracing to the system, and its relation ρ, in a sense, is what constitutes the system.

In a comprehensive representation of a normative system as reconstructed within our framework, its structure is conceived of as a net-like pattern where the *Bqo*'s represent the nodes and the joinings represent the links between nodes. Minimal joinings play an essential part for characterizing a normative system as so represented and for effectively describing changes made in such a system.

Fig.1 above might provide a first glimpse of the net-like pattern, supposing that the nodes are thought of as *Bqo*'s and the links between nodes as joinings. As appears from the middle part of Fig.1, systems of legal concepts internal to the system can be thought of as *Bqo*'s belonging to various strata of the normative system, these *Bqo*'s being interrelated by joinings of the system (lines between the nodes). Some of the elements of the *Bqo*'s in this middle part of the system can be called "intervenients", in the sense of that which "comes in between" the concepts for facts and the deontic concepts.

If a condition m is an intervenient between two strata $\mathcal{B}_i$ and $\mathcal{B}_j$, then there is an intro-condition for m in $\mathcal{B}_i$ and an elim-condition for m in $\mathcal{B}_j$. The intro-condition is the weakest ground in $\mathcal{B}_i$ for m, and the elim-condition is the strongest consequence in $\mathcal{B}_j$ of m. Thus one can derive m from conditions in $\mathcal{B}_i$ only *via* the intro-condition, and one can derive conditions in $\mathcal{B}_j$ from m only *via* the elim-condition. The concept of minimality for joinings between two strata, as well as the concept of intervenient between two strata, is defined in terms of weakest ground and strongest consequence. Thus the concepts of minimal joining and intervenient are interrelated. The present paper is much devoted to investigating these interrelations.

In the formal part of [11], we focused on systems consisting of one algebra of grounds and one algebra of consequences and a system of intervenients between these algebras. In the present paper, we further analyze the definitional requirements on intervenients, prove a number of results, and exhibit a rudimentary typology. Also, we take a step towards extending the theory so as to incorporate series of systems of intervenients. In such a series, the consequence-structure in one system can be the ground-structure in another, and the intervenients in one system can be grounds or consequences in another.

3 Development of the Theory

3.1 A Legal Example

In this section and the next we will present a number of results on weakest grounds, strongest consequences, minimality, and intervenients. These will be

[13] In a *sBa* the partial ordering determined by the Boolean algebra is a subset of ρ.

illustrated by specific figures 3A-J, but as well by a legal example concerning *ownership* and *trusteeship,* pictured as a rudimentary network (see Fig.3). The legal rules in this example are expressed in terms of joinings between *Bqo*'s $\mathcal{B}_1$, $\mathcal{B}_2$, $\mathcal{B}_4$, $\mathcal{B}_5$ for ownership, and between $\mathcal{B}_3$, $\mathcal{B}_4$ and $\mathcal{B}_5$ for trusteeship.[14] Both of $\mathcal{B}_2$ and $\mathcal{B}_4$ are intermediate structures, where B_4 is supposed to contain the intervenients ownership and trusteeship and B_2 the intervenients *purchase, barter, inheritance*, *occupation*, *specification*, *expropriation* (for public purposes or for other reasons), which are grounds for ownership. B_1 contains grounds for the conditions in B_2, such as making a contract for purchase or barter respectively, having particular kinship relationship to a deceased person, appropriating something not owned, creating a valuable thing out of worthless material, getting a verdict on disappropriation of property, either for public purposes or for other reasons. B_3 contains different grounds for trusteeship. B_5 contains the legal consequences of ownership and trusteeship, respectively, in terms of powers, permissions and obligations.

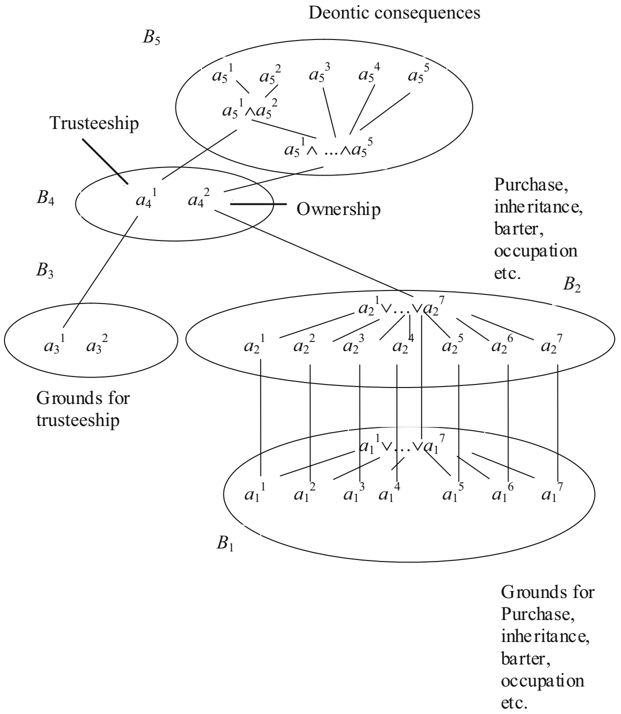


Fig. 3.

[14] Trust is where a person (trustee) is made the nominal owner of property to be held or used for the benefit of another. Trusteeship is the legal position of a trustee.

3.2 Weakest Grounds and Strongest Consequences

Definition 1. *A Bqo* $\mathcal{B}_i = \langle B_i, \wedge, ' , R_i \rangle$ lies within *an sBa* $\langle B, \wedge, ' , \rho \rangle$ *if* $\langle B_i, \wedge, ' \rangle$ *is a subalgebra of* $\langle B, \wedge, ' \rangle$ *and* $\rho | B_i = R_i$*. A Bjs* $\langle \mathcal{B}_1, \mathcal{B}_2, J \rangle$ lies within *an sBa* $\mathcal{S}$ *if* $\mathcal{B}_1$ *and* $\mathcal{B}_2$ *lie within* $\mathcal{S}$*,* $B_1 \cap B_2 \subseteq \{\top, \bot\}$ *and* $\rho | (B_1 \times B_2) = J$*.*

Suppose that $\mathcal{S} = \langle B, \wedge, ' , \rho \rangle$ is an *sBa* and that $\langle \mathcal{B}_1, \mathcal{B}_2, J_{1,2} \rangle$ is a *Bjs* lying within $\mathcal{S}$. That $a_1 \in B_1$ is one of the *weakest grounds* in B_1 of a_2 with respect to $\mathcal{S}$ is denoted $\mathrm{WG}_{\mathcal{S}}(a_1, a_2, B_1)$, and that $a_2 \in B_2$ is one of the *strongest consequences* of a_1 in $\mathcal{B}_2$ with respect to $\mathcal{S}$ is denoted $\mathrm{SC}_{\mathcal{S}}(a_2, a_1, \mathcal{B}_2)$. When there is no risk of ambiguity, we omit the subscript $\mathcal{S}$.

Definition 2

$$\mathrm{WG}_{\mathcal{S}}(a_1, a_2, B_1) \;\; \textit{iff} \;\; \langle a_1, a_2 \rangle \in J_{1,2} \;\; \& \; \forall b_1 \in B_1 : \langle b_1, a_2 \rangle \in J_{1,2} \longrightarrow b_1 R_1 a_1.$$
$$\mathrm{SC}_{\mathcal{S}}(a_2, a_1, B_2) \;\; \textit{iff} \;\; \langle a_1, a_2 \rangle \in J_{1,2} \;\; \& \; \forall b_2 \in B_2 : \langle a_1, b_2 \rangle \in J_{1,2} \longrightarrow a_2 R_2 b_2.$$

A weakest ground $a_1$of a_2 is *degenerated* if $a_1 \rho \bot$. A strongest consequence a_2 of a_1 is *degenerated* if $\top \rho a_2$.

Proposition 1. *(i) (See Fig.4A.) Suppose that* $\langle \mathcal{B}_1, \mathcal{B}_2, J \rangle$ *is a Bjs. Suppose further that* $a_1 R_1 b_1$*,* $\langle b_1, b_2 \rangle \in J$ *and* $b_2 R_2 a_2$*. Then* $\langle a_1, a_2 \rangle \in J$*.*

(ii) (See Fig.4B.) Suppose that $\langle \mathcal{B}_1, \mathcal{B}_2, J \rangle$ *is a Bjs and* $\mathrm{WG}(a_1, a_2, B_1)$ *and* $\mathrm{WG}(b_1, b_2, B_1)$*. If* $a_2 R_2 b_2$ *then* $a_1 R_1 b_1$*.*

(iii) If $\langle \mathcal{B}_1, \mathcal{B}_2, J \rangle$ *is a Bjs,* $\mathrm{WG}(a_1, a_2, B_1)$ *and* $\mathrm{WG}(b_1, a_2, B_1)$*, then* $a_1 Q_1 b_1$*.*

(iv) (See Fig.4C.) Suppose that $\langle \mathcal{B}_1, \mathcal{B}_2, J \rangle$ *is a Bjs and, that furthermore,* $\mathrm{SC}(a_2, a_1, B_2)$ *and* $\mathrm{SC}(b_2, b_1, B_2)$*. If* $a_1 R_1 b_1$ *then* $a_2 R_2 b_2$*.*

(v) If $\langle \mathcal{B}_1, \mathcal{B}_2, J \rangle$ *is a Bjs,* $\mathrm{SC}(a_2, a_1, B_2)$ *and* $\mathrm{SC}(b_2, a_1, B_2)$ *then* $a_2 Q_2 b_2$*.*

(vi) If $\mathrm{WG}(a_1, a_2, B_1)$ *and* $\mathrm{WG}(b_1, b_2, B_1)$ *then* $\mathrm{WG}(a_1 \wedge b_1, a_2 \wedge b_2, B_1)$*.*

(vii) If $\mathrm{SC}(a_2, a_1, B_2)$ *and* $\mathrm{SC}(b_2, b_1, B_2)$ *then* $\mathrm{SC}(a_2 \vee b_2, a_1 \vee b_1, B_2)$*.*

Some of the statements in the proposition above are exemplified below.

Example 1. (i) (See Fig.3.) We have $\langle a_1^1 \vee ... \vee a_1^7, a_2^1 \vee ... \vee a_2^7 \rangle \in J_{1,2}$, and $a_1^7 R_1 a_1^1 \vee ... \vee a_1^7$, and $a_2^7 R_2 a_2^1 \vee ... \vee a_2^7$. Hence $\langle a_1^7, a_2^7 \rangle \in J_{1,2}$.

(iii) (See Fig.3.) We have $\mathrm{WG}\langle a_1^1 \vee ... \vee a_1^7, a_2^1 \vee ... \vee a_2^7, B_1 \rangle$, and $\mathrm{WG}(a_1^7, a_2^7, B_1)$, and $a_2^7 R_2 a_2^1 \vee ... \vee a_2^7$. Hence $a_1^7 R_1 a_1^1 \vee ... \vee a_1^7$.

(v) (See Fig.3.) We have $\mathrm{SC}(a_2^1 \vee ... \vee a_2^7, a_1^1 \vee ... \vee a_1^7, B_2)$, and $\mathrm{SC}(a_2^7, a_1^7, B_2)$, and $a_1^7 R_1 a_1^1 \vee ... \vee a_1^7$. Hence $a_2^7 R_2 a_2^1 \vee ... \vee a_2^7$.

3.3 Minimality

In this section we study how the concept of minimal joining is related to the concepts of weakest ground and strongest consequence and prove some results on minimal joinings.

Theorem 1. *(See Fig.4D.) Suppose that* $\langle \mathcal{B}_1, \mathcal{B}_2, J \rangle$ *is a Bjs. Then* $\langle a_1, a_2 \rangle \in \min J$ *iff* $\mathrm{WG}(a_1, a_2, B_1)$ *and* $\mathrm{SC}(a_2, a_1, B_2)$*.*

Definition 3. *A Bjs* $\langle \mathcal{B}_1, \mathcal{B}_2, J \rangle$ *satisfies* connectivity *if whenever* $\langle c_1, c_2 \rangle \in J$ *there is* $\langle b_1, b_2 \rangle \in \min J$ *such that* $c_1 R_1 b_1$ *and* $b_2 R_2 c_2$*.*

Proposition 2. *Suppose that* $\langle \mathcal{B}_1, \mathcal{B}_2, J \rangle$ *is a Bjs which satisfies connectivity. Suppose further that* $\mathrm{WG}\,(a_1, a_2, B_1)$*. Then there is* $b_2 \in B_2 : \langle a_1, b_2 \rangle \in \min J$ *and* $b_2 R_2 a_2$*.*

The above proposition shows that a weakest ground of an element is the bottom of a minimal joining.

Proposition 3. *Suppose that* $\langle \mathcal{B}_1, \mathcal{B}_2, J \rangle$ *is a Bjs which satisfies connectivity. Suppose further that* $\mathrm{SC}\,(a_2, a_1, B_2)$*. Then there is* $b_1 \in B_1 : \langle b_1, a_2 \rangle \in \min J$ *and* $a_1 R_1 b_1$*.*

The above proposition shows that a strongest consequence of an element is the top of a minimal joining.

Proposition 4. *(See Fig. 4E.) Suppose that* $\langle \mathcal{B}_1, \mathcal{B}_2, J \rangle$ *is a Bjs that satisfies connectivity and* $\langle a_1, a_2 \rangle \in \min J$*. If* $\langle a_1, b_2 \rangle \in J$ *then* $a_2 R_2 b_2$ *and if* $\langle b_1, a_2 \rangle \in J$ *then* $b_1 R_1 a_1$*.*

Corollary 1. *(i) Suppose that* $\langle \mathcal{B}_1, \mathcal{B}_2, J \rangle$ *is a Bjs that satisfies connectivity. If* $\langle a_1, a_2 \rangle, \langle b_1, b_2 \rangle \in \min J$ *then* $a_1 R_1 b_1$ *iff* $a_2 R_2 b_2$*.*
(ii) (See Fig. 4F.) Suppose that $\langle \mathcal{B}_1, \mathcal{B}_2, J \rangle$ *is a Bjs that satisfies connectivity and* $\langle a_1, a_2 \rangle \in \min J$*. If* $\langle a_1, b_2 \rangle \in \min J$ *then* $a_2 Q_2 b_2$ *and if* $\langle b_1, a_2 \rangle \in \min J$ *then* $a_1 Q_1 b_1$*.*

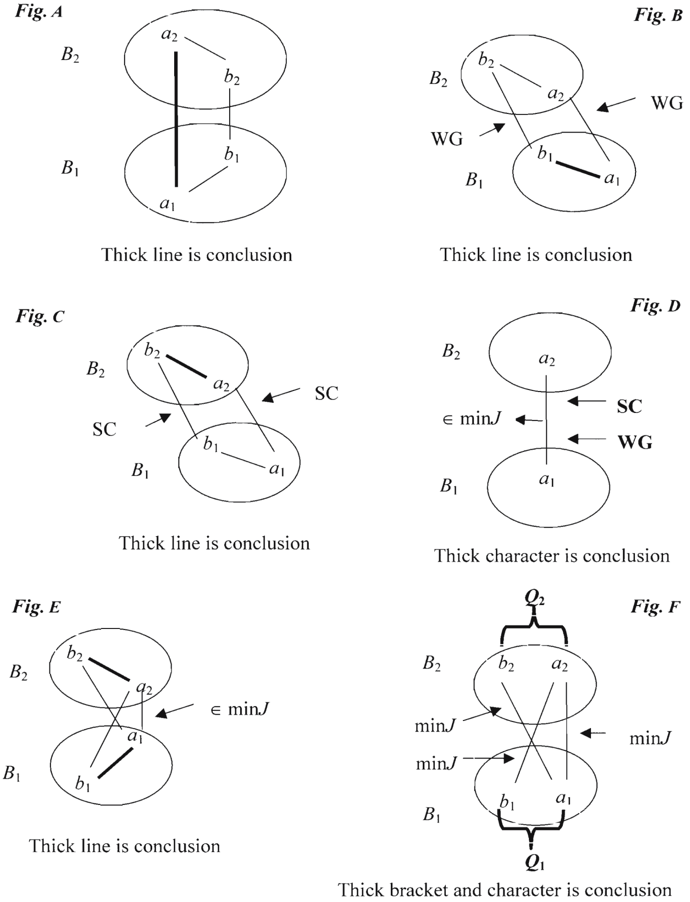
Theorem 2. *Suppose that* $\langle \mathcal{B}_1, \mathcal{B}_2, J \rangle$ *is a Bjs that satisfies connectivity. If* $\langle a_1, a_2 \rangle, \langle b_1, b_2 \rangle \in \min J$*, then there is* $c_2 \in B_2$*,* $d_1 \in B_1$ *such that* $\langle a_1 \wedge b_1, c_2 \rangle \in \min J$ *and* $\langle d_1, a_2 \vee b_2 \rangle \in \min J$*, and, furthermore, it holds that* $c_2 R_2 a_2 \wedge b_2$ *and* $a_1 \vee b_1 R_1 d_1$*.*

3.4 Intervenients

In this section we assume that $\mathcal{S} = \langle B, \wedge, ', \rho \rangle$ is an *sBa* and that $\langle \mathcal{B}_1, \mathcal{B}_2, J_{1,2} \rangle$, $\langle \mathcal{B}_2, \mathcal{B}_3, J_{2,3} \rangle$ and $\langle \mathcal{B}_1, \mathcal{B}_3, J_{1,3} \rangle$ are *Bjs* lying within $\mathcal{S}$ and satisfying connectivity. Suppose further that $J_{1,3} = J_{1,2} | J_{2,3}$. This means that $\langle a_1, a_3 \rangle \in J_{1,3}$ iff there is $a_2 \in B_2$ such that $\langle a_1, a_2 \rangle \in J_{1,2}$ and $\langle a_2, a_3 \rangle \in J_{2,3}$. We are interested in the relation between the minimal elements in the joining spaces $J_{1,2}$, $J_{2,3}$ and $J_{1,3}$ and, furthermore, the interplay between being an intervenient in B_2 and being a component in the minimal elements in the three joining spaces.

Definition 4. *We say that* a_2 *is an intervenient from* B_1 *to* B_3 *in* $\mathcal{S}$ *corresponding to* $\langle a_1, a_3 \rangle \in J_{1,3}$*, denoted* $a_2\ B_1 \curvearrowright^{\mathcal{S}} B_3 : \langle a_1, a_3 \rangle$*, if* $\mathrm{WG}_{\mathcal{S}}\,(a_1, a_2, B_1)$ *and* $\mathrm{SC}_{\mathcal{S}}\,(a_3, a_2, B_3)$ *and* a_1 *is a non-degenerated weakest ground and* a_3 *is a non-degenerated strongest consequence of* a_2*.*

In many situations the sets B_1 and B_3 and the system $\mathcal{S}$ are given by the context and we can simply write $a_2 \curvearrowright \langle a_1, a_3 \rangle$. Let $\mathrm{Iv}_{\mathcal{S}}\,(B_2, B_1, B_3)$ denote the set of elements in B_2 which are intervenients from B_1 to B_3 in $\mathcal{S}$.



Thick line is conclusion

Thick line is conclusion

Thick line is conclusion

Thick character is conclusion

Thick line is conclusion

Thick bracket and character is conclusion

Fig. 4.

Proposition 5. *If* $a_2 \curvearrowright \langle a_1, a_3\rangle$ *and* $a_2 \curvearrowright \langle b_1, b_3\rangle$ *then* $a_1 Q_1 b_1$ *and* $a_3 Q_3 b_3$.

Theorem 3. *(See Fig.5G.)* $\min J_{1,2} | \min J_{2,3} \subseteq \min J_{1,3}$.

Example 2. (See Fig.3.) We have $\langle a_2^1 \vee ... \vee a_2^7, a_4^2\rangle \in \min J_{2,4}$, and $\langle a_4^2, a_5^1 \wedge ... \wedge a_5^5\rangle \in \min J_{4,5}$. Hence, $\langle a_2^1 \vee ... \vee a_2^7, a_5^1 \wedge ... \wedge a_5^5\rangle \in \min J_{2,5}$.

Proposition 6. *(See Fig.5H.) Suppose* $\langle a_1, a_2\rangle \in \min J_{1,2}$, $\langle a_2, a_3\rangle \in \min J_{2,3}$, *not* $a_1 R_1 \bot$ *and not* $\top R_3 a_3$. *Then* $a_2 \curvearrowright \langle a_1, a_3\rangle$.

It does not hold generally that $\min J_{1,2} | \min J_{2,3} = \min J_{1,3}$. Thus:

Theorem 4. *(See Fig.5I.) Suppose that* $\langle a_1, a_3\rangle \in \min J_{1,3}$ *then there is* $a_2, b_2 \in B_2$ *such that* $\langle a_1, a_2\rangle \in \min J_{1,2}$ *and* $\langle b_2, a_3\rangle \in \min J_{2,3}$ *and* $a_2 R_2 b_2$.

Corollary 2. *(See Fig.5J.) Suppose that* $\langle a_1, a_3\rangle \in \min J_{1,3}$, *not* $a_1 R_1 \bot$ *and not* $\top R_3 a_3$. *Then there is* $a_2 \in B_2$ *such that* $a_2 \curvearrowright \langle a_1, a_3\rangle$.

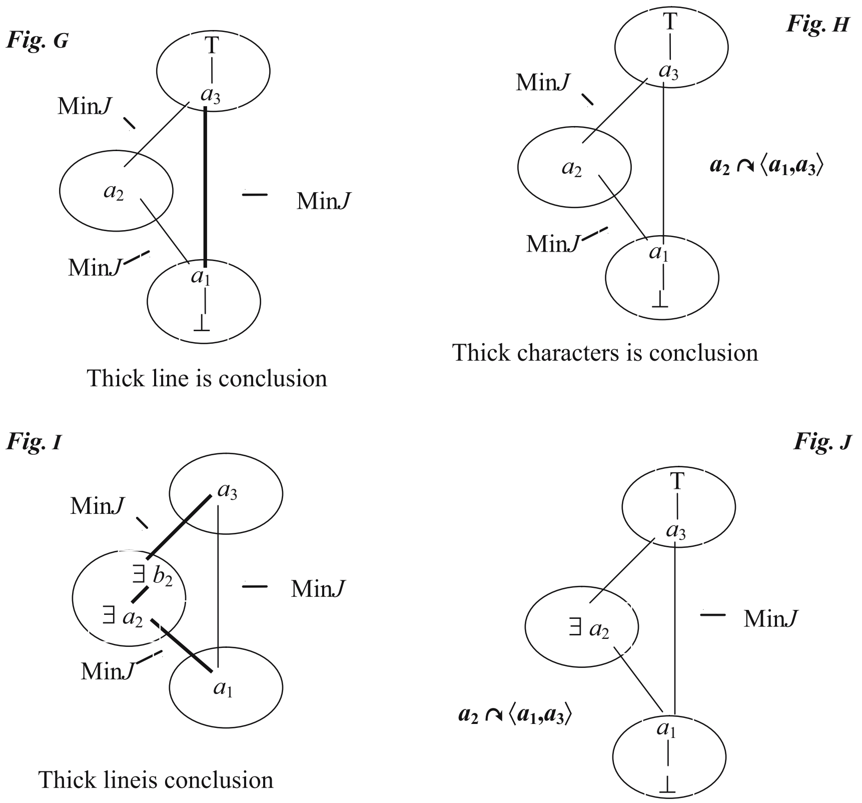


Fig. 5.

Proposition 7. *Suppose that* $a_2 \curvearrowright \langle a_1, a_3 \rangle \in \min J_{1,3}$ *and* $b_2 \curvearrowright \langle b_1, b_3 \rangle \in \min J_{1,3}$ *and not* $a_1 \wedge b_1 R_1 \bot$ *and not* $\top R_3 a_3 \vee b_3$. *Then the following holds:*

1. *If* $\langle a_1 \wedge b_1, a_3 \wedge b_3 \rangle \in \min J_{1,3}$ *then* $a_2 \wedge b_2 \curvearrowright \langle a_1 \wedge b_1, a_3 \wedge b_3 \rangle$.
2. *If* $\langle a_1 \vee b_1, a_3 \vee b_3 \rangle \in \min J_{1,3}$ *then* $a_2 \vee b_2 \curvearrowright \langle a_1 \vee b_1, a_3 \vee b_3 \rangle$.

Proposition 8. *If* $a_2 \curvearrowright \langle a_1, a_3 \rangle \in \min J_{1,3}$ *and* $b_2 \curvearrowright \langle b_1, b_3 \rangle \in \min J_{1,3}$ *and, furthermore, not* $a_1 \wedge b_1 R_1 \bot$ *and not* $\top R_3 a_3 \vee b_3$. *Then there are* $c_2, d_2 \in B_2$, $c_3 \in B_3$ *and* $d_1 \in B_1$ *such that*

$$c_2 \curvearrowright \langle a_1 \wedge b_1, c_3 \rangle \in \min J_{1,3} \textit{ and } d_2 \curvearrowright \langle d_1, a_3 \vee b_3 \rangle \in \min J_{1,3}.$$

3.5 Chains of More Than Three *Bqo*'s

A step towards analyzing more general structures in the law is taking into account chains of four or more *Bqo*'s. Let us pay regard to *Bjs*'s involving four *Bqo*'s $\mathcal{B}_1$, $\mathcal{B}_2$, $\mathcal{B}_3$, $\mathcal{B}_4$ such that $a_2 \curvearrowright \langle a_1, a_3 \rangle$ and $a_3 \curvearrowright \langle a_2, a_4 \rangle$ (See Fig. 6). Within our formal framework, we represent the intro-condition of a concept as

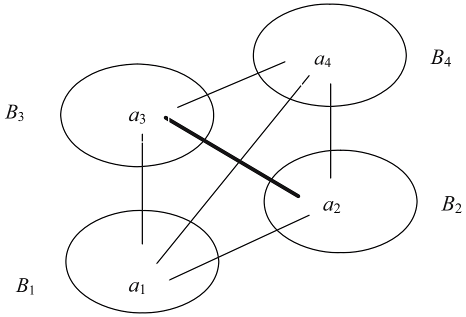


Fig. 6.

its weakest ground and the elim-condition as its strongest consequence. Thus a_2, a_3 are intervenients such that elim-condition of a_2 is a_3 and the intro-condition of a_3 is a_2. Expressed in terms of WG and SC this means that WG$\,(a_2, a_3, B_2)$ & SC$\,(a_3, a_2, B_3)$. According to theorem 1, this conjunction is equivalent to $\langle a_2, a_3\rangle \in \min J_{2,3}$. This is illustrated by the thick line in the figure.

Note that a chain of four *Bqo*'s can be continued at any length by adding $\mathcal{B}_5$, $\mathcal{B}_6$, and so on. In a chain of four or more *Bqo*'s, the previous results for three *Bqo*'s and intervenients will of course hold for any pair $\langle\langle \mathcal{B}_i, \mathcal{B}_j, J_{i,j}\rangle, \langle \mathcal{B}_j, \mathcal{B}_k, J_{j,k}\rangle\rangle$ of *Bjs*'s chosen from the chain.

A legal example of the result just mentioned is obtained in the case of ownership shown in Fig.3 above. In this example it can plausibly be assumed that (1) $a_2^1 \vee ... \vee a_2^7 \curvearrowright \langle a_1^1 \vee ... \vee a_1^7, a_4^2\rangle$ and (2) $a_4^2 \curvearrowright \langle a_2^1 \vee ... \vee a_2^7, a_5^1 \wedge ... \wedge a_5^5\rangle$. From (1) we derive $\mathrm{SC}_{\mathcal{S}}\left(a_4^2, a_2^1 \vee ... \vee a_2^7, B_4\right)$ and from (2) $\mathrm{WG}_{\mathcal{S}}\left(a_2^1 \vee ... \vee a_2^7, a_4^2, B_2\right)$. The conjunction of these two statements is equivalent to $\langle a_2^1 \vee ... \vee a_2^7, a_4^2\rangle \in \min J_{2,4}$. Thus we derive that the joining from *purchase or barter or inheritance ...* etc. to *ownership* is minimal.

3.6 The Different Kinds of Intervenient-Minimality

The previous sections illustrate the role of intervenient concepts in the representation of a normative system. Of special interest is where intervenients exhibit different kinds of *minimality*. Thus in a previous paper [11] (see Section 4) we underlined the central role of minimal joinings. One aspect of this importance relates to the effective formulation of a system; another aspect relates to greater facility when it comes to changing the system. Also, transitions from one normative system to another can be studied by an investigation of the formal structure of the set of minimal joinings.

The previous sections provide tools for distinguishing between different kinds of intervenient minimality. If $a_2 \in \mathrm{Iv}_{\mathcal{S}}\,(B_2, B_1, B_3)$ and $a_2 \curvearrowright \langle a_1, a_3\rangle$, we say that,

1. a_2 is *correspondence-minimal* if $\langle a_1, a_3\rangle \in \min J_{1,3}$,
2. a_2 is *ground-minimal* if $\langle a_1, a_2\rangle \in \min J_{1,2}$,
3. a_2 is *consequence-minimal* if $\langle a_2, a_3\rangle \in \min J_{2,3}$.

Combining these three cases with their negations ¬1, ¬2, ¬3, a total of eight (2^3) cases is obtained. In the case ¬1&¬2&¬3, the intervenient a_2 will be called *non-minimal*.

Not all eight cases are possible to realize. According to theorem 3, 1 is implied by 2&3. Hence, the case ¬1&2&3 is impossible to realize.

Due to the importance of minimality emphasized above, note that the following holds given the assumptions in section 3.4: Suppose C_2 is a subset of B_2 consisting of correspondence-minimal intervenients from B_1 to B_3 and it holds that if $\langle a_1, a_3 \rangle \in \min J_{1,3}$ then there is $c_2 \in C_2$ such that $c_2 \curvearrowright \langle a_1, a_3 \rangle$. Then

$$J_{1,3} = \{\langle a_1, a_3 \rangle \in B_1 \times B_3 \mid \exists c_2 \in C_2 : \langle a_1, c_2 \rangle \in J_{1,2} \text{ and } \langle c_2, a_3 \rangle \in J_{2,3}\}\,.$$

Hence, a set of correspondence-minimal intervenients can be a convenient way for characterizing a set of joinings.[15] However, intervenients can be useful even if they are not correspondence-minimal.[16]

4 Conclusion

In the paper, continuing previous work on intermediate concepts, we have proved a number of results on weakest grounds, strongest consequences, minimal joinings and intervenients. We have taken a step towards extending the theory so as to incorporate networks of structures of intervenients. We have distinguished different kinds of intervenient-minimality, thereby establishing a rudimentary typology.

Important tasks for future work is to incorporate a number of issues, dealt with in earlier papers, into the extended framework of networks of intervenients. These issues are: “openness” and extendability of intervenients, Boolean combinations of intervenients, *gic*-systems, bases of intervenients for a system. A theory covering these issues is necessary for a satisfactory representation of the network of concepts in a legal system.

Acknowledgement

Financial support was given by the Knowledge Foundation, The Swedish Research Council, the University of Gävle, and Harald and Louise Ekmans Forskningsstiftelse.[17]

[15] Cf. the fact that, since $\langle \mathcal{B}_1, \mathcal{B}_2, J_{1,3} \rangle$ satisfies connectivity, it holds that

$$J_{1,3} = \{\langle a_1, a_3 \rangle \in B_1 \times B_3 \mid \exists \langle b_1, b_3 \rangle \in \min J_{1,3} : \langle b_1, b_3 \rangle \trianglelefteq \langle a_1, a_3 \rangle\}$$

[16] A type worth considering is ¬1&2&¬3, i.e., where a_2 is ground-minimal but neither correspondence-minimal nor consequence-minimal. For instance, murder and high treason can have the same legal consequence (life imprisonment) notwithstanding that these crimes have different grounds. See also [11], Section 3.2 for the case of “Boche” in the “Boche-Berserk” example. “Boche” and “Berserk” have different grounds but the same consequence.

[17] The paper, as well as our earlier joint papers, are the result of wholly joint work where the order of appearance of our author names has no significance.

References

1. Aristotle.: The Nichomachean Ethics, transl. by Ross, W. D., Book V, http://classics.mit.edu/Aristotle/nicomachaen.5.v.html
2. Grossi, D.: Designing Invisible Handcuffs. Formal Investigations in Institutions and Organizations for Multi-agent Systems. SIKS Dissertation Series No. 2007-16 (2007)
3. Lindahl, L.: Position and Change. A Study in Law and Logic. Reidel, Dordrecht (1977)
4. Lindahl, L.: Deduction and Justification in the Law: The Role of Legal Terms and Concepts. Ratio Juris 17, 182–202 (2004)
5. Lindahl, L., Odelstad, J.: Intermediate Concepts as Couplings of Conceptual Structures. In: Prakken, H., McNamara, P. (eds.) Norms, Logics and Informations Systems. New Studies on Deontic Logic and Computer Science. IOS Press, Amsterdam (1999)
6. Lindahl, L., Odelstad, J.: An Algebraic Analysis of Normative Systems. Ratio Juris 13, 261–278 (2000)
7. Lindahl, L., Odelstad, J.: Normative Systems and Their Revision: An Algebraic Approach. Artificial Intelligence and Law 11, 81–104 (2003)
8. Lindahl, L., Odelstad, J.: Normative Positions within an Algebraic Approach to Normative Systems. Journal Of Applied Logic 2, 63–91 (2004)
9. Lindahl, L., Odelstad, J.: Intermediate Concepts in Normative Systems. In: Goble, L., Meyer, J.-J.C. (eds.) DEON 2006. LNCS (LNAI), vol. 4048. Springer, Heidelberg (2006a)
10. Lindahl, L., Odelstad, J.: Open and Closed Intermediaries in Normative Systems. In: van Engers, T.M. (ed.) Legal Knowledge and Information Systems (Jurix 2006). IOS Press, Amsterdam (2006b)
11. Lindahl, L., Odelstad, J.: Intermediaries and Intervenients in Normative Systems. Journal of Applied Logic (June 29, 2007); (Article in Press, Corrected Proof), doi:10.1016/j.jal.2007.06.010
12. Odelstad, J., Lindahl, L.: The Role of Connections as Minimal Norms in Normative Systems. In: Bench-Capon, T., Daskalopulu, A., Winkels, R. (eds.) Legal Knowledge and Information Systems. IOS Press, Amsterdam (2002)
13. Sartor, G.: The Nature of Legal Concepts: Inferential Nodes or Ontological Categories? EUI working paper LAW No. 2007/08. European University Institute. Department of Law (2007)

A Deontic Logic for Socially Optimal Norms

Jan Broersen, Rosja Mastop, John-Jules Ch. Meyer, and Paolo Turrini

Universiteit Utrecht

> “To assume that the ranking [of a fixed pair of alternative social states] does not change with any changes in individual values is to assume [...] that there exists an objective social good defined independently of individual desires. [...] Such a philosophy could be and was used to justify government by the elite, secular or religious, although we shall see below that the connection is not a necessary one.”
>
> (Kenneth Arrow, *Social Choice and Individual Values*, p.22)

Abstract. The paper discusses the interaction properties between preference and choice of coalitions in a strategic interaction. A language is presented to talk about the conflict between coalitionally optimal and socially optimal choices. Norms are seen as social constructions that enable to enforce socially desirable outcomes.

Introduction

One fundamental issue of social choice theory [1] is how to aggregate the preferences of individual agents in order to form decisions to be taken by society as a whole. However, once we want to take into account the capabilities of agents, as we do in Multi Agent Systems, mere social choice functions are not enough to explain how and (especially) why individual interests are aggregated in the way they are. In this context, norms should be seen as social constructions that enable us to enforce socially desirable outcomes [4].

In particular there are situations in which individual preferences are not compatible and coalitions compete to achieve a given social order. A typical case is that of an agent’s capability to positively or negatively affect the realization of other agents’ preferences. In our paper we will view the enactment of norms as aimed at the regulation of such interactions. By enacting a norm we mean *the introduction of a normative constraint on individual and collective choices in a Multi Agent System.*

We are specifically concerned with cases where the collective perspective is at odds with the individual perspective. That is, cases where we think that letting everybody pick their own best action regardless of others’ interest gives a non-optimal result. The main question we are dealing with is then: how do we determine which norms, if any, are to be imposed?

To answer this question, the paper presents a language to talk about the conflict between coalitionally optimal and socially optimal choices, and it expresses deontic notions referring to such circumstances.

R. van der Meyden and L. van der Torre (Eds.): DEON 2008, LNAI 5076, pp. 218–232, 2008.

Table 1. Lying or not lying

Column Row	Truth	Lie
Truth	$(3,3)$	$(0,4)$
Lie	$(4,0)$	$(1,1)$

Table 2. Clothing Conformity

Column Row	White Dress	Black Dress
White Dress	$(3,3)$	$(0,0)$
Black Dress	$(0,0)$	$(3,3)$

Motivating Examples. Let us consider a situation (Table 1) in which two players have the possibility of passing believed (truth) or disbelieved information (lie). If both players do not lie, they share their information, being both better off. If they both lie, they do not receive any advantage. But the worst case for a player is the one in which he does not lie and the opponent does.

In this situation, a legislator that wants to achieve the socially optimal state (players do not lie), should declare that lying is forbidden, thereby labeling the combinations of moves (lie, lie), (lie, truth) and (truth, lie) as violations.

The lying matrix is nothing but a Prisoner Dilemma [8], that is an interactive situation in which the advantages of cooperation are overruled by the incentive for individual players to defect. In Prisoner Dilemmas, individually rational players have no incentive to cooperate, because defecting is better for a player considering all possible answers of the opponent. Note that cooperation is in the interest of the players themselves, since they would be better off than if they had pursued the unique Nash Equilibrium [8], ending up in the (lie, lie) state. However it is by no means clear that players should not pursue their own interest. In fact once we reach a state in which one player lies and the other does not lie, we cannot move to any other state without one of the two players being worse off.

Other interesting examples concern conventional norms. A type of these norms are those in which players should conform to the other (i.e. *When in Rome do as the Romans do*). Suppose you have the usual two players who have to decide what to wear, with the goal of being conformant to the others' choice (see Table 2).

This setting boils down to a classical 'coordination' game. In this game the outcomes are good for both in case both make the same choice (e.g. they both decide to wear white), they are bad for both otherwise (e.g. one decides for white the other one for black). The preferred outcomes of the players are in fact the same. A norm helping players to reach an optimal outcome would be one that labels as violations combinations of discordant choices. However, in this kind of game Row will never know what the best thing to choose is, since the choice of Column is independent from his. In order to solve the problem a legislation should go beyond individual choice, by forcing the coalition made of Row and Column together to form and choose an efficient outcome.

A deontic logic for efficient interactions. Once we view a deontic language as regulating a Multi Agent System, we can say that a set of commands promote a certain interaction, prohibiting certain others. Following this line of reasoning it is possible, given a notion of optimality or efficiency, to construct a deontic language that requires this notion to hold.

What we do then is to provide a deontic language for all possible interactions, based on an underlying notion of optimality. This is quite a difference from the legal codes that we can find in a certain society, where norms are either explicitly and specifically formulated and written down in law books, contracts, etc., or are left implicit in the form of promises, values or mores [4]. The obligations and prohibitions in our system result from one general norm saying that all actions of sub-groups that do not take into account the interests of the society as a whole, are forbidden. Then, one way to use our logic is to derive obligations, permissions and prohibitions from conflicting group preferences, and use these as *suggestions* for norm introduction in the society.

We do not claim that the meaning of these operators, as studied in deontic logic, corresponds to our semantics, but rather we claim that when people make new norms they should choose those norms on the basis of the economical order behind them.

In order to represent abilities of agents we employ coalition logic [10], and we model an agent's preferences as a preorder on the domain of discourse. To model optimal social norms we introduce a generalization of the economical notion of Strong Pareto Efficiency (see for instance [8]), described as those sets of outcomes from which the grand coalition (i.e. the set of all agents taken together) has no interest to deviate. Our generalization consists of the fact that we do not make the assumption that these outcomes are singletons. In particular (unless specified) we do not make the assumption of playability described in [10], according to which the set of all agents can bring about any realizable outcome of the system. We consider then the elements of the complement of the efficient choices, i.e. all those that are not optimal, and we build the notion of obligation, prohibition and permission on top of them.

We postpone to future work all considerations about the effectivity of the norm, that is, all considerations about how, to what extent and in what way, the norm influences the behaviour of the agents involved.

As system designers, our aim is then to construct efficient social procedures that can guarantee a socially desirable property to be reached. We think that normative system design is at last a proper part of the Social Software enterprise [9].

The paper is structured as follows: In the first section we introduce the notions of effectivity and preference, discuss its relevant properties with respect to the problem of finding optimal social norms, and introduce the notion of domination, Pareto Efficiency and violation. In the second part we describe the syntax, the structures and the interpretation of our language. In the third part we discuss the deontic and collective ability modalities and their properties, and compare them with classical deontic and agency logics; moreover we discuss the introduction of further constraints in the models, in particular playability of the effectivity function. We show some examples to give the flavour of the situations we are able to capture with our formalism. A discussion of future work will follow and a summary of the present achievements will conclude the paper.

1 Effectivity and Preference

We start by defining some concepts underlying the deontic logic of this paper. They concern the *power* and the *preferences* of collectives. We begin with the first of these, by introducing the concept of a dynamic effectivity function, adopted from [10].

1.1 Effectivity

Definition 1 (Dynamic Effectivity Function)
Given a finite set of agents Agt and a set of states W, a dynamic effectivity function *is a function $E : W \to (2^{Agt} \to 2^{2^W})$.*

Any subset of Agt will henceforth be called a *coalition*. For elements of W we use variables $u, v, w, \ldots$; for subsets of W we use variables $X, Y, Z, \ldots$; and for sets of subsets of W (i.e., subsets of 2^{2^W}) we use variables $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \ldots$. The elements of W are called 'states' or 'worlds'; the subsets of Agt are called 'coalitions'; the sets of states $X \in E(w)(C)$ are called the 'choices' of coalition C in state w. The set $E(w)(C)$ is called the 'choice set' of C in w. The complement of a set $\overline{X}$ or of a choice set $\overline{\mathcal{X}}$ are calculated from the obvious domains.

A dynamic effectivity function assigns, in each world, to every coalition a set of sets of states. Intuitively, if $X \in E(w)(C)$ the coalition is said to be able to *force* or *determine* that the next state after w will be some member of the set X. If the coalition has this power, it can thus prevent that any state *not* in X will be the next state, but it might not be able to determine *which* state in X will be the next state. Possibly, some other coalition will have the power to refine the choice of C.

Many properties can be attributed to dynamic effectivity functions. An extensive discussion of them can be found in [10]. For what follows we do not need all the properties that may be considered reasonable for effectivity. However the following properties seem to be minimally required:

1. coalition monotonicity: for all X, w, C, D, if $X \in E(w)(C)$ and $C \subseteq D$, then $X \in E(w)(D)$;
2. regularity: for all X, w, C, if $X \in E(w)(C)$, then $\overline{X} \notin E(w)(\overline{C})$;

3. outcome monotonicity: for all X, Y, w, C, if $X \in E(w)(C)$ and $X \subseteq Y$, then $Y \in E(w)(C)$;
4. inability of the empty coalition: for all w, $E(w)(\emptyset) = \{W\}$

If a dynamic effectivity function has these properties, it will be called *coherent.*

The first property says that the ability of a coalition is preserved by enlarging the coalition. In this sense we do not allow new members to interfere with the preexistent capacities of a group of agents. The second property says that if a coalition is able to force the outcome of an interaction to belong to a particular set, then no possible combinations of moves by the other agents can prevent this to happen. We think that regularity is a key property to understand the meaning of ability. If an agent is properly able to do something this means that others have no means to prevent it. The third property says that if a coalition is able to force the outcome of the interaction to belong to a particular set, then that coalition is also able to force the outcome to belong to all his supersets. Outcome monotonicity is a property of all effectivity functions in coalition logic, which is therefore a monotonic modal logic [10]. The last condition is the "Inability of the Empty Coalition". As notice also by [2] such properties forces the coalition modality for the empty coalition to be universal: intuitively the empty coalition cannot bring about non-trivial consequences.

Proposition 1. *If the effectivity function is coherent then all coalitional effectivity functions are nonempty and do not contain the empty set.*

The last property ensures that the choice of the empty coalition is always the largest possible one. This property imposes that for all $C, w, E(w)(C) \neq \emptyset$ (by coalition monotonicity) and that $\emptyset \notin E(w)(C)$ (by regularity).

1.2 Preference

Once we have defined the notion of effectivity, we also need to make reference to the preferences of coalitions. The notion of preference in strategic interaction can be understood and modeled in many ways [13]. However we believe that in strategic reasoning players need to have preferences over the possible outcomes of the game. Thus those are the preferences that constitute our main concern. Nevertheless we know from the properties of effectivity as described above, that coalitions may have different abilities at different states, in particular the grand coalition of agents may gain or lose power while changing a state: the effectivity is actually a dynamic effectivity.

The claim is thus that agents do have a fixed ordering over the domain of discourse (what we call *desires*), and that they generate their strategic preference considering where the game may end (called *domination*). We are going to define both, discussing their properties.

We start from a preference relation for individuals over states working our way up to preferences for coalitions over sets. To do so, we start from an order on singletons, and we provide some properties to lift the relation to sets.

Definition 2 (Individual desires for states). *A desire ordering* $(\geq_i)_{i \in Agt}$ *consists of a partial order (reflexive, transitive, antisymmentric)* $\geq_i \subseteq W \times W$ *for all agents* $i \in Agt$*, where* $v \geq_i w$ *has the intuitive reading that* v *is 'at least as nice' as* w *for agent* i*. The corresponding strict order is defined as usual:* $v >_i w$ *if, and only if,* $v \geq_i w$ *and not* $w \geq_i v$*.*

Definition 3 (Individual desires for sets of states). *Given a desire ordering* $(\geq_i)_{i \in Agt}$*, we lift it to an ordering on nonempty sets of states by means of the following principles.*

1. $\{v\} \geq_i \{w\}$ *iff* $v \geq_i w$*;* *(Singletons)*
2. $(X \cup Y) \geq_i Z$ *iff* $X \geq_i Z$ *and* $Y \geq_i Z$*;* *(Left weakening)*
3. $X \geq_i (Y \cup Z)$ *iff* $X \geq_i Y$ *and* $X \geq_i Z$*.* *(Right weakening)*

Proposition 2. *The lifting preserves the partial order.*

The proof is omitted for reasons of space. We do not give a comprehensive specification of the logical properties of preference relations for coalitions, because this would not be relevant for the remainder of this paper. Different types of interaction may warrant the assumption of different properties for such a relation. Nevertheless, these are some properties that seem minimally required for calling some relation a preference relation. The first ensures that desires are copied to possible choices. The properties of left and right weakening ensure a lifting from singletons to sets.

The lifting enables us to deal with desire under uncertainty or indeterminacy. The idea is that if an agent were ever confronted with two choices X, Y he would choose X over Y provided $X >_i Y$. Desires do not consider any realizability condition, they are simply basic aspirations of individual agents, on which to construct a more realistic order on the possible outcomes of the game, which are by definition dependent on what all the agents can do together.

Out of agents' desires, we can already define a classical notion of Pareto Efficiency.

Definition 4 (Strong Pareto efficiency). *Given a choice set* $\mathcal{X}$*, a choice* $X \in \mathcal{X}$ *is* Strongly Pareto efficient *for coalition* C *if, and only if, for no* $Y \in \mathcal{X}$*,* $Y \geq_i X$ *for all* $i \in C$*, and* $Y >_i X$ *for some. When* $C = Agt$ *we speak of* Strong Pareto Optimality.

We will use the characterization of Pareto Efficiency and Optimality to refer to the notions we have just defined, even though the classical definitions (compare with [8]) are weaker.

The last definition is clearer when we consider the case $\mathcal{X} = E(w)(C)$, but it is formulated in a more abstract way in order to smoothen the next two definitions.

Proposition 3. *Given the preference relation over choices* $\geq_i$*, and taking* A, B *in a choice set* $\mathcal{X}$ *of a coaltion* C*, with* $PE(A)$ *to indicate that the choice* A *is Strongly Pareto Efficient in* $\mathcal{X}$*, Strong Pareto Efficiency is monotonic, that is* $A \subseteq B$ *implies that* $PE(B)$ *whenever* $PE(A)$*.*

Proof. Suppose $A \subseteq B$ and $PE(A)$ and suppose it is not the case that $PE(B)$. This means that there is a choice X in the choice set $\mathcal{X}$ of C such that $X >_i B$ for some $i \in C$. But being $A \subseteq B$ this would imply that $X >_i A$, contradicting the assumption that $PE(A)$.

Pareto Efficiency is usally defined disregarding the strategies of the players. Nevertheless, once we claim that the outcome of an interaction need not be a singleton, we need to adapt our evaluation of efficiency to such an assumption.

We now construct a preference relation on choices. To do so we first need to look at the interaction that agents' choices have with one another.

Definition 5 (Subchoice). *If E is an effectivity function, and $X \in E(w)(\overline{C})$, then the X-*subchoice set *for C in w is given by $E^X(w)(C) = \{X \cap Y \mid Y \in E(w)(C)\}$.*

As an example, let us take Table 1. Consider expressions of the form (Lie_C) to be intended as the set of worlds that make the proposition Lie_C true, with the obvious reading. In our example we have for instance the following cases:

- $E^{(Lie_C)}(w)(R) = \{(Lie_C \wedge Lie_R), (Lie_C \wedge Truth_R)\}$
- $E^{(Truth_C)}(w)(R) = \{(Truth_C \wedge Lie_R), (Truth_C \wedge Truth_R)\}$

Subchoices allow us to reason on a restriction of the game and to consider possible moves looking from a coalitional point of view, i.e. what is best for a coalition to do provided the others have already moved.

When agents interact therefore they make choices on the grounds of their own preferences. Nevertheless the moves at their disposal need not be all those that the grand coalition has. We can reasonably assume that preferences are filtered through a given coalitional effectivity function. That is we are going to consider what agents prefer among the things they can do.

Definition 6 (Domination). *Given an effectivity function E, X is* undominated *for C in w (abbr. $X \rhd_{C,w}$) if, and only if, (i) $X \in E(w)(C)$ and (ii) for all $Y \in E(w)(\overline{C})$, $(X \cap Y)$ is Pareto efficient in $E^Y(w)(C)$ for C.*

The idea behind the notion of domination is that if X' and X'' are both members of $E(w)(C)$ then, in principle, C will not choose X'', if X' dominates X''. This property ensures that a preference takes into account the possible moves of the other players. This resembles the notion of Individual Rationality in Nash solutions [8], according to which an action is chosen reasoning on the possible moves of the others.

Continuing our example, we have the following cases:

- $(Lie_R) \rhd_{R,w}$ for any w.
- $(Lie_C) \rhd_{C,w}$ for any w.
- not $(Lie_C, Lie_R) \rhd_{Agt,w}$

The preceding three definitions capture the idea that 'inwardly' coalitions reason Pareto-like, and 'outwardly' coalitions reason strategically, in terms of strict domination. A coalition will choose its best option given all possible moves of the opponents. Looking at the definition of Optimality we gave, we can see that undomination collapses to individual rationality when we only consider individual agents, and to Pareto efficiency when we consider the grand coalition of agents.

Proposition 4
$X \rhd_{Agt,w}$ iff X is a standard Pareto Optimal Choice in w.
$X \rhd_{i,w}$ iff X is a standard Dominating Choice in w for i.

Proof. For the first, notice that since $E(w)(\emptyset) = \{W\}$ (i.e., the empty coalition has no powers), then X is undominated for Agt in w iff it is Pareto efficient in $E(w)(Agt)$ for Agt (i.e., it is Pareto optimal in w). The second is due to the restriction of undomination to singleton agents.

Nevertheless, in our framework, domination is a relation between the choice sets of a given coalition. This approach looks different from the standard one (see for instance Osborne and Rubinstein [8]) that considers instead domination as a property of states.

Proposition 5. *Game-theoretical domination is expressible in our framework.*

It is possible to rewrite a domination of a state x over a state y as the domination of the choice $\{x\}$ over $\{y\}$, making it a particular case of our definition.

Proposition 6. *Undomination is monotonic, that is for X in $E(C)(w)$ for some C, w, if $X \subseteq Y$ and $X \rhd_{C,w}$ then $Y \rhd_{C,w}$.*

which follows from monotonicity of Pareto Optimality for choices and outcome monotonicity.

Violation. The fundamental idea of this work is that an efficient way to impose normative constraints in a Multi Agent System is to look at the optimality of the strategic interaction of such system. In particular the presence of possible outcomes in which agents could not unanimously improve (Pareto Efficient) can be a useful guide line for designing a new set of norms to be imposed.

Following this line we define a set of violation sets as the set of those choices that are not a Pareto Efficient interaction.

Definition 7 (Violation). *If E is an effectivity function and $C \subseteq C'$, then the choice $X \in E(w)(C)$ is a violation by C towards C' in w ($X \in VIOL_{C,C',w}$) iff there is a $Y \in E(w)(C' \setminus C)$, s.t. $(X \cap Y)$ is not undominated for C' in w.*

In words, X is a violation if it is not safe for the other agents, in the sense that not all the moves at their disposal yield an efficient outcome.

We indicate with $VIOL_{C,w}$ the set violations by C at w towards Agt[1].

Proposition 7. *If $C{=}C'$ then a violation is a dominated choice; If $C{=}C'{=}Agt$ a violation is a Pareto inefficient choice.*

If we consider the Prisoner Dilemma of Table 1 the following holds:

- $(Lie_R) = VIOL_{R,w}$ for any w, since $(Lie_R \wedge Lie_C)$ is not Agt- undominated;
- $(Lie_C \wedge Lie_R) = VIOL_{Agt,w}$ for any w, since not Pareto Efficient.

If we consider instead the Coordination Game of Table 2 the following holds:

- $(White_R) \in VIOL_{R,w}$, since $(White_R \wedge Black_C)$ is not Agt-undominated;
- $(Black_R) \in VIOL_{R,w}$, since $(White_C \wedge Black_R)$ is not Agt-undominated.

We can observe here that any choice made by a single agent is a violation. The reason why it is so has to be found in the form of the game, that requires the grand coalition to form for an efficient outcome to be forced.

2 Logic

We now introduce the syntax of our logic, an extension of the language of coalition logic [10] with modalities for permission, prohibition and obligation, and a modality for rational choice.

2.1 Language

Let Agt be a finite set of agents and $Prop$ a countable set of atomic formulas. The syntax of our logic is defined as follows:

$$\phi ::= p|\neg\phi|\phi \vee \phi|[C]\phi|P(C,\phi)|F(C,\phi)|O(C,\phi)|[rational_C]\phi$$

where p ranges over $Prop$ and C ranges over the subsets of Agt. The other boolean connectives are defined as usual. The informal reading of the modalities is: "Coalition C can choose ϕ", "It is permitted (/forbidden/obligated) for coalition C to choose ϕ", "It is rational for coalition C to choose ϕ".

[1] One interesting question is whether given any dynamic effectivity function and preference relation (with the above defined properties) we can always find a coalitionally dominated action (and hence a Pareto Efficient interaction). The acquainted reader will have noticed the resemblance of this problem with that of nonemptiness of the Core [8]. We leave though to further work the analysis of this relation. In case there is none, we may consider a satisfactory notion of optimal choice - as done for instance by Horty [5] - that looks at the relation between the choices in the choice sets of each coalition.

2.2 Structures

Definition 8 (Models). *A* model *for our logic is a quadruple*

$$(W, E, \{\geq_i\}_{i \subseteq Agt}, V)$$

where:

- *W is a nonempty set of states;*
- *$E : W \longrightarrow (2^{Agt} \longrightarrow 2^{2^W})$ is a coherent effectivity function, that associates to each state and each coalition a set of choices.*
- *$\geq_i \subseteq W \times W$ for each $i \in Agt$, is the desire relation, that associates to each agent a set of pairs of states. Out of this preference relation we define the undomination relation $\rhd \subseteq 2^{Agt} \times W \times 2^W \times 2^{2^W}$ as previously specified.*
- *$V : W \longrightarrow 2^{Prop}$ is a valuation function, that associates to every state a set of atomic propositions, with the intended meaning that the atoms associated to a state are all and only those true in that world.*

2.3 Semantics

The satisfaction relation of the formulas with respect to a pointed model M, w is defined as follows:

$$\begin{aligned}
M, w \models p &\text{ iff } p \in V(w) \\
M, w \models \neg\phi &\text{ iff } M, w \not\models \phi \\
M, w \models \phi \wedge \psi &\text{ iff } M, w \models \phi \text{ and } M, w \models \psi \\
M, w \models [C]\phi &\text{ iff } [[\phi]]^M \in E(w)(C) \\
M, w \models [rational_C]\phi &\text{ iff } \forall X (X \rhd_{C,w} \;\Rightarrow\; X \subseteq [[\phi]]^M) \\
M, w \models P(C, \phi) &\text{ iff } \exists X \in E(w)(C) \text{ s.t. } X \in \overline{VIOL}_{C,w} \text{ and } X \subseteq [[\phi]]^M \\
M, w \models F(C, \phi) &\text{ iff } \forall X \in E(w)(C)(X \subseteq [[\phi]]^M \Rightarrow X \in VIOL_{C,w}) \\
M, w \models O(C, \phi) &\text{ iff } \forall X \in E(w)(C)(X \in \overline{VIOL}_{C,w} \Rightarrow X \subseteq [[\phi]]^M)
\end{aligned}$$

In this definition, $[[\phi]]^M =_{def} \{w \in W \mid M, w \models \phi\}$.

The modality for coalitional ability is standard from Coalition Logic [10]. The modality for rational action requires for a proposition ϕ to be rational (wrt a coalition C in a given state w) that all undominated choices (for C in w) be in the extension of ϕ. This means that there is no safe choice for a coalition that does not make sure that ϕ will hold. Notice that it is still possible for a coalition to pursue a rational choice that may be socially not rational.

The deontic modalities are defined in terms of the coalitional abilities and preferences. A choice is permitted whenever it is safe, forbidden when it may be unsafe (i.e. when it contains an inefficient choice), and obligated when it is the only choice that is safe.

3 Discussion

The definition of strong permission does not allow for a permitted choice of an agent to be refined by the other agents towards a violation. In fact we define permission for ϕ as “a ϕ-choice guarantees safety from violation”. A more standard diamond modality would say “ doing ϕ is compatible with no violation”. A “safety” definition of permission has also been studied in [14], [11], [7], [3].

3.1 Properties

It is now interesting to look at what we can say and what we cannot say within our system.

Some Validities	
1	$P(C,\phi) \rightarrow \neg O(C,\neg\phi)$
2	$F(C,\phi) \leftrightarrow \neg P(C,\phi)$
3	$P(C,\phi) \vee P(C,\psi) \rightarrow P(C,\phi \vee \psi)$
4	$O(C,\phi) \rightarrow ([C]\phi \rightarrow P(C,\phi))$
5	$[rational_C]\phi \wedge [rational_{Agt}]\neg\phi \rightarrow F(C,\phi)$
6	$O(C,\phi) \vee O(C,\psi) \rightarrow O(C,\phi \vee \psi)$
7	$O(C,\top)$
8	$F(C,\phi) \wedge F(C,\psi) \rightarrow F(C,\phi \wedge \psi)$
9	$[rational_C]\phi \wedge (\phi \rightarrow \psi) \rightarrow [rational_C]\psi$

Some non-Validities	
10	$\neg O(C,\neg\phi) \rightarrow P(C,\phi)$
11	$P(C,\phi \vee \psi) \rightarrow P(C,\phi) \vee P(C,\psi)$
12	$O(C,\phi) \leftrightarrow \neg O(C,\neg\phi)$
13	$[rational_C]\phi \leftrightarrow [rational_{Agt}]\phi$
14	$O(C,\phi) \rightarrow P(C,\phi)$
15	$O(C,\phi \vee \psi) \rightarrow O(C,\phi) \vee O(C,\psi)$

The first validity says that the presence of permission imposes the absence of contrasting obligations, but the converse in not necessarily true. The second that prohibition and permission are interdefinable. The third says that the permission of ϕ or the permission of ψ implies the permission of ϕ or ψ. The fourth that the obligation to choose ϕ for an agent plus the ability to do something entails the permission to carry out ϕ. The validity number 5 says that the presence of a safe state that is rational for the grand coalition of agents is a norm for every coalition, even in case of conflicting preferences, i.e. in case of conflict the interest of the grand coalition prevails. The sixth one that obligation for ϕ or obligation for ψ implies the obligation for ϕ or ψ. Validity 7 says that there are no empty normative systems. The next validity says that prohibition is conjunctive. The last validity says that rational moves are monotonic. This has interesting implications on the choices of the agents, since refraining, i.e. choosing the biggest possible outcome, is always rational.

It is also useful to look at the non-validities: Number 10 says that if an agent is not obliged to choose something then it is permitted to do the contrary. But of course an agent may not be able to do anything, or may be not able to refine the choices "until the optimal". The next non-validity says that a permission of choice is not equivalent to a choice of permission. Number 11 says that a coalition can be obliged to do contradictory choices. This situation happens when a coalition is powerless or optimality is not possible. The next non validity says that the rational action for a certain coalition does not necessarily coincide with that of the grand coalition. Number 14, that ought does not imply can. The last does not allow to detach specific obligation from obligatory choices.

Further Assumptions

Playability. Our notion of agency is more general than that of game theory. In particular we assume that even the grand coalition of agents may not determine a precise outcome of the interaction.

This is due to the abandonment of the property of playability of the effectivity function, that requires, together with regularity, outcome monotonicity, coalition monotonicity that:

- $X \notin E(C)$ implies $\overline{X} \in E(\overline{C})$, that is any choice excluded to a coalition is possible for the rest of the agents (maximality);
- For all X_1, X_2, C_1, C_2 such that $C_1 \cap C_2 = \emptyset$, $X_1 \in E(C_1)$ and $X_2 \in E(C_2)$ imply that $X_1 \cap X_2 \in E(C_1 \cup C_2)$ (superadditivity)

Playability is a very strong property but it is needed to talk about games. As proved in [10] [Theorem 2.27], strategic games correspond exactly to playable effectivity functions [2]. With playable effectivity functions, the grand coalition can determine the exact outcome of the game and the dynamic effectivity function for the grand coalition of agents is the same in any state.

$$M, w \models [Agt]\phi \Leftrightarrow M \models [Agt]\phi$$

Moreover the fact that desires do not change, induces the following stronger invariance:

$$M, w \models [rational_{Agt}]\phi \Leftrightarrow M \models [rational_{Agt}]\phi$$

So not only is every outcome reachable, but any situation shares the same social optimality. Notice that this is independent of the solution concept we may consider.

[2] The proof involves the definition of strategic game as a tuple $\langle N, \{\Sigma_i | i \in N\}, o, S\rangle$ where N is a set of players, each i being endowed with a set of strategies σ_i from Σ_i, an outcome function that returns the result of playing individual strategies at each of the states in S; the definition of α effectivity function for a nonempty strategic game G, $E^\alpha_G : \wp(N) \rightarrow \wp\wp(S)$ defined as follows: $X \in E^\alpha_G \exists\sigma_C \forall\sigma_{\overline{C}} o(\sigma_C;\sigma_{\overline{C}}) \in X$. The above mentioned theorem establishes that $E^\alpha_G = E$ in case E is playable and G is a nonempty strategic game.

Finite Domain. Another interesting assumption can be made about the finitness of the domain of discourse. With finite W, for $\mathcal{C}$ being the class of our models, we have that

Proposition 8. $\models_{\mathcal{C}} [rational_{Agt}]\phi$ *implies that there exists an efficient outcome in the class of coalition models* $\mathcal{C}$

Proof. Take an arbitrary world w and an arbitrary model M in the class $\mathcal{C}$. $[rational_{Agt}]\phi$ means that all undominated choices Y are such that $Y \subseteq [[\phi]]^M$. Suppose that there are no such Y (if there are, the proposition is trivially true). We know that the effectivity function of Agt is nonempty. We also know $E(Agt)$ has cardinality $k \leq |2^W|$, that by assumption is finite. There must then be a choice Y' that is not undominated. This means that there is a choice Y'' dominating Y' that is in turn not undominated. But being the preference relation $>$ transitive and irreflexive, there must be an infinite chain of undomination within the effectivity function of Agt, contradicting the assumption.

Another property is the following:

$$\models_{\mathcal{C}} [rational_{Agt}]\neg\phi \wedge [C]\phi \rightarrow F(C, \phi) \tag{REG}$$

which says that any coalition has to refrain from a choice that is against an optimal state independently of its own preferences. A corresponding propery for obligation is instead the following:

$$\models_{\mathcal{C}} [rational_{Agt}]\neg\phi \wedge [C]\neg\phi \rightarrow O(C, \neg\phi) \tag{REG'}$$

Coalitionally optimal norms. The logic can be extended to treat norms that do not lead to a socially optimal outcome, but a coalitionally optimal outcome. That is it is possible to construct a deontic logic that pursues the interests of a particular coalition, independently of the other players' welfare. This extension is related to the work of Kooi and Tamminga on conflicting moral codes [6].

We limit the description to the obligation operator, the others are straightforward.

$M, w \models O^{C'}(C, \phi)$ iff $\forall X(X \rhd_{C,w}$ and $X \in \overline{VIOL}_{C,C',w} \Rightarrow X \subseteq [[\phi]]^M)$
where $VIOL_{C,C',w}$ is a C violation towards C', with $C \subseteq C'$.
For this operator it holds that

$$\models_{\mathcal{C}} O^C(C, \phi) \leftrightarrow [rational_C]\phi$$

that is playing for oneself boils down to rational action, and

$$\models_{\mathcal{C}} O^{Agt}(C, \phi) \leftrightarrow O(C, \phi)$$

that is, with the new operator we can express our original obligation operator.

3.2 Examples

Norms of Cooperation. To consider forbidden all non optimal choices may seem a very strong requirement. Nevertheless, take the example in Table 1.

It is interesting to notice how $VIOL$ is not equivalent to the situations that each player is forbidden to choose. This is due to the fact that each player can only refine the choices of the other players, but cannot determine alone the outcome of the game: a permitted choice cannot be refined by permitted choices towards an inefficient outcome. Moreover $M \models [R]\neg(T_R) \wedge [rational_{R,C}](T_R)$, that by (REG) allows to conclude $F(R, \neg(T_R))$.

No agent is in fact obligated not to lie, but only permitted. Why is it so? Because no agent can alone reach a singleton state that is only good. But of course as a coalition $\{R, C\}$ has the obligation to end up in the optimal state.

Norms of Conformity. Consider now the game of Conformity described in Table 2. What is interesting here is that players are individually permitted nothing:

$M \models \neg P(R, white_R) \wedge \neg P(R, black_R) \wedge \neg P(C, black_C) \wedge \neg P(C, black_C)$.

But as coalition they are:

$M \models P(\{R, C\}, white_{R,C}) \wedge M \models P(\{R, C\}, black_{R,C})$.

No precise indication of the choices is given by the resulting obligation:

$M \models O(\{R, C\}, (white_R, C) \vee (black_{R,C}))$

Notice that we have no way of detaching from this choice a more precise command, for $M \not\models O(\{R, C\}, (white_{R,C}) \vee (black_{R,C})) \rightarrow O(\{R, C\}, (white_{R,C})) \vee O(\{R, C\}, (black_{R,C}))$ (as witness for invalidity 15). This is revealing of the form of the game: no equilibrium can be achieved by the agents acting independently, but only as a coalition.

Both the Prisoner Dilemma and the Coordination Game have rules that say something about how coalitions should choose, and most interestingly how coalitions should form. While in the first game individual agents can make a permitted choice, in the second game individually players are permitted nothing: what the deontic statements in fact claim is that in the one case individual choice can lead to optimality, while in the second case the grand coalition is obligated to form, that being the only way of taking an efficient decision.

3.3 Future Work

The work here described allows for several developments. Among the most interesting ones is the study of the relation between imposed outcomes and steady states that describe where the game will actually end up (i.e. Nash Solution, the Core etc.). Conversely another feature that is worth studying is those structures in which Pareto Efficiency is not always present. Agents will reckon some actions as optimal even though there is no social equilibrium that can ever be reached. One more feature concerns the possibility of an inconsistent normative system. Further work could be done looking at the factual obedience of the norm, and how a norm affects preferences of agents (see for instance the work in [12]).

4 Conclusion

In this paper we proposed a deontic logic for optimal social norms. We described the concept of social optimality, explicitly linking it with the economical concept of Pareto Efficiency. Moreover we generalized the notion of Pareto Efficiency to capture those strategic interactions in which even the grand coalition of agents is not able to achieve every outcome. Technically we did not assume playability of the coalitional effectivity functions. It is an important question in itself to understand the class of interactions to which such effectivity function corresponds. On top of the notion of Optimality we constructed a deontic language to talk about a normative system resulting from the imposition of such norms. We analyzed the properties of the language and discussed in details various examples from game theory and social science.

References

1. Arrow, K.: Social Choice and Individual Values. Yale University Press (1970)
2. Borgo, S.: Coalitions in action logic. In: Proc. of IJCAI, pp. 1822–1827 (2007)
3. Broersen, J.: Modal action logics for reasoning about reactive systems. PhD-thesis Vrije Universiteit Amsterdam (2003)
4. Coleman, J.: Foundations of Social Theory. Belknap Harvard (1990)
5. Horty, J.: Deontic Logic and Agency. Oxford University Press, Oxford (2001)
6. Kooi, B., Tamminga, A.: Conflicting obligations in multi-agent deontic logic. In: Goble, L., Meyer, J.-J.C. (eds.) DEON 2006. LNCS (LNAI), vol. 4048, pp. 175–186. Springer, Heidelberg (2006)
7. Mastop, R.: What can you do? imperative mood in semantic theory. PhD thesis, ILLC Dissertation Series (2005)
8. Osborne, M., Rubinstein, A.: A course in Game Theory. MIT Press, Cambridge (1994)
9. Parikh, R.: Social software. Synthese 132(3), 187–211 (2002)
10. Pauly, M.: Logic for Social Software. ILLC Dissertation Series (2001)
11. van Benthem, J.: Minimal deontic logics. Bulletin of the Section of Logic 8(1), 36–42 (1979)
12. van Benthem, J., Liu, F.: Dynamic logic of preference upgrade. Journal of Applied Non-Classical Logics 14 (2004)
13. von Wright, G.: The logic of preference. Edimburgh University Press (1963)
14. Wright, G.V.: Deontic logic and the theory of conditions. In: Hilpinen, R. (ed.) Deontic Logic: Introductory and Systematic Readings, pp. 3–31 (1971)

Praise, Blame, Obligation, and Beyond: Toward a Framework for Classical Supererogation and Kin

Paul McNamara

Department of Philosophy
University of New Hampshire
Durham, NH 03824-3574 USA
paulm@cisunix.unh.edu

Abstract. Continuing prior work ([1, 2]), I integrate a simple system for personal obligation with a rich system for aretaic (agent-evaluative) appraisal. I then explore various relationships between definable aretaic statuses such as praiseworthiness and blameworthiness and deontic statuses such as obligatoriness and impermissibility. I focus on partitions of the normative statuses generated (cf. "normative positions" but without explicit representation of agency). In addition to representing and exploring traditional questions in ethical theory about the connection between blame, praise, permissibility and obligation, this allows me to carefully represent schemes for supererogation and kin. These controversial concepts have provided challenges to both ethical theory and deontic logic, and are among deontic logic's test cases.

Keywords: Supererogation, Offense, Praise, Blame, Obligation, Aretaic, Deontic, Neutral, Indifferent.

Introduction

I have delineated a framework called DWE ("Doing Well Enough") in [3-7] modeling the logical structure of fundamental but neglected features of common sense morality (CSM). I have focused on *deontic notions:* notions used to evaluate either the things done by an agent or the things brought about by what an agent does. The particular focus in DWE is on a set of notions in the logical neighborhood of that of *exceeding the moral minimum*. However, DWE contains no resources for representing *aretaic notions* (after the Greek term, *arête*): notions used primarily to evaluative agents (e.g. praiseworthiness and blameworthiness), especially for the way in which their agency reflects their worth as persons. And deontic logic generally has long neglected such *aretaic concepts*. Thus there is a gap between this work and that of the traditional work on supererogation and kin from he mid-twentieth century forward. With no representation of praiseworthiness or blameworthiness, there is no way to represent the standard analysis of supererogation, much less that of "offense" (or "suberogation"—the purported mirror image of supererogation).

A first step to rectify my neglect was taken in [1], and I will integrate some of that material with work on personal obligation from [2]. Having an integrated representation of both aretaic and deontic concepts will allow for the representation of a diversity

R. van der Meyden and L. van der Torre (Eds.): DEON 2008, LNAI 5076, pp. 233–247, 2008.

of positions in the ethical theory literature on connections often asserted to hold between these conceptual domains. Derivatively, the resulting framework allows for the representation of the main prior approaches to supererogation and offense other than my own, for example, the Chisholm-Sosa approach, the earlier Meinong-Schwartz approach, what Mellema calls the standard account, and Mellema's own extension of the standard account. I will take up the latter three here a bit. The logical framework also sheds light on various issues in the traditional literature on supererogation. In the end I will show that the classical analysis of supererogation is fundamentally flawed in a way that is both illuminating and ironic.

In section I, I introduce a simple modal logic for what is predetermined for an agent, and then generating a simple classical deontic fragment via a propositional constant for *morality's demands have been met*. In II, I introduce a modified version of the framework for aretaic appraisal in [1], which allows for the definition of a variety of aretaic notions. In III, I characterize some partitions these notions generate given our logics, leading to a 7-fold partition of normative positions generated by some of the aretaic concepts. In IV, I combine this aretaic partition with the deontic fragment, which generates a potentially 21-fold partition of normative statuses combining aretaic and deontic concepts. I explore two prima facia plausible bridging principles, that what is praiseworthy for an agent is permissible and that what is blameworthy for her is not obligatory, and identify the eliminations of normative positions these reductive principles would entail, but I also raise substantive doubts about these principles. In V, I turn to what I call the classical analysis of supererogation and offense, and to Mellema's addition of the notions of quasi-supererogation and quasi-offense, and I identify the six places where these fall on the prior aretaic-deontic partitions. I also briefly consider the Meinong-Schwartz aretaic-deontic ranking thesis, as well as Meinong's "laws of omission". I then examine some additional simple aretaic-deontic bridging principles, that whatever is praiseworthy is obligatory, and that whatever is blameworthy is impermissible, and show how these lead to further reductions in our 21-fold partition, especially with an eye to how they effect the classical analyses of supererogation and kin. I go on to suggest that one simple particular thesis might be behind the widespread skepticism about offenses even among friends of supererogation, but I also argue that the thesis is mistaken, however plausible it is at first blush. Indeed, I think the reflections challenge one long standing line of argument for rejecting offenses, to the benefit of ethical theory. In VI, I show that the traditional analysis of supererogation is fundamentally flawed. I then argue that the classical conception of supererogation presupposes the concept from DWE of doing more than one has to do.

In all this, my intention is two-fold: to model neglected normative statuses, and correlatively to counter the deeply entrenched objection (bias?) from ethical theorists that deontic logic is utterly irrelevant to their enterprise and is a dismal failure in that regard. Indeed, I would submit that some of the work herein is *ahead* of the ethical theory curve in that regard, and that frameworks like the simple one below place questions in sharp relief that ought to clearly benefit ethical theorists.

1 A Modal Framework for Predetermination and Obligation

The main operator in our framework for predetermination is just an interpreted classical necessity operator: PRp: It is (as of now) predetermined (for Jane Doe) that p. We use standard Kripke structures for modeling “PR”: $CO \subseteq W \times W$. COij iff what happens at j is *co*nsistent with our agent's current abilities and disabilities at i. The truth condition for PR is the usual one: $M \models_i$ PRp iff $\forall j(COij \rightarrow M \models_j p)$. We introduce the dual, “it is consistent with our agent's abilities that p”: COp $=_{df} \neg$PR$\neg$p, and its derived truth condition: $M \models_i$ COp iff $\exists j(COij \ \&\ \models_j p)$. We add a single constraint: *CO*-RFLX: *CO*ii. The worlds consistent with our agent's abilities at a given world, i, might then be thought of as the *i-accessible worlds*: $CO^i = \{j: <i,j> \in CO\}$. It will also prove convenient to introduce a notation for *the set of all propositions consistent with our agent’s abilities*: $CO_i = \{X: X \cap CO^i \neq \emptyset\}$. CO^i contains every world consistent with our agent’s abilities at i, whereas CO_i contains the set of propositions true at some such world. Note that the existence of a p-world *consistent with* my abilities does not entail that p *is within* my abilities. Just consider any tautology, or any independent action someone else may or may or not perform. The well known normal modal logic, KT, for PR (PR-KT) is determined by the class of *CO*-reflexive models.

We now add an Andersonian-Kangerian constant, *d*, for "The *d*emands on Jane Doe are all met" (or "Jane Doe’s responsibilities are all met"). We represent the extension of "*d*" as a set of worlds, DEM, DEM $\subseteq$ W, and we give "*d*"'s truth-conditions accordingly: M $\models_i d$ iff i $\in$ DEM. We define our *non-agential but personal obligation* operator:

$$\text{OBp} =_{df} \text{PR}(d \rightarrow \text{p}) ,$$

and read it as follows: OBp: it is obligatory *for Jane Doe* to be such that p.[1] We add an axiom, d, governing our deontic constant: $\vdash$ CO*d* (i.e. $\neg$PR$\neg$*d*). "CO*d*" says *d*'s truth is *consistent with* Jane Doe's abilities, but it does not say it is *within* her abilities, for good reason.[2] Axiom d is validated by the condition that satisfying Doe’s responsibilities is consistent with her abilities: $\forall i \exists j(COij \ \&\ j \in \text{DEM})$. Call the resulting system "PR-KT*d*". It is characterized by the class of all models satisfying this constraint ([7]).

Standard Deontic Logic (SDL) is part of the pure deontic fragment of PR-KTd:[3]

SL:	All Tautologies
OB-K:	OB(p $\rightarrow$ q) $\rightarrow$ (OBp $\rightarrow$ OBq)
OB-NC:	OBp $\rightarrow \neg$OB$\neg$p
MP:	If $\vdash$ p and $\vdash$ p $\rightarrow$ q then $\vdash$ q
OB-NEC:	If $\vdash$ p then $\vdash$ OBq.

[1] The intended reading of "OB", is developed and defended in [2]. It doesn't express the *impersonal* notion *it is obligatory that p*. It expresses a *personal* obligation our agent is under. Nonetheless, it does not require that she be *the agent of* p. We take the form of a personal obligation as an obligation to be such that p, and we then take an agential obligation to be a special case of a personal obligation, one to the effect that our agent has to be such that s*he herself brings it about that* p, and thus to be a compound of a personal obligation operator and an agency operator. We pass over agency and agential obligations here and allow the personal obligation operator and our person-relative modal notions to serve.

[2] Jane Doe may have delegated the last step in her project to her assistant, and it may now be pre-determined for Doe that her project will be completed only if the assistant completes it, which she will. The project‘s completion is no longer within Doe’s ability, but it is still consistent with her ability. Now just add that the project’s completion is equivalent to *d*. For more on the distinction, see [1].

[3] In fact the stronger system that results from adding “OB(OBp $\rightarrow$ p)” to SDL corresponds to KTd. See [7].

Plainly we are engaged in considerable idealization, but this simple familiar system allow first steps toward a more comprehensive integration of aretaic notions with deontic ones.

2 Preliminary Framework for the Aretaic Appraisal of an Agent

Aretaic Preference and Aretaic Appraisal

We take some states of affairs to reflect favorably on people, others unfavorably, some more favorably than others, and some neutrally. I sketch a simple framework here that allows for this, simplifying and slightly modifying that in [1], which gives more details.

We first define a world-relative ordering function, which will yield a weak or quasi-ordering relation, $\geqslant_i$,:

$$\geqslant: W \rightarrow \mathcal{P}(W) \text{ x } \mathcal{P}(W)) \quad [\text{i.e. } \geqslant_i \subseteq \mathcal{P}(W) \text{ x } \mathcal{P}(W)].$$

For each world i, and proposition pair, X and Y, $X \geqslant_i Y$ if and only if *X reflects as well on our agent as Y (X is aretaically as good as Y) from the standpoint of i.* We introduce a corresponding operator:

$$M \models_i p \geq q\text{: } ||p||^M \geqslant_i ||q||^M.$$

We evaluate agents for their actions, results of their actions, motives for acting, intentions in acting, traits of character, etc. To allow for this variety, the relata of our ordering relation excludes only propositions inconsistent with our agent's abilities:

$$\geqslant\text{-}CO_i \text{ Confinement: } \forall i(\geqslant_i \subseteq CO_i \text{ x } CO_i)$$

This validates: ≥-CO Confinement: $\vdash p \geq q \rightarrow (COp \;\&\; COq)$, among other things.

We will also assume that all propositions consistent with our agent's abilities are self-comparable, and we will assume transitivity as well:

Reflexive: $\forall i \forall X(X \in CO_i \rightarrow X \geqslant_i X)$
Transitive: $\forall i \forall X \forall Y \forall Z[(X \geqslant_i Y \;\&\; Y \geqslant_i Z) \rightarrow X \geqslant_i Z]$[4]

We do *not* endorse $\geqslant$-Connectivity, $\forall i \forall X \forall Y[X,Y \in CO_i \rightarrow (X \geqslant_i Y \vee Y \geqslant_i X)]$, as a basic constraint. It is not obvious that *any* two propositions consistent with our agent's abilities must be aretaically comparable since they may involve very different grounds for praise or blame. However, we will need it later to explore supererogation. We consider further constraints below, in the context of discussing the concepts of neutral, positive and negative aretaic appraisal of an agent.

The following basic schemata and rules are validated:

CO-Rflx(≥):	$\vdash COp \rightarrow p \geq p$
Trans(≥):	$\vdash p \geq q \;\&\; q \geq r) \rightarrow p \geq r$
≥-RE1:	If $\vdash p \leftrightarrow q$ then $\vdash r \geq p \rightarrow r \geq q$
≥-RE2:	If $\vdash p \leftrightarrow q$ then $\vdash p \geq r \rightarrow q \geq r$

[4] Clearly, confinement and reflexivity imply $\forall i \forall X(X \in CO_i \leftrightarrow \exists Y(X \geq_i Y \vee Y \geqslant_i X)$. We might thus designate the *aretaically evaluable* propositions, as those comparable with some proposition or other, those self-comparable, or those consistent with our agent's abilities. In turn, we might designate the propositions consistent with our agent's abilities as those that are aretaically evaluable.

A *strong preference* relation and an *equi-ranking* relation are definable in familiar ways: $X >_i Y =_{df} X \gtrsim_i Y \ \& \ \neg(Y \gtrsim_i X)$, $X =_i Y =_{df} X \gtrsim_i Y \ \& \ Y \gtrsim_i X$. Similarly for the corresponding operators: $p > q =_{df} p \geq q \ \& \ \neg(q \geq p)$ and $p \approx q =_{df} p \geq q \ \& \ q \geq p$. Derivative truth-conditions for these operators are: $M \vDash_i p > q$: $\|p\|^M >_i \|q\|^M$ and $M \vDash_i p \approx q$: $\|p\|^M =_i \|q\|^M$. From these axioms, rules, and our definitions, various familiar properties follow for $>$ and $\approx$, as well as: $COp \leftrightarrow p \geq p$; $COp \rightarrow p \approx p$; $p \approx q \rightarrow (COp \ \& \ COq)$; and $p > q \rightarrow (COp \ \& \ COq)$.

Neutrality

We take a neutral proposition as one consistent with an agent's abilities but reflecting no positive or negative merit all-in-all on our agent, perhaps because it involves no positive or negative aretaic components at all or because it has an equal balance of positive and negative aretaic value, "neutralizing" the two opposing values. We take all tautological propositions to reflect neutrally on agents, and define aretaic neutrality accordingly:

$$ANp =_{df} p \approx \top.$$

Since we are interested in the positive and negative aretaic appraisal of things consistent with our agent's ability, paving the way for linking such appraisal with the deontic appraisal of propositions consistent with our agent's abilities, we endorse:

$$\text{AN-CO: } \vdash ANp \rightarrow COp,$$

allowing us to derive: $\vdash \neg AN\bot$ and $\vdash ANp \rightarrow (ANq \leftrightarrow p \approx q)$. If $\vdash p \leftrightarrow q$ then $\vdash ANp \leftrightarrow ANq$, follows readily from the earlier RE principles for $\underset{.}{\geq}$ and our definitions.

Note however that neutrality is not indifferent to negation, $\nvdash ANp \leftrightarrow AN\neg p$, for this would entail $AN\bot$, given $AN\top$, and thus $CO\bot$. But even where COp and $CO\neg p$, neutrality is still not indifferent to negation: that *I do <u>not</u> bring it about that I now do some wonderful thing* might be consistent with my ability, as might its negation, and it might very well be neutral (e.g. there is nothing special about this opportunity to do good), but that *I <u>do</u> bring about something wonderful right now* (p) may not be neutral.

Indifference

Some propositions will be *aretaically indifferent* for our imagined agent.—whether true or false they will not reflect positively or negatively on her. We define this notion as:

$$AIp =_{df} ANp \ \& \ AN\neg p.$$

As stated earlier, aretaic indifference should be stronger than mere neutrality, and by definition, $AIp \leftrightarrow (p \approx \top \ \& \ \neg p \approx \top)$, so given RE for $\approx$, we get the mark of a true indifference notion: $AIp \leftrightarrow AI\neg p$. Also derivable are: $AIp \rightarrow ANp$; $AIp \ \& \ AIq \rightarrow p \approx q$; $AIp \rightarrow p \approx \neg p$; $\neg AI\top$; $\neg AI\bot$; $AIp \rightarrow COp$; and if $\vdash p \leftrightarrow q$ then $\vdash AIp \leftrightarrow AIq$.

Praiseworthiness and Blameworthiness

We take the praiseworthy (blameworthy) propositions as those ranked aretaically higher (lower) than neutral propositions, and this idea is captured by these concise definitions:

$$PWp =_{df} p > \top \quad \text{and} \quad BWp =_{df} \top > p$$

The derivative truth conditions are: $M \models_i \mathrm{PWp}$ iff $||\mathrm{p}||^M >_i \mathrm{W}$; $M \models_i \mathrm{BWp}$ iff $\mathrm{W} >_i ||\mathrm{p}||^M$. The following principles are validated and derivable:

$$\begin{array}{rl}
\text{PW/BW-CO:} & \vdash \mathrm{PWp} \vee \mathrm{BWp} \rightarrow \mathrm{COp} \\
\text{PW-BW EXCL:} & \vdash \mathrm{PWp} \rightarrow \neg\mathrm{BWp} \\
\text{PW-AN/AI EXCL:} & \vdash \mathrm{PWp} \rightarrow \neg(\mathrm{ANp} \vee \mathrm{AIp}) \\
\text{BW-AN/AI EXCL:} & \vdash \mathrm{BWp} \rightarrow \neg(\mathrm{ANp} \vee \mathrm{AIp}) \\
\text{PW>AN>BW:} & \vdash (\mathrm{PWp}\ \&\ \mathrm{ANq}\ \&\ \mathrm{BWr}) \rightarrow (\mathrm{p} > \mathrm{q}\ \&\ \mathrm{q} > \mathrm{r})
\end{array}$$

The following indifference exclusion principle is also derivable:

$$\text{AI-EXCL:} \vdash \mathrm{AIp} \rightarrow (\mathrm{COp}\ \&\neg\mathrm{BWp}\ \&\ \neg\mathrm{BW}\neg\mathrm{p}\ \&\ \neg\mathrm{PWp}\ \&\ \neg\mathrm{PW}\neg\mathrm{p})$$

The only thing that blocks the converse of AI-EXCL,

$$\text{AI-EXCL':} (\mathrm{COp}\ \&\ \neg\mathrm{BWp}\ \&\ \neg\mathrm{BW}\neg\mathrm{p}\ \&\ \neg\mathrm{PWp}\ \&\ \neg\mathrm{PW}\neg\mathrm{p}) \rightarrow \mathrm{AIp},$$

is incomparability. There may be propositions consistent with our agent's ability (and thus each comparable to itself), that are nonetheless not comparable to $\top$, and so not "placed" above, below, or among the neutrals. Thus p might satisfy the left side of AI-EXCL' merely because it is incomparable with $\top$. Such propositions would presumably contain conflicting positive and negative aretaic components separately pulling above and below the neutral line in a way that doesn't allow for resolution. Likewise, we don't have: $\mathrm{COp} \rightarrow (\mathrm{ANp} \vee \mathrm{PWp} \vee \mathrm{BWp})$. If, however, we add $\gtrsim$-Connectivity: $\forall i \forall X \forall Y[(X,Y \in \mathit{CO}_i) \rightarrow (X \gtrsim_i Y \vee Y \gtrsim_i X)]$, we validate this and comparability:

$$\text{CO-COMP:} \quad \models (\mathrm{COp}\ \&\ \mathrm{COq}) \rightarrow (\mathrm{p} \geq \mathrm{q} \vee \mathrm{q} \geq \mathrm{p}).$$

Given connectivity and CO-COMP, the following are validated and derivable:

$$\begin{array}{ll}
\text{CO-COMP':} & \vdash (\mathrm{COp}\ \&\ \mathrm{COq}) \rightarrow (\mathrm{p} > \mathrm{q} \vee \mathrm{q} > \mathrm{p} \vee \mathrm{p} \approx \mathrm{q}) \\
\text{CO-DEF'':} & \vdash \mathrm{COp} \leftrightarrow (\mathrm{p} \geq \top \vee \top \geq \mathrm{p}) \\
 & \vdash \mathrm{COp} \leftrightarrow (\mathrm{p} > \top \vee \top > \mathrm{p} \vee \mathrm{p} \approx \top) \\
 & \vdash \mathrm{COp} \leftrightarrow (\mathrm{ANp} \vee \mathrm{PWp} \vee \mathrm{BWp}) \\
\text{AN-DEF':} & \vdash \mathrm{ANp} \leftrightarrow (\mathrm{COp}\ \&\ \neg\mathrm{BWp}\ \&\ \neg\mathrm{PWp}) \\
\text{AI-EXCL':} & \vdash (\mathrm{COp}\ \&\ \neg\mathrm{BWp}\ \&\ \neg\mathrm{BW}\neg\mathrm{p}\ \&\ \neg\mathrm{PWp}\ \&\ \neg\mathrm{PW}\neg\mathrm{p}) \rightarrow \mathrm{AIp} \\
\text{AI-DEF':} & \vdash \mathrm{AIp} \leftrightarrow (\mathrm{COp}\ \&\ \neg\mathrm{BWp}\ \&\ \neg\mathrm{BW}\neg\mathrm{p}\ \&\ \neg\mathrm{PWp}\ \&\ \neg\mathrm{PW}\neg\mathrm{p})
\end{array}$$

AI-DEF' is essential to the classical framework for supererogation that it is one of our central aims to explore here, so we will assume it henceforth.

Do praiseworthiness and blameworthiness satisfy no conflicts principles:

$$\begin{array}{ll}
\text{PW-NC:} & \mathrm{PWp} \rightarrow \neg\mathrm{PW}\neg\mathrm{p}; \\
\text{BW-NC:} & \mathrm{BWp} \rightarrow \neg\mathrm{BW}\neg\mathrm{p}?
\end{array}$$

These seem plausible for "all-in-all" readings. For suppose you would be praiseworthy (all in all) for being kind or for saving the drowning child. It does not seem right to say that it is also possible that you would be praiseworthy (all in all) for the negation of these very things. Also, PW-NC and BW-NC are clearly presupposed in the classical conceptions of supererogation and offense. So we add these constraints:

$$\begin{array}{l}
\mathit{PW}\text{-NC'}: \forall i \forall X(X >_i \top \rightarrow \neg(W\text{-}X) >_i \top)); \\
\mathit{BW}\text{-NC'}: \forall i \forall X(\top >_i X \rightarrow \neg(\top >_i W\text{-}X)).
\end{array}$$

PW-NC' tells us that for any world i, and proposition X, if X is ranked higher than $\top$, then the negation of X is not ranked higher than $\top$. *BW-NC'* gives the mirror image. These two validate the following *upper and lower exclusion* principles:

$$\top\!> \quad \text{EXCL}: \vdash p > \top \rightarrow \neg(\neg p > \top); \qquad \top\!< \text{EXCL}: \vdash \top > p \rightarrow \neg(\top > \neg p).$$

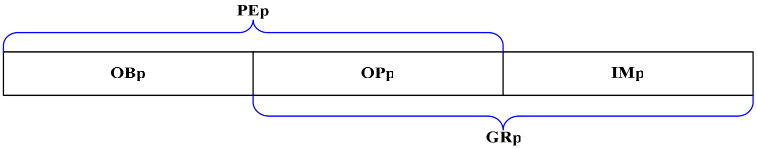
These are then derivable:

$$\text{PW-NC}: \vDash \text{PW}p \rightarrow \neg\text{PW}\neg p\ ; \qquad \text{BW-NC}: \vDash \text{BW}p \rightarrow \neg\text{BW}\neg p.$$

We will assume these stronger principles are operating, as they are essential for understanding the standard account of supererogation and they will facilitate our simple applications to show the fruitfulness of potential a mixed deontic-aretaic scheme.

3 Aretaic Partitions

It is well known that in the traditional deontic systems, all propositions are partitioned into three mutually exclusive and exhaustive classes, those *obligatory*, those *impermissible* and those *optional* (often mislabeled "indifferent"):

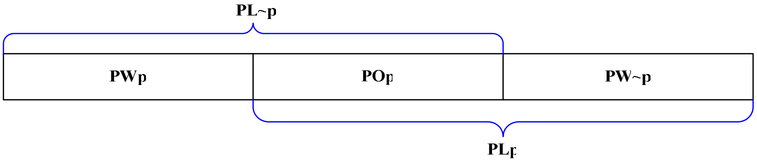


Similar relationships hold for PW and BW, but first let us introduce these definitions:

$\text{PL}p =_{\text{df}}\ \neg\text{PW}p$:	It is Praise-Less that p.
$\text{PO}p =_{\text{df}} \neg\text{PW}p\ \&\ \neg\text{PW}\neg p$:	It is Praise Optional that p
$\text{BL}p =_{\text{df}}\ \neg\text{BW}p$:	It is Blame-Less that p
$\text{BO}p =_{\text{df}} \neg\text{BW}p\ \&\ \neg\text{BW}\neg p$:	It is Blame-Optional that p

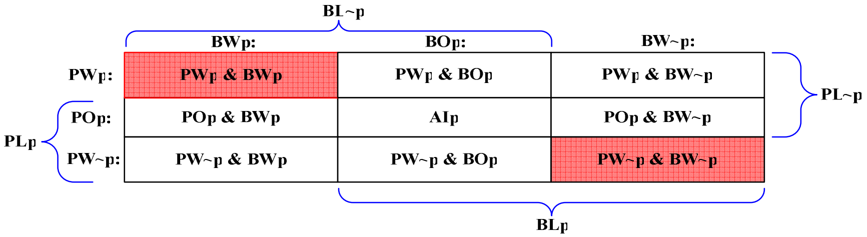
Given COMP, the following hold, $\text{AI}p \leftrightarrow (\text{CO}p\ \&\ \text{BL}p\ \&\ \text{BL}\neg p\ \&\ \text{PL}p\ \&\ \text{PL}\neg p)$; $\text{AI}p \leftrightarrow (\text{CO}p\ \&\ \text{BO}p\ \&\ \text{PO}p)$; and $\text{CO}p \rightarrow [\text{AI}p \leftrightarrow (\text{BO}p\ \&\ \text{PO}p)]$.

For both PW and BW we get an exact analogs call it the PW-P (for "PW-Partition"):



PW-P is this conjunction: a) $\text{PW}p \vee \text{PW}\neg p \vee (\neg\text{PW}p\ \&\ \neg\text{PW}\neg p)$ & b) $\neg(\text{PW}p\ \&\ \text{PW}\neg p)$ & c) $\neg(\text{PW}p\ \&\ (\neg\text{PW}p\ \&\ \neg\text{PW}\neg p))$ & d) $\neg(\text{PW}\neg p\ \&\ (\neg\text{PW}p\ \&\ \neg\text{PW}\neg p))$. Clearly, a), c) and d) are just truth-functional tautologies. Only b) is not, and it is just No PW Conflicts again: PW-NC: $\vdash \text{PW}p \rightarrow \neg\text{PW}\neg p$. So PW-P is equivalent to PW-NC: $\vdash$ PW-P $\leftrightarrow$ PW-NC. The BW Partition (BW-P) is perfectly analogous and similarly reduces to No BW Conflicts principle: $\vdash$ BW-P $\leftrightarrow$ BW-NC.

What happens when we consider compounding these two partitions and classify options in terms of *both* the positive and the negative operators above? This:



	BWp:	BOp:	BW~p:
PWp:	PWp & BWp	PWp & BOp	PWp & BW~p
POp:	POp & BWp	AIp	POp & BW~p
PW~p:	PW~p & BWp	PW~p & BOp	PW~p & BW~p

Nine possible combinations are indicated. The two eliminations in shaded boxes follow from our earlier theorem, PWp → ¬BWp, which derives from our definitions and the plausible thesis, p > q → ¬(q > p). Furthermore, the standard conception of supererogation and offense presuppose the exclusiveness of all-in-all praiseworthiness and blameworthiness, for else an act that was supererogatory and thus praiseworthy to do might nonetheless be blameworthy to do, which jars. Call the resulting 7-fold Partition "PW-BW P".[5] We turn now to integrating this aretaic framework with a standard deontic one.

4 Aretaic-Deontic Partitions and Some Underlying Issues

As we noted earlier, SDL entails an OB-Partition. What happens when we combine that set of deontic categories with the preceding seven aretaic ones? Ignoring the shading, text in parenthesis and in brackets for now, we get this 21-fold partition:

	PWp & BOp	PWp & BW~p	POp & BWp	AIp	POp & BW~p	PW~p & BWp	PW~p & BOp
OBp	PWp & BOp & OBp	PWp & BW~p & OBp	POp & BWp & OBp [elim by b)]	AIp & OBp	POp & BW~p & OBp	PW~p & BWp & OBp [elim by b)]	PW~p & BOp & OBp [elim by a)]
OPp	PWp & BOp & OPp (SU′p) [elim by c)]	PWp & BW~p & OPp (QSp) [elim by c)/d)]	POp & BWp & OPp (OF′p) [elim by d)]	AIp & OPp (FIp)	POp & BW~p & OPp (OFp) [elim by d)]	PW~p & BWp & OPp (QOp) [elim by c)/d)]	PW~p & BOp & OPp (SU′p) [elim by c)/d)]
IMp	PWp & BOp & IMp [elim by a)]	PWp & BW~p & IMp [elim by a)]	POp & BWp & IMp	AIp & IMp	POp & BW~p & IMp [elim by b)]	PW~p & BWp & IMp	PW~p & BOp & IMp

[5] The partition inherits the mutual exclusiveness and exhaustiveness of the two three-fold schemes that generated it. The exhaustiveness of the BW-P partition entails that if PWp, then p must satisfy that label as well as one of the three column labels, and thus find a place in at least one box in the top row. Similarly for if POp, or if PW¬p. But now by the exhaustiveness of the PW-P partition, every p must satisfy at least one of these three antecedent conditions—it must satisfy one of the three row labels. So by a 3-part version of Constructive Dilemma, every p must fit into one of the nine boxes, and since no p fits into the two red boxes in our framework, it follows that every p fits into at least one of the 7 white boxes. Similar reasoning about inheritance will show that no p can satisfy more than one of the labels in the boxes, for that would be inconsistent with the non-exclusiveness of the parent partitions, PW-P and BW-P.

We have already eliminated (BWp & PWp) and (BW¬p & PW¬p) in our framework, so there are no labels for those combinations. On top of the seven columns we have the seven cell labels from the aretaic partition, and left of the rows, we have the three prior deontic cell labels. As with the 7-fold partition, this 21-fold partition inherits the exhaustiveness and exclusiveness of the parent partitions, OB-P and PW-BW P.

Note that we can define and identify a variety of moral concepts of interest in this framework. *Culpable obligations*--obligations one would be blameworthy to violate ($=_{df}$ OBp & BW¬p) appear in the top row second column and fifth column, and given IMp $=_{df}$ OB¬p, in the same columns of the third row as well. Similarly for non-culpable obligations ($=_{df}$ OBp & BL¬p). We can then raise interesting questions such as "Can there be obligations of either of these kinds?", or even "Can there be obligations that are *blameworthy* for Jane Doe to *fulfill* or *praiseworthy* to *violate*?, and we can identify how positive answers would fit in the above scheme and explore the eliminative implications of negative answers. Which sorts of normative acts can exist is of fundamental importance in ethical theorizing, so this is a place where deontic logic has a greater chance of being of genuine aid. We will illustrate this in the remainder of this section and the next by considering some simple theses connecting aretaic and deontic concepts and their impact on the above partition.

Many would endorse two basic bridging theses at first glance:

a) No IM-PW Conflicts: $\vdash \neg(\text{IMp} \mathbin{\&} \text{PWp})$ [i.e. $\text{PWp} \rightarrow \text{PEp}$]
b) No OB-BW Conflicts: $\vdash \neg(\text{OBp} \mathbin{\&} \text{BWp})$ [i.e. $\text{BWp} \rightarrow \text{PE}\neg\text{p}$]

These are reductive theses, since it is east to see that they eliminate the possibility of certain normative positions. The six respective eliminations (three each) these entail are indicated in the shaded boxes in the top and bottom rows. (There is no impact on the middle row.) The result would be a reduction of deontic-aretaic statuses to a 15-fold partition. However, having illustrated the reductions a) and b) imply, and despite the fact that these are often taken for granted by friends of supererogation and ethical theorists generally, there are reasons to be less sure on reflection. *Doubts about a)*: It is widely thought that one can have an obligation that p and not realize it. For example, I can think I paid you back $20 already, but be mistaken. Furthermore, I can be non-culpably ignorant of certain facts and as a result not realize that p is impermissible. Now add that were it not for these facts of which I am ignorant, it would be very good that p, and I bring about p motivated by just such a belief. It seems that in this case, I am praiseworthy all things considered for the fact that p even though p is impermissible. For example, suppose I give to a charity shortly after, unbeknownst to me, my savings have been lost in a stock market crash. As a result, I'm obligated to not give to charity, since my family will need every penny I have, but given my blameless ignorance at the moment, and my very good intentions at the time, I am all things considered praiseworthy for giving to the charity, even though doing so was impermissible, unbeknownst to me. *Doubts about b)*: Similarly, suppose that I am subject to the same non-culpable ignorance of my sudden loss of savings and, so unbeknownst to me, it is obligatory for me to hang onto every cent I have for my family.

Now add that I could help out a friend who has helped me before by giving her $20, and as far as I know, I could do this permissibly, and at trivial cost. Yet I refuse to do so for the most selfish and callous of reasons. It then seems that all things considered I am blameworthy for refusing even though, unbeknownst to me, my familial obligations make it overridingly obligatory to refuse.[6]

Often friends of supererogation tacitly endorse a) and b) in the way they define such acts, but we will be more cautious. I turn to supererogation and kin now.

5 These Partitions and the Classical Analysis of Supererogation

The *classical analyses* of supererogation and of offense/suberogation are:

$SU^{a}p$: $PWp \mathbin{\&} \neg BW\neg p \mathbin{\&} OPp$
$OF^{\mathbf{a}}p$: $BWp \mathbin{\&} \neg PW\neg p \mathbin{\&} OPp$

Something is supererogatory (for Jane) iff it is praiseworthy, its negation is not blameworthy and it is (deontically) optional. In contrast something is an offense (suberogatory) if it is blameworthy, its negation is not praiseworthy, and it is optional.

[8, 9] proposes extending the classical scheme by adding acts of "quasi-supererogation" and "quasi-offense", and argues for their possible instantiation:

QSp: $PWp \mathbin{\&} BW\neg p \mathbin{\&} OPp$
QOp: $BWp \mathbin{\&} PW\neg p \mathbin{\&} OPp$

Something is quasi-supererogatory (for Jane) iff it is praiseworthy, its negation *is* blameworthy and it is optional; something is a quasi-offense (quasi-suberogatory) if it is blameworthy, its negation *is* praiseworthy, and it is optional.

Let us introduce only one more mixed concept--FI, for "Fully Indifferent":

$FIp =_{df} OPp \mathbin{\&} AIp.$

These five new operators concepts are easily accommodated and are present already in our prior 21-fold partition. Since they entail deontic optionality they occur only in the middle row of that partition and are indicated in parenthesis near the bottom of each cell in that row. The result suggests that lingering behind the classical conception of supererogation is a framework with 21 potential categories, far more than previously articulated.[7] Some of their logical features are also revealed at a glance, for example that the five new operators are mutually exclusive and that if

[6] Although beyond the scope of the current paper, the two principles above perhaps look plausible at first glance because we tend to conflate them with the genuinely plausible principles we get if we replace the partially agent-evaluative notions of blameworthiness and praiseworthiness with the more purely state of affairs evaluative notions of *goodness* and *badness* (when we call acts or results of acts “good” or “permissible” we evaluate them independent of our evaluation of the agent’s motives, etc., so it is plausible to expect stronger links here. I explore this elsewhere.

[7] [9] identifies nine, and unlike here, that is nine dependent partially on introducing action concepts in the scope of operators, but he also indicates he makes no claim the scheme is complete.

something is optional and not fully indifferent then it will satisfy one of the first four operators, and as alluded to earlier when we endorsed PW-BW Exclusion, no supererogatory or quasi-supererogatory option is blameworthy, and similarly, no offense or quasi-offense is praiseworthy.[8]

In the late twentieth century, supererogation was a hotly contested concept, with many arguing against its existence. For example, [10] argues for the rejection of supererogation by roughly endorsing the following aretaic-deontic bridging principle:

$$\text{c) PWp} \rightarrow \text{OBp.}$$

Clearly if we add this scheme, we get $\vdash \neg(\text{SU}^{\text{a}}\text{p} \vee \text{QSp} \vee \text{QOp})$. Only offenses remain. Note c) also entails a) $\text{PWp} \rightarrow \text{PEp}$, given $\text{OBp} \rightarrow \text{PEp}$, so those earlier possible eliminations would follow from c) as well. Furthermore, it is highly unlikely that anyone accepting PW-OB would not also endorse the following bridging principle:

$$\text{d) BWp} \rightarrow \text{IMp.}$$

Indeed I believe that d) is more widely endorsed than c). If we add this scheme, it yields $\vdash \neg\text{OF}^{\text{a}}\text{p}$, $\vdash \neg\text{QSp}$, and $\vdash \neg\text{QOp}$, and since d) entails our b) $\text{BWp} \rightarrow \text{PE}{\sim}\text{p}$, those prior eliminations follow as well. So under either c) or d), the quasi-notions are eliminated, and if both c) and d) are endorsed, all of the middle row save the central category of aretaic indifference is eliminated (as well as those shaded in the top and bottom row, leaving us with 9 cells). This reduction of the middle row to one cell is one version of what I have called Moral Rigor (MR): $\vdash \text{AIp} \leftrightarrow \text{OPp}$. Correlatively, the eliminations in the partition also reflect a version of Strong Exhaustion (SEX), that every option is either obligatory, impermissible or (fully) indifferent: $\vdash \text{OBp} \vee \text{IMp} \vee \text{FIp}$. It is a far cry in its substantive import from the Traditional Exhaustion (TEX) formula: $\vdash \text{OBp} \vee \text{IMp} \vee \text{OPp}$. MR and SEX rules out not only supererogation but all the non-indifferent optional concepts associated with it. If one thinks of the options that are neither obligatory nor impermissible as *indifferent*, one is a short step from tacitly ruling out all non-indifferent options.

Turning back now to c) and d), I think c) has little plausibility on reflection despite its explicit or tacit endorsement in much early literature hostile to supererogation. It often rests on conflating "ought" with "must". I think d) is more attractive and still recently endorsed by some (e.g. [11]), but we have already tacitly rejected it in rejecting the prior bridging principle, b). For there I argued that I might be blameworthy for refusing to help a friend, even though my circumstances have changed unbeknownst to me in such a way that it is indeed not only permissible but obligatory for me to refuse to help the friend out (else my family may starve). Here refusing appears to be obligatory and so permissible here, yet blameworthy.

Still I think d) may be behind the widespread rejection of offenses. Although supererogation has been a marginalized concept in ethical theory and deontic logic, it is nonetheless much less controversial than that of offenses (suberogation). Even the

[8] I know of no friend of supererogation that ever felt the need to add to the definition of supererogation that it was also not blameworthy to do, and similarly for the definition of offense with respect to praiseworthiness. There is a strong presupposition in favor of PW-BW Exclusion in the ethical theory tradition focused on these notions.

staunches defenders of supererogation (see [12] and [9]) raise serious doubts about offenses and argue against the alleged symmetry between offense and supererogation. One standard line of rejection is that if you allow for a full mirror image of supererogation, then not only can an option be blameworthy yet permissible, but it can be *blameworthy to the highest degree* and be permissible, since a supererogatory option can be praiseworthy to the highest degree and be permissible to skip. But then a downright reprehensible action could turn out to be permissible, which is deemed surely false. The objection is that once you open the door to allowing permissible blameworthy options like offenses, it seems hard to find any principled way to put any limit on the degree of blameworthiness that might be permissible in some circumstances. [9] offers two additions (QS and QO) to mitigate against the complaints of some anti-supererogationists who also invariably reject offenses, and he himself thinks there are better reasons to doubt offenses than to doubt supererogation, but he there overlooks the fact that the appeal of a very simple aretaic-deontic bridge principle like d) may be motivating the especially widespread rejection of offenses, one that would at once lead to a rejection of both of his quasi notions. In later work, he is clearer about this (e.g. [13]). However, our earlier reflections on b) suggest that a person can indeed be extremely blameworthy for an action that might not only be permissible but downright obligatory. Thus our reflections appear to suggest that this pathway for arguing against offenses (or Mellema's two quasi-notions) may very well rest on a simple though widely endorsed false presupposition.

Schwartz and Meinong both endorsed theses that are natural errors when first reflecting on supererogation. The main one I will call "the Ranking Thesis" (RT)[9]:

$$(SU^{a}p \;\&\; OBq \;\&\; OF^{\mathbf{a}}r \;\&\; IMs) \rightarrow p > q > r > s.$$

[14] makes clear the pro4blems with this view, which we here summarize by noting that a small favor like lending a book you've already read to a friend might be supererogatory and nice but not warranting high praise, whereas sometimes our obligations are truly arduous to fulfill and very tempting to shirk. Thus it will often be more praiseworthy to do something obligatory than to do something supererogatory. Similarly, if there are offenses, then there is no reason to think that these cannot sometimes be of such a caliber as to reflect much more poorly on an agent than shirking some very small obligation. [14] also rejects Meinong's "laws of omission", which we can partially illustrate here by:

$$SU^{a}p \leftrightarrow OF^{\mathbf{a}}\neg p.$$

If it is supererogatory for me to jump on a grenade to save my equally situated comrades, it is not ther4eby blameworthy for me to not do so. Similarly, if it is an offense for me to not say "hello" when passing you quickly in the hall, it is not supererogatory for me to do so. Behind these laws of Meinong are probably more fundamental aretaic mistakes (e.g. $PWp \leftrightarrow BW\neg p$) but we will study Schwartz', Meinong's, and Chisholm's schemes and the continuity requirement elsewhere.

In the next section we show that the classical analysis of the most plausible and widely accepted of the first four notions defined above, is fundamentally flawed.

[9] Really a version of what is called the "continuity requirement" [12, 9].

6 A Basic Flaw in the Classical Analysis of Supererogation

As mentioned, my prior work in DWE does not involve any aretaic operators, but I want to now suggest that this gap cuts both ways: that the classical conception of supererogation presupposes one of the deontic non–aretaic concepts of DWE.

First, note that an action can be obligatory and highly praiseworthy. Consider a solider on point. She stands her appointed ground faithfully in face of a sudden enemy attempt to overrun the main camp, despite extreme danger to those who stay on point. Second, notice that if something is obligatory, then doing the least you can do involves doing that thing, and sometimes the minimum you can permissibly do is the same as what is obligatory--there are no graded options to speak of. It follows from the preceding case and this DWE principle that it can be praiseworthy to do the minimum. Now the crucial question: *Can it be praiseworthy to do the minimum even when one can also go beyond the call by doing more than the bare minimum?* "Yes". Consider a minor variant of our prior case. Suppose there are now a first and second position on point, the second being a slightly safer fallback position but also slightly riskier for the camp, so the first is better all in all. Now suppose it is permissible to pick either spot to make a stand (for whoever is on point, by agreement of the group, etc.). Our soldier in good faith picks and holds the second position, again at great risk of death. Here the least she can do is hold the second position. Obviously she "could" also retreat, hide, or play dead, but not permissibly so. Still the temptation to take the latter sort of impermissible option might be very intense and such that many would do that. It can then surely be praiseworthy for her to hold even the second position in such circumstances. But now notice two other things abut this case: 1) It is deontically *optional for her to hold the second position*: for she can also hold the first position instead, thereby going beyond the call. 2) It is also *not blameworthy for the agent to not hold the second position*, for then if she went beyond the call by holding the first position for the best of reasons, she would thereby be blameworthy.

So here we have an action that satisfies all three conditions of the classical analysis of supererogation, yet it is not beyond the call; indeed the action is *the minimum required*. When we can do more good than we have to, doing the minimum will always be optional, and it can't be automatically blameworthy to not do the minimum, for then going beyond the call would invariably entail blameworthiness. And sometimes our permissible choices are arduous enough that even taking the minimally permissible one is praiseworthy. The soldier's holding the 2nd point is not intuitively supererogatory, and fails to fall under the classical *conception* of supererogation, so the classical analysis despite its pervasiveness is flawed: it does not give sufficient conditions for its target class of acts. In particular, the *deontic* condition is too weak. Focusing only on actions that are aretaically praiseworthy to do and not blameworthy to omit, and then merely adding deontic optionality is insufficient to guarantee they are of the intended kind. Put another way, such an act can be optional for the wrong reasons: because it is a *surpassable minimum*.

There is some irony in this. Supererogationists from Urmson forward accused deontic logic of ruling out supererogatory actions. I have argued that this is due to a conflation of deontic *optionality* with *indifference* by both deontic logicians and ethicists, but our reflections today suggests that the classical analysis that was intended to break free from the constraints of deontic logic and its deontic notions can't capture its intended class of actions without returning to deontic logic and *increasing* the deontic notions it relies on. Of course it needs more that the deontic concept of optionality that SDL can provide. It also needs, at the very least, the deontic concept of *doing the minimum that morality demands*. This concept is not expressible in SDL, but it is in DWE, which expands considerably upon the expressive resources of SDL, allowing for the expression of a person's exceeding the minimum that morality demands, and not entailing that this need be praiseworthy, for in fact it need not be. One can do more than the minimum for the wrong reasons or even for bad reasons and not be praiseworthy at all. I would suggest that the more objective and act-evaluative notion of doing more than the minimum is both more fundamental to our moral scheme and more important to it. This suggests that in addition to developing the framework within, and considering weakening some of the principles for aretaic appraisal that we indicated were less than self-evident, linking it with DWE is in order. The result will be a considerably enriched system, and one which will distinguish between two closely related and often conflated concepts: *supererogation* and *doing more good than one has to do*.

References

1. McNamara, P.: Toward A Framework For Agency, Inevitability, Praise And Blame. Nordic Journal of Philosophical Logic 5(2), 135–160 (2000)
2. McNamara, P.: Agential Obligation as Non-Agential Personal Obligation plus Agency. Journal of Applied Logic 2(1), 117–152 (2004)
3. McNamara, P.: Doing Well Enough: Toward a Logic for Commonsense Morality. Studia Logica 57(1), 167–192 (1996)
4. McNamara, P.: Must I Do What I Ought (Or Will the Least I Can Do Do?). In: Brown, M.A.J.C. (ed.) Deontic Logic, Agency and Normative Systems, pp. 154–173. Springer, New York (1996)
5. McNamara, P.: Making Room for Going Beyond the Call. Mind 105(419), 415–450 (1996)
6. Mares, E.D., McNamara, P.: Supererogation in Deontic Logic: Metatheory for DWE and Some Close Neighbours. Studia Logica 59(3), 397–415 (1997)
7. McNamara, P.: Doing Well Enough in an Andersonian-Kangerian Framework. In: McNamara, P., Prakken, H. (eds.) Norms, Logics and Information Systems: New Studies in Deontic Logic and Computer Science, pp. 181–198. IOS Press, Washington (1999)
8. Mellema, G.: Quasi-Supererogation. Philosophical Studies 52, 141–150 (1987)
9. Mellema, G.: Beyond the Call of Duty: Supererogation, Obligation and Offence, Albany: SUNY Pr (1991)
10. Pybus, E.M.: Saints and Heroes. Philosophy 57, 193–200 (1982)

11. Widerker, D.: Frankfurt on ’ought implies can’ and alternative possibilities. Analysis 51, 222–224 (1991)
12. Heyd, D.: Supererogation: Its Status in Ethical Theory. Cambridge Studies in Philosophy. Cambridge University Press, Cambridge (1982)
13. Mellema, G.: Supererogation, Blame, and the Limits of Obligation. Philosophia 24(1-2), 171–182 (1994)
14. Chisholm, R.M.: Supererogation and Offence: A Conceptual Scheme for Ethics. Ratio 5, 1–14 (1963)

Introducing Grades in Deontic Logics

Pilar Dellunde[1,2] and Lluís Godo[2]

[1] Univ. Autònoma de Barcelona
08193 Bellaterra, Spain
pilar.dellunde@uab.cat
[2] IIIA - CSIC
08193 Bellaterra, Spain
godo@iiia.csic.es

Abstract. In this paper we define a framework to introduce gradedness in Deontic logics through the use of fuzzy modalities. By way of example, we instantiate the framework to Standard Deontic logic (SDL) formulas. Given a deontic formula $\Phi \in SDL$, our language contains formulas of the form $\overline{r} \to N\Phi$ or $\overline{r} \to P\Phi$, where $r \in [0, 1]$, expressing that the preference or probability degree respectively of a norm Φ is at least r. We present sound and complete axiomatisations for these logics.

Keywords: Deontic Logic, Fuzzy Logic, Norms, Institutions.

1 Introduction

In their article [4], Tom R. Burns and Marcus Carson describe how agents adhere to and implement rule and normative systems to varying degrees. Agents conform to rule and normative systems to varying degrees, depending on their identity or status, their knowledge of the rules, the interpretations they attribute to them, the sanctions a group or organization imposes for noncompliance, the structure of situational incentives, and the degree competing or contradictory rules are activated in the situation, among other factors. Actually, the claim that obligations come in degrees goes back to W. D. Ross in his system of ethics, when dealing with the possibility of conflicting moral obligations (for a reference see [23]).

In hierarchical normative systems not every norm may have the same importance. In such a case, it seems interesting that agents can attach a level or degree of importance to each of these norms. These importance or preference degrees may be in turn useful for resolving conflicts among norms that may arise due to different reasons. Within a Multi-Agent System (MAS), normative conflicts may arise due to the dynamic nature of the MAS and simultaneous agents' actions. In a normative structure, one action can be simultaneously forbidden and obliged. Ensuring conflict freedom of normative structures at design time is computationally intractable as shown in [9], and thus real-time conflict resolution is required. In multi-institutional contexts, different institutions could have contradictory norms and therefore agents that participate in these institutions

R. van der Meyden and L. van der Torre (Eds.): DEON 2008, LNAI 5076, pp. 248–262, 2008.

should decide which norm they follow. Attaching a preference degree to norms could help agents in order to take this kind of decisions.

Moreover, in hierarchical normative multi-agent systems, even if a set of norms may have a same rank, there might be different expectations about their compliance or violation by agents. In such situations, it may also be useful to represent and reason about the probability of compliance of norms.

In this paper we would like to define a logical framework able to capture different graded aspects of norms. Taking Standard Deontic Logic (SDL) as the basic formalism to model normative systems as way of exemple, we present in this paper preliminary steps towards defining Graded Deontic logics. We are aware that SDL suffers from a number of paradoxes, mostly inherent in the normal modal Kripke semantics of its operators. Thus, our proposal is not to represent graded normative reasoning in MAS over the logic SDL. But we believe that begining the study in this basic logic, graded SDL, could led us to a better understanding of the main characteristics of graded normative systems in general.

Our fuzzy modal approach has been already used to define a number of uncertainty logics (probability, possibility, belief functions [13,10] or even graded BDI agent architectures [7]). More specifically, we define fuzzy modal languages over SDL to reason about preference (understood as necessity, in the possibilistic sense) and probability of deontic propositions. To this end we introduce two fuzzy modal-like operators N and P that apply over SDL , in such a way that e.g. the truth-degree of a formula $NO\varphi$ o $PO\varphi$ is respectively interpreted as the necessity degree or probability degree of φ being obliged. Then we use suitable fuzzy logics to reason about these intermediate truth-degrees, truth-degrees which are of neither of propositions φ nor $O\varphi$ (which remain two-valued) but of *fuzzy* propositions $NO\varphi$ and $PO\varphi$. Namely, the language of *Necessity-valued Standard Deontic Logic* NSDL will result from the union of the language of the logic $\mathrm{G}_{\Delta}(C)$ (Gödel Logic expanded with the Δ operator and a finite set $C \subset [0, 1]$ of truth-constants) and the language of SDL extended with the fuzzy unary operator N. On the other hand, the language of *Probability-valued Standard Deontic Logic* PSDL will result from the union of the language of Rational Pavelka Logic RPL (Łukasiewicz Logic expanded with rational truth-constants) and the language of SDL extended with a fuzzy unary operator P.

The main features of the Graded Deontic Logics we introduce in this paper are:

1. they are conservative extensions of SDL
2. they have a finite and recursive set of axioms
3. they keep classical semantics for formulas of SDL, in particular their truth-values remain always 0 or 1
4. they have as semantics extensions of the standard Kripke frames for SDL with necessity and probability measures over worlds respectively.
5. they contain formulas of the form $\overline{r} \to N\Phi$ or $\overline{r} \to P\Phi$, where $r \in [0, 1]$, expressing that the necessity or probability degree respectively of a norm Φ is at least r, where Φ is any closed formula of SDL (not only a propositional one).

The main objective of this article is to present the above mentioned four variants of Graded Deontic Logics and prove soundness and completeness results. This constitutes a purely logical study of these formalisms. Note that our purpose is not to *fuzzify* Deontic Logic by providing a different interpretation to its modalities in the sense of having fuzzy deontic modalities, see Section 6 for a discussion. Instead we have fuzzy, many-valued modalities (of necessity and probability) applying over classical deontic formulas.

This paper is structured as follows. In Section 2 we present some rather long preliminaries on the $G_\Delta(C)$ and RPL fuzzy logics that will be needed later. In Section 3, Necessity-valued and Probability-valued Deontic logics are defined over Standard Deontic logic and in Section 4 we present two small examples of application of the two graded logics. Finally Section 5 is devoted to related and future work.

2 Preliminaries on the $G_\Delta(C)$ and RPL Fuzzy Logics

Probably the most studied and developed many-valued systems related to fuzzy logic are those corresponding to logical calculi with the real interval $[0, 1]$ as set of truth-values and defined by a conjunction & and an implication → interpreted respectively by a (left-continuous) t-norm $*$ and its residuum $\Rightarrow$[1], and where negation is defined as $\neg\varphi = \varphi \to \overline{0}$, with $\overline{0}$ being the truth-constant for falsity. In the framework of these logics, called *t-norm based fuzzy logics*, each (left continuous) t-norm $*$ uniquely determines a semantical (propositional) calculus $PC(*)$ over formulas defined in the usual way from a countable set of propositional variables, connectives $\wedge$, & and → and truth-constant $\overline{0}$ [13]. Evaluations of propositional variables are mappings e assigning each propositional variable p a truth-value $e(p) \in [0, 1]$, which extend univocally to compound formulas as follows:

$$\begin{aligned} e(\overline{0}) &= 0 \\ e(\varphi \wedge \psi) &= \min(e(\varphi), e(\psi)) \\ e(\varphi \& \psi) &= e(\varphi) * e(\psi) \\ e(\varphi \to \psi) &= e(\varphi) \Rightarrow e(\psi) \end{aligned}$$

Note that, by definition of residuum, $e(\varphi \to \psi) = 1$ iff $e(\varphi) \leq e(\psi)$, in other words, the implication → captures the ordering. Further connectives are defined as follows:

$$\begin{aligned} \varphi \vee \psi &\text{ is } ((\varphi \to \psi) \to \psi) \wedge ((\psi \to \varphi) \to \varphi), \\ \neg\varphi &\text{ is } \varphi \to \bar{0}, \\ \varphi \equiv \psi &\text{ is } (\varphi \to \psi) \& (\psi \to \varphi). \end{aligned}$$

Note that, from the above defintions, $e(\varphi \vee \psi) = \max(e(\varphi), e(\psi))$, $\neg\varphi = e(\varphi) \Rightarrow 0$ and $e(\varphi \equiv \psi) = e(\varphi \to \psi) * e(\psi \to \varphi)$. A formula φ is a said to be a 1-tautology

[1] Defined as $x \Rightarrow y = \max\{z \in [0, 1] \mid x * z \leq y\}$, which always exists provided $*$ is left-continuous.

of $PC(*)$ if $e(\varphi) = 1$ for each evaluation e, and will be denoted as $\models_* \varphi$. The associated consequence relation is defined as usual: if T is a theory (set of formulas), then $T \models_* \varphi$ whenever $e(\varphi) = 1$ for all evaluations e such that $e(\psi) = 1$ for all $\psi \in T$. Two outstanding examples of *continuous* t-norm based fuzzy logic calculi are:

Gödel logic calculus: defined by the operations

$$x *_G y = \min(x, y)$$
$$x \Rightarrow_G y = \begin{cases} 1, & \text{if } x \leq y \\ y, & \text{otherwise.} \end{cases}$$

Łukasiewicz logic calculus: defined by the operations

$$x *_Ł y = \max(x + y - 1, 0)$$
$$x \Rightarrow_Ł y = \begin{cases} 1, & \text{if } x \leq y \\ 1 - x + y, & \text{otherwise.} \end{cases}$$

Actually, in these two calculi (and in general when $*$ is continuous) the min operation is also definable from $*$ and $\Rightarrow$ as :

$$\min(x, y) = x * (x \Rightarrow y)$$

and hence the connective $\wedge$ can be also considered as definable. These two fuzzy logic calculi turn out to correspond to the well-known infinitely-valued Łukasiewicz and Gödel logics[2], already studied much before fuzzy logic was born (see e.g. [13] for references there). If we denote by $\vdash_Ł$ and $\vdash_G$ the provability relations in Lukasiewicz and Gödel logics respectively, the following *standard* completeness hold:

$\vdash_Ł \varphi$ iff $\models_Ł \varphi$
$\vdash_G \varphi$ iff $\models_G \varphi$

where, for the sake of simpler notation, we have written $\models_Ł$ and $\models_G$ instead of $\models_{*_Ł}$ and $\models_{*_G}$ respectively. Interestingly enough, both Łukasiewicz and Gödel logics have been shown to be axiomatic extensions of Hájek's Basic fuzzy logic BL [13] which axiomatizes the set of all common tautologies to every calculus PC(*) with $*$ being a continuous t-nom. As a matter of fact, Łukasiewicz logic is the extension of BL by the axiom (Ł) $\neg\neg\varphi \rightarrow \varphi$,

forcing the negation to be involutive, and Gödel logic is the extension of BL by the axiom

(G) $\varphi \rightarrow (\varphi \& \varphi)$.

[2] Gödel logic is also known as Dummett logic and is the axiomatic extension of Intuitionistic logic with the pre-linearity axiom $(\varphi \rightarrow \psi) \vee (\psi \rightarrow \varphi)$.

forcing the conjunction to be idempotent. The above mentioned completeness for theorems extend to deductions from arbitrary theories in case of Gödel logic and only to deductions from finite theories in case of Łukasiewicz logic:

$T \vdash_{Ł} \varphi$ iff $T \models_{Ł} \varphi$, if T is finite
$T \vdash_G \varphi$ iff $T \models_G \varphi$

In a sense, due to the residuation property of implications, a t-norm based fuzzy logic L as defined above can be considered as a logic of *comparative truth*. In fact, a formula $\varphi \to \psi$ is a logical consequence of a theory T, i.e. if $T \vdash_L \varphi \to \psi$, if the truth degree of φ is at most as high as the truth degree of ψ in any interpretation which is a model of the theory T. Therefore, implications indeed implicitly capture a notion of comparative truth. This is fine, but in some situations one might be also interested to explicitly represent and reason with *partial degrees* of truth. One convenient way to allow for an explicit treatment of degrees of truth is by introducing truth-constants into the language. In fact, if one introduces in the language new constant symbols $\overline{\alpha}$ for suitable values $\alpha \in [0,1]$ and stipulates that $e(\overline{\alpha}) = \alpha$ for all truth-evalutations e, then a formula of the kind $\overline{\alpha} \to \varphi$ becomes 1-true under any evaluation e whenever $\alpha \leq e(\varphi)$.

This approach actually goes back to Pavelka [21] who built a propositional many-valued logical system PL which turned out to be equivalent to the expansion of Łukasiewicz Logic by adding into the language a truth-constant $\overline{r}$ for each *real* $r \in [0,1]$, together with a number of additional axioms. The semantics is the same as Łukasiewicz logic, just expanding the evaluations e of propositional variables in $[0,1]$ to truth-constants by requiring $e(\overline{r}) = r$ for all $r \in [0,1]$. Pavelka proved that his logic is complete for arbitrary theories in a non-standard sense. Namely, he defined the *truth degree* of a formula φ in a theory T as

$$\|\varphi\|_T = \inf\{e(\varphi) \mid e \text{ is a PL-evaluation model of } T\},$$

and the *provability degree* of φ in T as

$$|\varphi|_T = \sup\{r \in [0,1] \mid T \vdash_{\mathrm{PL}} \overline{r} \to \varphi\}$$

and proved that these two degrees coincide, i.e. $\|\varphi\|_T = |\varphi|_T$. This kind of completeness is usually known as Pavelka-style completeness, and strongly relies on the continuity of Łukasiewicz truth functions. Note that $\|\varphi\|_T = 1$ is not equivalent to $T \vdash_{PL} \varphi$, but only to $T \vdash_{PL} \overline{r} \to \varphi$ for all $r < 1$. Later, Hájek [13] showed that Pavelka's logic PL could be significantly simplified while keeping the completeness results. Indeed, he showed that it is enough to extend the language only by a countable number of truth-constants, one for each *rational* in $[0,1]$, and by adding only to the logic the two following additional axiom schemata, called *book-keeping axioms*:

$$\overline{r} \& \overline{s} \leftrightarrow \overline{r *_L s}$$
$$\overline{r} \to \overline{s} \leftrightarrow \overline{r \Rightarrow_L s}$$

for all $r \in [0,1] \cap \mathbb{Q}$, where $*_L$ and $\Rightarrow_L$ are the Łukasiewicz t-norm and its residuum respectively. He called this new system Rational Pavelka Logic, RPL for short. Moreover, he proved that RPL is strong standard complete for finite theories.

On the other hand, Hájek also shows that Gödel logic can be expanded with a *finite* set of truth constants together with a new unary connective Δ while preserving the strong standard completeness. Namely, let $C \subseteq [0,1]$ a finite set containing 1 and 0, and introduce into the language a truth-constant $\overline{r}$ for each $r \in C$, together with the so-called Baaz's projection connective Δ. Truth-evaluations of Gödel logic are extended in an analogous way to RPL as it regards to truth constants and adding the clause

$$e(\Delta\varphi) = \begin{cases} 1, & \text{if } e(\varphi) = 1 \\ 0, & \text{otherwise} \end{cases}$$

Note that despite φ is many-valued, $\Delta\varphi$ is a two-valued formula that is to be understood as a kind of presicification of φ. The introduction of the Δ is due to technical reasons to avoid clashes with the truth-constants. Finally, the axioms and rules of this new logic, denoted $G_\Delta(C)$ are those of Gödel logic G plus the above book-keeping axioms for truth-constants from C and the following axioms for Δ

(Δ1) $\Delta\varphi \vee \neg\Delta\varphi$
(Δ2) $\Delta(\varphi \vee \psi) \to (\Delta\varphi \vee \Delta\psi)$
(Δ3) $\Delta\varphi \to \varphi$
(Δ4) $\Delta\varphi \to \Delta\Delta\varphi$
(Δ5) $\Delta(\varphi \to \psi) \to (\Delta\varphi \to \Delta\psi)$

plus the bookeping axioms

$$\Delta\overline{r} \to_G \overline{0} \text{ for each } r \in C \setminus \{1\}$$

and the *Necessitation* rule for Δ: from φ derive $\Delta\varphi$. Then the following strong completeness result holds: $T \vdash_{G_\Delta(C)} \varphi$ iff $T \models_{G_\Delta(C)} \varphi$, for any theory T and formula φ.

Notation: in the rest of the paper we will write connectives with subindexes G or L, like $\wedge_G$, $\to_G$, $\to_L$, $\neg_L$, etc., to differentiate whether they are from Gödel or Łukasiewicz logics.

3 Graded Standard Deontic Logics

As already mentioned, in this section we are going to define two logics to reason about necessity and probability of Standard Deontic Logic formulas. Necessity and probability measures are two outstanding families of plausibility measures [14]. Given a Boolean algebra F, $\mu : F \to [0,1]$ is a *plausibility measure* if the following holds:

1. $\mu(\emptyset) = 0$
2. $\mu(W) = 1$
3. If $X, Y \in F$ and $X \subseteq Y$, then $\mu(X) \leq \mu(Y)$

A plausibility measure μ is a *necessity* measure if in addition μ satisfies

$$\mu(X_1 \cap X_2) = \min(\mu(X_1), \mu(X_2)), \text{ for all } X_1, X_2 \subseteq F$$

and μ is a (finitely additive) probability measure if it satisfies

$$\mu(X_1 \cup X_2) = \mu(X_1) + \mu(X_2) \text{ when } X_1 \cap X_2 = \emptyset, \text{ for all } X_1, X_2 \subseteq F$$

Necessity mesures are purely qualitative in the sense that the order is what matters, and they have been widely used to model a notion of ordinal preference [3,17]. We will mainly assume this last interpretation as intended semantics although we do not exclude other possibilities.

3.1 Necessity-Valued Standard Deontic Logic

We define a fuzzy modal language over Standard Deontic Logic SDL to reason about the necessity degree of deontic propositions. The language of *Necessity-valued Deontic Logic* (NSDL) results from the union of the language of the logic $\mathrm{G}_\Delta(C)$ (Gödel logic extended with the Δ operator and a finite set $C \subset [0, 1]$ of truth-constants) and the language of Standard Deontic Logic (SDL), extended with a fuzzy unary operator N. Formulas of $NSDL$ are of two types:

- *Deontic formulas:* they are the formulas of SDL, built in the usual way with the obligation deontic modality O. $\top$ and $\bot$ denote the truth-constants *true* and *false* respectively. It is said that a formula of SDL is *closed* if every propositional variable is in the scope of a modality.
- *N-formulas:* they are built from elementary N-formulas $N\varphi$, where φ is a closed SDL-formula, and truth-constants $\overline{r}$, for each rational $r \in C \subset [0, 1]$, using the connectives of Gödel many-valued logic:

  - If $\varphi \in SDL$ is closed, then $N\varphi \in NSDL$
  - If $r \in C \subset [0, 1]$ then $\overline{r} \in NSDL$
  - If $\Phi, \Psi \in NSDL$ then $\Phi \rightarrow_G \Psi \in NSDL$ and $\Phi \wedge_G \Psi \in NSDL$ (where $\wedge_G$ and $\rightarrow_G$ correspond to the conjunction and implication of Gödel logic)
  - If $\Phi \in NSDL$ then $\Delta\Phi \in NSDL$

  Other $\mathrm{G}_\Delta(C)$ logic connectives for the N-formulas can be defined from $\wedge_G$, $\rightarrow_G$ and $\overline{0}$ in the way described in Section 2.

Since in Gödel Logic $G_\Delta(C)$ the formula $\Phi \rightarrow_G \Psi$ is 1-true iff the truth value of Ψ is greater or equal to that of Φ, formulas of the type $\overline{r} \rightarrow_G N\varphi$ (where φ is a closed formula of SDL) express that the necessity degree of the norm φ is at least r.

In this language we can express with the formula $\neg_G \neg_G N\varphi$, that the necessity degree of the norm φ is positive[3], and with the formula $\varphi \equiv_G \overline{r}$, that is exactly of degree r. Comparisons of degrees are done by means of formulas of the form $N\varphi \rightarrow_G N\psi$.

NSDL Semantics. The semantics of our language is given by means of *Necessity-valued Deontic Kripke models* of the following form: $K = (W, R, e, \mu)$, where (W, R, e) is an standard Kripke model of SDL, and μ is a *necessity measure* on some Boolean subalgebra $F \subseteq 2^W$ such that the sets $\{w \mid e(w, \psi) = 1\}$, for every closed SDL-formula ψ, are μ-measurable. Remember that in every standard Kripke model (W, R, e) of SDL, R is a serial binary relation on W (that is, for every $w \in W$ there is $t \in W$ such that $(w, t) \in R$).

The truth value $e(w, \varphi)$ of a SDL formula φ in a world w is defined as usual (either 0 or 1). The truth-value of atomic N-formulas $N\psi$ in the model K is defined as

$$\|N\psi\|_K = \mu(\{w \mid e(w, \psi) = 1\})$$

Then the truth-value $\|\Phi\|_K$ of compound N-formulas Φ is defined by using $G_\Delta(C)$ truth-functions. If Φ is a N-formula, we will write $\models_{NSDL} \Phi$ when $\|\Phi\|_K = 1$ for any model K, and if T is a set of N-formulas, $T \models_{NSDL} \Phi$ when $\|\Phi\|_K = 1$ for all models K such that $\|\Psi\|_K = 1$ for $\Psi \in T$.

NSDL Axioms and Rules. Axioms of NSDL are:

1. Axioms of SDL (for SDL-formulas)
2. Axioms of $\mathrm{G}_\Delta(C)$ (for N-formulas)
3. Necessity Axioms (where φ and ψ are closed SDL formulas):
   (a) $N(\varphi \rightarrow \psi) \rightarrow_G (N\varphi \rightarrow_G N\psi)$
   (b) $N(\varphi \wedge \psi) \equiv_G N(\varphi) \wedge_G N(\psi)$
   (c) $\neg_G N(\bot)$
   (d) $N\psi$, for every SDL-theorem

Deduction rules for NSDL are Modus Ponens (both for → of SDL and for $\rightarrow_G$ of $\mathrm{G}_\Delta(C)$) necessitation for the obligation deontic modality O (from φ derive $O\varphi$, if $\varphi \in SDL$) and necessitation for Δ (from Φ derive $\Delta\Phi$, for N-formulas). Alternatively, instead of Necessity Axiom 3 (d), one may add the rule "from ϕ infer $N\phi$" for a closed SDL-formula, in this way one can obtain a system with finitely-many axiom schemes and rules. We have introduced a recursive Hilbert-style axiom system since provability in SDL is decidable. We will denote by $\vdash_{NSDL}$ the usual notion of proof from the above axioms and rules

It is worth pointing out that Necessity Axiom 3 (a) ensures that N preserves SDL logical equivalence. Observe that the formula

$$N(\varphi \vee \psi) \equiv_G N(\varphi) \vee_G N(\psi)$$

[3] Notice that in Gödel Logic, $(x \Rightarrow_G 0) \Rightarrow_G 0 = 1$ iff $x > 0$.

is neither sound nor provable from the above axioms. On the other hand the following formulas are indeed provable:

1. $N(\varphi \wedge \neg\varphi) \equiv_G \overline{0}$
2. $N(\varphi \vee \neg\varphi) \equiv_G \overline{1}$

Soundness and Completeness Theorems of NSDL

Definition 1. *A set of formulas T is a N-theory if all the formulas in T are N-formulas.*

By definition of the NSDL axioms and rules it is easy to check that for every set of SDL of formulas Σ and every SDL-formula ϕ,

$$\Sigma \vdash_{NSDL} \phi \text{ iff } \Sigma \vdash_{SDL} \phi.$$

Therefore, NSDL is a conservative extension of SDL. Moreover, observe that every SDL-formula provable in a N-theory is a SDL-theorem.

Following Theorem 8.4.9 of [13], a N-theory T can be represented as a theory over the propositional logic $\mathrm{G}_{\Delta}(C)$. For each closed SDL-formula ϕ we introduce a propositional variable p_ϕ, corresponding to the formula $N\phi$. We define the following translation: $(N\phi)^* = p_\phi$, $(\overline{r})^* = \overline{r}$, for each rational $r \in C \subset [0,1]$ and for every N-formula ϕ and φ, $(\phi \wedge_G \varphi)^* = \phi^* \wedge_G \varphi^*$ and $(\phi \rightarrow_G \varphi)^* = \phi^* \rightarrow_G \varphi^*$. Let T^* be the following set of $\mathrm{G}_{\Delta}(C)$ formulas:

– Propositional variables p_ϕ, for each closed formula ϕ, theorem of SDL.
– formulas of the form φ^*, for each Necessity Axiom φ.
– α^*, for each formula $\alpha \in T$

Lemma 2. *If T is a N-theory and ϕ a N-formula, then*

$$T \vdash_{NSDL} \phi \textit{ iff } T^* \vdash_{G_{\Delta}(C)} \phi^*$$

Proof. Assume that $T^* \vdash_{G_{\Delta}(C)} \phi^*$. Let $\alpha_1^*, \ldots, \alpha_k^*$ be a $\mathrm{G}_{\Delta}(C)$-proof of ϕ^* in T^*. Then the sequence $\alpha_1, \ldots, \alpha_k$ can be converted in a NSDL-proof of ϕ in T by adding for each formula of the form p_ψ that occurs in $\alpha_1^*, \ldots, \alpha_k^*$, a proof of ψ in SDL and then applying the rule of necessitation for N-formulas.

Conversely, assume $T \vdash_{NSDL} \phi$. Then a $\mathrm{G}_{\Delta}(C)$-proof of ϕ^* in T^* can be obtained by taking the translation of the formulas of one NSDL-proof of ϕ in T, once the SDL-formulas are deleted. Use the fact that every SDL-formula provable in a N-theory is a SDL-theorem.

From the fact that the Necessity Axioms are 1-true in every Necessity-valued Deontic Kripke model follows the Soundness Theorem:

Lemma 3. ***(Soundness)*** *For every N-theory T over NSDL and every N-formula ϕ, $T \vdash_{NSDL} \phi$ implies $T \models_{NSDL} \phi$.*

Theorem 4. ***(Completeness)*** *For every N-theory T over NSDL and every N-formula ϕ:*

$T \models_{NSDL} \phi$ implies $T \vdash_{NSDL} \phi$.

Proof. By Lemma 2 and the Completeness Theorem of the Logic $\mathrm{G}_\Delta(C)$ it is enough to prove that

$$T \models_{NSDL} \phi \text{ implies } T^* \models_{G_\Delta(C)} \phi^*.$$

Assume $T^* \not\models_{G_\Delta(C)} \phi^*$. Let E be a model of T^*, with evaluation v of the propositional variables p_ψ such that $v(\phi^*) < 1$. We show that there is a model K of T that is not a model of ϕ.

Let (W, R, e) be the canonical model of SDL. Observe that for every formula $\phi \in SDL$, the canonical model satisfies:

$$\psi \text{ is valid in } (W, R, e) \text{ iff } \psi \text{ is a theorem of SDL.}$$

Consider now the following Boolean subalgebra $F \subseteq 2^W$:

$$F = \{\{w \mid e(w, \psi) = 1\} : \psi \text{ is a closed formula of SDL}\}$$

let us denote by X_ψ the set $\{w \mid e(w, \psi) = 1\}$. We define a function μ on F in the following way: $\mu(X_\psi) = v(p_\psi)$. Then we can show:

(i) μ is a necessity measure on F.

1. μ is a well-defined function. Proof: if $X_\alpha = X_\beta$, for α and β, closed SDL-formulas, then $X_{\alpha\equiv\beta} = W$ and $\alpha \equiv \beta$ is valid in the canonical model. Consequently, $\alpha \equiv \beta$ is a theorem of SDL. Since E is a model of T^*, $v(p_{\alpha\equiv\beta}) = 1$. By using the translation by the *-operation of Necessity Axiom 3 (a), $N(\alpha \to \beta) \to_G (N\alpha \to_G N\beta)$, we have $v(p_\alpha) = v(p_\beta)$. Thus, we can conclude that $\mu(X_\alpha) = \mu(X_\beta)$.
2. It is easy to check with the same kind of argument as before that $\mu(\emptyset) = 0$ and $\mu(W) = 1$.
3. For every α and β, closed SDL-formulas
$$\mu(X_\alpha \cap X_\beta) = \min(\mu(X_\alpha), \mu(X_\beta))$$
Proof: Since E is a model of T^*, E is also a model of the *-translation of the Necessity Axiom 3 (b), $N(\alpha \wedge \beta) \equiv_G N(\alpha) \wedge_G N(\beta)$. Therefore E is a model of the formula $p_{\alpha\wedge\beta} \equiv_G p_\alpha \wedge_G p_\beta$ and thus
$$v(p_{\alpha\wedge\beta}) = \min(v(p_\alpha), v(p_\beta))$$
we can conclude that
$$\mu(X_\alpha \cap X_\beta) = \mu(X_{\alpha\wedge\beta}) = \min(\mu(X_\alpha), \mu(X_\beta))$$

(ii) For every N-formula $\Phi \in NSDL$, $\|\Phi\|_K = v(\Phi^*)$.
Proof: For this, it is enough to show that for every closed formula $\varphi \in SDL$, $\|N\varphi\|_K = v(p_\varphi)$. It is easy to check by induction on the complexity of the N-formulas and by definition of μ.

Let us denote by M_v the model (W, R, e, μ). We have just proved that M_v is a necessity-valued deontic Kripke model of T and $\|\phi\|_{M_v} = v(\phi^*) < 1$.

3.2 Probability-Valued Deontic Logic

In a quite similar way to NSDL, we define now a fuzzy modal language over Standard Deontic Logic to reason about the probability degree of deontic propositions. The language of *Probability-valued Deontic Logic* PSDL is defined as follows. Formulas of PSDL are of two types:

- *Deontic formulas:* Formulas of SDL.
- *P-formulas:* they are built from elementary P-formulas $P\varphi$, where φ is a closed SDL-formula, and truth-constants $\overline{r}$, for each rational $r \in [0,1]$, using the connectives of Rational Pavelka Logic.

The semantics of our language is given by means of *Probability-valued Deontic Kripke models* of the following form: $K = (W, R, e, \mu)$, where (W, R, e) is an standard Kripke model of SDL, and μ is a finitely additive probability on some Boolean subalgebra $F \subseteq 2^W$ such that the sets $\{w \mid e(w,\psi) = 1\}$, for every closed SDL-formula ψ, are μ-measurable.

The truth value of an atomic P-formula $P\psi$ in a model K is defined as

$$\|P\psi\|_K = \mu(\{w \mid e(w,\psi) = 1\})$$

and the truth-value of compound P-formulas are computed from the atomic ones using the truth-functions of Łukasiewicz logic. Now, given a P-theory T (a set of P-formulas) one defines the *truth-degree* of a P-formula Φ over T as the value

$$\| \phi \|_T = \inf\{\| \phi \|_K \mid K \text{ is a PSDL-model of } T\}$$

where K is a PSDL-model of T when $\|\Psi\|_K = 1$ for every $\Psi \in T$. We define the *provability degree* of Φ over T as

$$| \phi |_T = \sup\{r \mid T \vdash_{PSDL} \overline{r} \to \phi\}$$

We introduce now a sound and recursive axiom system for PSDL. Axioms of PSDL are:

1. Axioms of SDL (for SDL-formulas)
2. Axioms of RPL (for P-formulas)
3. Probability Axioms (where φ and ψ are closed SDL-formulas):
   (a) $P(\varphi \to \psi) \to_L (P\varphi \to_L P\psi)$
   (b) $P(\varphi \vee \psi) \equiv P\varphi \oplus (P\psi \ominus P(\varphi \wedge \psi))$
   (c) $\neg_L P(\bot)$
   (d) $P\psi$, for every SDL-theorem

where $\Phi \oplus \Psi$ is a shorthand for $\neg_L \Phi \to_L \Psi$ and $\Phi \ominus \Psi$ is a shorthand for $\neg_L(\Phi \to_L \Psi)$[4]. Deduction rules for PSDL are Modus Ponens (both for $\to$ of SDL and for $\to_L$ of RPL) and necessitation for the obligation modality O.

We can prove in an analogous way we did with the logic NSDL that PSDL is a conservative extension of SDL. Although a completeness theorem analogous to the one for NSDL does not hold for PSDL, similar techniques allows us to prove that the Pavelka-style completeness of RPL extends to PSDL.

[4] Note that in Łukasiewicz Logic $(x \Rightarrow_L 0) \Rightarrow_L y = \min(1, x+y)$ and $(x \Rightarrow_L y) \Rightarrow_L 0 = \max(0, x-y)$.

Theorem 5. *[Pavelka Completeness] Let T a P-theory and Φ a P-formula. Then it holds that $\|\Phi\|_T = |\Phi|_T$.*

4 Examples

Norm preferences. Necessity-valued graded deontic logics allow to attach preference degrees to norms and this may be used to deal with conflicts among norms, in a kind of defeasible logic approach. Consider the following example adapted from [6]. Suppose Company A has the following norms: premium costumers of Company A are entitled to get a discount, but costumers which place special orders are not allowed to get such a discount. Such a policy can be described by the theory

$$\begin{aligned}\Gamma = \{ & \, PremiumCostumer(x) \rightarrow ODiscount(x), \\ & \, SpecialOrder(x) \rightarrow O\neg Discount(x) \ \}\end{aligned}$$

John is a premium costumer but has placed a special order. In this case, assuming both norms to have the same priority level, the above policy clashes: clearly, $\Gamma \cup \{PremiumCostumer(John), SpecialOrder(John)\} \vdash_{SDL} \bot$.

Let us further assume that the norm regarding special orders has a higher priority than the norm regarding premium costumers. In this case, we can describe the company policy by the following NSDL theory

$$\begin{aligned}\Gamma^* = \{ & \, \overline{r_1} \rightarrow N(PremiumCostumer(x) \rightarrow ODiscount(x)), \\ & \, \overline{r_2} \rightarrow N(SpecialOrder(x) \rightarrow O\neg Discount(x)) \ \}\end{aligned}$$

where $r_1 < r_2$. Now we have

$$\Gamma^* \cup \{NPremiumCostumer(John), NSpecialOrder(John)\} \vdash_{NSDL} \overline{r_1} \rightarrow N(\bot).$$

But in terms of the non-monotonic consequence relation associated to a necessity-valued logic[5] [3], this leads to

$$\Gamma \cup \{PremiumCostumer(John), SpecialOrder(John)\} |\!\sim O\neg Discount(John),$$

this is to say, the norm about special orders prevails.

Norm Compliance. We illustrate the use of Probability-valued Deontic Logics by means of an example in which an agent i evaluates the probability of achieving a certain goal g. Let agent i represent a person with disabilities. Agent i wants to buy a certain product and he can choose between two supermarkets, A and B, for buying the desired product. In order to take this decision, on the one hand, agent i calculates, for each supermarket, the probability of the norm compliance by other clients of the provision of parking places for people with disabilities. On the other hand, he calculates the probability of norm compliance by supermarkets A and B, of the Disability Discrimination Act (1995, extended in 2005), regarding buildings accessibility. We could formalise in SDL sentences expressing these norms, in a simplified way:

[5] Let Γ^* be a theory over NSDL and let $\alpha = \sup\{r \mid \Gamma^* \vdash_{SDL} \overline{r} \rightarrow \bot\}$. Then define $\Gamma |\!\sim \varphi$ when $\Gamma^* \vdash_{NSDL} \overline{r} \rightarrow \varphi$ with $r > \alpha$.

- It is obligatory for supermarket A (B) to have at least one accessible entrance route $OAcc_A$, ($OAcc_B$, respectively).
- It is prohibited to park in a place reserved for people with disabilities in supermarket A (B) $O\neg Park_A$ ($O\neg Park_B$, respectively)

A graded deontic language will allow us to reason about probabilities in this context of uncertainty. Consider the following sentences of SDL:

1. $OAcc_A \wedge O\neg Park_A$,
2. $OAcc_B \wedge O\neg Park_B$

1. is equivalent to $O(Acc_A \wedge \neg Park_A)$ and 2. is equivalent to $O(Acc_B \wedge \neg Park_B)$. Let T be a set of premises of the graded logic PSDL representing the information we have about norm compliance in this two supermarkets. Then, if the following holds:

$$T \vdash_{PSDL} PO(Acc_A \wedge \neg Park_A) \rightarrow_L PO(Acc_B \wedge \neg Park_B)$$

then agent i would take the decision of going to buy to supermarket B. Observe that the formula of PSDL

$$PO(Acc_A \wedge \neg Park_A) \rightarrow_L PO(Acc_B \wedge \neg Park_B)$$

is 1-true in a model iff the probability degree of the norm compliance of $O(Acc_B \wedge \neg Park_B)$ is greater or equal than the degree of norm compliance of $O(Acc_A \wedge \neg Park_A)$.

5 Related and Future Work

The paper [6] provides a logical analysis of conflicts between informational, motivational and deliberative attitutes. The resolution of conflicts is based on Thomason's idea of prioritization, which is considered in the BOID logic [5] as the order of derivations from different types of attitudes. Thomason's BDP logic (see [24]) is based on Reiter's default logic and extended in the BOID logic with conditional obligations and intentions.

In [11] the authors follow the BOID architecture to describe agents and agent types in Defeasible Logic. Reasoning about agents can be embedded in frameworks based on non-monotonic logics, as one the most interesting problems concerns the cases where the agent's mental attitudes are in conflict or when they are incompatible with obligations and other deontic provisions.

BOID specifies 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. Intentions and beliefs are viewed as constituting the internal constraints of an agent while obligations are its external constraints.

More recently, in [20], M. Nickles proposes a logic-based approach based on the notion of behavioral expectation. In his paper he presents a quantification of the norm adherence of an agent using the measurement of norm deviance. Normative expectations are defined via the degree of resistance to social dynamics in the course of time. Our approach differs from previous ones because our purpose is not to fuzzify deontic logic giving a different interpretation to its modalities. We want to provide a reasoning model for an agent in order to represent how he attaches necessity or probability degrees to norms of a given institution.

Up to now we have applied our framework to SDL. However, our purpose is to provide a general way to define a necessity-valued or probability fuzzy logic over any given deontic logic, allowing to attach a grade to the norms described in the deontic language. Future work will include working with other logics such as Dynamic Logic (see [22] and [15]) Dyadic Deontic Logic (see [29] or [26]), the KARO formalism [27], B-DOING Logic [8], the Logic of "Count-as" [12], Normative ATL [25] or Temporal Logic of Normative Systems [1]. Since different definitions of norm adherence or of probability of norm compliance would give rise to a variety of formal systems, changing or adding new axioms to the basic axiomatization we have introduced here, future work will be devoted to the study of these notions in a multi-institutional setting.

Acknowledgements. The authors wish to express their thanks to Carles Sierra and the reviewers of this paper for their helpful comments. Research partially supported by the Generalitat de Catalunya, under the grant 2005-SGR-00093 and the Spanish projects "Agreement Technologies" (CONSOLIDER CSD2007-0022, INGENIO 2010) and TIN2006-15662-C02-01.

References

1. Ågotnes, T., van der Hoek, W., Rodríguez-Aguilar, J.A., Sierra, C., Wooldridge, M.: On the Logic of Normative Systems. In: Twentieth International Joint Conference on AI, IJCAI 2007, pp. 1175–1180. AAAI Press, Menlo Park (2007)
2. Aqvist, L.: Deontic Logic. In: Gabbay, et al. (eds.) Handbook of Philosophical Logic, vol. 8, pp. 147–265 (2001)
3. Benferhat, S., Dubois, D., Prade, H.: Towards a possibilistic logic handling of preferences. Applied Intelligence 14, 303–317 (2001)
4. Burns, T.R., Carson, M.: Actors, Paradigms, and Institutional Dynamics: The Theory of Social Rule Systems Applied to Radical Reforms. In: Hollingsworth, R., et al. (eds.) Advancing Socio-Economics: An Institutionalist Perspective, pp. 109–147. Rowman and Littlefield Publishers, New York (2002)
5. Broersen, J., Dastani, M., Hustijin, J., Torre, L.W.: Goal generation in the BOID architecture. Cognitive Science Quaterly 2, 428–447 (2002)
6. Broersen, J., Dastani, M., Torre, L.W.: Resolving Conflicts between Beliefs, Obligations, Intentions, and Desires. In: Benferhat, S., Besnard, P. (eds.) ECSQARU 2001. LNCS (LNAI), vol. 2143, pp. 568–579. Springer, Heidelberg (2001)
7. Casali, A., Godo, L., Sierra, C.: Graded BDI Models for Agent Architectures. In: Leite, J.A., Torroni, P. (eds.) CLIMA 2004. LNCS (LNAI), vol. 3487, pp. 126–143. Springer, Heidelberg (2005)
8. Dignum, F., Morley, D., Sonenberg, E.A., Cavedon, L.: Towards Socially Sophisticated BDI Agents. In: Proc. 4th Int. Conf. on Multi-Agent Systems ICMAS-2000, Boston, MA, pp. 111–118 (2000)

9. García-Camino, A., Rodríguez-Aguilar, J.A., Vasconcelos, W.: Distributed Norm Management in Regulated Multi-agent Systems. In: Procs. of 6th Int'l Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2007), Hawai'i, pp. 624–631 (2007)
10. Godo, L., Hájek, P., Esteva, F.: A Fuzzy Modal Logic for Belief Functions. Fundamenta Informaticae, pp. 1001–1020 (2001)
11. Governatori, G., Rotolo, A.: BIO Logical Agents: Norms, Beliefs, Intentions in Defeasible Logic. In: Boella, G., van der Torre, L., Verhagen, H. (eds.) Normative Multi-agent Systems. Dagstuhl Seminar Proceedings, 07122 (2007), http://drops.dagstuhl.de/opus/volltexte/2007/912
12. Grossi, D., Meyer, J.-J.Ch., Dignum, F.: Modal logic investigations in the semantics of counts-as. In: ICAIL 2005: Proceedings of the 10th international conference on Artificial intelligence and law, Bologna, Italy, pp. 1–9. ACM, New York (2005)
13. Hájek, P.: Metamatematics of Fuzzy Logic. Trends in Logic, vol. 4. Kluwer Academic Publishers, Dordrecht (1998)
14. Halpern, J.Y.: Reasoning about Uncertainty. The MIT Press, Cambridge Massachusetts (2003)
15. Harel, D.: Dynamic logic. In: Gabbay, D., Guenthner, F. (eds.) Handbook of Philosophical Logic, vol. II, pp. 497–604 (1984)
16. Knijnenburg, P.M.W., van Leeuwen, J.: On Models for Propositional Dynamic Logic. Theor. Comput. Sci. 91(2), 181–203 (1991)
17. Liau, C.J.: On the possibility theory-based semantics for logics of preference. International Journal of Approximate Reasoning 20, 173–190 (1999)
18. Meyer, J.-J.Ch., Veltman, F.: Intelligent Agents and Common-Sense Reasoning. In: Blackburn, P., et al. (eds.) Handbook of Modal Logic. Studies in Logic and Practical Reasoning, vol. 3, pp. 991–1030 (2006)
19. 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 29(1), 109–136 (1998)
20. Nickles, M.: Towards a Logic of Graded Normativity and Norm Adherence. In: Boella, G., van der Torre, L., Verhagen, H. (eds.) Normative Multi-agent Systems. Dagstuhl Seminar Proceedings, 07122 (2007), http://drops.dagstuhl.de/opus/volltexte/2007/926
21. Pavelka, J.: On fuzzy logic I, II, III. Zeitschrift für Mathematische Logik and Grundlagen der Mathematik 25, 45–52, 119–134, 447–464 (1979)
22. Pratt, V.R.: Semantical considerations on Floyd-Hoare logic. In: Proceedings of the 17th IEEE Symposium on the Foundations of Computer Science, pp. 109–121 (1976)
23. Ross, W.D.: Foundations of Ethics. Oxford University Press, London (1939)
24. Thomason, R.H.: Desires and Defaults: A Framework for Planning with Inferred Goals. In: Proceedings of the KR 2000, pp. 702–713. Morgan Kaufmann, San Francisco (2000)
25. van der Hoek, W., Wooldridge, M.: On obligations and normative ability: towards a logical analysis of the social contract. Journal of Applied Logic 3, 396–420 (2005)
26. van der Torre, L., Tan, Y.-H.: Contrary-to-duty reasoning with preference-based dyadic obligations. Annals of Mathematics and Artificial Intelligence 27, 49–78 (1999)
27. van Linder, B.: Modal Logics for Rational Agents, PhD. Thesis, Utrecht University (1996)
28. von Wright, G.H.: Deontic Logic. Mind 60, 1–15 (1951)
29. von Wright, G.H.: A New System of Deontic Logic. Danish Yearbook of Philosophy 1 (1964)

Pushing Anderson's Envelope: The Modal Logic of Ascription

Davide Grossi

Individual and Collective Reasoning Group
Computer Science and Communications
University of Luxembourg
davide.grossi@uni.lu

Abstract. The paper proposes a formal analysis of the ascriptive view of norms as resulting from pulling together Anderson's reductionist approach, the analysis of counts-as, and a novel modal approach to the formal representation of languages in logic. This unifying attempt results in the definition of a new form of reduction of deontic logic based on counts-as statements. Such result is discussed also in the light of Jørgensen's dilemma.

Keywords: Modal logic, Anderson's reduction, counts-as, ascription, Jørgensen's dilemma.

1 Introduction

The present paper intends to pull together independent threads which have been thus far followed by the (formal) studies of norms. Such threads are the reductionist approach to norms started with [1,2,3,19], the study of counts-as initiated in [26,27] and first pursued with formal means in [17], and the ascriptive view of norms first put forth in [24] and more recently developed, among others, in [16]. According to this latter, norms are actually ascriptions of deontic properties to actions or states of affairs. In short, to state norms means to create new properties, which are somehow inexistent in reality (e.g., Anderson's "violation"), to create new words to name them, and consequently to predicate them of the relevant states of affairs or actions. The paper proposes a formal analysis of this view of norms which builds, in the first place, on Anderson's reduction, in the second place, on the formal analysis of counts-as developed in [11,13,12,14,10] and, in the third place, on a formal characterization of the language creation aspect of the ascriptive view of norms. As a result, a comprehensive formal theory of norms is presented and formalized in modal logic.

The paper is structured as follows. Section 2 summarizes Anderson's reduction approach and provides a contextual version of it. Part of the section consists of a summary of the results presented in [11, 13, 12, 14, 10] and provides the ground for a counts-as based view of Anderson's reduction. At the end of the section the ascriptive view of norms is exposed in more details to introduce Section 3. There, a language-based notion of indistinguishability between propositional models is introduced and a modal logic, first studied in [20,21], is exposed for reasoning about it. This language-based notion of indistinguishability will be the key for capturing the phenomenon of language

R. van der Meyden and L. van der Torre (Eds.): DEON 2008, LNAI 5076, pp. 263–277, 2008.

creation inherent in the ascriptive view of norms. A simple example is used throughout the exposition of the formalism. Section 4 applies the formalisms presented in Sections 2 and 3, providing a formal characterization of the ascriptive view of norms in the guise of a new notion of counts-as. The section ends with some remarks concerning the relation between the formal analysis presented and Jørgensen's dilemma. Finally, Section 5 draws some conclusions and sketches future research lines.

2 Anderson's Reduction Revisited

By "Anderson's reduction" the present paper intends, in general, the approach to deontic logic which is based on the reduction of deontic notions to evaluative ones (e.g., 'good', 'ideal, 'bad', 'violation'). Such approach was first systematically developed in Anderson's work [1,3,2]. In that work, the reduction of deontic statements to alethic ones is based on the intuition according to which the fact that ϕ is obligatory means that $\neg\phi$ "necessarily" implies a violation, in symbols: $\Box(\neg\phi \rightarrow \mathtt{V})$, where $\mathtt{V}$ is a specific atom for which it is valid that $\Diamond\neg\mathtt{V}$, i.e., that the violation is not "necessary". The nature of the reduction lies in how this reference to a "necessity" is formally modeled. In the original proposal of Anderson the system chosen for the reduction was **K**[1]. Various alternative versions of Anderson's reduction are studied, for instance, in [7,20,22].

2.1 Terminological Necessities

We start considering the form of reduction based on system **S5** such as the ones studied in [7,20]. By interpreting the $\Box$ operator occurring in the reduction expression as an **S5** necessity, formulae $\Box(\neg\phi \rightarrow \mathtt{V})$ could be soundly rephrased as: *the negation of ϕ unconditionally implies a violation*. Notice that the **S5**-based interpretation of the reduction is in line with Anderson's intuition [4] that the occurrence of a violation follows *analytically* from the fact that an obligation is not fulfilled.

It is well-known that **S5** is the modal logic of universal quantification since the so-called universal modality (i.e., the modality interpreted on the $W \times W$, where W is the model's domain) is an **S5** modality [6]. Now, viewing the $\Box$ modality in Anderson's reduction as the universal modality, which we denote by $[u]$, conveys a key semantic hint:

$$\mathcal{M}, w \models [u](\neg\phi \rightarrow \mathtt{V}) \text{ iff } \forall w' \in W : \mathcal{M}, w' \models \neg\phi \rightarrow \mathtt{V} \quad (1)$$

$$\text{iff } \mathcal{I}(\neg\phi) \subseteq \mathcal{I}(\mathtt{V}) \quad (2)$$

where $\mathcal{M}$ is a model for the modal language with universal modality $[u]$, W is its domain and $\mathcal{I}$ its evaluation function. Formulae 1 and 2 show a very precise interpretation of Anderson's reduction: ϕ is obligatory means that all states (i.e., possible worlds) are

[1] It might be instructive to recall that Kanger independently developed an analogous reduction based on a constant $\mathtt{Q}$ denoting normative ideality, or the absence of violation [19]. In this case, the fact that ϕ is obligatory means that ϕ "necessarily" follows from ideality, in symbols: $\Box(\mathtt{Q} \rightarrow \phi)$.

such that either ϕ is true or, if ϕ is false, then a violation is also true. In this view, deontic statements amount to set-theoretic relations concerning the interpretation $\mathcal{I}(\mathtt{V})$ of the atom $\mathtt{V}$.

If the deontic statements of a normative system can be represented by modal formulae involving the universal modality and the violation atom, what happens if we want to consider, under the same formalism, deontic statements belonging to several different normative systems? Technically speaking, we then look for operators that can "locally" behave like a universal modality, but that can "globally" behave in a weaker way allowing for the representation of different and possibly inconsistent deontic statements at the same time. We should find a multi-modal logic such that: a) the logic enables as many modalities as the normative systems we intend to represent; b) these modalities retain as many characteristics of $[u]$ as possible; c) the logic allows for the satisfiability of expressions such as: $[i](\neg\phi \rightarrow \mathtt{V}) \wedge \neg[j](\neg\phi \rightarrow \mathtt{V})$. To put it roughly, we look for a modal logic by means of which to express *contextual* terminological necessity.

2.2 A Modal Logic of Context

In logic, contexts have been studied as sets of models [9]. Now, if the models considered are models of propositional languages, then contexts can be studied as sets of possible worlds [28]. The present section exposes a logic based on this view[2]. The result is a contextual version of Anderson's reduction.

Syntax of Cxt$^{\mathbf{u}}$. The syntax of **Cxt$^{\mathbf{u}}$** is the syntax of a multi-modal language $\mathcal{L}^{Cxt}$ [6] where n is the cardinality of the set $\mathtt{C}$ of contexts and u the index of the universal modality. The alphabet of $\mathcal{L}^{Cxt}$ contains: an at most countable set $\mathbf{P}$ of propositional atoms p; the set of boolean connectives $\{\neg, \wedge, \vee, \rightarrow\}$; a finite non-empty set $\mathtt{C}$ of context indexes containing the context index u. Metavariables $i, j, \ldots$ are used to denote elements of $\mathtt{C}$. The set of well-formed formulae ϕ of $\mathcal{L}^{Cxt}$ is defined by the following BNF:

$$\phi ::= \top \mid p \mid \neg\phi \mid \phi_1 \wedge \phi_2 \mid \phi_1 \vee \phi_2 \mid \phi_1 \rightarrow \phi_2 \mid [i]\phi \mid \langle i \rangle \phi.$$

where i denotes elements in $\mathtt{C}$.

Semantics of Cxt$^{\mathbf{u}}$. Languages $\mathcal{L}^{Cxt}$ are given a semantics via the class of $\mathrm{C}{\scriptstyle\mathrm{XT}}^{\top}$ frames $\mathcal{F} = \langle W, \{W_i\}_{i \in \mathtt{C}} \rangle$ such that $W \in \{W_i\}_{i \in \mathtt{C}}$. Leaving technicalities aside, these frames consist of the domain W and of a finite number $n = |\mathtt{C}|$ of subsets of W among which W itself[3]. Such subsets straightforwardly model the notion of context sketched above.

[2] Readers are again referred to [10] for a more detailed exposition.

[3] Notice that such structures are multi-sets, or bags, rather than frames. However, it is proven that they represent secondarily universal frames which also contain one universal relation [10, Ch. 4]. An alternative more general semantics to **Cxt$^{\mathbf{u}}$** can be given via the class $\mathfrak{IE}^{\sim}$ of frames satisfying the following properties: they are i-j transitive (if wR_iw' and $w'R_jw''$ then wR_jw''), i-j euclidean (if wR_iw' and wR_jw'' then $w'R_jw''$), and they contain an equivalence relation R_u such that for all $i \in \mathtt{C}$, $R_i \subseteq R_u$. It is proven, however, that classes $\mathrm{C}{\scriptstyle\mathrm{XT}}^{\top}$ and $\mathfrak{IE}^{\sim}$ are modally equivalent (see [10, Appendix]).

Notice that the domain W represents the global, or universal, context. Models are, as usual, structures $\langle \mathcal{F}, \mathcal{I} \rangle$ where $\mathcal{F}$ belongs to the class $\textsc{Cxt}^{\top}$ and $\mathcal{I}$ is the valuation function $\mathcal{I} : \mathbf{P} \longrightarrow \mathcal{P}(W)$.

Definition 1. (Satisfaction based on $\textsc{Cxt}^{\top}$ frames)
Let $\mathcal{M}$ be a model built on a $\textsc{Cxt}^{\top}$ frame.

$$\mathcal{M}, w \models [i]\phi \;\; \mathit{iff} \;\; \forall\, w' \in W_i : \mathcal{M}, w' \models \phi$$

where u is the universal context index, $W_u = W$ and i ranges on the context indexes in $\mathtt{C}$. The obvious clauses for the Boolean connectives and the dual of $[i]$ are omitted.

Notice that the $[u]$ is the universal operator. Notice also that, while in standard modal logic the truth of $[i]$ and $\langle i \rangle$ formulae depends on the evaluation state, the truth of such formulae interpreted within $\textsc{Cxt}^{\top}$ frames abstracts therefrom: in other words truth implies validity. This is what we would intuitively expect for the contexts of normative systems: what holds in the context of a given normative system is not determined by the point of evaluation but just by the system as such, i.e., by its own norms.

Axiomatics of $\mathbf{Cxt^u}$. Logic $\mathbf{Cxt^u}$ results from the union of the modal logic $\mathbf{K45^{ij}_n}$, which axiomatizes contexts [10], with an **S5** logic axiomatizing the behavior of the global context u, plus the interaction axiom $\subseteq .ui$, which just states that u is the biggest context.

(P) all tautologies of propositional calculus

($\mathtt{K}^i$) $[i](\phi_1 \rightarrow \phi_2) \rightarrow ([i]\phi_1 \rightarrow [i]\phi_2)$

(4^{ij}) $[i]\phi \rightarrow [j][i]\phi$

(5^{ij}) $\neg[i]\phi \rightarrow [j]\neg[i]\phi$

($\mathtt{T}^u$) $[u]\phi \rightarrow \phi$

($\subseteq .ui$) $[u]\phi \rightarrow [i]\phi$

(Dual) $\langle i \rangle \phi \leftrightarrow \neg[i]\neg\phi$

(MP) IF $\vdash \phi_1$ AND $\vdash \phi_1 \rightarrow \phi_2$ THEN $\vdash \phi_2$

($\mathtt{N}^i$) IF $\vdash \phi$ THEN $\vdash [i]\phi$

where i, j denote elements of the set of indexes $\mathtt{C}$ and u denotes the universal context index in $\mathtt{C}$. The interaction axiom $\subseteq .ui$ states something quite intuitive concerning the interaction of the $[u]$ operator with all other context operators: what holds in the global context, holds in every context. Soundness and completeness of this axiomatization w.r.t. $\textsc{Cxt}^{\top}$ frames are proven in [10].

2.3 Anderson's Reduction Contextualized

Everything has been put into place to provide a contextualization of the version of Anderson's reduction sketched in Section 2.1. The fact that ϕ is ideally the case in context i can be formalized as $[i](\neg\phi \rightarrow \mathtt{V})$ and read as: *the negation of ϕ necessarily implies* *a violation within context i*. It becomes thus possible to express that ϕ is obligatory in the context i of a given normative system, while $\neg\phi$ is permitted in the context j of a different normative system: $[i](\neg\phi \rightarrow \mathtt{V}) \wedge \langle j \rangle (\neg\phi \wedge \neg\mathtt{V})$.

In [10] such reduction has been called a "counts-as reduction of deontic logic". Counts-as is the problematic locution introduced in [26, 27], and formally investigated for the first time in [17], which Searle takes as the basic syntax of constitutive rules, that is, of the building blocks of social reality. From a semantic point of view, such locution can acquire several different meanings, some of which have been systematically analyzed in [11, 13, 12, 14, 10]. One of these senses —the classificatory counts-as— is there formalized as the strict implication in $\mathbf{Cxt^u}$:

$$\neg\phi \Rightarrow_i^{cl} \mathtt{V} := [i](\neg\phi \rightarrow \mathtt{V}) \tag{3}$$

Intuitively, the negation of ϕ counts as a violation in context i, meaning that the negation of ϕ is classified as a violation in context i.

Such reduction can be straightforwardly strengthened by considering stronger senses of counts-as. One of these is the proper classificatory counts-as, also formalizable in $\mathbf{Cxt^u}$:

$$\neg\phi \Rightarrow_i^{cl+} \mathtt{V} := [i](\neg\phi \rightarrow \mathtt{V}) \wedge \neg[u](\neg\phi \rightarrow \mathtt{V}) \tag{4}$$

Intuitively, the negation of ϕ counts as a violation in context i, meaning that the negation of ϕ is classified as a violation in context i (first conjunct of the right-hand side of Formula 4), but the negation of ϕ is not always classified as a violation (second conjunct of the right-hand side of Formula 4).

2.4 Norms as Ascriptions

The reduction of deontic to counts-as statements of the type displayed in Formula 4 stresses that a state of affairs properly determines a violation only within a context, since outside the context that would not necessarily be the case. In [12] and [10], the rationale behind this formal characterization was taken from Searle's words themselves:

> [...] where the rule (or system of rules) is constitutive, behaviour which is in accordance with the rule can receive specifications or descriptions which it could not receive if the rule did not exist [26, p. 35].

Constitutive rules add something to what is already the case and proper contextual classification is a way to capture this intuition. However, there is also another way to look at the novelty introduced by constitutive rules. In a sense, what they do is to literally introduce new concepts, rather than just validating classifications which would otherwise not be valid. They create new terms to be used for a further conceptualization of reality. Such view of rules as ascriptions has a long history, starting with Pufendorf's notion of "impositio" [24, pp.100–101] and have been advanced in more recent times for instance in [16]. As a matter of fact, Searle's thesis according to which institutional facts are construed upon brute ones [27] is an instance of this ascriptive view of social reality.

Now, the central aspect of ascription is language creation. In order for an ascription to take place, a new term needs to be created, which can then be used for denoting the desired property. If we take an ascriptive view of Anderson's reduction, this means that the term "violation" is introduced in order to separate desired or ideal actions or states of affairs from their undesired or sub-ideal counterparts. Interestingly enough, this exact view is neatly formulated in Jørgensen's paper which introduced his dilemma [18]:

> How is a sentence of the form "Such and such is to be so and so" to be verified? How is it for instance to be verified that all promises are to be kept? To this question I know of no other answer than the following: The phrase "is to be etc." describes not a property which an action or a state of affairs either has or not, but a kind of quasi-property which *is ascribed* to an action or a state of affairs when a person is willing or commanding the action to be performed, resp. the state of affairs to be produced [18, pp. 292–293]

The following sections develop a formal analysis of this ascriptive view of norms. The primary technical difficulty resides in providing a suitable formal ground for the representation of language creation. From a propositional point of view, language creation means that new propositional atoms are somehow introduced in the language and consequently evaluated in the models. Therefore, in order to model language creation in logic, we should first be able to model, within the same logical framework, different languages. This is an aspect which, at first, might look hard to capture in a standard logical framework since evaluation functions are typically not partial, i.e. they evaluate all the atoms in the language.

3 "In the Beginning Was the Word"

The present section shows how modal logic offers an elegant way to represent different languages within one same formalism, without resorting to non-standard tools such as partial evaluation functions.

3.1 Adam and Eve

Consider the propositional language $\mathcal{L}$ built from the alphabet $\mathbf{P}$ of propositional atoms: $\mathtt{eat\_apple}$ ("the apple has been eaten"), $\mathtt{V}$ ("a violation has occurred"). We have of course four possible models such that: $w_1 \models \mathtt{eat\_apple} \wedge \mathtt{V}$, $w_2 \models \mathtt{eat\_apple} \wedge \neg\mathtt{V}$, $w_3 \models \neg\mathtt{eat\_apple} \wedge \mathtt{V}$ and $w_4 \models \neg\mathtt{eat\_apple} \wedge \neg\mathtt{V}$. That is, we have the state in which the apple is eaten and there is a violation (w_1), the state in which the apple is eaten but there is no violation (w_2), the state where the apple is not eaten and there is a violation (w_3), and finally the state where no apple is eaten nor there is a violation (w_4).

Obviously, all these states can be distinguished from each other. But suppose now to compare the models ignoring atom $\mathtt{V}$. Models w_1 and w_2 would not be distinguishable any more, nor would states w_3 and w_4. Which is just another way to say that, had we used a sublanguage $\mathcal{L}_i$ of $\mathcal{L}$ containing only atom $\mathtt{eat\_apple}$, we would have been able to distinguish only states w_1 from w_3 and w_2 from w_4. This latter can be considered to be the language at disposal of Adam & Eve in their pre-moral stage, before hearing God commanding "you shall not eat of the fruit of the tree that is in the middle of the garden"—rather than before actually eating the apple. In fact, after hearing God's command they were already endowed with the possibility to discern good ($\neg\mathtt{eat\_apple}$)

from evil (`eat_apple`), that is, their language was enriched and they got to distinguish also states w_1 from w_2 and w_3 from w_4, thanks to the newly introduced notion of violation (`V`).

3.2 Propositional Sublanguage Equivalence

The intuitions sketched in the previous section are here made formal. Take two propositional models m and m' for a propositional language $\mathcal{L}$. Models m and m' are equivalent if they satisfy the same formulae expressible in $\mathcal{L}$: $m \models \phi$ iff $m' \models \phi$. If m and m' are equivalent ($m \sim m'$) then there is no set Φ of formulae of $\mathcal{L}$ whose models contain m but not m', or vice versa. That is to say, the two models are indistinguishable for $\mathcal{L}$. However, two models which are not equivalent with respect to a given alphabet (a given set of atomic propositions), may become equivalent if only a sub-alphabet (a subset of the atomic propositions) is considered.

Definition 2. *(Propositional sublanguage equivalence)*
Two models m and m' for a propositional language $\mathcal{L}$ are equivalent w.r.t. sublanguage $\mathcal{L}_i$ if they satisfy the same set of formulae expressible using the alphabet of $\mathcal{L}_i$. For any $\phi \in \mathcal{L}_i$: $m \models \phi$ iff $m' \models \phi$. If m and m' are equivalent w.r.t. $\mathcal{L}_i$ ($m \sim_i m'$) then they cannot be distinguished by any set Φ of formulae of $\mathcal{L}_i$.

The definition makes precise the idea of two propositional models agreeing up to what is expressible on a given alphabet. To put it another way, it formalizes the idea that two models m and m' are equivalent modulo the alphabet in the complement $-\mathcal{L}_i$ (i.e., $\mathcal{L}\backslash\mathcal{L}_i$) of the sublanguage considered: m is indistinguishable from m' if we disregard the alphabet of $-\mathcal{L}_i$. Notice that if $m \sim_i m'$ and $\mathcal{L}_i = \mathcal{L}$ (i.e., the maximal element in $\mathfrak{Sub}(\mathcal{L})$) then $\sim_i = \sim$, that is, $\sim_i$ is the standard equivalence between propositional models.

Proposition 1. *(Properties of $\sim_i$)*
Let m and m' be two models for the propositional language $\mathcal{L}$. The following holds:

1. *For every sublanguage $\mathcal{L}_i$ of $\mathcal{L}$, relation $\sim_i$ is an equivalence relation on the set of all models of language $\mathcal{L}$.*
2. *For all sublanguages $\mathcal{L}_i$ and $\mathcal{L}_j$ of $\mathcal{L}$: if $\mathcal{L}_i \subseteq \mathcal{L}_j$ then $\sim_j \subseteq \sim_i$. It follows that for every sublanguage $\mathcal{L}_i$ of $\mathcal{L}$: $\sim \subseteq \sim_i$, that is, standard equivalence implies sublanguage equivalence.*

Proof. Claim (1) is straightforwardly proven. It is easy to see that: identity is a subrelation of $\sim_i$ for any sublanguage $\mathcal{L}_i$; and that $\sim_i \circ \sim_i$ and $\sim_i^{-1}$ are subrelations of $\sim_i$ for any sublanguage $\mathcal{L}_i$. Claim (2) is proven by considering that, if $\mathcal{L}_i$ is a sublanguage of $\mathcal{L}_j$ and $m \sim_j m'$, then for all propositions $\phi \in \mathcal{L}_i$: $m \models \phi$ iff $m' \models \phi$. Hence, $m \sim_i m'$.

3.3 Release Logic

Propositional release logics (**PRL**) have been first introduced and studied in [20, 21] in order to provide a modal logic characterization of the notion of irrelevancy. Irrelevancies are, in short, those aspects which we can choose to ignore. Irrelevancy is represented via modal release operators, specifying what is relevant to the current situation and what can instead be ignored. Release operators are indexed by an abstract 'issue' denoting what is considered to be irrelevant for evaluating the formula in the scope of the operator: $\Delta_I \phi$ means 'formula ϕ holds in all states where issue I is irrelevant', or 'ϕ holds in all states modulo issue I' or 'ϕ necessarily holds while releasing issue I'; $\nabla_I \phi$ means 'formula ϕ holds in at least one of the states where issue I is irrelevant', or 'ϕ possibly holds while releasing issue I'.

Issues can be in principle anything, but their essential feature is that they yield equivalence relations which cluster the states in the model. An issue I is conceived as something that determines a partition of the domain in clusters of states which agree on everything but I, or which are equivalent modulo I. Release operators are interpreted on these equivalence relations. As such, propositional release logic can be thought of as a "logic of controlled ignorance" [20]. They represent what we would know, and what we would ignore, by choosing to disregard some issues.

Syntax of PRL. The syntax of **PRL** is the syntax of a standard multi-modal language $\mathcal{L}^{Prl}$ [6] where n is the cardinality of the set $\mathtt{Iss}$ of releasable issues. The alphabet of $\mathcal{L}^{Prl}$ contains: an at most countable set $\mathbf{P}$ of propositional atoms p; the set of boolean connectives $\{\neg, \wedge, \vee, \rightarrow\}$; a finite non-empty set $\mathtt{Iss}$ of issues. Metavariables $I, J, ...$ are used for denoting elements of $\mathtt{Iss}$. The set of well formed formulae ϕ of $\mathcal{L}^{Prl}$ is defined by the usual BNF:

$$\phi ::= \top \mid p \mid \neg\phi \mid \phi_1 \wedge \phi_2 \mid \phi_1 \vee \phi_2 \mid \phi_1 \rightarrow \phi_2 \mid \Delta_I \phi \mid \nabla_I \phi.$$

where I denotes elements in $\mathtt{Iss}$.

One last important feature of **PRL** should be addressed before getting to the semantics. We have seen that modal operators are indexed by an issue denoting what is disregarded when evaluating the formula in the scope of the operator. The finite set $\mathtt{Iss}$ of these issues is structured as a partial order, that is to say, $\langle \mathtt{Iss}, \leqq \rangle$ is a structure on the non-empty set $\mathtt{Iss}$, where $\leqq$ ("being a sub-issue of") is a binary relation on $\mathtt{Iss}$ which is reflexive, transitive and antisymmetric. The aim of the partial order is to induce a structure on the equivalence relations denoting the release of each issue in $\mathtt{Iss}$: if $I \leqq J$ then the clusters of states obtained by releasing J contain the clusters of states obtained by releasing I. Intuitively, if I is a sub-issue of J then by disregarding J, I is also disregarded. This aspect is made explicit in the models which, for the rest, are just Kripke models.

Semantics of PRL. The semantics of **PRL** is given via the class $\mathfrak{Prl}$ of frames $\mathcal{F} = \langle W, \{R_I\}_{\mathtt{Iss}} \rangle$ such that W is a non-empty set of states and $\{R_I\}_{\mathtt{Iss}}$ is a family of equivalence relations such that: if $I \leqq J$ then $R_I \subseteq R_J$. Models are, as usual, structures $\mathcal{M} = \langle \mathcal{F}, \mathcal{I} \rangle$ where $\mathcal{I}$ is an evaluation function $\mathcal{I} : \mathbf{P} \longrightarrow \mathcal{P}(W)$ associating to each atom the set of states which make it true. **PRL** models are therefore just $\mathbf{S5}_n$ models with the further constraint that the granularity of the equivalence relations follows the partial order defined on the set of issues: the $\leqq$-smaller is the issue released, the more granular is the partition obtained via the associated equivalence relation. The satisfaction relation is standard. Boolean clauses are omitted.

Definition 3. *(Satisfaction for* **PRL** *models)*
Let $\mathcal{M}$ be a **PRL** *model.*

$$\mathcal{M}, w \models \Delta_I \phi \textit{ iff } \forall\, w', wR_I w' : \mathcal{M}, w' \models \phi$$
$$\mathcal{M}, w \models \nabla_I \phi \textit{ iff } \exists w', wR_I w' : \mathcal{M}, w' \models \phi.$$

where $I \in \mathtt{Iss}$. As usual, a formula ϕ is said to be valid in a model $\mathcal{M}$, in symbols $\mathcal{M} \models \phi$, iff for all w in W, $\mathcal{M}, w \models \phi$. It is said to be valid in a frame $\mathcal{F}$ ($\mathcal{F} \models \phi$) if it is valid in all models based on that frame. Finally, it is said to be valid on a class of frames F *(*F $\models \phi$*) if it is valid in every frame $\mathcal{F}$ in* F.

Axiomatics of PRL. Finally, the axiomatics amounts to a multi-modal **S5** plus the PO (partial order) axiom:

$$\begin{array}{rl}
(\mathtt{P}) & \text{all tautologies of propositional calculus} \\
(\mathtt{K}) & \Delta_I(\phi_1 \to \phi_2) \to (\Delta_I \phi_1 \to \Delta_I \phi_2) \\
(\mathtt{T}) & \Delta_I \phi \to \phi \\
(4) & \Delta_I \phi \to \Delta_I \Delta_I \phi \\
(5) & \nabla_I \phi \to \Delta_I \nabla_I \phi \\
(\mathtt{PO}) & \Delta_I \phi \to \Delta_J \phi \quad \text{IF } J \leqq I \\
(\mathtt{Dual}) & \nabla_I \phi \leftrightarrow \neg \Delta_I \neg \phi \\
(\mathtt{MP}) & \text{IF } \vdash \phi_1 \text{ AND } \vdash \phi_1 \to \phi_2 \text{ THEN } \vdash \phi_2 \\
(\mathtt{N}^I) & \text{IF } \vdash \phi \text{ THEN } \vdash \Delta_I \phi
\end{array}$$

where $I, J \in \mathtt{Iss}$. A proof of the soundness and completeness of this axiomatics w.r.t. to the semantics presented in Definition 3 is exposed in [21].

4 Modal Aspects of Ascriptivism

This section puts logics **Cxt^u** and **PRL** at work together. Their fusion [8] **Cxt^u** ⊗ **PRL** on language $\mathcal{L}^{Cxt} \otimes \mathcal{L}^{Prl}$ is all we need to get the axiomatics and semantics we are interested in. Notice that completeness will be preserved by the fusion of the axiom systems exposed in Sections 2.2 and 3.3 w.r.t. to the fusion $\mathfrak{TE}^{\sim} \otimes \mathfrak{Prl}$ of their classes of frames[4].

4.1 Propositional Sublanguage Equivalence as Release

Reasoning about propositional sublanguage equivalence is an instance of reasoning in release logic.

[4] Notice that the fusion $\mathfrak{TE}^{\sim} \otimes \mathfrak{Prl}$ considers the semantics of **Cxt^u** given in terms of i-j transitive and i-j euclidean frames containing an equivalence relation including all contexts (see Footnote 3). This is necessary because CXT$^{\top}$ frames are not closed under disjoint unions, which is a prerequisite for preserving Kripke completeness in fusions. See [8, Ch. 4].

Proposition 2. *(Sublanguage equivalence models)*
Consider a propositional language $\mathcal{L}$ on the set of atoms $\mathbf{P}$*, and a set of states W. Any evaluation function $\mathcal{I} : \mathbf{P} \longrightarrow \mathcal{P}(W)$ determines a* **PRL** *model $m = \langle W, \{\sim_{-i}\}_{i \in \mathfrak{Sub}(\mathcal{L})}, \mathcal{I} \rangle$.*

Proof. It follows from the properties of $\sim_i$ proven in Proposition 1.

Notice that the release issues $\mathtt{Iss}$ are the complements $-\mathcal{L}_i$ of the sublanguages in $\mathfrak{Sub}(\mathcal{L})$. In fact, what is released is just what cannot be expressed. The accessibility relations should therefore be taken to be the sublanguage-equivalence relations $\sim_{-i}$. Notice also that the set $\mathtt{Iss}$ is ordered by set-theoretic inclusion $\subseteq$ between sublanguages of $\mathcal{L}$[5].

To put it roughly, what the theorem says is that **PRL** is the logic to reason about scenarios like the Adam & Eve one sketched in Section 3.1. Let us get back to that example. Now it is possible to represent both the pre- and post- God's commandment situations, within the same formalism, by making use of the release operators of **PRL**. Suppose Adam & Eve to be at state w_1 in the model with domain $W = \{w_1, w_2, w_3, w_4\}$ and evaluation $\mathcal{I}$ as in Section 3.1. Recall that the language was built on atoms $\mathbf{P} = \{\mathtt{eat\_apple}, \mathtt{V}\}$. So let us denote with $\{\mathtt{V}\}$ and $\{\mathtt{eat\_apple}\}$ the sublanguages containing only atom $\mathtt{V}$ and, respectively, atom $\mathtt{eat\_apple}$. These sublanguages represent the releasable issues together with the empty language 0 and the full language $1 = \mathbf{P}$. Let $\mathcal{M} = \langle W, \{\sim_{\{\mathtt{V}\}}, \sim_{\{\mathtt{eat\_apple}\}}, \sim_0, \sim_1\}, \mathcal{I} \rangle$ be the resulting release model. We have that:

$$\mathcal{M}, w_1 \models \mathtt{eat\_apple} \land \mathtt{V} \tag{5}$$

$$\mathcal{M}, w_1 \models \Delta_0(\mathtt{eat\_apple} \land \mathtt{V}) \tag{6}$$

$$\mathcal{M}, w_1 \models \Delta_{\{\mathtt{V}\}}\mathtt{eat\_apple} \land \neg\Delta_{\{\mathtt{V}\}}\mathtt{V} \tag{7}$$

So Formula 5 just states what holds in w_1, which is the actual state where Adam & Eve eat the apple committing a violation. Formula 6 does the same by saying that, if you evaluate $\mathtt{eat\_apple}$ and $\mathtt{V}$ after releasing nothing, i.e., by using the full descriptive power of the language, then both $\mathtt{eat\_apple}$ and $\mathtt{V}$ necessarily hold. In fact, in the model at issue the set of states reachable from w_1 via $\sim_0$ coincides with w_1 itself, since there are no other states in W which are equivalent with w_1 if all available atoms are used in the comparison. Hence, in the model at issue, Δ_0 refers to the current evaluation state, i.e., w_1. Formula 7 shows what the effects of releasing atom $\mathtt{V}$ are. In fact, by abstracting from $\mathtt{V}$, state w_1 is not distinguishable any more from state w_2: $w_1 \sim_{\mathtt{V}} w_2$. Hence there exists a state $w_2 \in W$ such that $\mathcal{M}, w_2 \models \mathtt{eat\_apple} \land \neg\mathtt{V}$.

Formulae 7 and 6 represent Adam & Eve's situation after and, respectively, before God's commandment "you shall not eat of the fruit of the tree that is in the middle of the garden". Such commandment introduces a further characterization of reality, exemplified here by the notion of violation, which was not available to Adam & Eve before the commandment was uttered.

[5] It is instructive to notice that although all models based on sublanguage equivalence relations are *PRL* models, the reverse does not hold. In a sense the characterization in terms of *PRL* is too liberal. Future work will try to find axiomatizations for characterizing exactly the models based on sublanguage equivalence relations (see Section 5).

4.2 Ascription Formalized

God's commandment not to eat the apple is a statement $\texttt{eat\_apple} \rightarrow \texttt{V}$. Let us now suppose the set of all God's commandments to be Γ. Such set naturally defines a context i whose extension W_i is just the set of states satisfying Γ[6]. Since Γ contains $\texttt{eat\_apple} \rightarrow \texttt{V}$, such statement can be studied as a classificatory counts-as statement pertaining to the context i of divine commands. It corresponds to the validity of strict implication $[i](\texttt{eat\_apple} \rightarrow \texttt{V})$ in the model. To represent this, we should add contexts to the **PRL** model $\mathcal{M}$ introduced in the previous section. Let it be $\mathcal{M}' = \langle W, \{W, W_i\}, \{\sim_{\{\texttt{V}\}}, \sim_{\{\texttt{eat\_apple}\}}, \sim_0, \sim_1\}, \mathcal{I}\rangle$. Clearly, $[i](\texttt{eat\_apple} \rightarrow \texttt{V})$ will be valid in $\mathcal{M}'$ only if W_i does not contain state w_2, since $\mathcal{M}', w_2 \models \texttt{eat\_apple} \wedge \neg\texttt{V}$. Leaving technicalities aside, stating $[i](\texttt{eat\_apple} \rightarrow \texttt{V})$ in the Adam & Eve scenario modeled in $\mathcal{M}'$ corresponds to setting the boundaries of the context i of divine norms Γ in such a way to rule out states in which eating the apple is compatible with the non occurrence of a violation.

We hope the simple example of Adam & Eve to have conveyed the basic ideas behind the study of norms presented here, which builds on Anderson's reductionist tradition, on the analysis of counts-as presented in [10], and on the notion of propositional sublanguage equivalence. If we now pull these threads together within logic $\mathbf{Cxt^u} \otimes \mathbf{PRL}$, a new form of reduction can be defined which is based on a sense of counts-as taking its ascriptive aspect into account.

Definition 4. (Ascription of violation: $\Rightarrow^{As}_i$)
"$\texttt{V}$ is ascribed to $\neg\phi$ in context i" is formalized in the logic $\mathbf{Cxt^u} \otimes \mathbf{PRL}$*, on a multimodal language $\mathcal{L}^{Cxt} \otimes \mathcal{L}^{Prl}$ containing atom $\texttt{V}$ and the set of issues $\texttt{Iss} = \mathfrak{Sub}(\mathcal{L})$, with $\mathcal{L}$ being the non-modal fragment of $\mathcal{L}^{Cxt} \otimes \mathcal{L}^{Prl}$, as follows:*

$$\neg\phi \Rightarrow^{As}_i V := [i](\neg\phi \rightarrow V) \wedge \neg[i]\Delta_{\{V\}}(\neg\phi \rightarrow V) \tag{8}$$

Intuitively, the ascription of violation amounts to a classificatory counts-as[7] (first conjunct of the right-hand side of Formula 8) with the further condition (second conjunct) that the predicated implication does not hold in context i any more if it is evaluated releasing its consequent (in this case the violation atom $\texttt{V}$). It goes without saying that Definition 4 can easily be generalized to cover a notion of ascriptive counts-as $\phi_1 \Rightarrow^{As}_i \phi_2$ between any two formulae ϕ_1 and ϕ_2, where what is released in the second conjunct of the definition is the alphabet of ϕ_2. The ascription of atom $\texttt{V}$ is just a special case of ascriptive counts-as. In the next section we briefly sketch some properties of this counts-as operator which adds on the formal analysis of counts-as developed in [11, 13, 12, 14, 10].

Definition 4 represents a strengthening of Anderson's reduction along the line of Formula 3 and 4. It is worth spending a few more words on the right-hand side of Formula 8. Its dual version better displays the key idea behind it: $\neg\langle i\rangle(\neg\phi \wedge \neg\texttt{V}) \wedge \langle i\rangle\nabla_{\{\texttt{V}\}}(\neg\phi \wedge \neg\texttt{V})$. By releasing the consequent $\texttt{V}$ of the ascription, it becomes impossible

[6] The definition of contexts by sets of norms has been thoroughly investigated in [14, 10] in relation with constitutive rules and with the warning raised in [23]: "no logic of norms without attention to a system of which they form part".

[7] Note that the stronger form of proper classificatory counts-as could also be used.

to distinguish states which satisfy $\mathtt{V}$ from states which falsify $\mathtt{V}$. Now, the definition says that, in order for an ascription to hold, there is a state belonging to context i from which another state w' outside of context i can be reached which is indistinguishable from w once $\mathtt{V}$ is released, and which falsifies the implicative content of the counts-as $(\neg\phi \wedge \neg\mathtt{V})$[8].

4.3 On the Properties of Ascriptive Counts-As

There is no space to exhibit a full structural analysis of the syntax of the new counts-as connective $\Rightarrow_i^{Asc}$. However, it is worth noticing that it is very similar, structurally speaking, to proper classificatory counts-as $\Rightarrow_i^{Cl+}$ [12, 10]. Like proper classificatory counts-as, it satisfies the core of the structural properties of counts-as isolated in [17] (i.e., left and right logical equivalence, disjunction of the antecedents and conjunction of the consequents) and it falsifies transitivity[9]. However, there are also two essential differences.

First of all, ascriptive counts-as requires non-empty contexts: $[i]\bot \rightarrow \neg(\phi_1 \Rightarrow_i^{As} \phi_2)$. The validity of the property is easily checked semantically. None of the senses of counts-as analyzed in [11, 13, 12, 14, 10] enjoys this property. This is not surprising since the ascription of a property to something should presuppose the existence of that something. Secondly, contraposition, i.e., $(\phi_1 \Rightarrow_i^{As} \phi_2) \rightarrow (\neg\phi_2 \Rightarrow_i^{As} \neg\phi_1)$, is not valid. It fails in all models where a state in context i can be found which falsifies $\phi_1 \rightarrow \phi_2$ by releasing ϕ_2 but no state in i can be found which falsifies $\phi_1 \rightarrow \phi_2$ by releasing ϕ_1. This typically happens in models where i validates also $\phi_1 \vee \phi_2$. The failure of contraposition is an interesting aspect of $\Rightarrow_i^{As}$ since contraposition was one of the problematic properties of the classificatory view of counts-as. Ascription seems therefore to be a fruitful development of the classificatory perspective pursued in the series of works [11, 13, 12, 14, 10]. Further investigations in the structure of $\Rightarrow_i^{Asc}$ and in its logical relationships with the other senses of counts-as is left for future research.

4.4 An Ascriptive Glance at Jørgensen's Dilemma

The first of the ten philosophical problems urging today's deontic logic according to [15] was the problem, already formulated in [23], concerning a suitable foundation of deontic logic in the face of Jørgensen's dilemma:

> How can deontic logic be reconstructed in accord with the philosophical position that norms are neither true nor false? [15, p. 3]

It is our claim that the ascriptive view of noms can provide the ground for such a reconstruction. Let us sketch how this would work in the case of Adam & Eve scenario. There, God's commandment does three different things at the same time. First, the commandment defines the context i of divine norms. As such, formula $\mathtt{eat\_apple} \rightarrow \mathtt{V}$

[8] Typically, state w satisfies $\neg\phi$, that is, the antecedent of the counts-as since w and w' differ only in the interpretation of atom $\mathtt{V}$. In the Adam & Eve scenario, for instance, $w = w_1$ and $w' = w_2$.

[9] The countermodel of transitivity for $\Rightarrow_i^{Cl+}$ (see [12, 10]) works also for $\Rightarrow_i^{Asc}$.

defines the "logical space" [25, p. 6] of the normative system at issue, i.e., the context of the system (states w_1, w_3, w_4). Notice that, as such, $\mathtt{eat\_apple} \to \mathtt{V}$ is properly speaking neither true nor false, but it is rather taken or assumed to be true, exactly like an axiom. Second, the commandment teaches Adam & Eve how to recognize, to say it with Searle [27], states with a certain "institutional" property ('violation') on the ground of a "brute" property ('eating the apple'). Third, the commandment increases the granularity of Adam & Eve's language so that they can distinguish state w_1 from state w_2 (and w_3 from w_4) by making use of suitable "institutional" terms. This is the aspect of language creation proper of the ascriptive view of norms. To sum up, a norm $\phi \to \mathtt{V}$ in a set of norms Γ works like an axiom defining the context i of the normative system Γ, and defines the violation term(s) $\mathtt{V}$ by ascribing it to term(s) ϕ built from some "brute" language.

With respect to the third point, notice that the statement $\mathtt{eat\_apple} \to \mathtt{V}$ is neither true nor false if a "brute" language is spoken, where the "institutional" term $\mathtt{V}$ is not used. In fact, in the scenario there are states in the model where neither $\Delta_{\{\mathtt{V}\}}(\mathtt{eat\_apple} \to \mathtt{V})$ nor $\Delta_{\{\mathtt{V}\}}\neg(\mathtt{eat\_apple} \to \mathtt{V})$ are true. That is why, to say it with Jørgensen, norms correspond to "quasi-properties" of reality [18, pp. 292–293]. Properties, or to use Searle's terminology again, "brute facts" hold independently of the human ascriptive activity, while"quasi-properties" or "institutional facts" hold only as a result of ascription, and in this sense they are in a way less true. Notice, however, that this notion of truth is not the technical one used in Kripke semantics: the notion of truth in Jørgensen's dilemma (i.e., truth as what is evaluated as true given the brute language) is not the Kripke notion of truth (i.e., truth as what is evaluated as true given the whole language). The logic presented here generalizes this distinction to any possible partition besides the "brute" vs. "institutional" one.

5 Conclusions and Future Work

By providing Anderson's reduction with sufficient modal means for supporting a notion of context and of linguistic indistinguishability, the paper has provided an original view of deontic statements as forms of ascriptions (Definition 4). This has been claimed to be a sound perspective for grounding a reduction-based deontic logic in the face of Jørgensen's dilemma (Section 4.4).

Future work will focus on three aspects: first, a more accurate axiomatic characterization of $\mathfrak{Prl}$ frames with sublanguage equivalence relations will be pursued (see Footnote 5); second, the logical relations between ascriptive counts-as and the other forms of counts-as characterized in [14, 10] will be investigated; finally, the dynamic aspect of ascription will be studied making use of some form of update logic in the spirit of, for instance, [5].

Acknowledgments. This study was supported by the *Ministère de la Culture, de L'Enseignement Supérieur et de la Recherche, Grand-Duché de Luxembourg* (grant BFR07/123). The author is very grateful to the anonymous referees of DEON'08, Prof. Leendert van der Torre, Prof. Guido Boella and Paolo Turrini. Thanks to their accurate comments the present paper has been considerably improved.

References

1. Anderson, A.: The formal analysis of normative concepts. American Sociological Review 22, 9–17 (1957)
2. Anderson, A.: The logic of norms. Logique et Analyse 2, 84–91 (1958)
3. Anderson, A.: A reduction of deontic logic to alethic modal logic. Mind 22, 100–103 (1958)
4. Anderson, A.: Some nasty problems un the formal logic of ethics. Noûs 1, 345–360 (1967)
5. van Benthem, J., van Eijck, J., Kooi, B.: Logics of communication and change. Information and Computation 204(11), 1620–1662 (2006)
6. Blackburn, P., de Rijke, M., Venema, Y.: Modal Logic. Cambridge University Press, Cambridge (2001)
7. d'Altan, P., Meyer, J.-J.C., Wieringa, R.: An integrated framework for ought-to-be and ought-to-do constraints. In: Tan, Y.H. (ed.) Working Papers of the Workshop om Deontic and Non-Monotonic Logics, Rotterdam, pp. 14–45 (1993)
8. Gabbay, D., Kurucz, A., Wolter, F., Zakharyaschev, M.: Many-dimensional modal logics. Theory and applications. Elsevier, Amsterdam (2003)
9. Ghidini, C., Giunchiglia, F.: Local models semantics, or contextual reasoning = locality + compatibility. Artificial Intelligence 127(2), 221–259 (2001)
10. Grossi, D.: Designing Invisible Handcuffs. Formal Investigations in Institutions and Organizations for Multi-agent Systems. PhD thesis, Utrecht University, SIKS (2007)
11. Grossi, D., Meyer, J.-J.C., Dignum, F.: Modal logic investigations in the semantics of counts-as. In: Proceedings of the Tenth International Conference on Artificial Intelligence and Law (ICAIL 2005), pp. 1–9. ACM, New York (2005)
12. Grossi, D., Meyer, J.-J.C., Dignum, F.: Classificatory aspects of counts-as: An analysis in modal logic. Journal of Logic and Computation 16(5), 613–643 (2006)
13. Grossi, D., Meyer, J.-J.C., Dignum, F.: Counts-as: Classification or constitution? An answer using modal logic. In: Goble, L., Meyer, J.-J.C. (eds.) DEON 2006. LNCS (LNAI), vol. 4048, pp. 115–130. Springer, Heidelberg (2006)
14. Grossi, D., Meyer, J.-J.C., Dignum, F.: The many faces of counts-as: A formal analysis of constitutive-rules. The Journal of Applied Logic(to appear)
15. Hansen, J., Pigozzi, G., van der Torre, L.: Ten philosophical problems in deontic logic. In: Boella, G., van der Torre, L., Verhagen, H. (eds.) Normative Multi-agent Systems, number 07122 in Dagstuhl Seminar Proceedings.Internationales Begegnungs- und Forschungszentrum fuer Informatik (IBFI), Schloss Dagstuhl, Germany (2007)
16. Hart, H.L.A.: The ascription of responsibility and rights. In: Flew, A. (ed.) Logic and Language. Blackwell, Malden (1951)
17. Jones, A.J.I., Sergot, M.: A formal characterization of institutionalised power. Journal of the IGPL 3, 427–443 (1996)
18. Jorgensen, J.: Imperatives and logic, Erkentniss, pp. 288–296 (1937)
19. Kanger, S.: New fondations for ethical theory. In: Hilpinen, R. (ed.) Deontic Logic: Introductory and Systematic Readings, pp. 36–58. Reidel Publishing Company (1971)
20. Krabbendam, J., Meyer, J.-J.Ch.: Contextual deontic logics. In: McNamara, P., Prakken, H. (eds.) Norms, Logics and Information Systems, pp. 347–362. IOS Press, Amsterdam (2003)
21. Krabbendam, J., Meyer, J.-J.Ch.: Release logics for temporalizing dynamic logic, orthogonalising modal logics. In: Barringer, M., Fisher, M., Gabbay, D., Gough, G. (eds.) Advances in Temporal Logic, pp. 21–45. Kluwer Academic Publisher, Dordrecht (2000)
22. Lomuscio, A., Sergot, M.: Deontic intepreted systems. Studia Logica 75(1), 63–92 (2003)
23. Makinson, D.: On a fundamental problem of deontic logic. In: McNamara, P., Prakken, H. (eds.) Norms, Logics and Information Systems. NewStudies in Deontic Logic and Computer Science. Frontiers in Artificial Intelligence and Applications, vol. 49, pp. 29–53. IOS Press, Amsterdam (1999)

24. Pufendorf, S.: De Jure Naturae et Gentium, 1934th edn., p. 1688. Clarendon Press
25. Ricciardi, M.: Constitutive rules and institutions. Paper presented at the meeting of the Irish Philosophical Club, Ballymascanlon (February 1997)
26. Searle, J.: Speech Acts. An Essay in the Philosophy of Language. Cambridge University Press, Cambridge (1969)
27. Searle, J.: The Construction of Social Reality. Free Press (1995)
28. Stalnaker, R.: On the representation of context. Journal of Logic, Language, and Information 7, 3–19 (1998)

Author Index